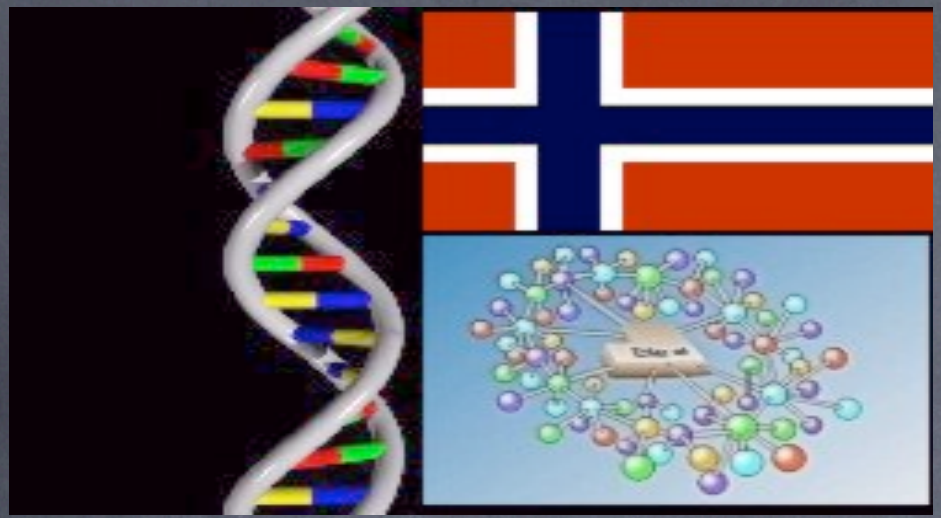


MRS for Sysadmins

George Magklaras PhD RHCE

Biotechnology Center of Oslo &

The Norwegian Center of Molecular Medicine

University of Oslo, Norway

http://www.biotek.uio.no

http://www.ncmm.uio.no

http://www.no.embnet.org

Talleres Internacionales de Bioinformática - UNAM- Enero 2012

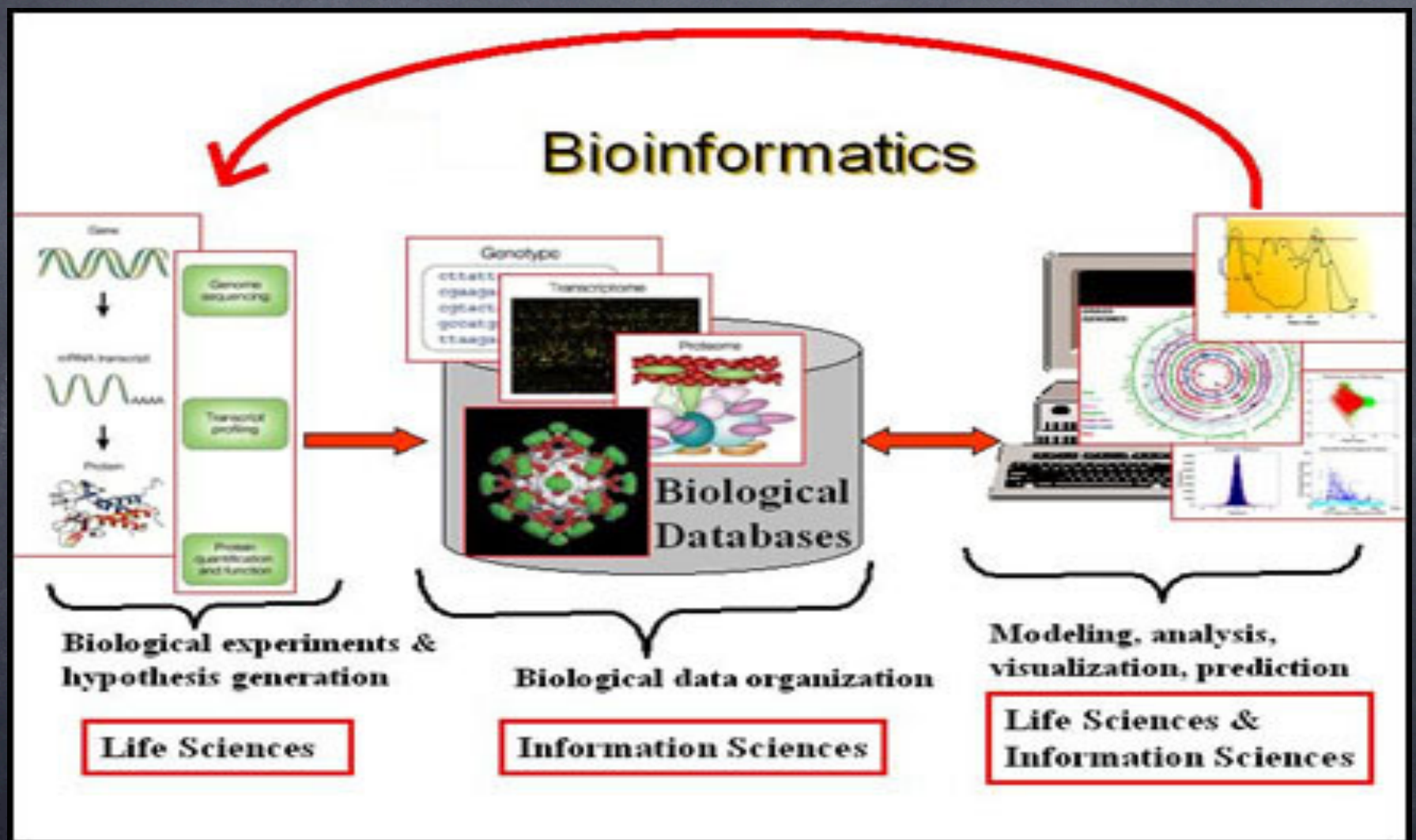
Agenda

- MRS: Who is it for and what is it?
- What do you need to run it (production)?
- A sample MRS production setup
- Installation
- Post install configuration and maintenance

What is it? (definitions)

- Biological and medical databases
- Flatfiles versus indexed flatfiles
- Web services
- The specifics of sequence retrieval

The Life Science information flow



Flat File "databases"

