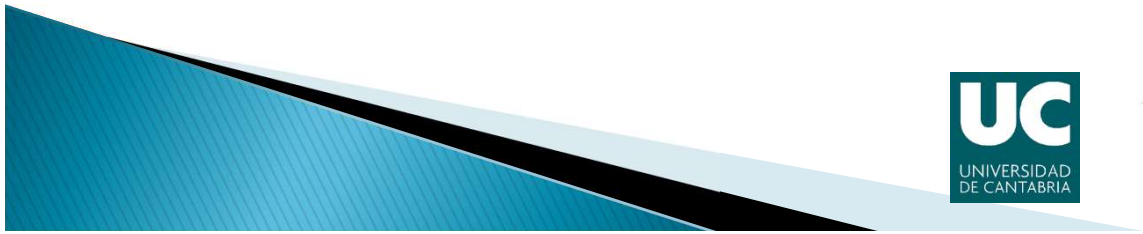
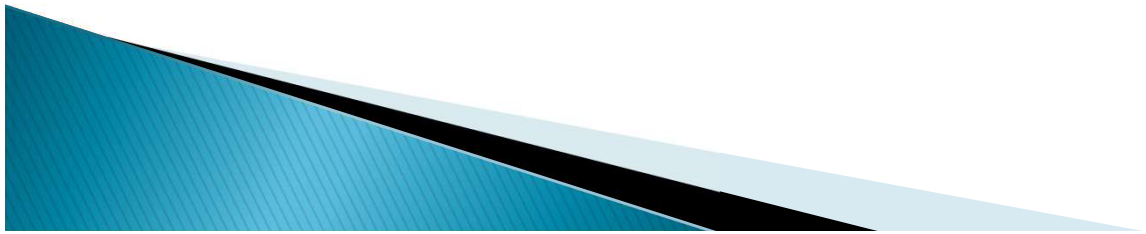
Data Science
Data Life Cycle
Data Management

- Computación Avanzada y e-Ciencia – IFCA-CSIC
  - Ibán Cabrillo Bartolomé
- Santander Enero del 2020



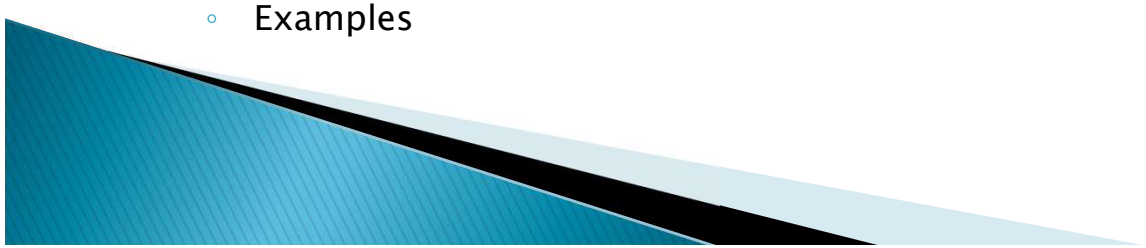
Summary

- Physical Storage Devices
- Network Storage Devices
- Data Storage
- Data Management
- Backup



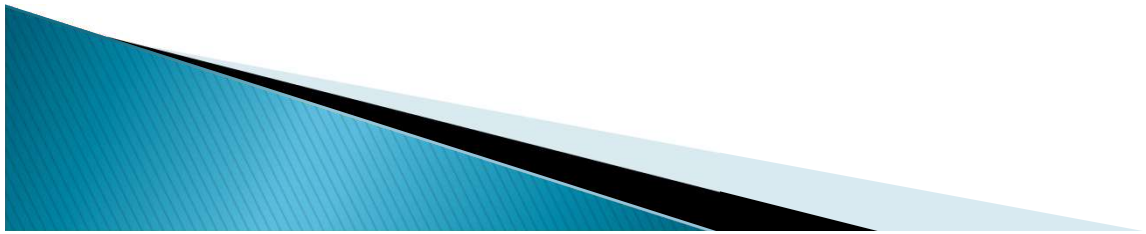
Schema

- Data Management
  - Data Access Modes
    - Sequential
    - Ramdom
  - Data Privacy
  - Data Integrity
    - Encrytion Algorithms
      - Symmetric and Asymmetric
      - Digital Signature
      - Digital Certificates
  - Data Security
    - Disk encryption
    - E-mail Encryption
    - Network Encryption
    - Harware protection
  - Data Life Cicle
  - Repositories
  - OAIS
  - DSA
  - Data Transfer/Login Services
    - Ftp
    - GRIDFtp
    - SSH
    - SCP
  - Examples



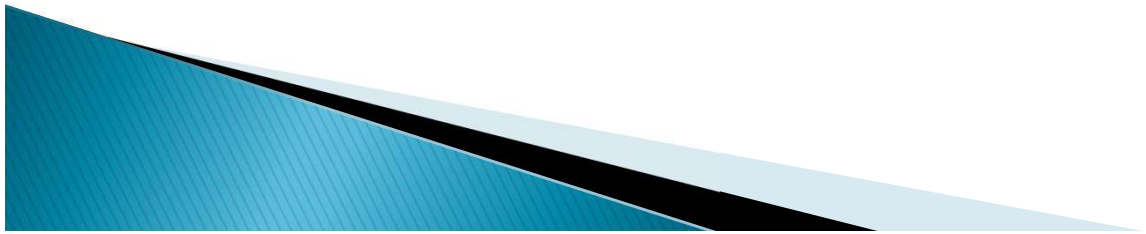
Data Management

- Data Resource Management is the development and execution of architectures, policies, practices and procedures that properly manage the full data lifecycle needs
- Data Access
  - Refers to software and activities related to storing, retrieving, or acting on data.
  - Data Access is simply the authorization you have to access different data files
    - Sequential Access
      - Accessing data sequentially is much faster than accessing it randomly because of the way in which the disk hardware works
      - workloads that have high I/O rates, consider using stripe sets because they add physical disks, increasing the system's ability to handle concurrent disk requests
      - Be carafull with disk fragmentation
    - Ramdom Access
      - Reading or Writing randomly involves a higher number of seek operations than does sequential reading
      - workloads that are predominantly random I/O, use a drive with faster seek time
      - For workloads of either random or sequential I/O, use drives with faster rotational speeds



Data Privacy

- Data Privacy
  - Article 18.4 of the Spanish constitution of 1978: "
The law will limit the use of information technology in order to guarantee honour, personal and family intimacy of citizens and all their rights."
  - Organisations or businesses collecting personal information are obliged to protect the data from unauthorised access or unauthorised alteration
  - Sensible Data
    - Criminal or justice proceedings
    - Healthcare records
    - Financial transactions
  - The challenge in data privacy is to share data while protecting personally identifiable information
    - Anonymice the personal Data
    - Avoid Data Access to unauthorized personnel

