ICQB

Introduction to Computational & Quantitative Biology (G4120)
Spring 2017
Oliver Jovanovic, Ph.D.
Columbia University
Department of Microbiology & Immunology

The Internet in 1969

The ARPANET in December 1969
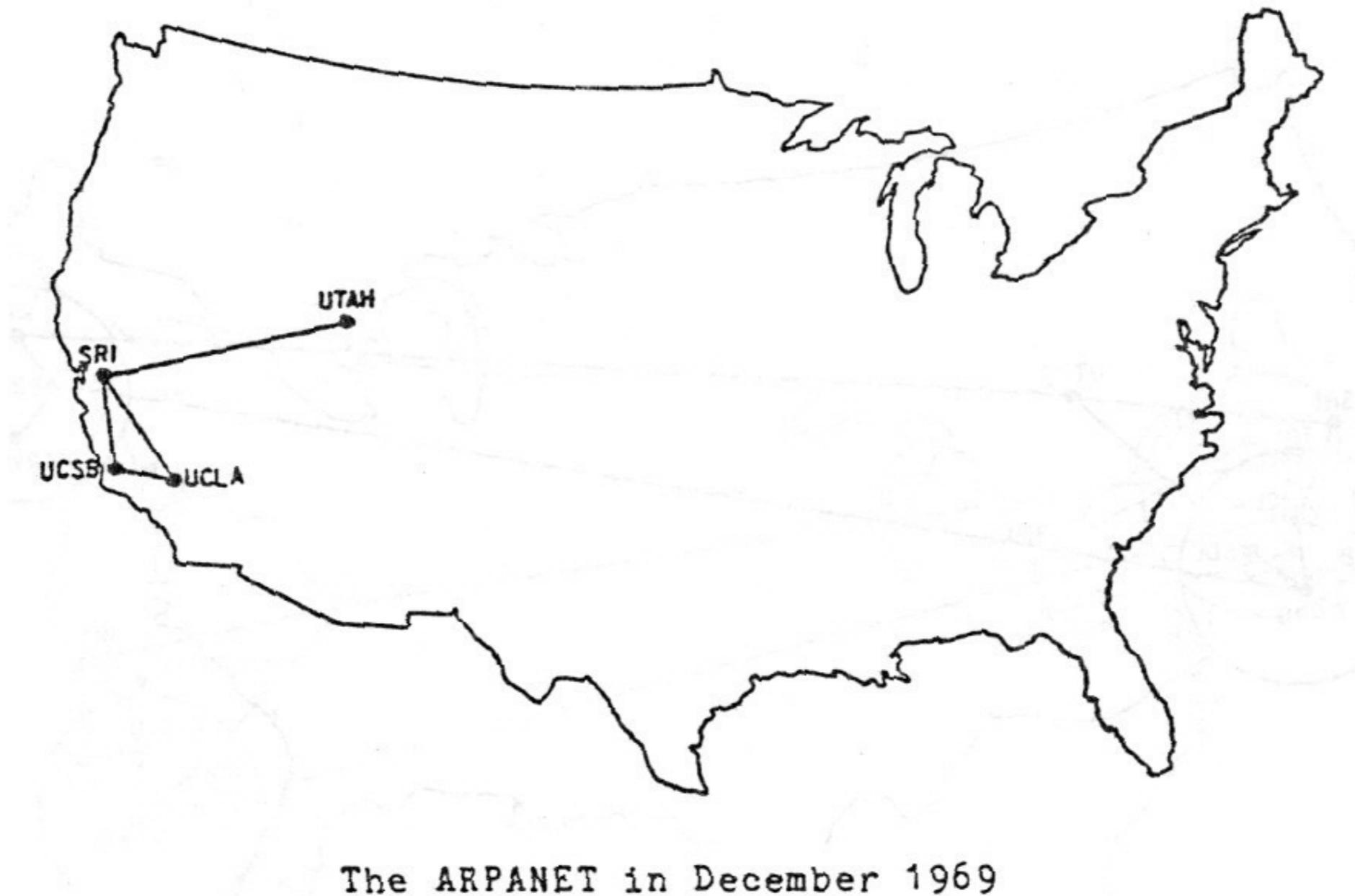
Growth of the Internet
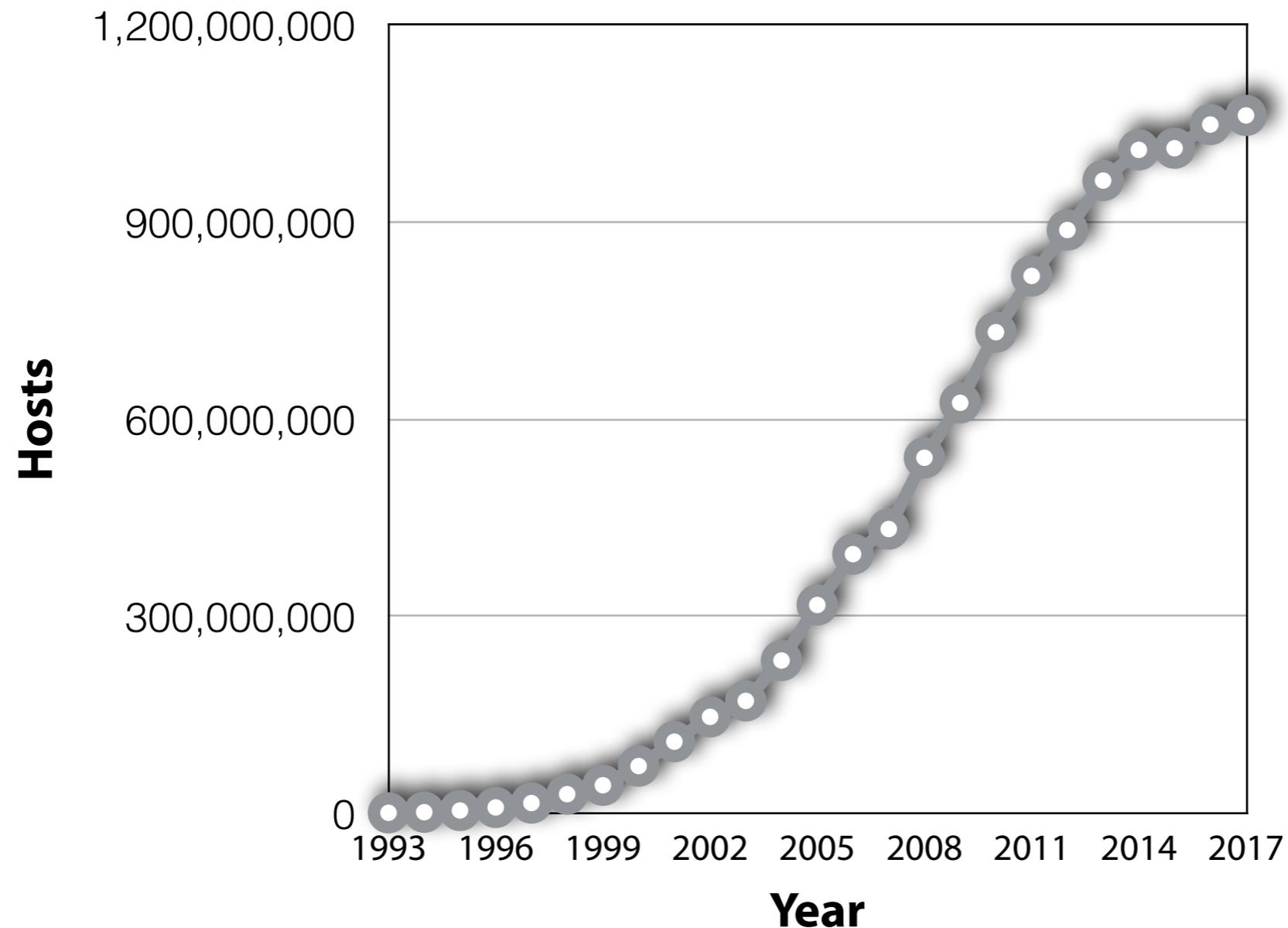
From 1.3 million Internet hosts in 1993 to over 1 billion in 2017

