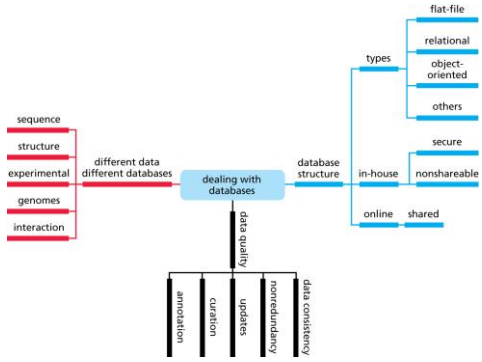
BINF 630: Bioinformatics Methods

Iosif Vaisman

Email: ivaisman@gmu.edu

Databases in Bioinformatics



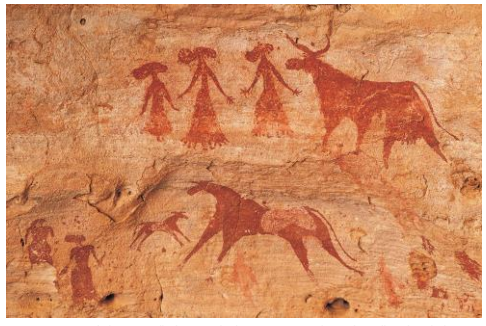
Data management and utilization

An early example of biological data depository:

Cave painting: Lascaux Grotto, near Montignac, France., ca. 15,000 BCE (Ralph Morse, Getty Images)

Data management and utilization

An early example of biological data depository:



Cave painting: Ennedi Plateau, Chad, ca. 7,000 BCE (Encyclopædia Britannica)

Genomic data

Number of species
$3 \times 10^6 - 1 \times 10^8$
80,000

Individual genome
$10^6 - 10^9$ nucleotides

Population size
humans – 7.4×10^9
bacteria – ??
300,000

Total number of cells in human body
$5 \times 10^{12} - 2 \times 10^{20}$
(Bianconi et al, Ann. Hum. Biol. 40:463, 2013)

Total gut microbiota
$3 \times 10^{13} - 4 \times 10^{14}$
(Suau et al, Appl Environ Microbiol. 65:4799, 1999)

Last month estimate
Human: 3×10^{13}
Bacterial: 4×10^{13}
(Sender et al, Cell 164:337, 2016)

Genomic data

Protein structures
PDB – **116,258** (Feb 2016)

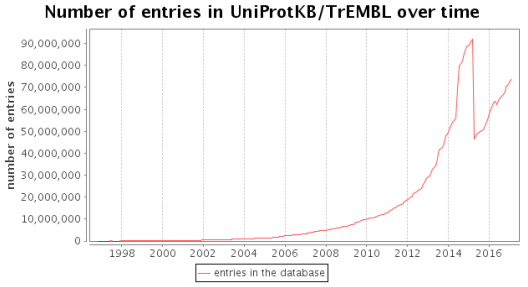
Protein sequences
Swiss-Prot – **550,552** (Feb 2016)
TrEMBL – **61 million** (Feb 2016)

Functional genomics
GEO (Feb 2016) –
DataSets **3,848**
Series **65,847**
Samples **1,738,931**

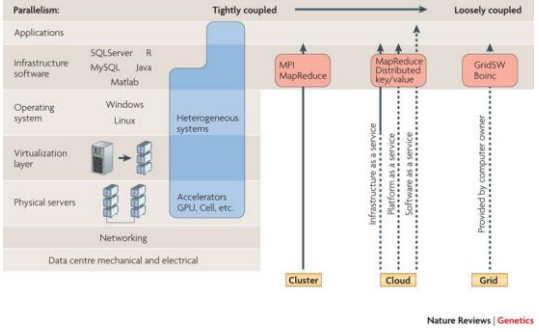
Variations in cancer
COSMIC (Feb 2016) –
Samples **1,192,776**
Coding Mutations **3,942,175**
Whole Genomes **25,133**
Copy Number Variants **1,064,039**
Gene Expression Variants **9,479,893**

Molecular Databases

Nucleic acid sequences:	GenBank	(198 million, Dec 2016)
	WGS	(395 million, Dec 2016)
Protein sequences:	UniProtKB:	
	Swiss-Prot	(550 thousand, Jan 2016)
	TrEMBL	(59 million, Jan 2016)
Protein structures:	PDB	(126 thousand, Feb 2017)