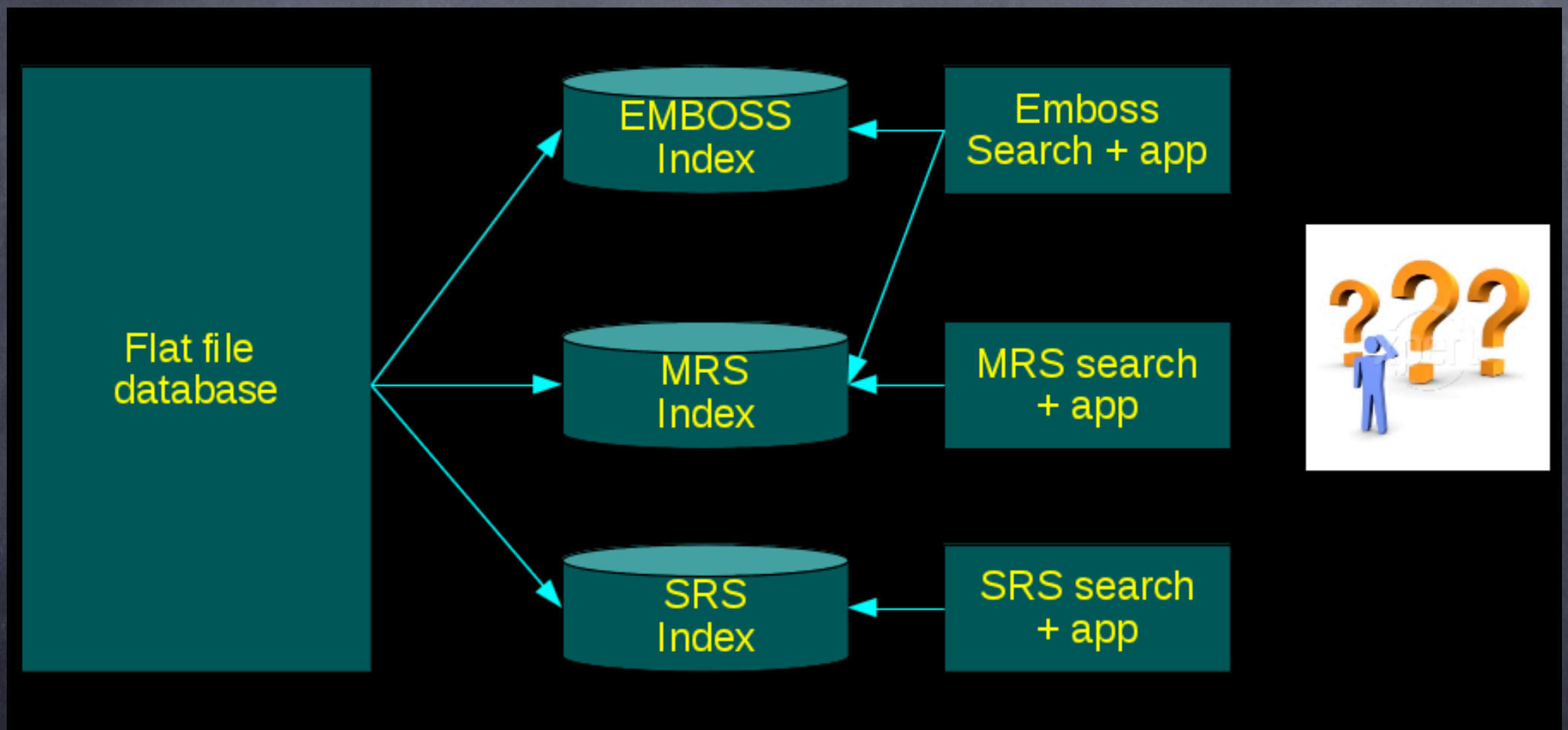
- The computer world has many different types of databases, for example:
  - Relational
  - Object oriented
- Biological sequences are often organized in "flatfiles": The source files that contain the sequences are simple human readable files, as opposed to unreadable binary files.

```
ID   Q06S78_9INFA            Unreviewed;       757 AA.
AC   Q06S78;
DT   31-OCT-2006, integrated into UniProtKB/TrEMBL.
DT   31-OCT-2006, sequence version 1.
DT   01-SEP-2009, entry version 14.
DE   RecName: Full=RNA-directed RNA polymerase catalytic subunit;
DE            EC=2.7.7.48;
OS   Influenza A virus (A/cat/Germany/606/2006(H5N1)).
OC   Viruses; ssRNA negative-strand viruses; Orthomyxoviridae;
OC   Influenzavirus A.
...
KW   Nucleotide-binding; Nucleotidyltransferase; RNA replication;
KW   RNA-directed RNA polymerase; Transferase.
SQ   SEQUENCE   757 AA;  86462 MW;  273457664D64BC0D CRC64;
     MDVNPTLLFL KVPVQNAIST TFPYTGDPPY SHGTGTGYTM DTVNRTHQYS EKGKWTTNTE
     TGAPQLNPID GPLPEDNEPS GYAQTDCVLE AMAFLEESHP GIFENSCLET MEIVQQTRVD