The Modern Internet
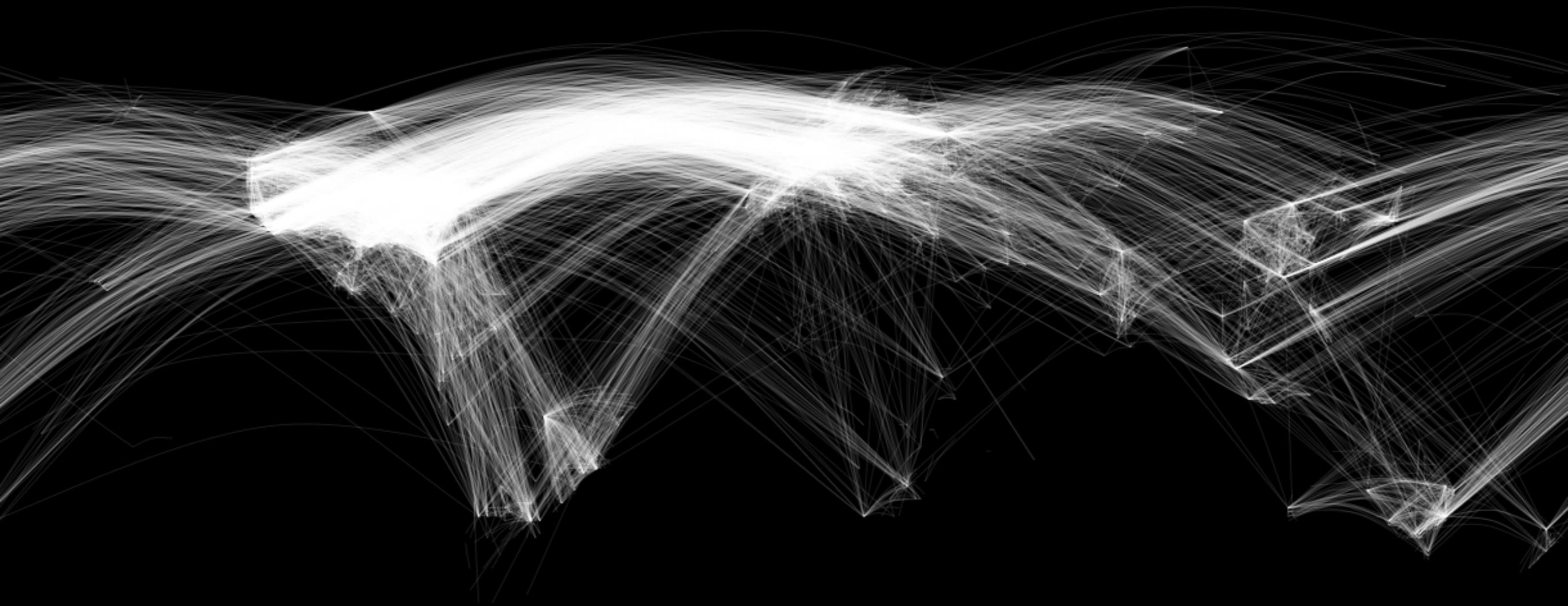
Lecture 3: Databases and Internet Resources
February 13, 2017