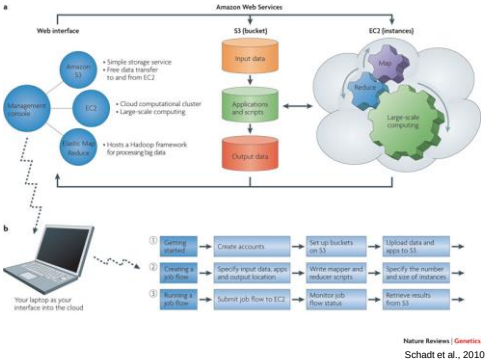
Infrastructure organization



Nature Reviews | Genetics

Schadt et al., 2010

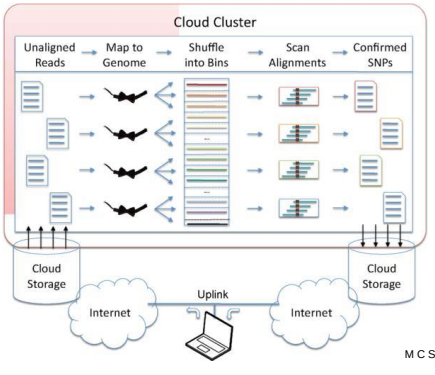
Cloud computing



Nature Reviews | Genetics

Schadt et al., 2010

Cloud computing and DNA sequencing



M C Schatz et al., 2010

Types of computational environments

Large-scale computing platform	Computing architectures	Advantages	Disadvantages	Example applications
Cluster computing	Multiple computers linked together, typically through a fast local area network, that effectively function as a single computer	Cost-effective way to realize supercomputer performance	Requires a dedicated, specialized facility, hardware, system administrators and IT support	• BLAST • Bayesian network reconstruction • Computing genetic associations in large-scale GWA studies
Cloud computing	Computing capability that abstracts the underlying hardware architectures (for example, servers, storage and networking), enabling convenient, on-demand network access to a shared pool of computing resources that can be readily provisioned and released (NIST, Technical Report)	The virtualization technology used results in extreme flexibility; good for one-off HPC tasks, for which persistent resources are not necessary	Privacy concerns; less control over processes; bandwidth is limited as large data sets need to be moved to the cloud before processing	• Searching sequence databases • Aligning raw sequencing reads to genomes • General purpose genomics tools (for example, GeneSifter from Geospiza) • Most applications running on a cluster can be transferred to a cloud
Grid computing	A combination of loosely coupled networked computers from different administrative centres that work together on common computational tasks. Typified by volunteer computing efforts (such as Folding@Home), which 'scavenge' spare computational cycles from volunteers' computers	Ability to enlist large-scale computational resources at low or no cost (large-scale volunteer-based efforts)	Big data transfers are difficult or impossible; minimal control over underlying hardware, including availability	• Protein folding (Folding@Home) • Proteome analysis • Protein prediction (Rosetta@Home) • Predicting interactions between small molecules and proteins (FightAIDS@Home) • Condor project
Heterogeneous computing	Computers that integrate specialized accelerators — for example, GPUs or reconfigurable logic (FPGAs) — alongside GPPs	Cluster-scale computing for a fraction of the cost of a cluster; optimized for computationally intensive fine-grained parallelism; local control of data and processes	Significant expertise and programmer time required to implement applications; not generally available in cluster- and cloud-based services	• Bayesian network learning • Protein folding (Folding@Home) • Molecular dynamics simulation (NAMD) • BLAST • CLUSTALW • HMMER • Reconstruction of evolutionary trees

The above categories are not exclusive. For example, heterogeneous computers are often used as the building blocks of cluster, grid or cloud computing systems; the shared computational clusters available in many organizations could be described as private Platform as a Service (PaaS) clouds. The main differences between the platforms are degree of coupling and tenancy — grid and cloud computers are designed for loosely coupled parallel workloads, with the grid resources allocated exclusively for a single user, whereas the underlying hardware resources in the cloud are typically shared among many users (multi-tenancy). Cluster computers are typically used for tightly coupled workloads and are often allocated to a single user. FPGA, field-programmable gate array; GPP, general purpose processor; GPU, graphics processing unit; GWA, genome-wide association; HPC, high-performance computing; NIST, National Institute of Standards and Technology.