     KLTQGRQTYD WTLNRNQPAA TALANTIEIF RSNGLTANES GRLIDFLKDV MESMDKEEME
     ITTHFQRKRR VRDNMTKKMV TQRTIGKKKQ RLNKKSYLIR ALTLNTMTKD AERGKLKRRA
     IATPGMQIRG FVYFVETLAR SICEKLEQSG LPVGGNEKKA KLANVVRKMM TNSQDTELSF
     TITGDNTKWN ENQNPRMFLA MITYITRNQP EWFRNVLSIA PIMFSNKMAR LGRGYMFESK
     SMKLRTQIPA EMLANIDLKY FNELTKKKIE KIRPLLIDGT ASLSPGMMMG MFNMLSTVLG
     VSILNLGQKR YTKTTYWWDG LQSSDDFALI VNAPNHEGIQ AGVDRFYRTC KLVGINMSKK
     KSYINRTGTF EFTSFFYRYG FVANFSMELP SFGVSGINES ADMSIGSTVI RNNMINNDLG
     PATAQMALQL FIKDYRYTYR CHRGDTQIQT RRSFELKKLW EQTRSKAGLL VSDGGPNLYN
     IRNLHIPEVC LKWELMDEDY QGRLCNPLNP FVSHKEIESV NNAVVMPAHG PAKGMEYDAV
     ATTHSWIPKR NRSILNTSQR GILEDEQMYQ KCCNLFEKFF PSSSYRRPVG ISSMVEAMVS
     RARIDARIDF ESGRIKKEEF AEIMKICSTI EELRRPK
//
```

Flat file index

- An **index** is a set of pointers to information in the flatfile.
- It speeds up information retrieval.
- It enables sequence and record mining/filtering.
- Example: Out of 100 million sequence entries, give me those that are human hemoglobin sequences.