Internet Protocols

TCP/IP (Transmission Control Protocol/Internet Protocol)
A multilayered protocol architecture that allows for the reliable transmission of data over networks. TCP provides connection oriented streaming with error detection and correction between two computers. IP governs the delivery of datagrams, packets of data containing the IP addressing information needed to switch them from one network to another until they arrive at their final destination. The current version is IPv4, to be succeeded by IPv6.

UDP (User Datagram Protocol)
Provides low overhead connectionless datagram delivery.

DHCP (Dynamic Host Configuration Protocol)
Automates the configuration of computers that use TCP/IP by automatically assigning Internet addressing information from a network DHCP server.

HTTP (HyperText Transfer Protocol) and HTTPS
The protocol behind the World Wide Web (WWW). The secure form is **HTTPS**, which provides an encrypted **SSL/TLS** connection that depends on a trusted certificate authority and public key infrastructure.

DNS (Domain Name Server)
A hierarchical naming system used to locate servers on the WWW by name instead of numerical IP address.

URL (Uniform Resource Locator)
Used to locate resources on the WWW. Various prefixes, usually HTTP, so **http://www.columbia.edu/index.html** but could be **ftp://ftp.ncbi.nlm.nih.gov** or **file://Macintosh%20HD/Documents/Read%20Me.pdf** (note, a blank space in a URL " " is replaced by a "**%20**"). The DNS part is not case sensitive, but the rest of the URL can be.

File Transfer Protocols

FTP (File Transfer Protocol)
Allows you to transfer files to or from a remote machine running FTP. Usually anonymous, but insecure. Use Fetch (for Mac, free to students at **fetchsoftworks.com/Licensing/edustore.application.html**) or FileZilla (Mac or Win, free at **filezilla-project.org**).

SFTP (Secure File Transfer Protocol)
Allows you to securely transfer files with the data encrypted. Use Fugu (for Mac, free at: **sourceforge.net/projects/fugussh/files**) or WinSCP (for Win, free at **winscp.net**) or FileZilla.

AFP (Apple File Protocol)
Allows you to transfer files to or from a Macintosh.

SMB (Server Message Block)
Allows you to transfer files to or from a PC.

SMTP (Simple Mail Transfer Protocol)
Used by email systems and clients.

SSH (Secure Shell)
Can use to securely login remotely or forward outgoing email. Can transfer files securely using secure copy (SCP) or FTP over an SSH tunnel.

Internet Addressing

IP Address (IPv4)
An IP address is a 32 bit number, written in the form of four decimal numbers in the range 0-255 separated by dots (e.g. **128.59.48.24**). Columbia University Medical Center (CUMC) public IP addresses will always have the format **156.111.x.x** or **156.145.x.x**. A private address, such as a printer or computer on the private VLAN, will have the format **10.x.x.x**.

Subnet Mask
A subnet mask allows for defining a local network, called a subnet, within a larger network. CUMC subnet masks are always in the format **255.255.255.0**.