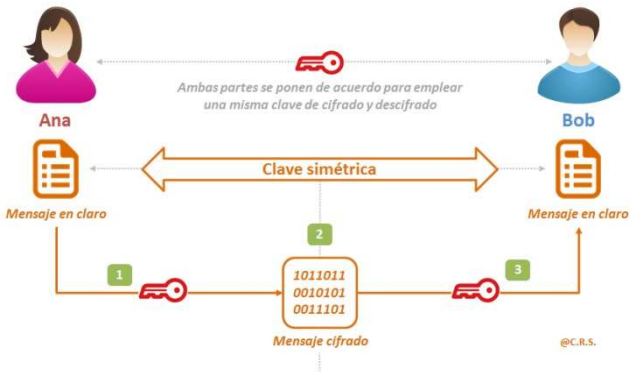


Data Integrity I

- Encryption
  - Cryptography is the science of mathematics that studies the information security and related actions, including encryption, authentication and access control
  - Entity is a user, program or machine
  - Credentials are data provides evidence of identity
  - Authentication is the verification of the identity of the entity
  - Authorization is the granting of privileges to an entity
  - Confidentiality is the encryption of information so that only the recipient can understand
  - Integrity is the assurance that the information has not been altered in the transaction
  - Non-repudiation is the inability to deny the authenticity

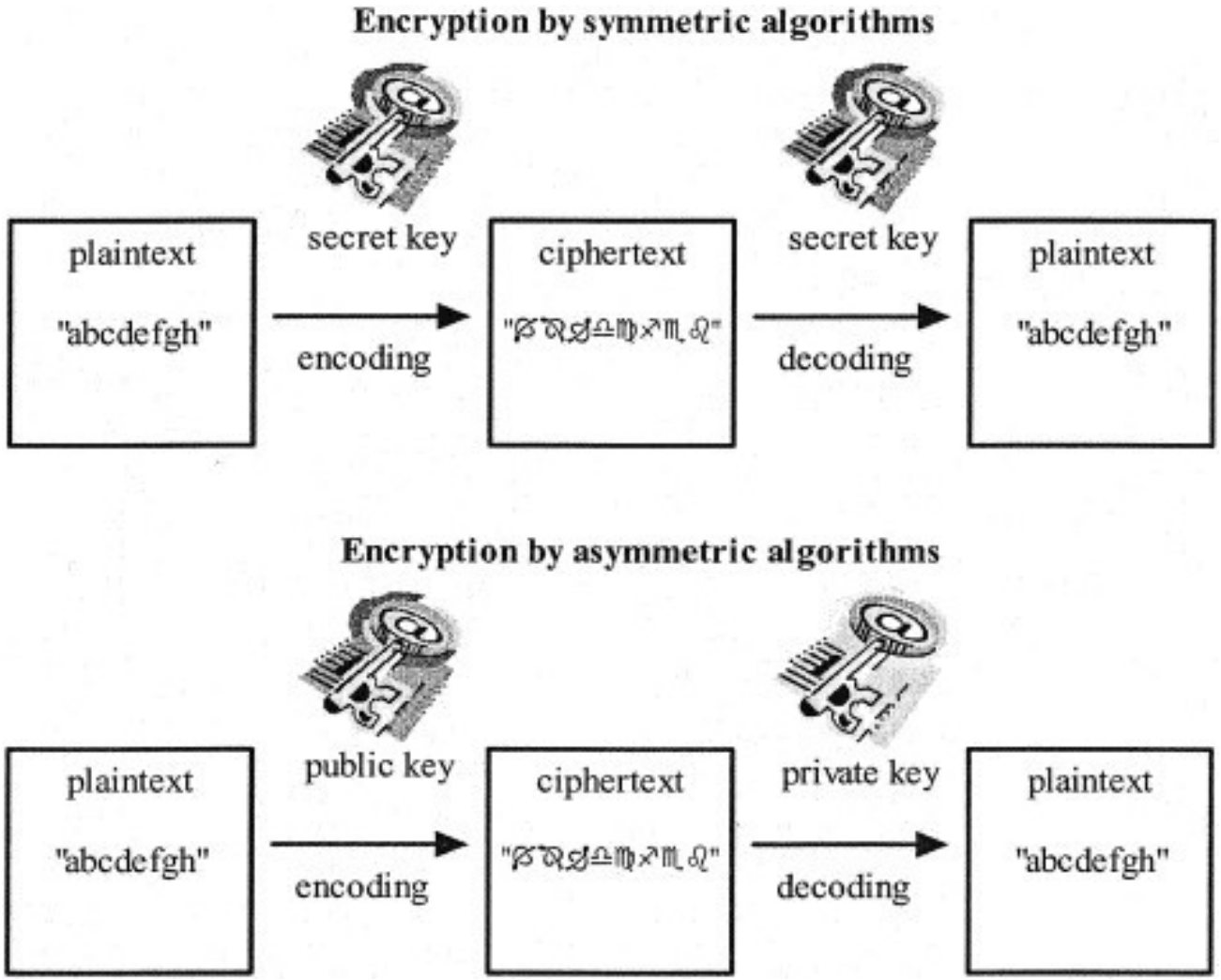
Data Integrity II

- Symmetric and Asymmetric Algorithms
  - Symmetric
    - Same key to encrypt and desencrypt
    - Fast algorithm
    - But how is the key distributed?
    - DES, 3DES, AES, Blowfish, kerberos



  - Asymmetric
    - Every user has two keys, plublic and private
    - Is not possible to obtain the private key from public one.
    - A mesage encrypted with one must be decrypted with the other one.
    - Is not necesary a secret change
    - Diffie-Helmann, RSA,DSA

