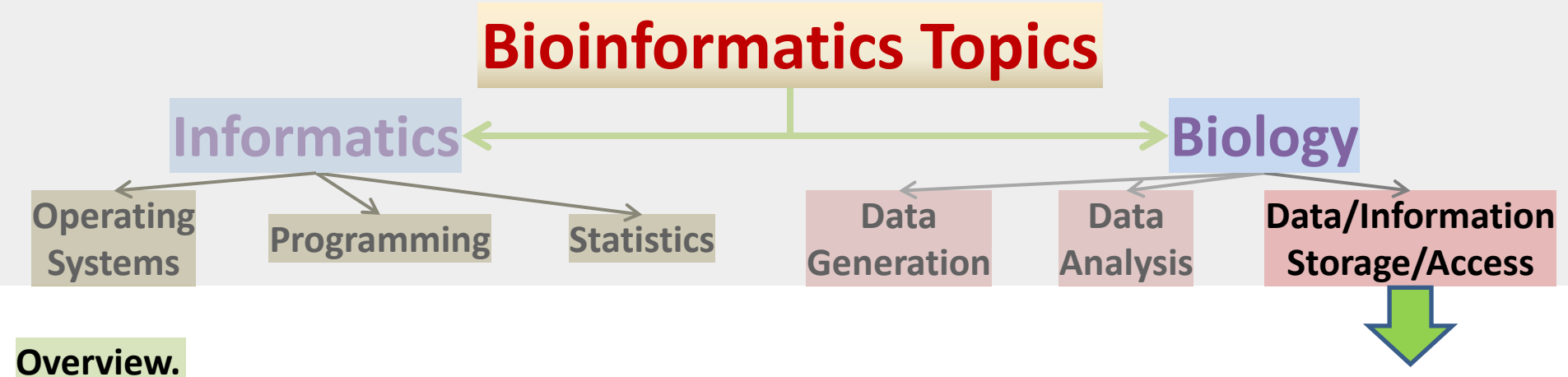


Overview.

Raw Experimental Data, can next be Annotated in the light of analytical revelation.

Data + Annotation = Information.

Information can now be stored in Databases that allow users easy and *unrestricted* access.

Primary DNA Sequence Databases

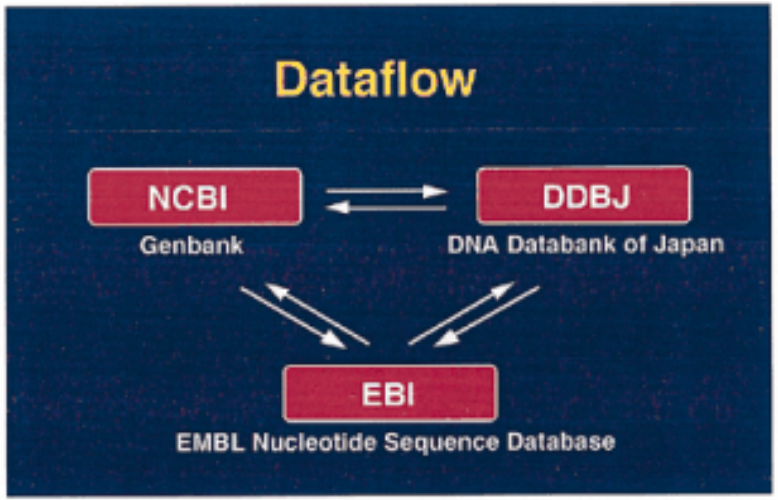
Original submission by experimentalists
Content controlled by the submitter

EMBL NUCLEOTIDE SEQUENCE DATABASE

NCBI GenBank

DDBJ DNA Data Bank of Japan

INSDC International Nucleotide Sequence Database Collaboration



Primary Protein Sequence Databases

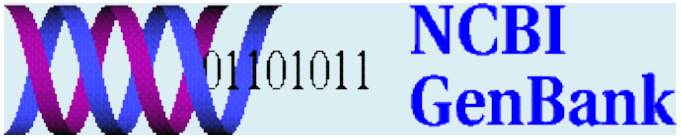
UniProtKB

an encyclopedia on proteins

composed of 2 sections
UniProtKB/TrEMBL and UniProtKB/Swiss-Prot
unreviewed and reviewed
automatically annotated and manually annotated

Derivative Sequence Databases

Built from primary data



RefSeq

Submission by experimentalists
Significant redundancy
Annotation inconsistent
DNA and RNA only

non-redundant
richly annotated
DNA, RNA, protein
diverse taxa

akin to the primary research literature

akin to the review literature

Derivative Databases for Protein Features

Collections of HMMs representing Protein Domains and/or Motifs derived from Protein sequence Databases.