Router
A device that routes packets of data between networks. A router sits between your computer and local area network and the networks beyond it. CUMC router IP addresses generally match the public address but end in a **.1**, so **156.111.x.1** or **156.145.x.1**.

DNS (Domain Name Server)
These specialized servers automatically translate an easy to remember domain name (e.g. **microbiology.columbia.edu**) into the appropriate IP address (e.g. **156.111.98.150**). CUMC DNS IP addresses are **10.70.235.40** and **10.168.1.20**.

Ethernet Address
A unique 48 bit number, usually written in the form of 12 hexadecimal digits in six groups of two digits each separated by colons (e.g. **00:03:93:bc:3c:18**), which is assigned to every piece of network hardware, including Ethernet cards and AirPort cards. It is also called a MAC (media access control) address. When registering a wired device, use the Ethernet MAC, not the wireless MAC.

IPv6
Internet Protocol version 6 will eventually replace 32 bit IPv4 addresses with 128 bit IPv6 addresses, to address the growing problem of IPv4 address exhaustion. An IPv6 address is represented by 32 hexadecimal digits in eight groups of four digits each separated by colons. Leading zeros can be omitted, groups of all zeros can be represented by a single zero, and the leftmost longest run of all zeros can be replaced by a double colon (e.g. **::ffff:803b:3018** instead of **0000:0000:0000:0000:0000:ffff:803b:3018**).

CUMC Internet

The CUMC campus network has redundant core routers linked to a router in each building. Each floor of a building then has its own router. The campus has its own 40 Gbps WAN with redundant 10 Gbps fiber optic links to the main campus and other sites, as well as partial wireless network coverage.

The CUMC network is walled off from the rest of the Internet by a firewall and is centrally administered by CUMC IT. See **http:// cumc.columbia.edu/it/** for details. All network traffic is monitored for threats and suspicious activity, including, but not limited to potential HIPPA, HITECH, PHI or PII violations. Third parties can track copyright violations to campus computers, resulting in fines and potentially loss of network access.

Computers personally owned by students must be registered through the Bradford Network Registration System at **https://sentry-mc.cuc.columbia.edu** and scanned to ensure they are malware free and have up to date operating systems. To register a lab owned computer for access, see **http://cumc.columbia.edu/it/getting_started/wired.html** for details, and use the New IP Request form. For wired access, you will need to provide your computer's Ethernet adapter hardware address, also known as Media Access Control (MAC) address. This is a 12 digit hexadecimal number (e.g. **00:00:af:a0:b1:89**). If you have problems getting connected, try calling the CUMC computer help line at **212-305-HELP** or see the help desk on the second floor of the HHSC library.

CUMC Network Settings

In general, you will want to use **DHCP** to acquire all your network setting information automatically.

IP Address: 156.111.x.x or 156.145.x.x or 10.x.x.x
Subnet Mask: 255.255.255.0
Router: 156.111.x.1 or 156.145.x.1 or 10.x.x.1

DNS Servers
10.70.235.40
10.168.1.20
(Please note that these recently changed, the old ones were 156.111.60.150 and 156.111.70.150 and no longer work).

Internet Resources

National Center for Biotechnology Information (NCBI)

• PubMed, PubMed Central, Books and other reference material
• GenBank, RefSeq, CDD, MMDB and other sequence and structure databases
• Prokaryotic genome data and browsers (over 23,000 microbial, 7,000 virus and 8,000 plasmid)
• Eukaryotic genome data and browsers (over 4,000 complete genomes, also in progress, maps, partial sequences)
• BLAST, PSI-BLAST, Primer-BLAST, DELTA-BLAST and VAST search tools, Cn3D visualization tool
http://www.ncbi.nlm.nih.gov/

Ensembl, EMBL-EBI and Biocatalogue

Ensembl provides access to genome data and browsers, including vertebrates (human, chimp, mouse, rat, etc.), metazoa (*C. elegans, D. melanoganster, A. gambia,* etc.), plants, fungi, protists and bacteria. EMBL-EBI provides database search and sequence analysis tools. Biocatalogue provides a biological web services directory.
http://www.ensembl.org/ and **http://www.ebi.ac.uk/** and **http://biocatalogue.org/**

UCSC Genome Bioinformatics and ENCODE

Genome data and browsers including mammals (human, chimp, gorilla, rhesus, mouse, rat, dog, cat, horse, cow and platypus among others), vertebrates (chicken, *X. tropicalis,* zebrafish, elephant shark and fugu among others), deutrosomes (lancelet, *C. intestinalis* and *S. purpuratus*) insects (*D. melanogaster, D. simulans* and *A. gambiae* among others), nematodes (*C. elegans* and *C. briggsae* among others), virus (Ebola) and other (*S. cerevisiae*). The Encyclopedia of DNA Elements (ENCODE) provides a database of functional elements in the human and mouse genomes and now seeks to characterize them in healthy and diseased human cells.
http://genome.ucsc.edu/ and **http://www.encodeproject.org**