Data Integrity III



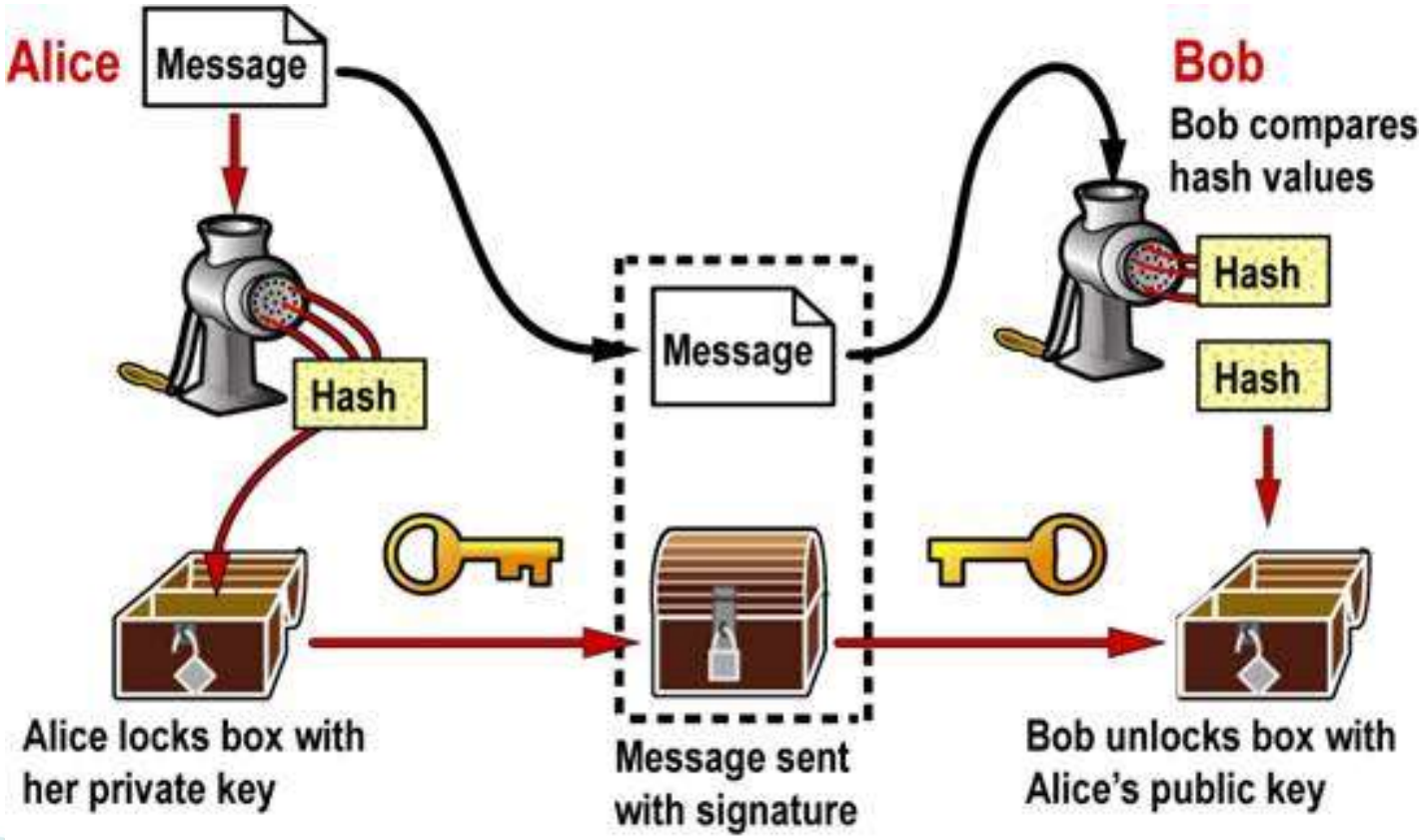
Data Integrity IV

- Digital Signature
  - A mathematical scheme for demonstrating the authenticity of a digital message or document
  - Employ a type of asymmetric cryptography
  - Hash
    - is any algorithm or subroutine that maps large data sets of variable length to smaller data sets of a fixed length.
    - The values returned by a hash function are called **hash values, hash codes, hash sums, checksums**
    - Is imposible to obtain data from hash output
    - Output leng is always the same for a given hash

Algorithm		output	Colissions	Performance
MD5		128	Found	255 MiB/s
SHA-0		160	Found	
SHA-1		169	theorical	153 MiB/s
SHA-2	*SHA-224* *SHA-256*	224 256	None	111 MiB/s
	SHA-384 SHA-512 SHA-512/224 SHA-512/256	384 512 224 256	None	99 MiB/s
SHA-3	224/256/384/512	1600	None	

Data Integrity V

- Digital signature is the result of calculating the hash of the message and the encrypting the hash with private key

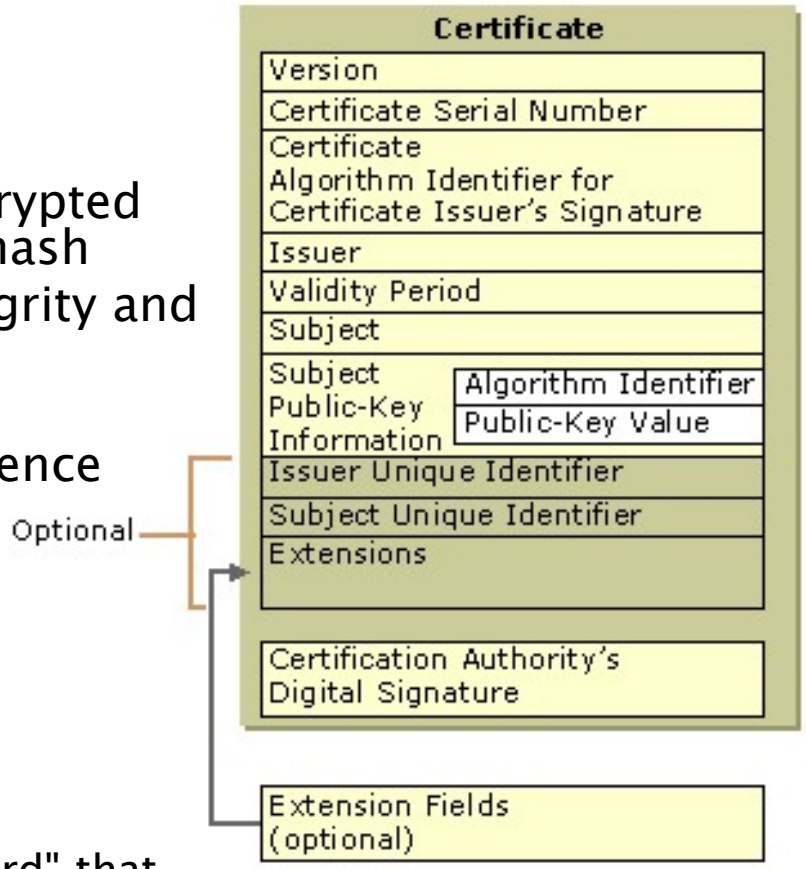


Data Integrity VI

- The receiver calculates the hash of the message without digital signature, decrypted with the public key and compares the hash
- if both hash are equal ensures the integrity and the principle of non-repudiation