Derivative Databases for Protein Features

It is generally wise to use more than one Feature Searching service.

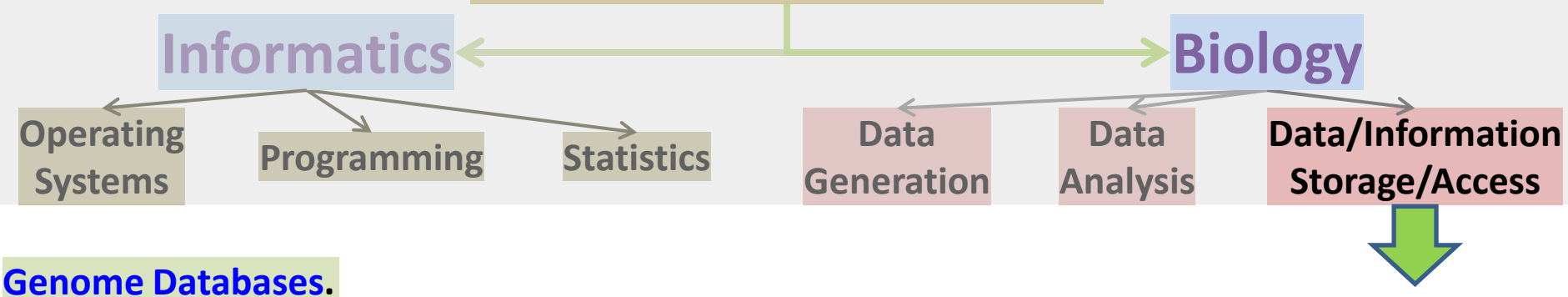
This can be tedious, involving many websites and different search tools.

is a consortium of member databases.

defines protein families, domains, regions, repeats and sites according to matches against member databases

enables any subset of member databases to be searched together

Bioinformatics Topics



Genome Databases.

Genome Databases store entire genome sequence(s) AND their interpretation.

Each new sequenced genome or significantly re-assembled existing genome is fully analysed.

The individual processes for manual analysis are the same as those for automatic analysis.
Most have been mentioned in this simple talk.

Analysing an individual gene can be done manually.
Analysing an entire genome is only practical using automated strategies.

Bioinformatics Topics

- Informatics
  - Operating Systems
  - Programming
  - Statistics
- Biology
  - Data Generation
  - Data Analysis
  - Data/Information Storage/Access

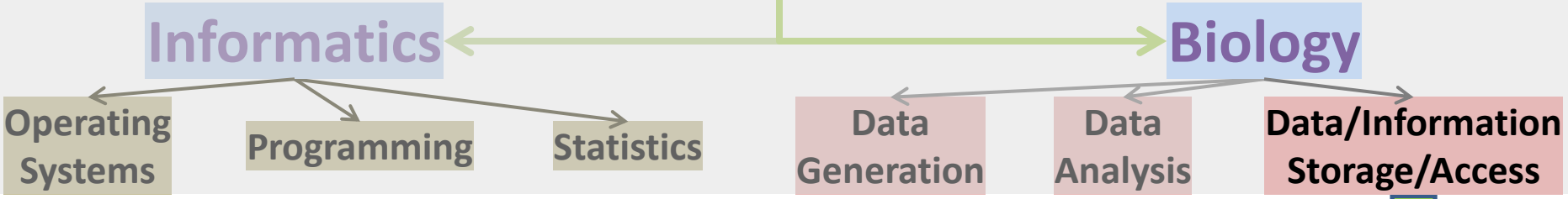
Genome Databases.

