

**Information Security and Privacy**

**Data Privacy, the need of the hour**



**Project Submitted to:**

**Prof. M. P. Sebastian**

**Department of Information Technology and Systems**

**IIMK**

**Submitted by:**

|  |  |
| --- | --- |
| **Name** | **Roll Number** |
| Anirban Roy Choudhury | EEPITM-01-001 |
| Aravinda A Hegde | EPGP-04B-014 |
| Gaurav Pratap Singh | EEPITM-01-004 |
| Satyajit Mohanty | EEPITM-01-020 |
| Saurav Sarkar | EPGP-04B-104 |



Contents

[Motivation and Project Summary 4](#_Toc358061605)

[Introduction, Understanding Data privacy and Personal Data 4](#_Toc358061606)

[Data breach 5](#_Toc358061607)

[Types of Data breach 6](#_Toc358061608)

[Well-meaning Insiders 6](#_Toc358061609)

[Targeted Attacks 7](#_Toc358061610)

[The Malicious Insider 8](#_Toc358061611)

[Stopping Data Breaches 8](#_Toc358061612)

[Examples of Security Breach/ Data Breach 9](#_Toc358061613)

[Legal aspects of Data privacy 12](#_Toc358061614)

[Financial Data 12](#_Toc358061615)

[Fair Credit Reporting Act 12](#_Toc358061616)

[Fair and Accurate Credit Transaction Act (2003) 13](#_Toc358061617)

[Right to Financial Privacy Act of 1978 13](#_Toc358061618)

[Health Information 13](#_Toc358061619)

[The American Recovery and Reinvestment Act of 2009, Title XIII, Subtitle D 13](#_Toc358061620)

[Electronic Communication Privacy Act (1986) 13](#_Toc358061621)

[Export of Personal Data 13](#_Toc358061622)

[Organization for Economic Co-operation and Development Fair Information Practices (1980) 13](#_Toc358061623)

[European Union Data Protection Directive 14](#_Toc358061624)

[German-Leach-Billey Act 15](#_Toc358061625)

[Freedom of Information Act (FOIA) (1966, amended 1974) 15](#_Toc358061626)

[Technology and Data Privacy 15](#_Toc358061627)

[Data encryption 15](#_Toc358061628)

[Data masking 15](#_Toc358061629)

[Steganography 16](#_Toc358061630)

[Online Privacy Certification Programs 16](#_Toc358061631)

[P3P 16](#_Toc358061632)

[Internet Content Rating Association (ICRA) 17](#_Toc358061633)

[Privacy-enhancing technologies (PET) 18](#_Toc358061634)

[Goals of PETs 18](#_Toc358061635)

[Existing PETs 19](#_Toc358061636)

[Future PETs 19](#_Toc358061637)

[Negotiation and Enforcement of Data Handling Conditions 20](#_Toc358061638)

[SmartScreen 20](#_Toc358061639)

[U-Prove 21](#_Toc358061640)

[Microsoft Office Document Inspector 21](#_Toc358061641)

[Microsoft Online Advertising Privacy 21](#_Toc358061642)

[Achieving k-anonymity Using Improved Greedy Heuristics (Very Large Relational Databases) 22](#_Toc358061643)

[Differential-Private Data Publishing Through Component Analysis 22](#_Toc358061644)

[Privacy Preserving Distributed DBSCAN Clustering 22](#_Toc358061645)

[Top Location Anonymization for Geosocial Network Datasets 23](#_Toc358061646)

[t-Plausibility: Generalizing Words to Desensitize Text 23](#_Toc358061647)

[Social awareness: 23](#_Toc358061648)

[Conclusion 23](#_Toc358061649)

[References 24](#_Toc358061650)

[Disclaimer 25](#_Toc358061651)

# 

# Motivation and Project Summary

The amount of digital data that is being churned all across the world is growing at an alarming rate. Today the amount of data that is in a newspaper is more than a person was likely to come across in a lifetime in the 18th century. It is estimated that 4 Exabyte of unique information will be generated every year. This digital data can be broadly classified as confidential, sensitive and public. Access to useful information from available data is something which gives the person having this information an edge. He/she can cause damage of unknown magnitude. This can affect governments, financial institutions, companies and individuals. Data security and Data privacy are therefore the need of the hour. In simple English Data security means physically protecting the data source from destructive forces and from unauthorized access. Data privacy or information privacy is the relationship between collection and dissemination of [data](http://en.wikipedia.org/wiki/Data), [technology](http://en.wikipedia.org/wiki/Technology), the public [expectation of privacy](http://en.wikipedia.org/wiki/Expectation_of_privacy), and the [legal](http://en.wikipedia.org/wiki/Legal) and [political](http://en.wikipedia.org/wiki/Political) issues surrounding them. In this project we would like to touch upon the various aspects of Data privacy and understand in details the need for Data privacy, ways to achieve data privacy and also understand the consequences of misuse of data.

# Introduction, Understanding Data privacy and Personal Data

During the second half of the last century, businesses, organizations and the government began using computers to store information about their customers, clients and staff in databases. This included information like names, address, contact information, employment records, medical information, convictions, credit history etc. The information of this type is usually confidential and sensitive and one would not want this information to be shared or made available to other. This data is ‘personal’ in nature. Now let us try to define what personal data. ‘Personal Data’ is data relating to a living or dead individual who is or can be identified either from the data or from the data in conjunction with other information that is in, or is likely to come into, the possession of the data controller. Either stored digitally or not, such data must not reach into the hands of malicious users. Digitally this data is stored in databases. Databases can be easily accessed, searched and edited. It’s also far easier to cross reference information stored in two or more databases than if the records were paper-based. The computers on which databases resided were often networked. This allows for organization-wide access to databases and offers an easy way to share information with other organizations. With more and more organisations using computers to store and process personal information there was a danger the information could be misused or get into the wrong hands

The following issues need to be looked into

* Who can access this information?
* Can this information be easily copied?
* How accurate was the information is?
* Is it possible to store information about a person without the knowledge or permission of that person?
* Was a record kept of any changes made to information typically by logs?