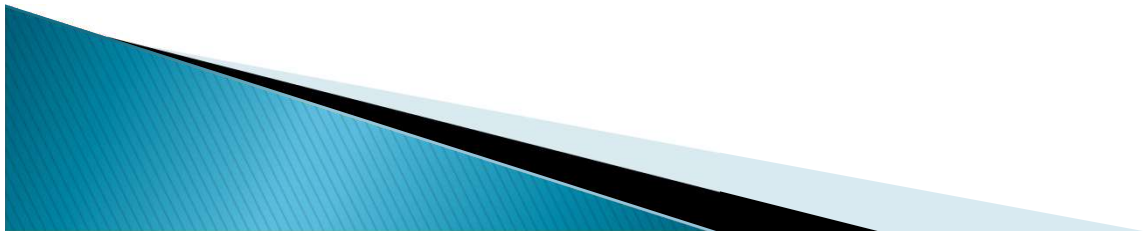
- A third party has to ensure correspondence between the public key and private
  - The CA (Certification Authority)
    - Issue digital certificates
    - Check applicants identity
    - Register authorities (RAs)

- Digital Certificates
  - A digital certificate is an electronic “ID card" that establishes your credentials when doing business or other transactions on the Web
  - contains your name, a serial number, expiration dates, a copy of the certificate holder's public key and the digital signature of the certificate-issuing authority so that a recipient can verify that the certificate is real



Data Security I

- Data Security
  - Protecting data from destructive forces and the unwanted actions of unauthorized users
  - Disk encryption
    - Technology which protects information by converting it into unreadable code that cannot be deciphered easily by unauthorized people
    - Whole Disk
      - Refers to the encryption of an entire physical or logical disk
      - encrypts the entire contents of a disk or volume and decrypts/encrypts it during use after a key has been given
      - not protect from situations like sending information over the network (e-mail, websites, etc)
        - PGP
        - Hard drives with integrated hardware encryption
        - EncFS
        - BitLocker

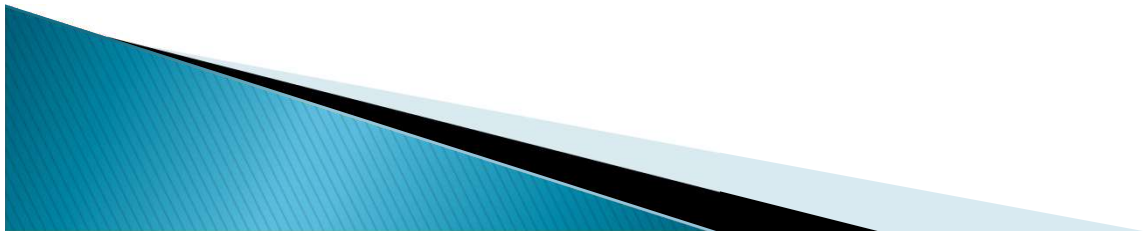


Data Security II

- **Single-user file/folder level**
  - Individual wishes to encrypt a single file or group of files
  - Many encryption programs have the ability to create an encrypted "virtual drive“
  - can protect against data disclosure on a lost or stolen computer
  - individual file can be encrypted and then sent as an e-mail attachment
    - TrueCrypt
    - PGP/GnuPG
    - EFS
- **Multi-user file/folder level**
  - Allowing multiple users to simultaneously access encrypted information
  - The encryption software must allow the use of either multiple keys or a shared key
    - PGP/GnuPG
    - EFS

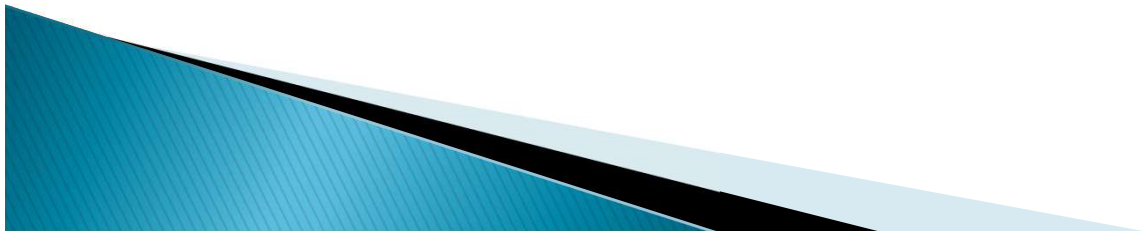
◦ **E-mail Encryption**

- encrypting just an attached file
  - Encrypting an attached file can be accomplished using any single-file encryption process
- encrypting an entire message
  - S/MIME requires users to have trusted certificates
  - PGP to encrypt e-mail requires installing software



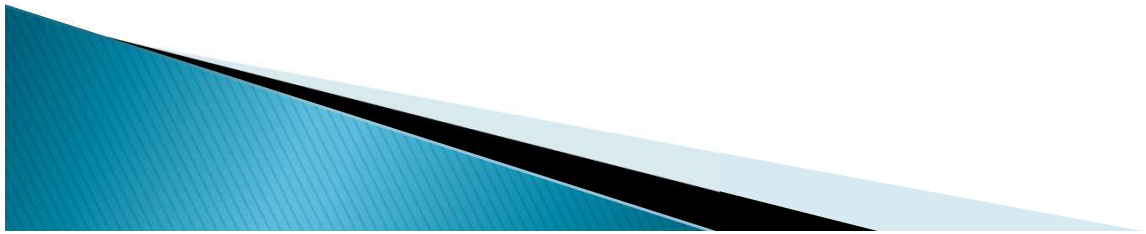
Data Security III

- Network Traffic Encryption
  - One of the most common, and important, uses of encryption
  - The most popular forms of this encryption is Secure Sockets Layer (SSL)/Transport Layer Security (TLS)
  - SSL/TLS can also be used for "tunneling" to encrypt other forms of network transmission
  - IP Security (IPSec)
  - Wireless networks
    - Wired Equivalent Privacy (WEP) Deprecated
    - WiFI Protected Access (WPA), replaced by WPA2 (soon will be replaced by WPA3)
- Harware protection
  - A hardware device allows a user to log in, log out and to set different privilege levels by doing manual actions
  - Biometric technology to prevent malicious users from logging in, logging out, and changing privilege levels
  - With hardware based protection, software cannot manipulate the user privilege levels, it is impossible for a hacker or a malicious program to gain access to secure data protected by hardware or perform unauthorized privileged operations



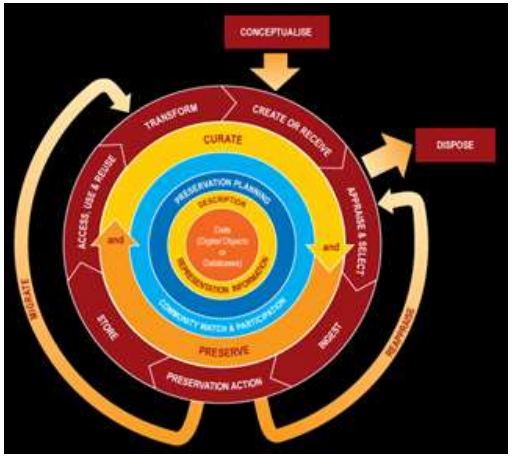
Data Security IV

- Backups
  - Aways do Backup!!
  - If it is possible more that one and in diferent physical locations
- Data masking
  - Obscuring/masking sensitive data
- Data erasure
  - Software overwriting that completely destroys all electronic data residing on a hard drive or other digital media
  - Harware that is able to delete by electromagnetic fields the data stored on magnetic device