The Three foremost Genome Database options

- Ensembl
- NCBI / NCBI Map Viewer
- UCSC Genome Browser

Ensembl and UCSC Browser software can be downloaded and used for private datasets.

Bioinformatics Topics



Protein Structure Databases.

RCSB PDB
PROTEIN DATA BANK

Worldwide
Protein Data Bank
Foundation

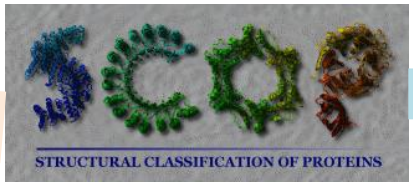
PDBj
Protein Data Bank Japan

PDBe
Protein Data Bank in Europe

Bioinformatics Topics

- Informatics
  - Operating Systems
  - Programming
  - Statistics
- Biology
  - Data Generation
  - Data Analysis
  - Data/Information Storage/Access

Protein Structure Databases.

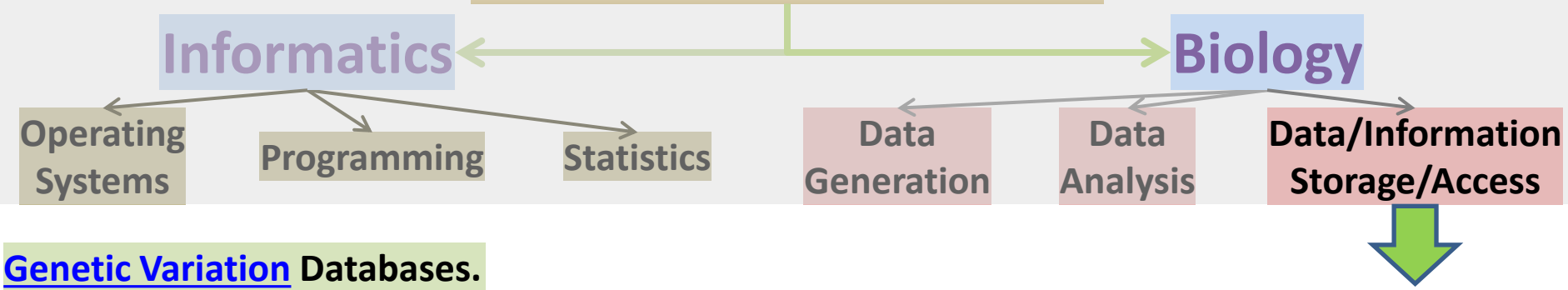


Superfamily
HMM library and genome assignments server

Worldwide
Protein Data Bank
Foundation

CATH
Gene3D

Bioinformatics Topics



Genetic Variation Databases.

Databases storing the many genetic variations that occur between individuals and species.

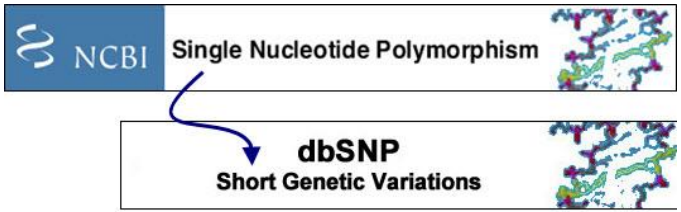
Widely incorporated into Genome Databases, such as Ensembl.

Since High Throughput Sequencing (HTS) has become standard, variation detection has become easier. Databases have developed dramatically.

Bioinformatics Topics

- Informatics
  - Operating Systems
  - Programming
  - Statistics
- Biology
  - Data Generation
  - Data Analysis
  - Data/Information Storage/Access

Genetic Variation Databases.



dbSNP is the largest general database for genetic variations.