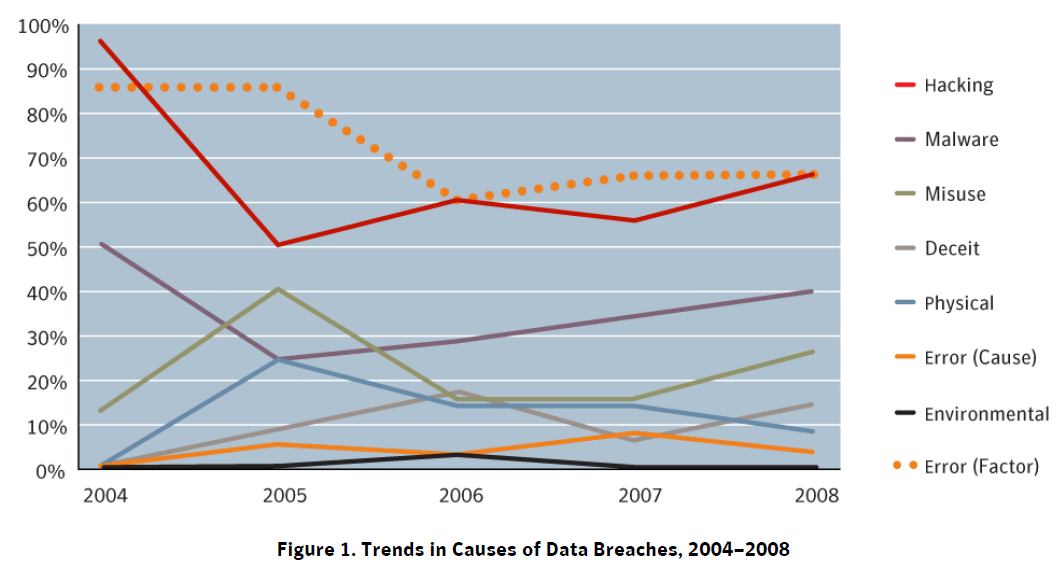
Personal Information in the wrong hands can lead to disastrous consequences. It could lead to identify theft where in information sold to criminals. Criminals commit identity frauds. It could also lead to information reaching the hands of competitors. For example, the customer information of company would be a priced possession for the competitor of that company. Through this project we try to understand the various ways in which the Data Privacy can be achieved and how technology can help to do the same.

# Data breach

A data breach is an incident in which sensitive, protected or confidential information has been stolen, viewed or used by an unauthorized user. This can happen intentionally or unintentionally. Data breaches may involve personal health information (PHI), personally identifiable information (PII), trade secrets or intellectual property.



A data breach incident may not need a hacker always to steal data from corporate networks i.e. If an unauthorized hospital employee observes patient’s health information over the shoulder of an authorized employee that also constitutes data breach. Today, strict governance of sensitive, protected or confidential data is required to avoid data breaches and to comply with government regulations. For example, there is PCI (Payment Card Industry) compliance for payment card industry as well as e-commerce sites which enables credit card payments. Also, the Health Insurance Portability and Accountability Act (HIPAA) regulate who may see and use PHI (Patient Health Information) such as name, date of birth, Social Security number and health history information. If a data breach results in identity theft or a violation of government or industry compliance mandates, the offending organization may face fines or other civil or criminal prosecution.

To stop data breaches we need to first understand ‘Why data breaches happen’? Third-party research into the root causes of data breaches, gathered from the Verizon Business Risk Team2 and the Open Security Foundation3, reveals three main types: well-meaning insiders, targeted attacks, and malicious insiders. In many cases, breaches are caused by a combination of these factors.

We will not see the various ways in which Data breaches are caused.

## Types of Data breach

### Well-meaning Insiders

Company employees who do not follow data security policies constitute the largest population of data breaches. According to the Verizon report, 67 percent of breaches in 2008 were aided by "significant errors" on the part of well-meaning insiders. In a 2008 survey of 43 organizations that had experienced a data breach, the Ponemon Institute found that over 88 percent of all cases involved incidents resulting from insider negligence. Breaches caused by this category can be further subdivided into five types.

**Data exposed on servers and desktops:** Corporate employees store, send or copy sensitive information in insecure system without encryption against standard data security policy. Systems that held sensitive data the organization did not know was stored on them accounted for 38 percent of all breaches in 2008—and 67 percent of the records breached.

**Lost or Stolen laptops:** Laptops carry lot of sensitive corporate information, customer emails, contacts etc. Not following corporate data security policy to encrypt sensitive information makes way for data breaches. The 2008 Ponemon Institute study found that lost laptops were the top cause of data breaches, representing 35 percent of organizations polled.

**Email, Web mail, and removable devices:** Employees copying sensitive data to memory stick, CD/DVD etc. or sending confidential data to personal email account against data security policy. Data is vulnerable to attack both during transmission as well as when storing in an unprotected device.

**Third-party data loss incidents:** Business partnership with third party vendors, outsourced processes, supply chain management etc. often requires exchange of sensitive information and when partners fail to implement data security policies then data breaches occur. The Verizon report implicated business partners in 32 percent of all data breaches.

**Business processes automate the spread of sensitive data:** Out of date business processes automatically distributing confidential data to unauthorized persons or systems and hence making it vulnerable to attack. Onsite risk assessments by Symantec find that in nearly half of these cases, outdated or unauthorized business processes are to blame for exposing sensitive data on a routine basis.