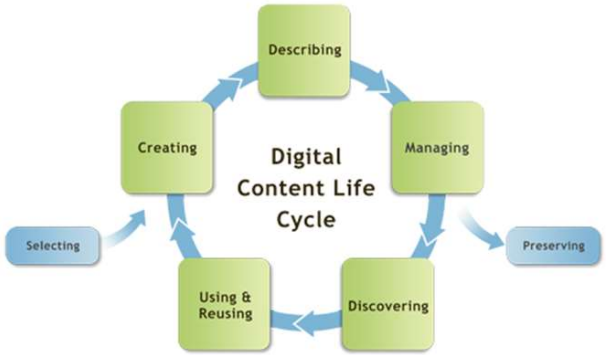
Data Life Cicle I

- Data, any information in binary digital form, is at the centre of the Curation Lifecycle. This includes:
  - Digital Objects: simple digital objects
    - text files, image files or sound files, along with their related identifiers and metadata)
    - complex digital objects (discrete digital objects made by combining a number of other digital objects, such as websites)
  - Databases: structured collections of records or data stored in a computer system



Data Life Cicle II

- Full Lifecycle Actions
  - Description and Representation Information
    - Assign administrative, descriptive, technical, structural and preservation metadata, using appropriate standards, to ensure adequate description and control over the long-term.
    - Collect and assign representation information required to understand and render both the digital material and the associated metadata.
  - Preservation Planning
    - Plan for preservation throughout the curation lifecycle of digital material.
    - Plans for management and administration of all curation lifecycle actions.
  - Community Watch and Participation
    - Maintain a watch on appropriate community activities, and participate in the development of shared standards, tools and suitable software.
  - Curate and Preserve
    - Be aware of, and undertake management and administrative actions planned to promote curation and preservation throughout the curation lifecycle



Data Life Cicle III

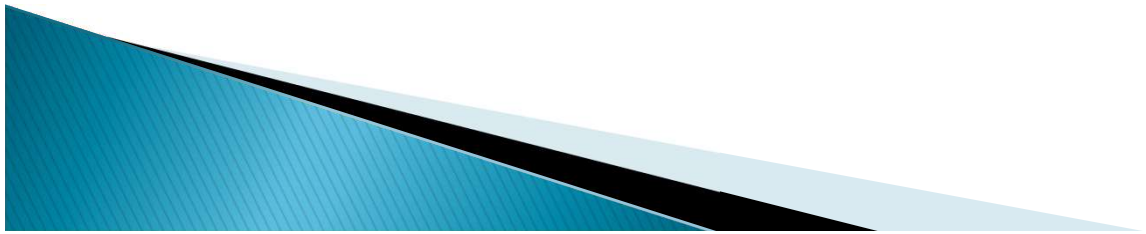
- Sequential Actions
  - Conceptualise
    - Conceive and plan the creation of data, including capture method and storage options.
  - Create or Receive
    - Create data including administrative, descriptive, structural and technical metadata. Preservation metadata may also be added at the time of creation.
  - Appraise and Select
    - Evaluate data and select for long-term curation and preservation. Adhere to documented guidance, policies or legal requirements.
  - Ingest
    - Transfer data to an archive, repository, data centre or other custodian. Adhere to documented guidance, policies or legal requirements.
  - Preservation Action
    - Undertake actions to ensure long-term preservation and retention of the authoritative nature of data.
      - Data remains authentic, reliable and usable while maintaining its integrity.
      - Data cleaning, validation, assigning preservation metadata,
      - Assigning representation information and ensuring acceptable data structures or file formats.
  - Store
    - Store the data in a secure manner adhering to relevant standards.
  - Access, Use and Reuse
    - Ensure that data is accessible to both designated users and reusers, on a day-to-day basis. This may be in the form of publicly available published information. Robust access controls and authentication procedures may be applicable.
  - Transform
    - Create new data from the original,
      - by migration into a different format
      - by creating a subset, by selection or query, to create derived results, perhaps for publication

Data Life Cicle IV

- Occasional Actions
  - Dispose
    - Data, which has not been selected for long-term curation and preservation in accordance with documented policies, guidance or legal requirements.
    - Data may be transferred to another archive, repository, data centre or other custodian.
    - Data is destroyed. The data's nature may, for legal reasons, necessitate secure destruction.
  - Reappraise
    - Return data which fails validation procedures for further appraisal and re-selection.
  - Migrate
    - Migrate data to a different format.
    - Ensure the data's immunity from hardware or software obsolescence.

Repositories I

- Data repository is a somewhat general term used to refer to a destination designated for data storage
- IT experts use the term more specifically to refer to a particular kind of setup within an overall IT structure
  - A group of databases and Software, where an enterprise or organization has chosen to keep various kinds of data.
- Each institution will have its own unique approach for establishing a repository that reflects their specific context and community
  - 1. Making the business case
    - Focusing on the benefits to the institution
    - Increase the usage and impact of its research effort
    - Maximize the visibility of its outputs and provide a management information
  - 2. Defining the purpose of the repository
    - Concentrating on supporting digital publishing initiatives
    - Aim for the preservation of content.
  - 3. Defining repository services