Protein Data Bank (PDB) and Expert Protein Analysis System (ExPASy)

PDB is a worldwide repository for 3D protein structure data and biochemical information. ExPASy is a bioinformatics resource portal with links to useful tools, software and references for genomics, proteomics and systems biology.
http://www.rcsb.org/ and **http://www.expasy.org/**

Software Resources

European Molecular Biology Open Software Suite (EMBOSS)

http://emboss.sourceforge.net/

TIGR/J. Craig Venter Institute Open Source Software Tools

http://www.jcvi.org/cms/research/software/

SourceForge

http://sourceforge.net

GitHub

http://github.com

The Bio Projects

www.bioconductor.org (R), www.biojava.com (Java), www.bioperl.org (Perl),
www.biopython.org (Python), www.bioruby.org (Ruby)

DNASTAR Lasergene Departmental Site License

www.microbiology.columbia.edu/dnastar.html

Information Security

CUMC is subject to HIPPA and HITECH regulations, so is particularly sensitive to Internet and computer security issues.

PHI and PII

Any Protected Health Information (PHI) or Personally Identifiable Information (PII) must be stored on fully disk encrypted computers, mobile devices or memory sticks. This information must also be protected in transit. Encrypting laptops, mobile devices and memory sticks is generally a good idea, since they can easily be stolen or lost, putting your personal data at risk. Using a reliable backup system allows you to easily recover your data from such a loss.

Symantec Endpoint Protection

Anti-virus and anti-malware software (Mac or Win, available free to faculty, staff or students at **https://cuit.columbia.edu/symantec**)

Encryption

Disk Encryption

Full disk encryption with pre-boot authentication is required for CUMC computers storing PHI or PII. This is automatically provided on Macs by FileVault using AES-XTS block cipher encryption (**System Preferences > Security & Privacy > FileVault**) and on Windows by commercial solutions such as Symantec Endpoint Encryption (formerly known as GuardianEdge). Disk encryption typically uses symmetric key algorithms, which use the same key to encrypt and decrypt the data. Care must be taken to use a memorable password and store the password (or a recovery key) in a secure manner. When backing up a fully encrypted disk, make sure the backups are encrypted as well.

File Encryption

Individual files or folders can be encrypted or stored on encrypted virtual volumes. This can be useful for securing or transporting sensitive personal data when full disk encryption is not available. Disk Utility on any Mac can create an encrypted disk image (.dmg) using up to 256 bit AES encryption. GnuPG is free software available for Mac, Windows and Unix at **www.gnupg.org** that supports file encryption with a variety of algorithms, including RSA, up to 256 bit AES, Blowfish and Twofish.

Public Key Encryption

Public key cryptography is often used to secure communication across open networks such as the Internet. It is based on the use of two keys, one private, which remains secret, the other public. The keys are related in a mathematically intractable manner. The public key is used to encrypt or verify a signature, the private key is used to decrypt or sign. An Enveloped Public Key Encryption (EPKE) method adds additional security. The RSA algorithm is the best known public key encryption system, and is based on the mathematical difficulty of factoring the product of two large prime numbers.

Backup

Time Machine

Macintosh computers come with a built in backup system called Time Machine that creates a complete incremental backup that allows for the restoration of accidentally deleted files or earlier version of files as well as the entire computer. It automatically backs up hourly to an external drive or Time Capsule and keeps hourly backups for 24 hours, daily backups for a month, and weekly backups for as long as there is space. Encryption is a built-in option.

Windows Backup

Windows has built in backup systems which vary slightly from version to version. Vista and Windows 7 let you use Backup and Restore (from Control Panel) to backup files or the entire computer, allowing you to select the backup drive and set up an automatic backup schedule. Windows 8 and 10 have a new backup system called File History (from Control Panel) which allows for partial Time Machine style incremental backups of user account files.

Cloud Backup

A number of commercial cloud backup services exist, including Carbonite and CrashPlan. Personal cloud storage through iCloud, Google Drive, or Dropbox can serve to backup data as well. Cloud backup offers the advantage of automatic offsite backup, but is not approved for use at CUMC, so is only appropriate for personal use with data that is not sensitive.

Cloud Computing

Cloud computing uses networks of remote servers to provide data storage and computing services across the Internet. Amazon and Google are two major cloud service providers.

Amazon Web Services (AWS)

Using Amazon's Elastic Compute Cloud (EC2), virtual servers with storage can be launched as needed (even thousands at once), billed in hourly increments, depending on server performance. Persistent Elastic Block Store (EBS) storage is included. Auto Scaling to adjust capacity is free. Elastic Load Balancing and detailed monitoring is available at additional cost. EC2 server instances can run either Linux or Windows, and you can either create your own Amazon Machine Images or use a variety of premade AMIs. Elastic MapReduce (EMR) costs extra, but allows easy launching of a server as a node in a Hadoop cluster (MapR or Amazon Hadoop).