### Targeted Attacks

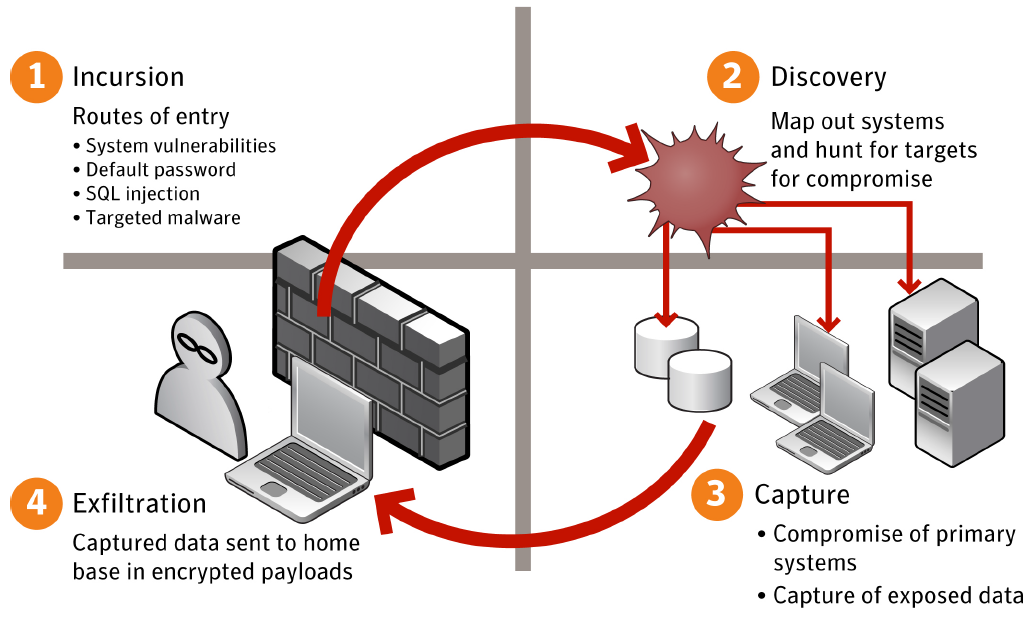
There is a constant rise in the rate of organized cybercrime and organizations always have some percent of data lying in the connected world. In 2008, Symantec created more than 1.6 million new malicious code signatures—more than in the previous 17 years combined—and blocked on average 245 million attempted malicious code attacks worldwide per month. There are four phases of targeted attacks such as incursion, discovery, capture, exfiltration.

**Phase1 - Incursion:** Hackers break into corporate network by exploiting system vulnerabilities, using default password violation, SQL injection, or targeted malware.

**Phase2 - Discovery:** Hacker scans for confidential data by mapping organization systems.

**Phase3 - Capture:** Exposed data stored in unprotected systems across the organization are captured immediately. Root kits are also installed in targeted systems and in network access points to capture confidential data flowing across the organization.

**Phase4 - Exfiltration:** Confidential data is sent back to hackers via email, encrypted packets or zipped files with password protection.

**

### The Malicious Insider

Data breaches caused by malicious insider i.e. insiders with the intention to steal confidential information are costlier then breaches caused by the negligence of insiders. This can be further categorized in following types.

**White collar crime:** The employee who knowingly steals data as part of an identity theft ring.

**Terminated Employees:** Data breaches caused by disgruntled former employees are common in today’s economic scenario where layoffs occur frequently. Employee uses the time window between notifications to network, exchange access removal to copy confidential data to memory sticks. A recent study of the effects of employee terminations on data security revealed that 59 percent of ex-employees took company data, including customer lists and employee records.

**Career building with company data:** Employees can store company data in their home systems to record their work samples for future references or career opportunities. The intention is NOT bad here but if the home system is compromised then it can damage the company image severely.

**Industrial espionage:** This is an unhappy employee who can send examples of work, product details, marketing plans, customer lists, financial data etc. to the competing company.

## Stopping Data Breaches

Given below are the various mechanisms to stop data breach

**Stop incursion:** Implement end point security to stop malware, intrusion prevention systems on servers to safeguard against attacks such as SQL Injection. Ensure default passwords are NOT active. Ensure security policies are enforced and not violated and centrally deploy policy & manage end points to apply timely patches, encryption, network access control & other security settings.

**Real time threat detection:** Security services that daily monitor millions of emails & systems worldwide can be leveraged to analyze internal event data to protect against emerging threats.

**Protect Information:** Find sensitive information located on servers to devices and protect it with policy based encryption. Inspect ALL outbound communications such as email, IM, Web, FTP, P2P, and generic TCP etc. and enforce policies to prevent confidential information from going out.

**IT compliance controls:** Automate IT compliance controls which are based on data security best practices & industry standards such as ISO 17799, COBIT, NIST SP800-53, Sarbanes-Oxley, PCI DSS, HIPAA, GLBA and others. Continuously measure on how well organization is meeting IT compliance controls & correct accordingly.

**Prevent data exfiltration:** Network software can be used to block confidential data from an outbound transmission phase. Identify transmissions to known hacker sites and alert security teams to prevent the exfiltration of confidential data.

**Prevention and response strategies:** It is very much necessary to integrate breach prevention and response plan into the day to day operation of the security team. Technology can be used to monitor & protect information and progressively reduce risk based on a constantly expanding knowledge of threats and vulnerabilities.

## Examples of Security Breach/ Data Breach

We will visit some popular data breach/privacy scandals in this section. These examples ranges from search to social media, e-mail to voice mail, mobile phones to Webcams to GPS but at the core of all these scandals are companies collecting personal data without users knowledge or consent and then either sharing with third part or not able to keep users private data safe.

**T-Mobile case:** UK’s fourth largest mobile company with 16.6 million customer base had the biggest data breach of its kind in which millions of T-Mobile customer data was sold to third party brokers. Customer data sold includes contract expiry details, name, address, and phone numbers. This data theft did not include financial data such as bank account or credit card details. This breach only targeted T-Mobile’s contract customers. This case shows that just fines are not enough to reduce this type of crimes but definitely jail sentences are required.

**Voice mail hacking:** It was a big story in 2011 which involved widespread hacking of mobile voice mail accounts of politicians, celebrities and crime victims for story publication in tabloids. Investigation into this illegal behavior is ongoing but already led to several high profile arrests & resignations of News Corps' weekly, U.K Publication executives.

**GM Case:** General motors ran into privacy issues with its OnStar GPS based system which may continue to track customers even after they cancel the service. Therefore, GM changed its privacy policy in 2011 indicating that it reserve the right to share the data with other companies it collected and this data includes vehicle's speed, location, odometer reading, seat belt usage and airbag deployment. GM said data sold would be anonymous and aggregated. Vehicle-based telematics systems are an emerging area for privacy concerns today.