Repositories II

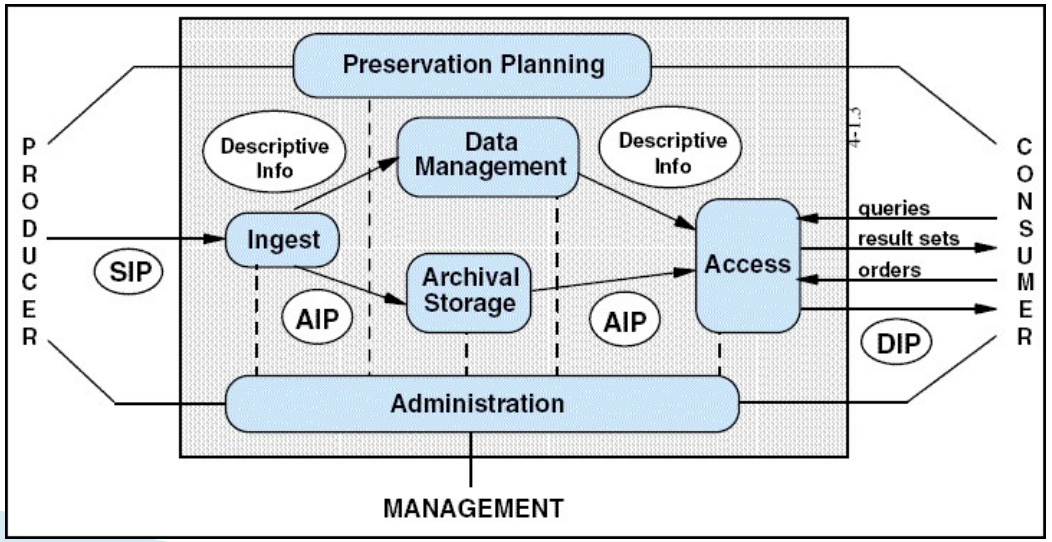
- 4. Choosing repository software
  - Commercial Software
    - Pay for the software and, optionally, any additional subscription or consulting fees. You own the use of the software and, with a subscription, get software upgrades
  - Open Source Software
    - Software is free to download, but usually requires some level of expertise to implement and maintain
    - CDSware, DSpace, EPrints, Fedora, Greenstone.
- 5. Developing repository policies
  - Collection
    - What types of materials will be accepted into the repository?
    - Whose work can be included in the repository?
    - Criteria for determining what constitutes a collection in the repository.
    - How will the repository be ?
    - Who will deposit content? (library staff or authors)
  - Management
    - General rights and responsibilities of libraries and those who create collections of digital content.
    - What types of metadata will be used.
    - What preservation activities will be undertaken.
  - Access
    - Privacy policy for registered users of the system.
    - Will the repository restrict access to content if requested by author?
    - Will the repository enable embargo periods for content?

Repositories III

- 6. Staffing
  - Repository Manager:manages the 'human' side of the repository.
    - Content policies, advocacy.
    - User training Repository
  - Repository Administrator: manages the technical implementation, customisation and management of repository software
    - Manages metadata fields and quality,
    - Creates usage reports and tracks the preservation issues.
- 7. Setting up communities
  - Enables repository staff to focus on seeding the repository with content,
  - Testing the software
  - Check the procedures and policies
  - Early adopter communities
- 8. Marketing the repository
  - Profiling Strategy
    - discuss the general benefits
  - Pull Strategy
    - Encourage authors to deposit their work in the repository
  - Push Strategy
    - Demonstrated the positive effects of the repository once the material has been deposited
    - Sssisting authors with their deposits
  - Consultation Strategy
    - Developing the repository to meet their needs (feedback)

OAIS

- Open Archival Information System (OAIS)
  - The Open Archive Information System (oais) reference model is an iso standard. Developed by the Consultative Committee for Space Data Systems (ccsds).
    - The model serves to define the processes for effective, long-term preservation of information
    - Ensuring access to information
    - Provides a common language for describing these objects and has been widely accepted in the digital preservation community



DSA I

- Data Seal of Approval (DSA)
  - Guideline for the application and verification of quali aspects with regard to creation, storage and (re-)use digital research data in the social sciences and humanities
    - Gives researchers the assurance that their research results will be stored in a reliable manner and can be reused
    - Provides research sponsors with the guarantee that research results will remain available for reuse
    - Enables researchers, in a reliable manner, to assess the repository where researchdata are held.
    - Allows data repositories to archive and distribute research data efficiently
  - Three stakeholders
    - The *data producer* is responsible for the quality of the digital research data.
    - The *data repository* is responsible for the quality of storage and availability of the data: data management.
    - The *data consumer* is responsible for the quality of use of the digital research data.

DSA II

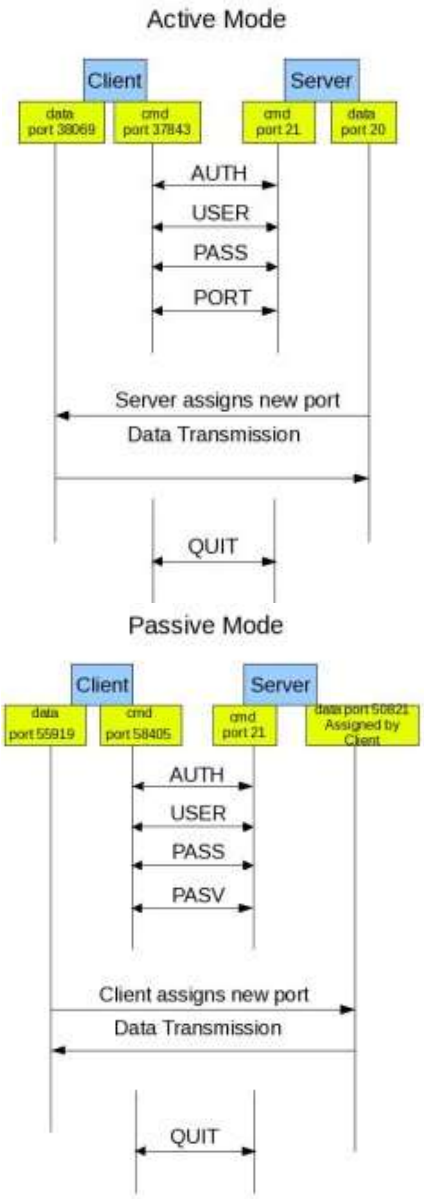
- 1. The *data producer* deposits the research data in a data repository with sufficient information for others to assess the scientific and scholarly quality of the research data and compliance with disciplinary and ethical norms.
- 2. The *data producer* provides the research data in formats recommended by the data repository.
- 3. The *data producer* provides the research data together with the metadata requested by the data repository.
- 4. The *data repository* has an explicit mission in the area of digital archiving and
- promulgates it.
- 5. The *data repository* uses due diligence to ensure compliance with legal regulations
- and contracts.
- 6. The *data repository* applies documented processes and procedures for managing
- data storage.
- 7. The *data repository* has a plan for long-term preservation of its digital assets.
- 8. Archiving takes place according to explicit workflows across the data life cycle.

DSA III

- 9. The *data repository* assumes responsibility from the data producers for access to and availability of the digital objects.
- 10. The *data repository* enables the users to utilize the research data and refer to
- them.
- 11. The *data repository* ensures the integrity of the digital objects and the metadata.
- 12. The *data repository* ensures the authenticity of the digital objects and the metadata.
- 13. The technical infrastructure explicitly supports the tasks and functions described in internationally accepted archival standards like OAIS.
- 14. The *data consumer* must comply with access regulations set by the data repository.
- 15. The *data consumer* conforms to and agrees with any codes of conduct that are
- generally accepted in higher education and research for the exchange and proper use of knowledge and information.
- 16. The *data consumer* respects the applicable licences of the data repository regarding the use of the research data.