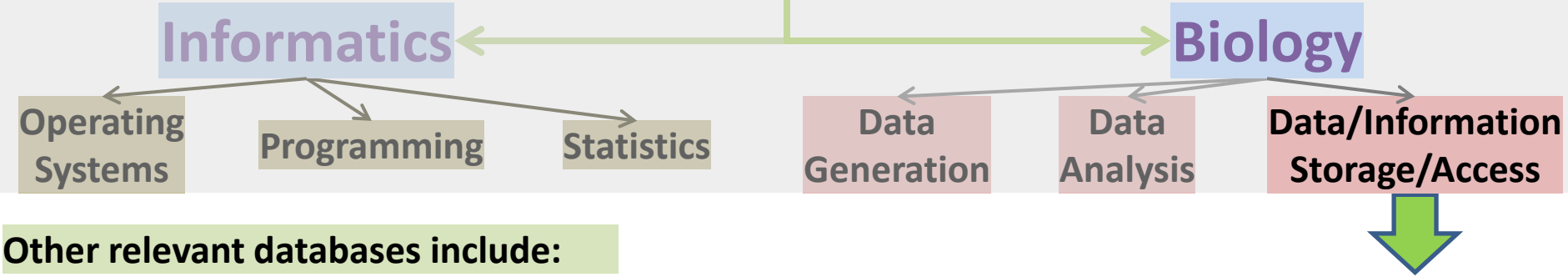
Originally just Single Nucleotide Polymophisms (SNPs).

Now includes other types of Short Genetic Variation.

dbSNP, originally focused on human variations, now covers many organisms.

dbSNP now records relationships between variation and phenotype.

Bioinformatics Topics



Other relevant databases include:

Microarray databases

There are a considerable number, both commercial and public domain.

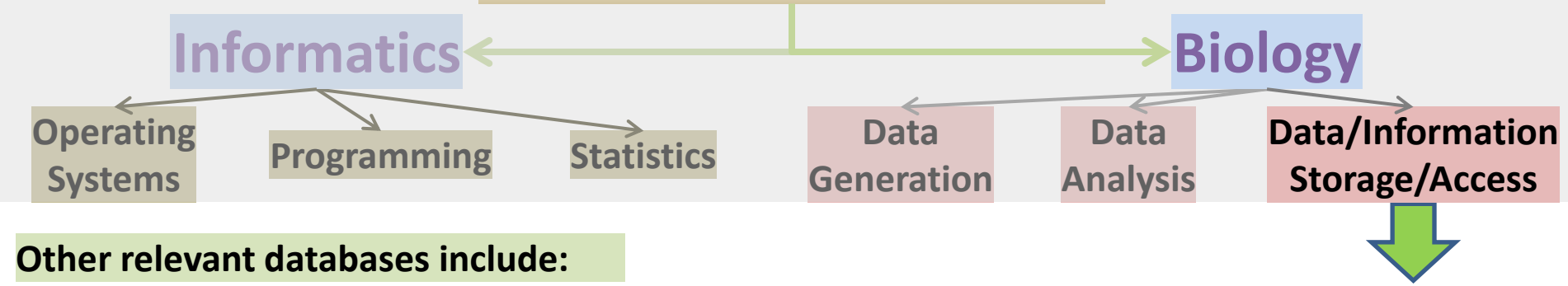
Two major Public Domain Microarray Databases are:

The Gene Expression Omnibus (GEO), maintained in America.

GEO Gene Expression Omnibus

ArrayExpress, maintained in Europe.