Google Compute Engine

Virtual Linux machines can be launched on demand as needed, billed by the minute (10 minute minimum), depending on performance. Persistent Google Cloud Storage is extra. BigQuery allows for interactive analysis of massive datasets at a cost that varies with the amount of data processed in an analysis.

Cloud Storage

Amazon S3

Simple Storage Service. Cloud storage with low monthly costs, transfer in is free, transfer out from S3 to the Internet is charged at a low cost. Can be used to host static web content or entire websites. CloudFront allows delivery from a global network of edge locations for rapid downloads in other areas of the world at an additional cost.

Amazon Glacier

Archival cloud storage with very low monthly costs (if for at least 3 months), transfer in free, retrieval of 5% of total storage per month is free, retrieval above that is at a low cost varying with the speed of retrieval. Transfer out from Glacier to the Internet is charged at a low cost.

Google Cloud Storage

Cloud storage with low monthly costs, free transfer in, and low cost transfer out. Can be used to host static web content or static web sites. Data can be stored in the U.S. or Europe.

Databases

Flat File Database (FFDB)

A collection of similar files made useful by ordering and indexing. All the information about one sequence would be stored in one structured text file, and you generally examine one file at a time.
Examples: GenBank, Excel, older versions of FileMaker or Access

Relational Database Management System (RDBMS)

All data is stored inside one or more tables of rows and column, with all operations done on the tables themselves or producing other tables as the result. All the information about one sequence would be stored in a collection of tables with other data, so you can easily look at just the information relating to that sequence, or how it relates to the database as a whole. Structured Query Language (SQL) is used to access data in a relational database.
Examples: MySQL, PostgreSQL, SQLite, Microsoft SQL Server, Oracle

Not Only SQL (NoSQL)

Data is stored and retrieved in a different manner than tabular relations, the details varying from one to the other. They are often used for big data or real time applications, and their data structures give them certain advantages and disadvantages handling particular data compared to an RDBMS.
Examples: Amazon DynamoDB (key-value) and SimpleDB (eventual consistency), Apache CouchDB (document) and HBase (column), Dynamo (key-value), MongoDB (document), OrientDB (graph)

PubMed and EndNote

PubMed

http://www.ncbi.nlm.nih.gov/pubmed
http://www.nlm.nih.gov/bsd/disted/pubmedtutorial/
http://www.ncbi.nlm.nih.gov/books

EndNote and PubMed

EndNote can act as a database client to directly connect to the PubMed database and search and retrieve references from it: **Tools > Online Search… > Favorites > PubMed (NLM)**
Alternatively, the PubMed Clipboard can be used to collect references, export them as a text file in MEDLINE format, then import them into EndNote using the **PubMed (NLM)** import filter.

EndNote and PDF Files

EndNote can also be used to organize PDF files, which are otherwise easy to lose track of. With a reference selected or open, clicking on the paperclip, or **References > File Attachments > Attach File…** inserts a link to the PDF file you select. Use **References > File Attachments > Open with Preview** to open the linked PDF in Preview (on a Mac) or double click the reference to view the PDF in EndNote. EndNote also offers the ability to import folders of PDFs, share libraries with other researchers, and back them up to the web.

Papers

Papers, available for Mac and Win at a discount to students at **http://www.papersapp.com** is another excellent option for importing PubMed references and managing references and PDFs.

Flat File

```
L27758. Birmingham IncP-a...[gi:508311]    Related Sequences, PubMed, Taxonomy

LOCUS       BIACOMGEN                60099 bp    DNA      linear    BCT 08-JUL-1994
DEFINITION  Birmingham IncP-alpha plasmid (R18, R68, RK2, RP1, RP4) complete
            genome.
ACCESSION   L27758
VERSION     L27758.1  GI:508311
KEYWORDS    complete genome.
SOURCE      Birmingham IncP-alpha plasmid (plasmid Birmingham IncP-alpha
            plasmid, kingdom Prokaryotae) DNA.
ORGANISM    Birmingham IncP-alpha plasmid
            broad host range plasmids.
REFERENCE   1  (bases 1 to 60099)
AUTHORS     Pansegrau,W., Lanka,E., Barth,P.T., Figurski,D.H., Guiney,D.G.,
            Haas,D., Helinski,D.R., Schwab,H., Stanisich,V.A. and Thomas,C.M.
TITLE       Complete nucleotide sequence of Birmingham IncP-alpha plasmids:
            compilation and comparative analysis
JOURNAL     J. Mol. Biol. 239, 623-663 (1994)
MEDLINE     94285211
FEATURES             Location/Qualifiers
     source          1..60099
                     /organism="Birmingham IncP-alpha plasmid"
                     /plasmid="Birmingham IncP-alpha plasmid"
                     /db_xref="taxon:35419"
BASE COUNT    10839 a  18681 c  18448 g  12131 t
ORIGIN
        1 ttccaccccg aaacacgagca cggcacccgc gaccactatg ccaagaatgc ccaaggtaaa
       61 aattgccggc cccgccatga agtccgtgaa tgccccgacg gccgaagtga agggcaggcc
      121 gccacccagg ccgccgccct cactgcccgg cacctggtcg ctgaatgtcg atgccagcac
      181 ctgcggcacg tcaatgcttc cgggcgtcgc gctcgggctg atcgcccatc ccgttactgc
      241 cccgatcccg gcaatggcaa ggactgccag cgccgcgatg aggaagcggg tgccccgctt
      301 cttcatcttc gcgcctcggg cctcgaggcc gcctacctgg gcgaaaacat cggtgtttgt
etc.
```