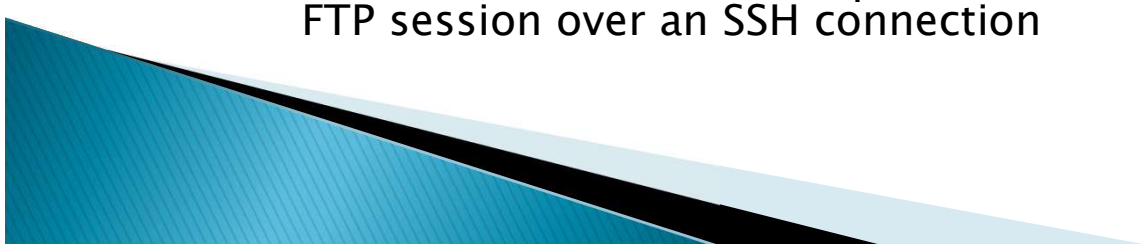
Data Transfers service

- File Transfer Protocol (FTP)
  - Standard network protocol used to transfer files from one host or to another host over a TCP-based network, such as the Internet.
  - client-server architecture and uses separate control and data connections between the client and the server
  - FTP is often secured with SSL/TLS ("FTPS")
  - FTP may run in *active* or *passive* mode, which determines how the data connection is established
    - Active Mode
      - The client creates a TCP control connection to the server and sends the server the client's IP address
      - Arbitrary client port number
    - Pasive Mode
      - the client uses the control connection to send a PASV command to the server and then receives a server IP address
      - Recives the server port number from the server



Data Transfers service

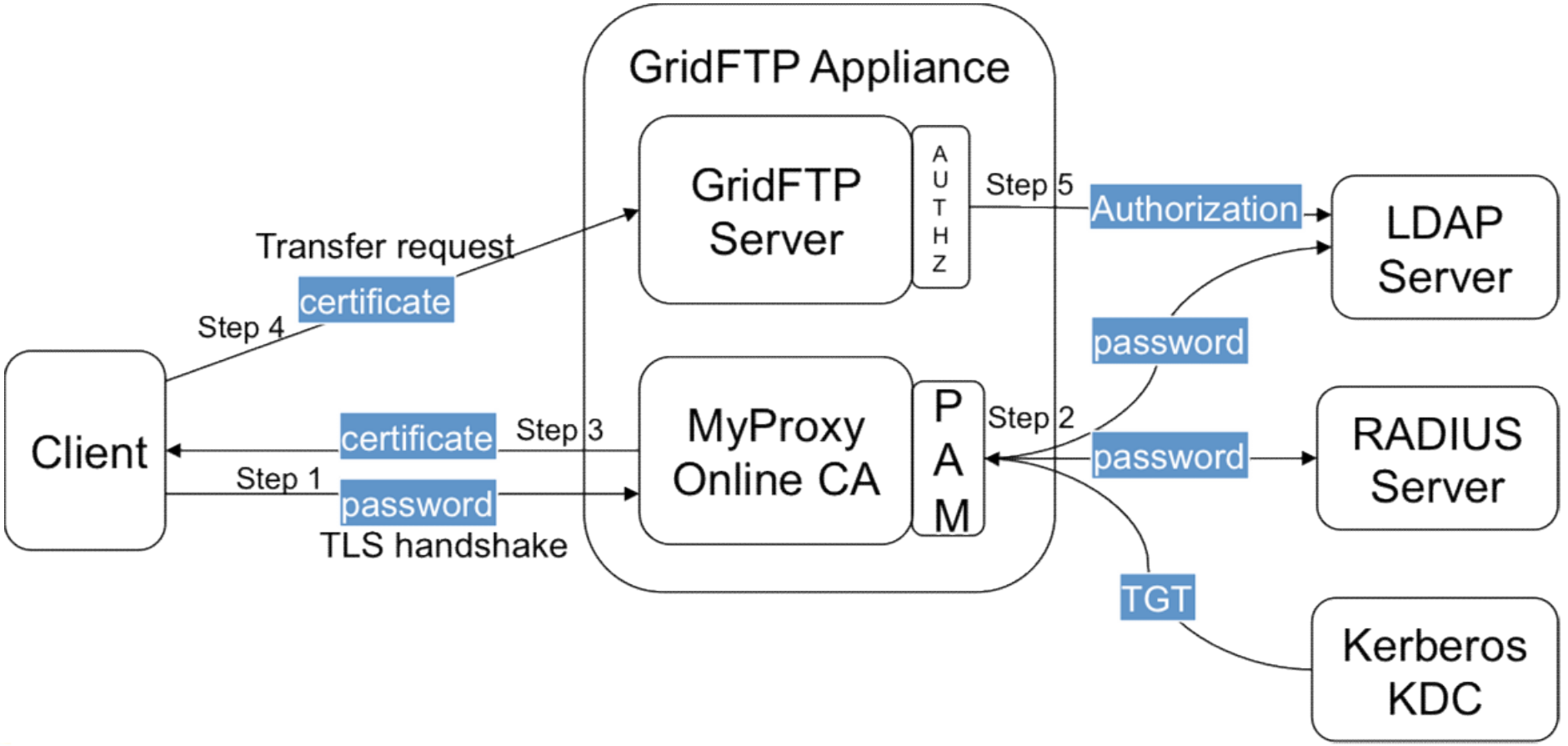
- Three Data Transfer modes
  - Stream mode: Data is sent as a continuous stream, relieving FTP from doing any processing
  - Block mode: FTP breaks the data into several blocks and then passes it on to TCP
  - Compressed mode: Data is compressed using a single algorithm
- FTP login utilizes a normal usernames and password scheme for granting access
  - The username is sent to the server using the USER command
  - The password is sent using the PASS command
  - A host that provides an FTP service may provide anonymous FTP access
    - 'anonymous' as user when prompted for user name
    - E-mail address instead of password
- FTP was not designed to be a secure protocol
  - FTPS is an extension to the FTP standard that allows clients to request that the FTP session be encrypted
  - FTP over SSH refers to the practice of tunneling a normal FTP session over an SSH connection



Data Transfer/Login services

- GridFTP
  - GridFTP is a high-performance, secure, reliable data transfer protocol optimized for high-bandwidth wide-area networks
  - Based on FTP
  - GridFTP is the answer to the problem of incompatibility between storage and access systems
  - Provides authentication and encryption to file transfers, with user specified levels of confidentiality and data integrity
  - GridFTP achieves much greater use of bandwidth by allowing multiple simultaneous TCP streams
    - Files can be downloaded in pieces simultaneously from multiple sources
    - separate parallel streams from the same source
    - Transfers can also be automatically restarted