**Carrier IQ Case:** Carrier IQ sells analytics software for mobile phones. This software is used in millions of smart phones. A security researcher found in his Smartphone that this software was capturing battery life, connections, text messages, emails and other actions. After this, Carrier IQ customers being taken to task for allegedly key logging, spying and tracking. But later it was found that this issue was mainly due to the confusion with the debug statement mistakenly left in the Android code by HTC (phone maker) programmers which was used by the original issue raiser. Finally, it turned out that carrier IQ was simply collecting performance data to optimize end user performance.

**Disney Violates Kid Data Rule:** In May, 2011, Disney's Playdom, Inc. had the dubious honor of paying the largest-ever COPPA fine, which was a $3 million civil penalty from the FTC for gathering and sharing personal information about hundreds of thousands of children without parental consent. Playdom which runs popular pony stars site, collected kids' ages and email addresses and allowed them to post their full names and locations. US web sites which targets children subscriptions & sales must comply with Children's Online Privacy Protection Act (COPPA) which has special rules including gathering permission from parents for children data.

**Sony PlayStation Case:** In April 2011, Sony announced that hackers had stolen personal data from 77 million PlayStation subscribers. Sony PlayStation network was breached to steal customer names, addresses, email addresses and birthdates. Sony was unsure of credit card data theft but warned customers to be on the lookout for identity theft. Security experts say Sony breach was one of the largest data breach on record. Sony estimated that the incident cost the company $171 million to rebuild its computers and purchase credit protection services for its customers.

**iPhone Case:** Apple was criticized for its user location data collection and storing who used their products such as iPhone & iPads for which their CEO Steve Jobs apologized in 2011. This issue came into focus when security researches found an unencrypted file inside the device which contained a cache of locations visited in last 12 months. Jobs emphasized that Apple was not tracking its customers: "Never have. Never will," he said, in response to the criticisms raised. However, this is not the only incident of gathering location data by mobile devices from Wi-Fi hotspots as Google & Microsoft later admitted they also gather similar data in their mobile operating systems.

**HealthNet Case:** Health Net a large health insurer announced a breach of sensitive information on nearly two million people. The breached information includes names, addresses, Social Security numbers, health and financial information. The data was not encrypted and stored on the hard drives of IBM’s data center a contractor for HealthNet. A nationwide class action suit was filed against HealthNet and IBM as a result of this incident.

**Facebook Case:** Facebook faced lot of data privacy related issues over the years. Many of the popular Facebook applications shared information by providing access to people's names and, in some cases, their friends' names—to dozens of advertising and Internet tracking companies. In October 2010, Facebook admitted that its top 10 most popular applications including FarmVille and Texas Hold`em shared user data. A Wall Street Journal investigation found that this breach impacted tens of millions of people including people who had very strict privacy settings. Previously also Facebook faced problems when they transmitted user id numbers to advertising companies when users clicked on ads. In November 2011, Facebook settled a case with the U.S. Federal Trade Commission about several incidents and agreed to 20 years of third-party privacy audits. FTC chairman Jon Leibowitz said "Facebook's innovation does not have to come at the expense of consumer privacy”.

**Google Street View Case:** In May 2007, when Google launched street view feature to Google Maps since then its struggling with privacy complains, paying fines and facing audits. Google Street View provides panoramic views of streets gathered by webcams. It raised privacy concerns for showing people leaving strip clubs, protesters at an abortion clinic, sunbathers in bikinis, cottagers at public parks, people picking up prostitutes and people engaging in activities visible from public property in which they do not wish to be photographed and have published online. Google says that pictures are taken from public property but Street View cameras take pictures from an elevated position, enabling them to look over hedges and walls designed to prevent some areas from being open to public view. Google allows users to flag inappropriate images for removal and also added the ‘blurring’ feature for faces & license plates. However, still Google ran into privacy battles with Switzerland, France, Belgium, Germany and South Korea, to name a few countries. In March 2011, France fined Google the equivalent of $142,000 related to Street Views, but in August 2011, in a review by the U.K. government gave Google positive marks for improving the privacy of Street View. But Google must undergo regular privacy audits mandated by the FTC for the next 20 years.

# Legal aspects of Data privacy

In today’s world, where anything to everything can be achieved over ICT, whether it is applying for your passport, paying electricity bills, opening an account in a bank, registering your details for a mobile phone or online shopping. This has opened up the business as the customer base have increased, reduced the cost of ownership etc., so with the good comes the devil i.e. it has given rise to lot of data privacy issues related to personal data, financial data, health data etc. Any process, people having access to any of these data , can play a very destructive role by taking the image of yours (impersonification) or selling these data in the open market or to your competitor to make quick money, then the owner of these datasets can use these to gain market share or any other illegal activities. Here is a scenario, where an insider from T-Mobile in Europe, while leaving the job at T-Mobile took the data of the customers and sold the same in the open market and to the competitor of T-Mobile to make quick money. Then the competitor of T-Mobile started making cold calls to these customers and asking them to switch to their network by providing them attractive gifts and discounts. This made a huge dent to the revenue of the T-Mobile as well as to the reputation and brand name of the company. This made to kick start some of the legal aspects by the European government as well as US government. Here we will highlight some of the legal aspects of data privacy that each company dealing with personal data must follow.

## Financial Data

### Fair Credit Reporting Act

The fair credit reporting act of 1970 regulates the operations of credit reporting bureaus including how they collect, store and use credit information. The act, enforced by the U.S federal trade commission is designed to ensure the accuracy, fairness and privacy of information gathered by credit reporting agencies and to check these systems that gather and sell information about people. The act outlines who may access your credit information, how you can find out what is in your file, how to dispute inaccurate data and how long data is retained. It also prohibits the credit reporting bureau/agencies from giving out information about you to your employer or potential employer without your written consent.

### Fair and Accurate Credit Transaction Act (2003)

The fair and accurate credit transaction act was passed in 2003 as an amendment to Fair Credit Reporting Act, and it allows consumers to request and obtain a free credit report once each year from each of the three primary consumer credit reporting companies (Equifax, Experian and Trans Union).