Bioinformatics Topics



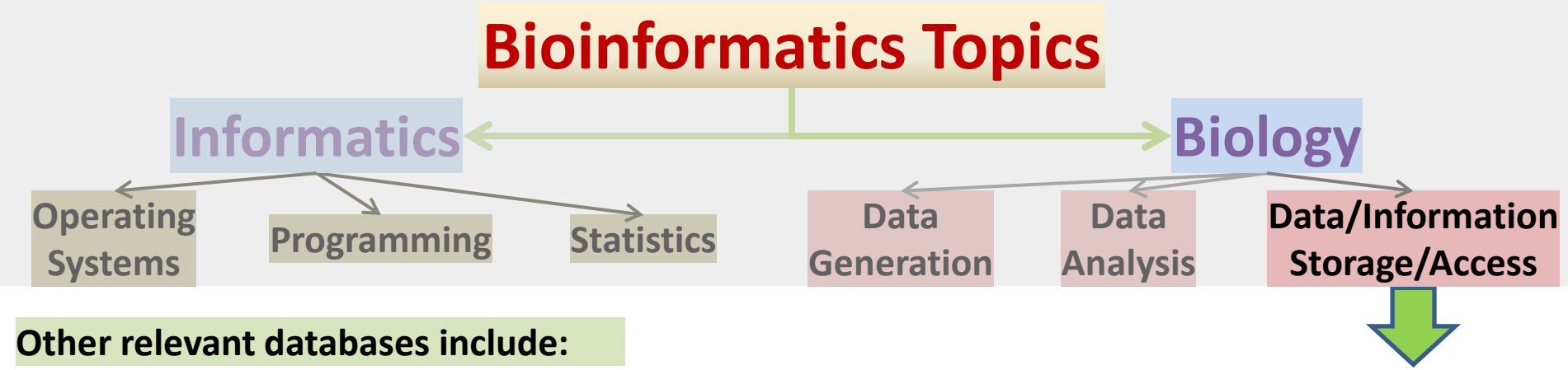
Other relevant databases include:

Microarray databases

High Throughput Sequencing (HTS) has become a viable option to the use of Microarrays.

Accordingly, both GEO and ArrayExpress now manage HTS data sets.

ArrayExpress regularly imports data from GEO.



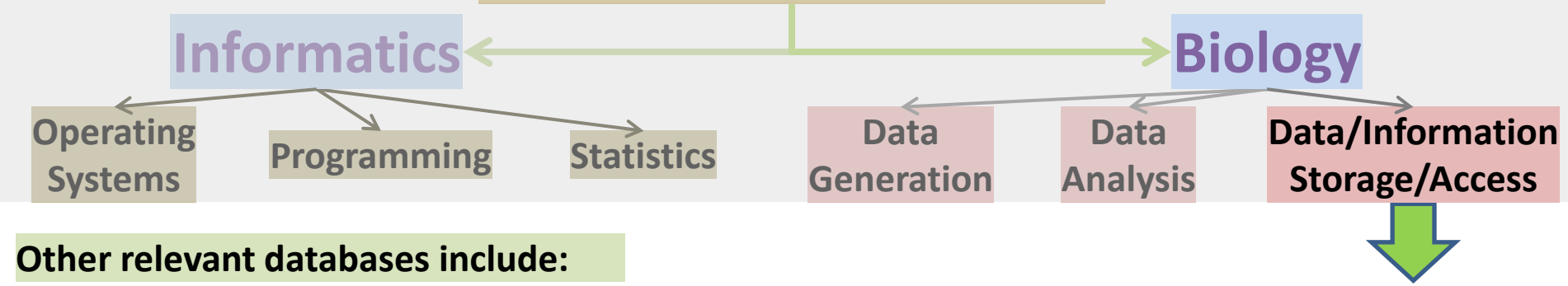
Other relevant databases include:

Literature databases

Many free literature search/access services are available via the INTERNET.

You will be introduced to, arguably, the best and most famous as a part of this course.

Bioinformatics Topics



Other relevant databases include:

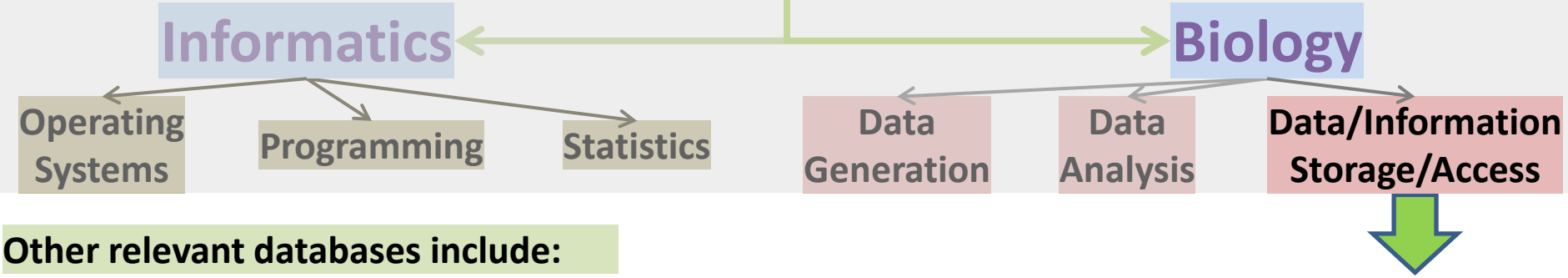
[Gene Ontology Database](#)

Early Primary Database annotation was poor.