Flat-file index and biocomputing applications



MRS index examples

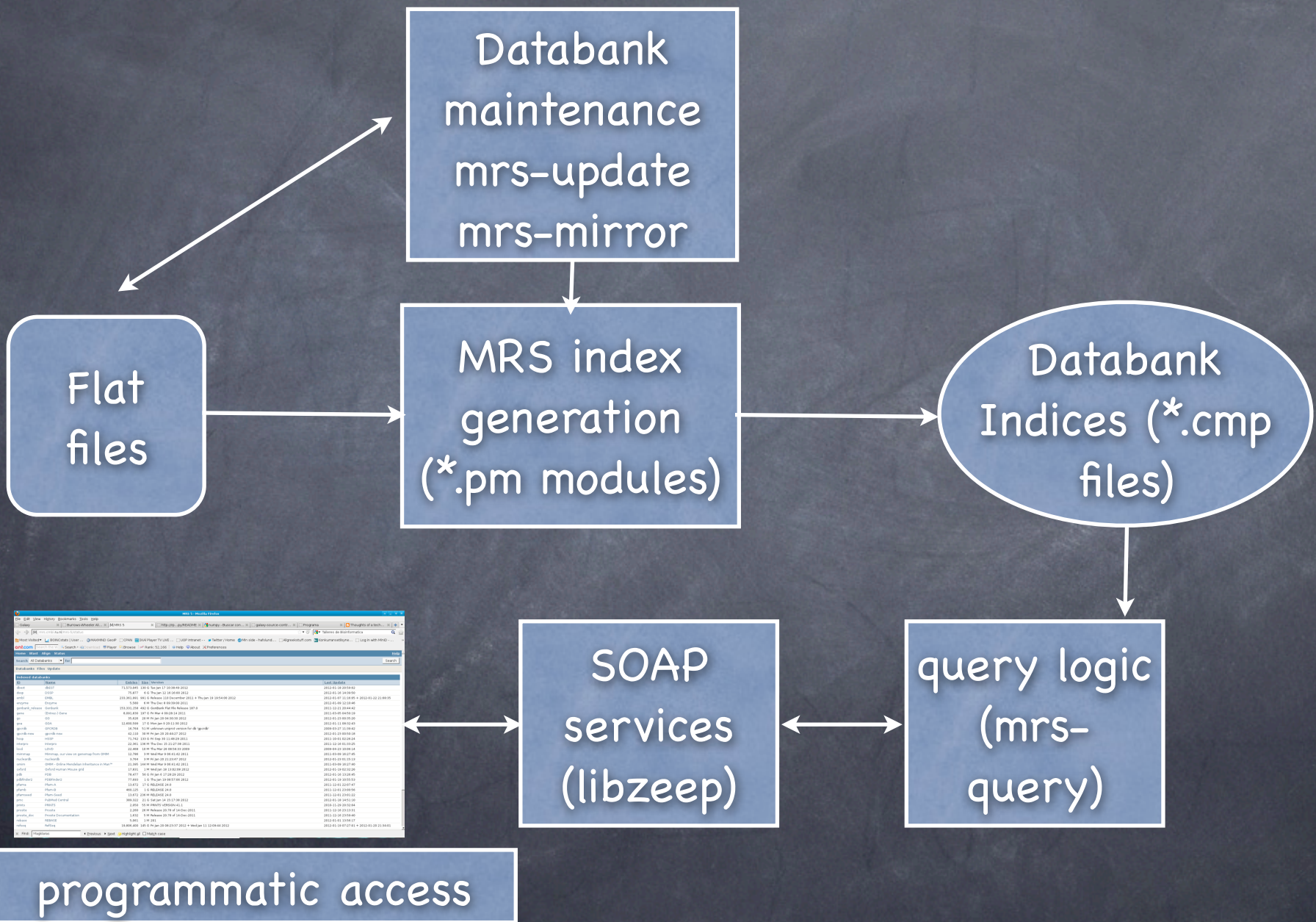
Indices for embl (EMBL)			
id	description	kind	count
__ALL_TEXT__	__ALL_TEXT__	FullText	546,837,627
ac	Accession number	FullText	233,414,321
as	as	FullText	540,943
cc	Comments and Notes	FullText	12,786,923
cd_date	cd_date	Date	8,104
co	Contigs	FullText	40,277,195
dc	Data class	FullText	18
de	Description	FullText	157,688,431
dr	Database cross-reference	FullText	23,530,976
ft	Feature table data	FullText	119,314,707
id	Identification	Unique	233,361,991
kw	Keywords	FullText	72,859
length	Sequence length	Number	385,369
oc	Organism classification	FullText	87,879
og	Organelle	FullText	6,265
os	Organism species	FullText	508,342
pr	Project	FullText	6,051
ref	Any reference field	FullText	1,106,500
sv	Sequence version	Number	99
td	Taxonomic division	FullText	30
topology	Topology (circular or linear)	FullText	4
up_date	up_date	Date	6,163
xref	xref	FullText	32,243,010