### Right to Financial Privacy Act of 1978

The right to Financial Privacy Act of 1978 protects the financial records of financial institution customers from unauthorized scrutiny by the federal government. Prior to passage of this act, financial institution customers were not informed if their personal records were being turned over for review by a government authority, nor could customers challenge government access to their records. Under this act, a customer must receive a written notice that a federal agency intends to obtain their financial records, along with an explanation of the purposed for which records are sought. The customers must also be given written procedures to follow if he or she does not wish the records to be made available. The financial institution cannot release a customer’s financial records until the government authority seeking the records certifies in writing that it has complied with the applicable provision of the act.

## Health Information

### The American Recovery and Reinvestment Act of 2009, Title XIII, Subtitle D

The American Recovery and Reinvestment Act 2009 is a wide-ranging act that authorized $787 billion in spending and tax cuts over a 10-year period. Title XIII, Subtitle D of this act included a strong privacy provisions for electronic health records, including banning the sale of health information, promoting the use of audit trails and encryption, and providing rights of access for patients. This section of the act also mandates that within 60 days after discovery of a data breach, each individual whose health information has been exposed must be notified, and if a breach involves 500 or more people, notice must be provided to prominent media outlets.

## Electronic Communication Privacy Act (1986)

The Electronic Communication Privacy Act 1986 (ECPA) deals with three main issues: (1) the protection of communications while in transfer from sender to receiver; (2) the protection of communication held in electronic storage; and (3) the prohibition of devices to record dialing, routing, addressing, and signaling information without a search warrant. ECPA was passed as an amendment to Title III of the Omnibus crime control and Safe Streets Act.

## Export of Personal Data

### Organization for Economic Co-operation and Development Fair Information Practices (1980)

One agency concerned with privacy is the Organization for Economic Co-operation and Development (OECD), an international organization consisting of 34 member countries including Australia, Canada, France, Germany, Italy, Japan, USA etc. The goal is to set policy and come to an agreement on topics for which multilateral consensus is necessary for individual countries to make progress in a global economy. The 1980 privacy guidelines set by OECD –also known as Fair Information Practices are often held up as the model of ethical treatment of consumer data for organizations to adapt. Some of the principles are highlighted below.

|  |  |
| --- | --- |
| **Principle** | **Guideline** |
| Collection Limitation | The collection of personal data must be limited, all such data must be obtained lawfully and fairly with the subject’s consent and knowledge |
| Data Quality | Personal data should be accurate, complete, current and relevant to the purpose for which it is used |
| Purpose Specification | The purpose for which personal data are collected should be specified and should not be changed |
| Use limitation | Personal data should not be used beyond the specified purpose without a person’s consent or by authority of law. |
| Security Safeguards | Personal data should be protected against unauthorized access, modification or disclosure |
| Individual participation | People should have right to review their data, to challenge its correctness and to have incorrect data changed |
| Accountability | A data controller should be responsible for ensuring that the above principles are met |

### European Union Data Protection Directive

The European Union Data Protection Directive requires any company doing business within the borders of 15 western European nations to implement a set of privacy directives on the fair and appropriate use of information. Basically, this directive requires member countries to ensure data transferred to non-European Union (EU) countries are protected. It also bars the export of data to countries that do not have data privacy standards comparable to the EU’s. The following list summarizes these seven European Data privacy principles:

* Notice – Tell all customers what is done with their information
* Choice – Give the customer a way to opt out of marketing
* Onward transfer – When data is transferred to suppliers of other business partners, companies must observe the notice and choice principles mentioned above and require all recipients of such data to provide at least the same level of protection of such data.
* Access – Give customers access to their information
* Security – Protect customer information from unauthorized access.
* Data Integrity – Ensure that information is accurate and relevant
* Enforcement – Independently enforce the privacy policy

### German-Leach-Billey Act

It establishes guidelines for the collection and disclosure of personal financial information; requires financial institutions to document their data security plan; and encourages institutions to implement safeguards against protexting.

### Freedom of Information Act (FOIA) (1966, amended 1974)

The Freedom Information Act (FOIA), passed in 1966 and amended in 1974, grants citizen the rights to access certain information and records of the federal government upon request. FOIA is a powerful tool that enables journalists and the public to acquire information that the government is reluctant to release. There are two basic requirements for filing a FOIA request: (1) the requestor must not require wide-ranging, unreasonable, or burdensome searches for records, and (2) the request must be made according to agency procedural regulations published in the Federal Register

# Technology and Data Privacy

In this section we will see various technology aids which we can use to achieve data privacy.

## Data encryption

This is one of the most common and dominant form of data security technique. In this the data is encrypted or converted into an unintelligible form. Data can be encrypted from the sender with the help of a key to produce a cipher or in the unintelligible form. The receiver then using a key decrypts the data to get the original text. In an organization data remain on hard drives, mobile devices, communicated through emails. It is imperative for an organization to protect data in all these channels. Data can be encrypted on a hard disk. This technique is known as disk encryption .If a person loses his hard disk or his mobile device and if his data is encrypted. Any unauthorized person won’t be able to decipher the data. The email sent could again be encrypted to stop unauthorized person reading the content of email, this is known as email encryption. Another important place to safeguard data is in the databases. Lot of database vendors such as Microsoft and Oracle provides database encryption along with the disk and backup disk encryption. This technology is known as transparent data encryption. Data Encryption can also be merged with stringent user authentication mechanisms which can be seen as a two factor authentication

## Data masking

This technique is used to hide the data by keeping the meaning or metadata of the data valid. Most of the software applications are developed in two or three system. One is the production system, which is the actual system on which the end users use. Other two could be development and testing which is used by developers for developing and testers for testing. Some organization may merge this development and testing system. Sometimes the production data is copied to these development and testing systems to give their database valid or production like data. This can again cause data security issue as these systems may not have that stringent access controls or can be accessed by an unauthorized person. In these cases it is important to mask the data such that the actual data is hidden but whatever properties of data are required for the application is retained. Some common techniques of data masking are Substitution, Shuffling and Masking out.