- Data type not distinct from record.
- Tremendous duplication of data.
- Mixed hierarchical and non-hierarchical data.
- Difficult to query in a sophisticated manner.
- Difficult to link to other data.

Relational Database

What distinguishes a relational database from a flat file or spreadsheet is the ability to use relational algebra to create sophisticated queries that provide many alternative views of the data or subsets of the data. The language used to do this is called Structured Query Language, or SQL, a declarative language for querying data in which you describe what you want to see.

- A relational database contains one or more tables.
- Tables are known as relations, and are identified by a unique name.
- A row is also known as a tuple, or record. No rows are identical. They are in no particular order.
- A column is also known as an attribute, or field. Each column is identified by a name and contains only one type of data, e.g. an integer or text.

Relational Databases in Biology

Using a relational database with biological data allows you to ask questions that would be difficult or impossible with a flat file format, add functional annotation and easily work with subsets of data to improve search strategy and sensitivity.

Structured Query Language

Structured Query Language, or SQL was developed in the 1970s at IBM by Donald Chamberlin and Raymond Boyce, and is now the standard for interacting with relational databases. It is a specialized programming language for managing data in a relational database management system, or RDBMS, with declarative and procedural features. It handles both data definition and data manipulation tasks, and supports relational algebra and calculus. SQL became an ANSI (American National Standards Institute) standard in 1978.

Note that SQL is not case sensitive, but SQL commands are often written in all caps. Some implementations require a semicolon at the end of any SQL statement, others only when chaining multiple statements together. Whitespace is generally ignored. An asterisk is used to specify all columns in a table.

Undefined data is considered to have a value of null, and two null values cannot be logically compared.

Example:
SELECT * FROM data

SQL Data Manipulation Language

CREATE DATABASE creates a new database
ALTER DATABASE modifies an existing database
CREATE TABLE creates a new table
ALTER TABLE modifies an existing table
DROP TABLE deletes a table
CREATE INDEX creates an index
DROP INDEX deletes an index
UPDATE updates existing data using **SET** and **WHERE**
DELETE deletes data from a database
INSERT INTO inserts new data into a database

Each column in a table is expected to have a name and a specified data type. Common data types include **TEXT, CHAR(n), VARCHAR(max), INTEGER, FLOAT, BOOLEAN, DATE, TIME** and **BLOB.**

Example:
CREATE TABLE protein (
protein_id INTEGER PRIMARY KEY AUTOINCREMENT
sequence TEXT
length INTEGER)

SQL Queries

SELECT extracts a column from a database
SELECT DISTINCT extracts distinct values in a column from a database

Clauses
FROM specifies one or more tables to retrieve data from
WHERE filters data that does not meet the specified criteria
ORDER BY specifies the order in which rows appear in the results
INNER JOIN returns all rows from two or more tables where there is at least one match in each table (other forms exist)

UNION combines the results of two or more **SELECT** statements
RESTRICT remove rows that do not meet the specified criteria
PRODUCT combines rows from two or more tables in all possible ways

Example:
SELECT protein_id FROM protein

SQL Functions and Operators

Functions
AVG()
COUNT()
COUNT (DISTINCT)
FIRST()
FORMAT()
GROUP BY ()
ORDER BY ()
LAST()
MAX()
MIN()
MIN()
SUM()

Operators
SQL also supports many standard mathematical operators, such as **=. >, >=. <, <=, BETWEEN, LIKE, IN, IS, IS NOT,** etc.% is a wildcard that matches one or more characters and _ is a a wildcard that matches any one character.

Example:
SELECT * FROM protein WHERE length BETWEEN 51 AND 100

SQL Optimization

Normalization

Normalization seeks to reduce redundancy, which increases storage efficiency, data consistency and scalability. This involves splitting large tables into smaller, less redundant tables, and defining their relationships such that a change of a field can be made in a single table and then propagated through the database.