Data Transfer/Login services

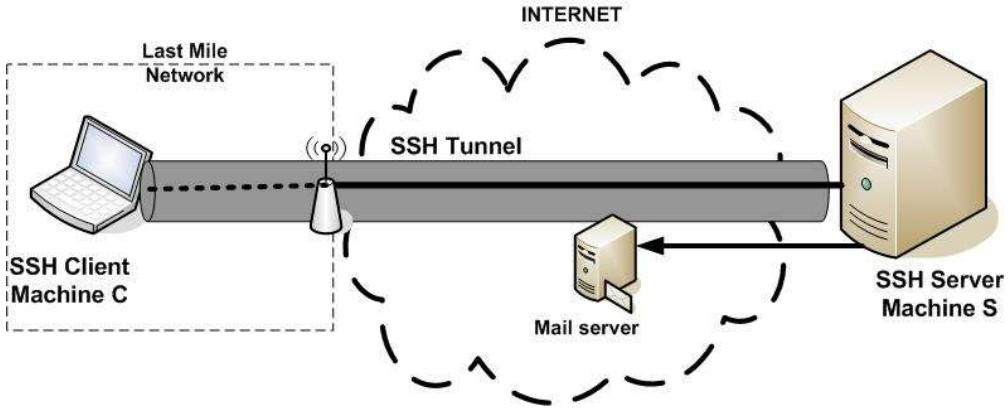


Data Transfer/login services

- Secure Shell (SSH)
  - Cryptographic network protocol for secure data communication, remote shell services or command execution and other secure network services between two networked computers that connects
  - Two major versions that are referred to as SSH-1 and SSH-2
  - The encryption used by SSH is intended to provide confidentiality and integrity of data over an unsecured network, such as the Internet.
  - Use the public-private key pair.
    - The public key is placed on all computers that must allow access to the owner of the matching private key
    - The owner keeps the private key secretç
  - The standard TCP port 22 has been assigned for contacting SSH servers
    - Uses
      - log into a remote machine and execute commands

        #ssh *user*@remote*host*

      - supports tunneling, forwarding TCP ports and X11 connections
      - In combination with rsync to back up, copy and mirror files efficiently and securely



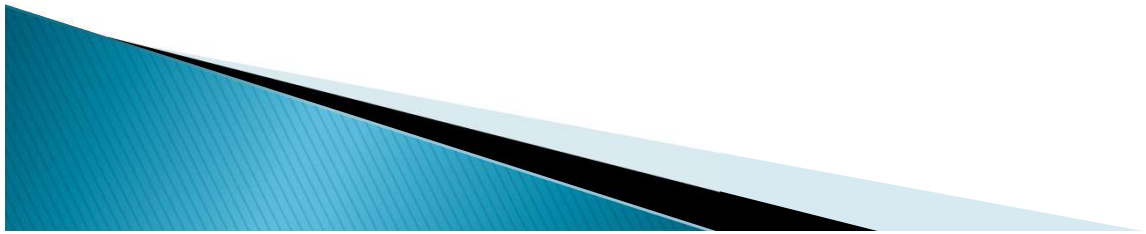
Data Transfer/login services

- Secure Copy (SCP)
  - is a network protocol, based on the BSD RCP protocol, which supports file transfers between hosts on a network
  - SCP uses Secure Shell (SSH) for data transfer and utilizes the same mechanisms for authentication
  - Ensure the authenticity and confidentiality of the data in transit
  - SCP runs over TCP port 22 by default
  - the most widely used SCP program is the command line scp program, which is provided in most SSH implementations
    - Coping a file to remote host:

      #scp *SourceFile user*@remote*host*:*directory*/*TargetFile*

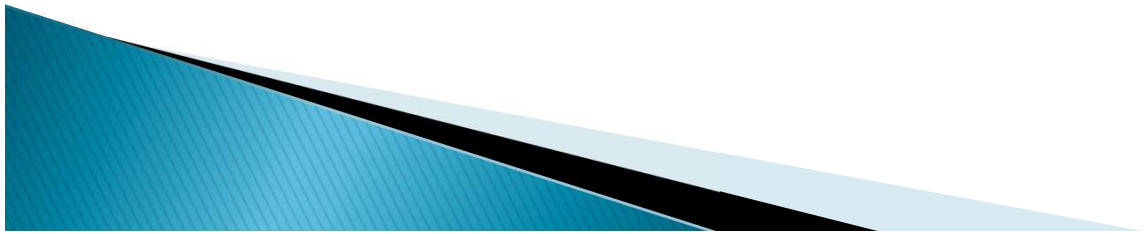
    - Coping a file from remote hots

      #scp *user*@remote*host*:*directory*/*SourceFile directory*/*TargetFile*



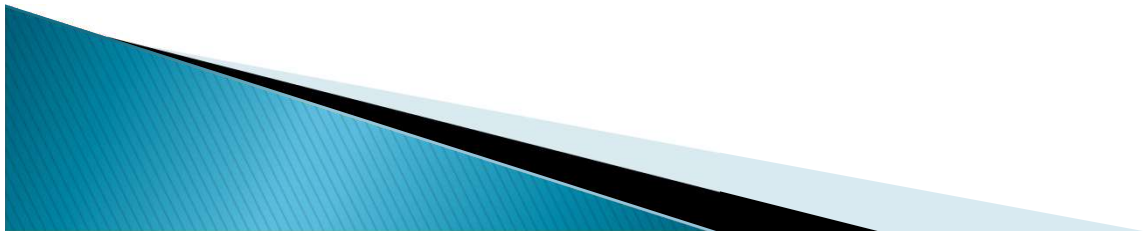
Examples

- Online Hash Calculator
- Using PGP/GnuPGP
- Directory Encription



Example pgp uses

- Download and Install pgp tool (https://gnupg.org/download/)
  - **sudo apt install gpg**
  - List keys (if there is no one we have to generate it)
    - **gpg --list-keys**
  - Send/share pub key to some keyserver
    - **gpg -keyserver pgp.mit.edu -send-key "XXXXXXXX"**
      - **pgp.rediris.es**
      - **pgp.surfnet.nl**
      - **pgp.uni-mainz.de**
  - Import public key from keyserver
    - **gpg -keyserver pgp.mit.edu -search "e-mail"**
  - Encryt a file with user's pub key
    - **gpg --output file.gpg --encrypt --recipient "e-mail" file**
  - Decrypt a file using the private key
    - **gpg --output doc --decrypt doc.gpg**
  - **Signing a files**
    - **gpg --clearsign file**
    - **gpg -output file.sig -detach-sig file**
    - **gpg --verify file.sig file**



Example Encryting a Directory

- Install encfs (It's not totally safe)
  - #sudo apt-get install encfs
- Create directores to encrypt/decrypt
  - #mkdir -p ~/encryted
  - #mkdir -p ~/decrypted
- Asign encfs directory relatioship
  - #encfs ~/encripted ~/decrypted
  - #Strong passwd
- Umount Directory to encrypt data
  - #fuserumount -t ~/decripted
- Check directories
  - #ls -la ~/decripted
  - #ls -la ~/encrypted
- This is a good option to encrypt out cloud data directory.
  - Dropbox
  - Owncloud