## Steganography

Steganography is the art of hiding the information. While in cryptography the eavesdropper knows that there is some secret information, in steganography he is unaware of any secrecy. Steganography has its roots from WW2 days when Germans had used Microdots and also the usage of invisible inks was prevalent during that time. Digital steganography refers to the usage of steganography techniques in the digital arena. Common usage of digital steganography is in data security, securing the stored data, peer to peer communication etc. Digital Watermarking is one the most common techniques of digital steganography where a document or an image is embedded with some other information. Watermarking can be visible or invisible. Copyright protection, certification and conditional access find the use of watermarking. Digital steganography along with Cryptography again can provide more robust security for the data

## Online Privacy Certification Programs

Online Certification or "Seal" programs are an example of industry self-regulation of privacy policies. Seal programs usually require implementation fair information practices as determined by the certification program and may require continued compliance monitoring. Online seal programs are

1. TRUSTe (first online privacy seal program, included more than 1800 members by 2007)
2. The Trust Guard Privacy Verified program,
3. eTrust
4. Webtrust

## P3P

P3P is a machine-readable language that helps to express a website’s data management practices. P3P manages information through privacy policies. When a website uses P3P, they set up a set of policies that allows them to state their intended uses of personal information that may be gathered from their site visitors. When a user decides to use P3P, they set their own set of policies and state what personal information they will allow to be seen by the sites that they visit. Then when a user visits a site, P3P will compare what personal information the user is willing to release, and what information the server wants to get – if the two do not match, P3P will inform the user and ask if he/she is willing to proceed to the site, and risk giving up more personal information.

The main content of a privacy policy is the following:

* Which information the server stores.
* Kind of information is collected (identifying or not, IP address, email address, name, etc.)
* Use of the collected information:
* Uses of Information (for regular navigation, tracking, personalization, telemarketing, etc.);
* Who will receive this information (Current company or Third Party)
* Permanence and visibility:
* How long information is stored;
* Whether and how the user can access the stored information (read-only, opt-in, and opt-out).

The privacy policy can be retrieved as an [XML](http://en.wikipedia.org/wiki/XML) file or can be included, in compact form, in the [HTTP](http://en.wikipedia.org/wiki/HTTP) header. The location of the XML policy file that applies to a given document can be:

1. Specified in the [HTTP](http://en.wikipedia.org/wiki/HTTP) header of the document
2. Specified in the [HTML](http://en.wikipedia.org/wiki/HTML) head of the document
3. If none of the above is specified, the well-known location /w3c/p3p.xml is used (for a similar location compare [/favicon.ico](http://en.wikipedia.org/wiki/Favicon))

## Internet Content Rating Association (ICRA)

ICRA created a content description system which allowed webmasters and digital content creators to self-label their content in categories such as [nudity](http://en.wikipedia.org/wiki/Nudity), [sex](http://en.wikipedia.org/wiki/Sex), [language](http://en.wikipedia.org/wiki/Language) (vulgar terms etc.), [violence](http://en.wikipedia.org/wiki/Violence), other potentially undesired material and online interactivity such as social networking and chat. There are context variables such as [art](http://en.wikipedia.org/wiki/Art), [medicine](http://en.wikipedia.org/wiki/Medicine) and [news](http://en.wikipedia.org/wiki/News)—for example, a piece of content or site can be described as having depictions of nudes, but they are in an artistic context. A key point is that ICRA does not rate internet content, nor do they make value judgments about sites – the content providers self-label, and then parents and other concerned adults make a decision as to what is or is not appropriate for themselves or their children. The labeling was done using a web-based questionnaire. The content creators checked which of the elements in the questionnaire are present or absent from their Web sites, and a small file is automatically generated using the [RDF](http://en.wikipedia.org/wiki/Resource_Description_Framework) format, which is then linked to the content on one or more domains. Formerly, the system was based on [PICS](http://en.wikipedia.org/wiki/Platform_for_Internet_Content_Selection). Users, especially parents of young children, could then use [content filtering software](http://en.wikipedia.org/wiki/Content_filtering_software) to allow or disallow various types of content. One such application, ICRA plus, was maintained by ICRA itself. The ICRA also intended to launch a service to verify the accuracy of ICRA labels and to provide this information to third-party tools and services, such as search engines.

Alternative labeling projects are:

1. Quatro (An [EU](http://en.wikipedia.org/wiki/European_Union)-funded project which integrates content labels with quality and trust marks)
2. Quatro Plus.

As of October 2010, the ICRA labeling engine has been discontinued by FOSI. Additionally, FOSI has withdrawn all support for the ICRA rating system and taken down all documentation for labeling websites with ICRA ratings. The reason is the ICRA label failed to gain widespread acceptance.

## Privacy-enhancing technologies (PET)

Privacy enhancing technologies (PET) is a general term for a set of computer tools, applications and mechanisms which when integrated with online services or applications or when used in conjunction with such services or applications - allow online users to protect the privacy of their personally identifiable information (PII) provided to and handled by such services or applications.

Privacy enhancing technologies can also be defined as:

“Privacy-Enhancing Technologies is a system of ICT measures protecting informational privacy by eliminating or minimizing personal data thereby preventing unnecessary or unwanted processing of personal data, without the loss of the functionality of the information system.”