For example, a unique primary key in a table can be helpful, but although a name that appears unique to an organism might appear to make a good primary key, it could later need to be changed or modified, which would require it to be changed throughout the entire database. A better primary key would be a unique number that never changes, stored in a table that also contains the organism name. The name would only have to be changed in that one table. Primary keys from one table used to provide a link to another table are called foreign keys.

Indexing

Indexing is a technique used to speed searching in indexed columns, it can greatly increase the speed of certain queries, but indexes will slow writing data to those columns and will occupy additional space. They are particularly useful when dealing with large tables and frequently queried columns.

Normalized Tables
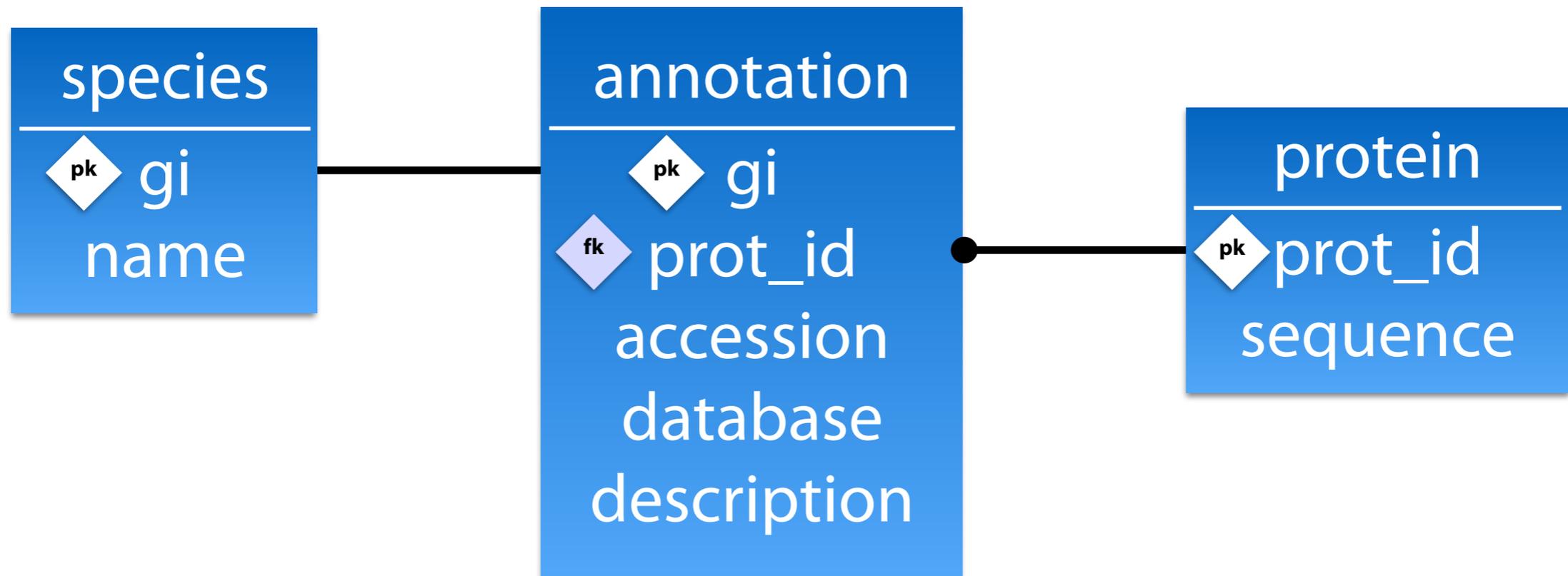
Cloud Databases

Amazon SimpleDB

Simple non relational data store. Limited to 10 GB of data per table and 25 writes per second. The first 25 SimpleDB Machine Hours and 1 GB of storage are free each month.

Amazon RDS

Relational Database Service. On demand database instances with MySQL, PostgreSQL, MariaDB, Oracle, Microsoft SQL Server or Amazon Aurora preinstalled. Cost depends on the database and server speed.

Amazon DynamoDB

NoSQL database service. Cost based on reserved performance capacity plus a fee for data storage.

Amazon Redshift

Petabyte scale data warehousing capable of massive parallel processing of very large datasets. Cost depends on server speed and type of node (Dense Compute or Dense Storage). Backups and data transfer are free.

Google Cloud SQL

MySQL database service residing in Google's cloud. Cost based on server performance.

Database References

Practical Computing for Biologists
by Steven Haddock and Casey Dunn

Databases Demystified, Second Edition
by Andrew Oppel

MySQL, Fifth Edition
by Paul Dubois

Databases in Depth: Relational Theory for Practitioners
by C.J. Date