Indices for omim (OMIM - Online Mendelian Inheritance in Man™)			
id	description	kind	count
__ALL_TEXT__	__ALL_TEXT__	FullText	324,904
av	Allelic variation	FullText	84,416
cd	Creation name	FullText	144
cd_date	Creation date	Date	4,676
cn	Contributor name	FullText	170
cs	Clinical Synopsis	FullText	16,870
ed	Edit history	FullText	45
ed_date	Edit history (date)	Date	5,720
id	Number	Unique	21,395
rf	References	FullText	167,712
sa	See also	FullText	6,794
ti	Title	FullText	45,692
tx	Text	FullText	168,274

MRS: A life science data mining platform

- The "Google of biological sequence databases".
- Allows life scientists to mine/query effectively flatfile databases.
- Query via web browser/programmatic access (command-line, web services)
- Emphasis on speed and computational efficiency.

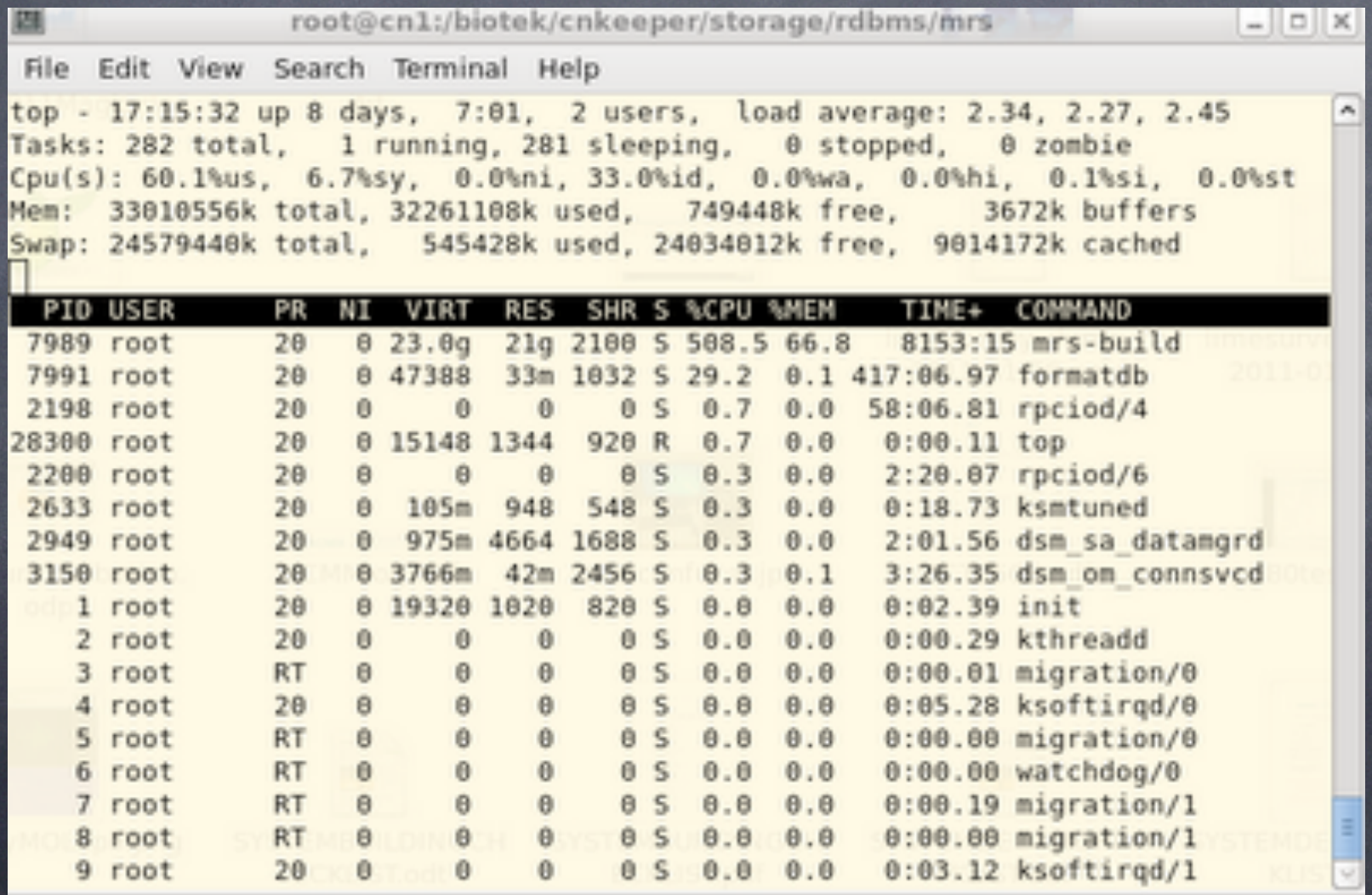
MRS system module view



MRS v5 Hardware requirements

Resources	Minimum (2012)	2012	2013	2014
Disk space	2 Tb	4 Tb	6 Tb	9 Tb
RAM	32 Gb	64 Gb	64 Gb	64 Gb
CPU cores	8	16	16	16

Indexing RAM consumption



Estimating storage space for an MRS databank

- Estimation rule: old MRS index + new flatfiles compressed + new MRS Index + temp files + BLAST indices
- Example for EMBL Release 110: 980 + 190 + 1100 + 100 + 50 = 2429 Gb = 2.4 Tb

Sample MRS production setup

DAS, FC
FCoE

MRS front end
32 cores-
256 Gb of RAM

NFS over
10G
Ethernet

MRS back end
8 cores-
64 Gb of RAM

http,
SOAP,
command
-line

MRS
users