### Goals of PETs

PETs aim at allowing users to take one or more of the following actions related to their personal data sent to, and used by, online service providers, merchants or other users:

* Increase control over their personal data sent to, and used by, online service providers and merchants (or other online users) (self-determination)
* Data minimization: minimize the personal data collected and used by service providers and merchants
* Choose the degree of anonymity (e.g. by using pseudonyms, anonymizers or anonymous data credentials)
* Choose the degree of unlink ability (e.g. by using multiple virtual identities)
* Achieve informed consent about giving their personal data to online service providers and merchants
* Provide the possibility to negotiate the terms and conditions of giving their personal data to online service providers and merchants (data handling/privacy policy negotiation). In Privacy Negotiations, consumers and service providers establish, maintain, and refine privacy policies as individualized agreements through the ongoing choice amongst service alternatives.
* Provide the possibility to have these negotiated terms and conditions technically enforced by the infrastructures of online service providers and merchants (i.e. not just having to rely on promises, but being confident that it is technically impossible for service providers to violate the agreed upon data handling conditions)
* Provide the possibility to remotely audit the enforcement of these terms and conditions at the online service providers and merchants (assurance)
* Data tracking: allow users to log, archive and look up past transfers of their personal data, including what data has been transferred, when, to whom and under what conditions
* Facilitate the use of their legal rights of data inspection, correction and deletion

### Existing PETs

* Communication anonymizers hiding the real online identity (email address, IP address, etc.) and replacing it with a non-traceable identity (disposable / one-time email address, random IP address of hosts participating in an anonymizing network, pseudonym, etc.). They can be applied to email, Web browsing, P2P networking, VoIP, Chat, instant messaging, etc.
* Shared bogus online accounts. One person creates an account for ABC, providing bogus data for Name, address, phone number, preferences, life situation etc. They then publish their user-ID and password on the Internet. Everybody can now use this account comfortably. Thereby the user is sure that there is no personal data about him in the account profile.
* Access to personal data: The service provider's infrastructure allows users to inspect, correct or delete all their data stored at the service provider.

### Future PETs

* **Wallets of multiple virtual identities**: ideally unlink able. Such wallets allow the efficient and easy creation, management and usage of virtual identities.
* **Anonymous credentials**: asserted properties/attributes or rights of the holder of the credential that don't reveal the real identity of the holder and that only reveal so much information as the holder of the credential is willing to disclose. The assertion can be issued by the user herself, by the provider of the online service or by a third party (another service provider, a government agency, etc.)
  + - An assertion of minimal age, issued by the state, proving that the holder is older than 23 (i.e. the actual age is not provided)
    - A driving license i.e. an assertion, issued by the motor vehicle control agency, that the holder is entitled to drive cars
    - A proof of insurance issued by the health insurance
    - Digital cash

## Negotiation and Enforcement of Data Handling Conditions

Before ordering a product or service online, the user and the online service provider or merchant negotiate the type of personal data that is to be transferred to the service provider. This includes the conditions that shall apply to the handling of the personal data, such as whether or not it may be sent to third parties (profile selling) and under what conditions (e.g. only while informing the user), or at what time in the future it shall be deleted (if at all).

* **Data transaction log**. Users can log what personal data they sent to which service provider, when and under what conditions. These logs are stored and allow users to determine what data they have sent to whom, or they can establish the type of data that is in possession by a specific service provider.
* **Filter with ActiveX.** You can turn on ActiveX Filtering to help prevent programs from accessing your computer without your consent
* **Prevent attacks from fraudulent websites.** Cross-site scripting (XSS) filter can help prevent attacks from websites that might attempt to steal your personal and financial information.
* **128-bit Secure Sockets Layer (SSL) connection.** An SSL connection helps Internet Explorer create an encrypted connection with websites run by banks, online stores, medical sites, or other organizations that handle your personal information.

## SmartScreen

Filter helps you avoid threats such as phishing sites and online fraud when you browse the web. Phishing is one of the most widespread online scams. Criminals set up a fake website and send emails designed to trick people into disclosing critical personal data, such as bank account or credit card numbers, passwords, or personally identifiable information.

When the filter encounters a suspected phishing or malicious software site, it displays a warning—"This Web site has been reported as unsafe"—and prompts the user to return to his or her home page to resume browsing. SmartScreen Filter also checks for malicious downloads. If you attempt to download a file that has been reported as unsafe, Internet Explorer displays a warning.

Unsafe sites are identified by a series of steps:

* SmartScreen Filter checks the website against a list of high-traffic sites that is stored in your computer.
* If the website or download file address (or URL) is not in that list, its URL is sent to Microsoft for comparison against a frequently updated list of sites and downloads that have been reported as unsafe or suspicious. All information sent to Microsoft is encrypted on the client computer and transmitted using Hypertext Transfer Protocol Secure (HTTPS).

## U-Prove

U-Prove is a cryptographic technology that allows organizations to verify statements about an individual, with just a minimal amount of information. These statements, known as claims, could include the name of an individual, their age, their affiliations, their school, and so forth.

U-Prove tokens can include privacy and security properties such as:

* Allowing you the option to minimize the amount of personal information you disclose when you access services.
* Allowing organizations to balance access to information and services with individual privacy, by allowing untraceable and unlikable transactions between parties.
* Allowing you to present claims to a service provider, without the claim issuer knowing where they were used.

Example: A student who receives an electronic ID card from his university. This card could contain a number of U-Prove tokens, each of which has information that could be used by the student to share claims about him. He might choose to access a countrywide network of online libraries, and the electronic books therein. U-Prove tokens can allow the participating libraries to verify the claim that he has a right to access the service, based on his status as a registered university student, without his having to reveal other information about his status or himself (such as his name, student ID, graduation date, age, and so on).

## Microsoft Office Document Inspector

Document Inspector lets you select the types of hidden content you want to retain in documents and check for sensitive or prohibited content you want to remove. Many documents go through lengthy approval processes that involve multiple reviewers and contributors. Microsoft Office, like many other productivity applications, allows you to associate a wide range of details with a document such as each reviewer's comments, name, organization, and the folder and server where the file is located stored within the file for reference.

## Microsoft Online Advertising Privacy

Every public Microsoft web page includes a link to the Microsoft Online Privacy Notice Highlights, which contains information about Microsoft's personalized advertising practices. The Online Privacy Notice Highlights page provides a link to an opt-out page where you can choose to stop receiving personalized ads on a specific computing device or browser.