Schadt et al., 2010

Types of computational environments

Environment	URL
Cloud computing	
Amazon Elastic Compute Cloud	http://aws.amazon.com/ec2
Bionimbus	http://www.bionimbus.org
NSF CluE	http://www.nsf.gov/cise/clue/index.jsp
Rackspace	http://www.rackspacecloud.com
Science Clouds	http://www.scienceclouds.org
Heterogeneous computing	
NVIDIA GPUs	http://www.nvidia.com
AMD/ATI GPUs	http://www.amd.com
Heterogeneous cloud computing	
SGI Cyclone Cloud	http://www.sgi.com/products/hpc_cloud/cyclone
Penguin Computing On Demand	http://www.penguincomputing.com/POD/Summary

GPU, graphics processing unit; NSF, US National Science Foundation.

Schadt et al., 2010

Defining Big Data

NOT JUST SIZE

The three Vs of Big Data: volume, variety and velocity

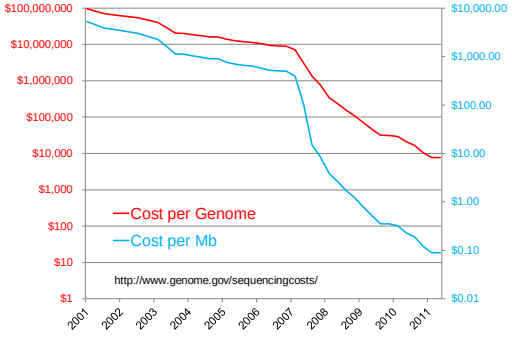
(D.Laney, 2001)

Elements of "Big Data" include:

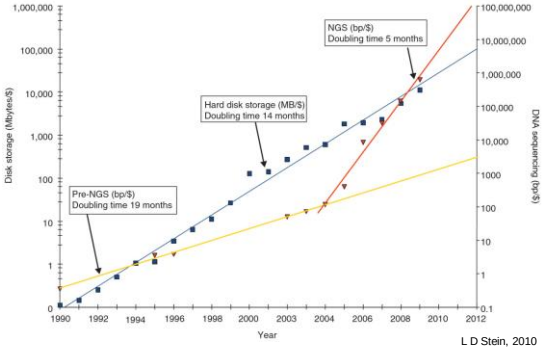
- The degree of complexity within the data set
- The amount of value that can be derived from innovative vs. non-innovative analysis techniques
- The use of longitudinal information supplements the analysis

http://mike2.openmethodology.org/wiki/Big_Data_Definition

Genome sequencing costs



Sequencing and storage cost



L D Stein, 2010

Data Mining

- Data mining is the exploration and analysis, by automatic or semiautomatic means, of large quantities of data in order to discover meaningful patterns and rules
- Common data mining tasks
  - Classification
  - Estimation
  - Prediction
  - Affinity Grouping
  - Clustering
  - Description

Knowledge Discovery

Knowledge is a.pattern that exceeds certain threshold of interestingness.

Factors that contribute to interestingness:

coverage
confidence
statistical significance
simplicity
unexpectedness
actionability

Knowledge Discovery

- Directed and Undirected KD
- Directed KD
  - Purpose: Explain value of some field in terms of all the others
  - Method: We select the target field based on some hypothesis about the data. We ask the algorithm to tell us how to predict or classify it
  - Similar to hypothesis testing (e.g., in regression modeling) in statistics

Knowledge Discovery

- Undirected KD
  - Purpose: Find patterns in the data that may be interesting
  - Method: clustering, affinity grouping
  - Closest to ideas of machine learning in artificial intelligence
- Comparison
  - UKD helps us to recognize relationships & DKD helps us to explain them

Classification

- Classifying observations into different categories given characteristics

Prediction

- Rules that explain how to predict a future value or classification, given characteristics

Estimation

- Rules that explain how to estimate a value given characteristics

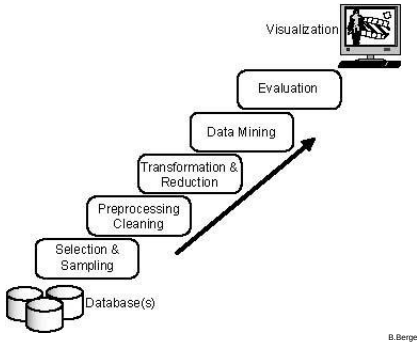
Affinity Grouping

- Grouping by relations (not by characteristics)

Clustering

- Segmenting a diverse population into more similar groups
- In clustering, there are no pre-defined classes and no examples. Records are grouped together by some similarity measure.

Knowledge Discovery



B.Bergeron, 2002