Annotation was left to the submitted and then not curated .

In consequence, Database Searching just by Keyword was far from reliable.

Bioinformatics Topics



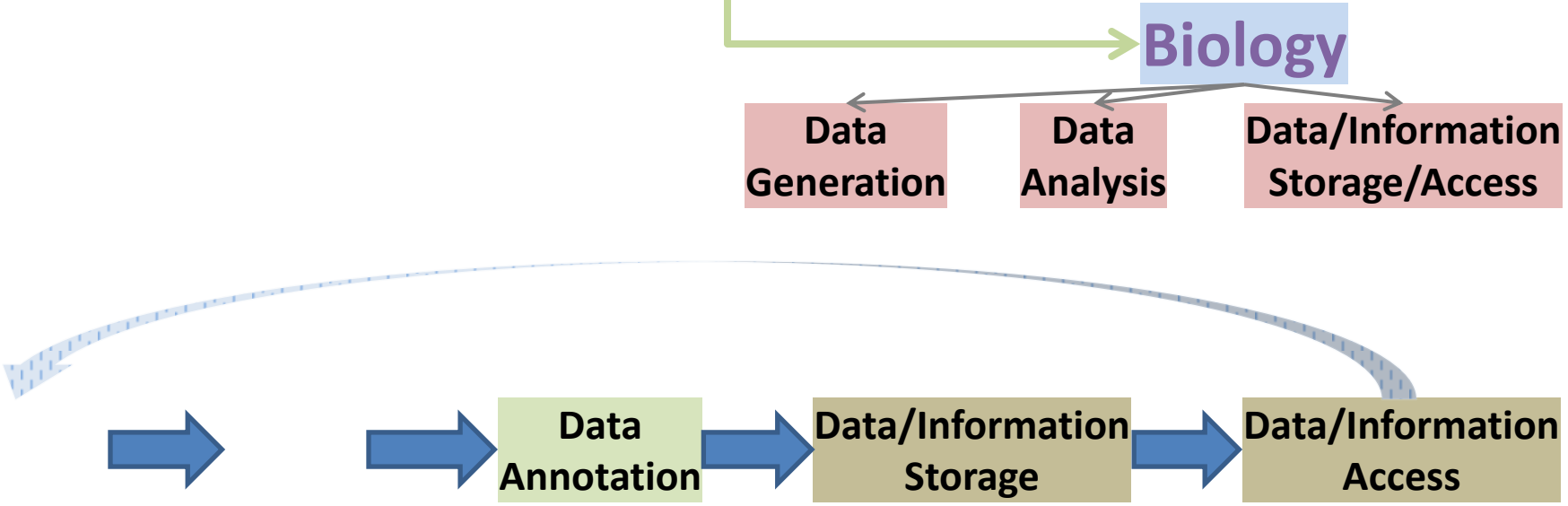
Other relevant databases include:

[Gene Ontology Database]

The [Gene Ontology] (GO) database provides a hierarchy of formally agreed terms to describe gene products accurately and unambiguously.

Searching with these terms radically improves the efficacy of annotation searching.

Bioinformatics Topics



A simplistic ordering for the Bioinformatics Topics discussed here

End of Part 2

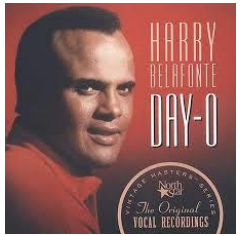
THE END

BREAK!

More to come I fear … but time for a swift cup of tea perchance?

Maybe time for a short jig? The whistling of a merry tune?

Or, mayhap, a delving into the melodic possibilities of youtube?
There be much good stuff there … I offer you a few of my favourites.



Once fully refreshed …. Click on mon braves!