## Achieving k-anonymity Using Improved Greedy Heuristics (Very Large Relational Databases)

Advances in data storage, data collection and inference techniques have enabled the creation of huge databases of personal information. Dissemination of information from such databases - even if formally anonymized, creates a serious threat to individual privacy through statistical disclosure. One of the key methods developed to limit statistical disclosure risk is k-anonymity. Several methods have been proposed to enforce k-anonymity notably Samarati's algorithm and Sweeney's Datafly, which both adhere to full domain generalization. Such methods require a tradeoff between computing time and information loss. The improved greedy algorithm was compared with the original methods. Metrics like information loss, computing time and level of generalization were deployed for comparison. Results show that the improved greedy algorithm maintains a better balance between computing time and information loss.

## Differential-Private Data Publishing Through Component Analysis

A reasonable compromise of privacy and utility exists at an 'appropriate' resolution of the data. We proposed novel mechanisms to achieve privacy preserving data publishing (PPDP) satisfying e-differential privacy with improved utility through component analysis. The mechanisms are Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). The differential PCA-based PPDP serves as a general-purpose data dissemination tool that guarantees better utility (i.e., smaller error) compared to Laplacian and Exponential mechanisms using the same “privacy budget”. Our second mechanism, the differential LDA-based PPDP, favors data dissemination for classification purposes. Both mechanisms were compared with state-of-the-art methods to show performance differences.

## Privacy Preserving Distributed DBSCAN Clustering

DBSCAN is a well-known density-based clustering algorithm, which offers advantages for finding clusters of arbitrary shapes compared to partitioning and hierarchical clustering methods. However, there are few papers studying the DBSCAN algorithm under the privacy preserving distributed data-mining model, in which the data is distributed between two or more parties, and the parties cooperate to obtain the clustering results without revealing the data at the individual parties. We first propose two protocols for privacy preserving DBSCAN clustering over horizontally and vertically partitioned data respectively and then extend them to arbitrarily partitioned data. We also provide performance analysis and privacy proof of our solution.

## Top Location Anonymization for Geosocial Network Datasets

Geosocial networks such as Foursquare have access to users' location information, friendships, and other potentially privacy sensitive information. We show that an attacker with access to a naively-anonymized geosocial network dataset can breach users' privacy by considering location patterns of the target users. We introduce k-anonymity-based properties for Geosocial network datasets, propose appropriate data models and algorithms, and evaluate our approach on both synthetic and real-world datasets.

## t-Plausibility: Generalizing Words to Desensitize Text

De-identified data has the potential to be shared widely to support decision-making and research. While significant advances have been made in anonymization of structured data, anonymization of textual information is in its infancy. Document sanitization requires finding and removing personally identifiable information. While current tools are effective at removing specific types of information (names, addresses, dates), they fail on two counts. The first is that complete text redaction may not be necessary to prevent re-identification, since this can affect the readability and usability of the text. More serious is that identifying information, as well as sensitive information, can be quite subtle and still be present in the text even after the removal of obvious identifiers. Observe that the diagnosis “tuberculosis” is sensitive, but in some situations it can also be identifying. Replacing it with the less sensitive term ``infectious disease'' also reduces identifiability. That is instead of simply removing sensitive terms, these terms can be hidden by more general but semantically related terms to protect sensitive and identifying information, without unnecessarily degrading the amount of information contained in the document.

## Social awareness:

Above all the methods, social awareness is the most important data breach prevention technique. It is important that all the members of the organization are aware of the security policies, guidelines and standards of the organizations. They should not allow any form of social engineering to happen. Employees must be given adequate training on this regard. Employees must follow and be aware of all the organization’s security policies like email, password policy etc.

# Conclusion

Data is everywhere and forms an important entity in any digital application or software. With increasing amount of volume of data and their varying degree of use, their lies a concern of privacy and breach of data. Companies like T-Mobile, Sony, Disney, and GM have lost lot of credibility and money when data breach happened in their systems. There are numerous techniques which should be adopted by the companies to make their data secure. These techniques must comply with different data privacy standards and regulations according to the industry and location in which a company operates. There are lot of tools which employ these techniques and algorithms to ensure data privacy and prevent data breach. Above all, privacy of data should not be a concern of organization; it must be practiced by an individual. All companies must ensure they have policies and standards regarding data privacy. All employees must be aware of those policies and should adhere to them.

# References

Data Encryption: <https://en.wikipedia.org/wiki/Disk_encryption>

Data Encryption: <http://en.wikipedia.org/wiki/Email_encryption>

Data Encryption: <http://en.wikipedia.org/wiki/Transparent_Data_Encryption>

Data Encryption: <http://www.mcafee.com/us/resources/solution-briefs/sb-gemalto.pdf>

Data Masking: <http://en.wikipedia.org/wiki/Data_masking>

Data Masking: <http://www.dataobfuscation.com.au/>

Steganography: <http://www.ijcaonline.org/volume9/number7/pxc3871887.pdf>

Steganography: <http://en.wikipedia.org/wiki/Steganography>

Tools: <http://www-01.ibm.com/software/data/optim/core/data-privacy-solution/>

Tools: <http://www.mcafee.com/in/products/endpoint-encryption.aspx>

Tools: <http://www.symantec.com/products-solutions/families/?fid=data-loss-prevention>

<http://www.tdp.cat/issues11/vol05n02.php>

<http://searchsecurity.techtarget.com/definition/data-breach>

<http://eval.symantec.com/mktginfo/enterprise/white_papers/b-anatomy_of_a_data_breach_WP_20049424-1.en-us.pdf>

<http://news.bbc.co.uk/2/hi/8364421.stm>

<http://www.guardian.co.uk/uk/2009/nov/17/t-mobile-phone-data-privacy>

<http://www.networkworld.com/news/2012/012612-privacy-scandals-255357.html>

<http://en.wikipedia.org/wiki/Google_Street_View_privacy_concerns>

<http://en.wikipedia.org/wiki/Information_privacy>

<http://en.wikipedia.org/wiki/Data_security>

<http://weis2010.econinfosec.org/papers/session1/weis2010_romanosky.pdf>

# Disclaimer

The following project report may contain confidential data of certain organizations. It may only be used for academic purpose in teaching business administration at IIM Kozhikode. Confidentiality has to be maintained about the content. Publications and duplications of the content in any form are forbidden. For data privacy reasons the certain names have been masked.