MRS Software prerequisites

- Before installing MRS you should have:
  - gcc 4.4.x compiler or more recent versions
  - PERL version 5.10 or more recent versions
  - perl-XML-LibXSLT module
  - The Boost C++ library versions >=1.42<= 1.48
  - The libarchive interface
  - A copy of "snarf".

Two ways to obtain the MRS sources

- https://launchpad.net/~hekkel/+archive/ppa/+packages
- Checkout sources via SVN:
  - svn co http://svn.cmbi.ru.nl/libzeep/trunk (libzeep sources)
  - svn co -r 1393 https://svn.cmbi.ru.nl/mrs/trunk (MRS sources)

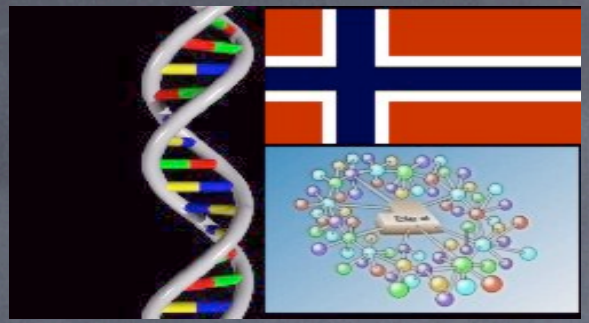
SVN: The latest and the greatest, easy to update, but can be unstable
Launchpad: Stable releases but not the latest features/bugfixes

Demo of an installation

- http://epistolatory.blogspot.com/2011/12/bioinformatics-sysadmin-craftmanship.html
- http://epistolatory.blogspot.com/2012/01/bioinformatics-sysadmin-craftmanship.html
- Keep these blog articles as a reference
- They are RHEL 6 oriented but similar steps/procedures apply to Ubuntu and other Linux distros.

Post installation steps

- Inspection of the (/usr/local/etc/mrs/)**mrs-config.xml** file: general MRS operational parameters
- Inspection of the (/usr/local/etc/mrs/) **databank.info** file: which databases are available
- Running the mrs-update script to update databanks.
- Programmatic access (MRS::Client)



Questions?

admin@embnet.uio.no

and

http://www.embnet.org/join/ContactRegistration