BIOINFORMATICS

Introduction

David Fazekas

fazekas@netbiol.elte.hu

Department of Genetics (ELTE, HU)

Earlham Institute (UK)

COURSE INFORMATION

LECTURERS

- Eszter Ari - chief trainer and administration
- David Fazekas
- Zsuzsa Dosztányi

TRAINERS

- Amanda Demeter
- Dániel Gerber
- Gábor Erdős

COURSE MATERIAL

https://genetics.elte.hu

username: genetika2016
password: genetika2016

SYLLABUS

1. Introduction
2. Data sources
3. Sequence alignment
4. Sequence databases and searching
5. Molecular phylogenetics I.
6. Molecular phylogenetics II.
7. Genomics and transcriptomics I.

   autumn break

8. Genomics and transcriptomics II.
9. Network and systems biology I. - exam I. (lecture 1-6)
10. Network and systems biology II.
11. Network and systems biology III.
12. Protein structure bioinformatics
13. exam I. (lecture 7-12)

EXAM

Lecture

- 2 written exam during the semester
- Average of those
- If either is 1, oral exam is required

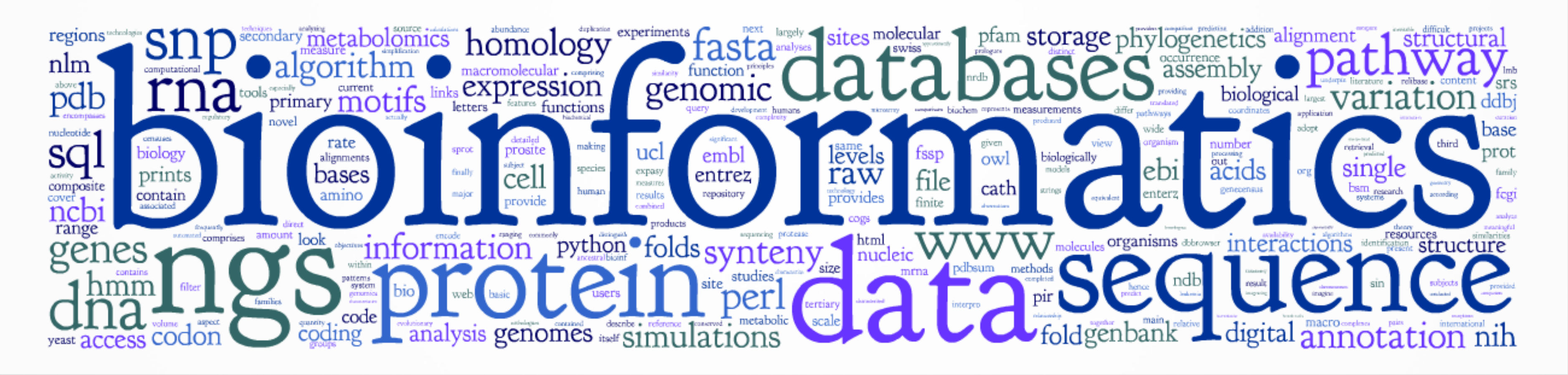
Practice

- Maximum 3 absenteeism
- Submit a project
- Group of 3 student

CHOOSE ONE

- RASK (KRAS)
- ERK1 (MAPK3)
- JAK1
- IGF1R
- GSK3B
- AXIN1
- SMAD2
- NOTCH1

WHAT IS BIOINFORMATICS?



Definition: Bioinformatics

“Research, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data.” “Bioinformatics applies principles of information sciences and technologies to make the vast, diverse, and complex life sciences data more understandable and useful.”

Working definition by the NIH Biomedical Information Science and Technology Initiative Consortium, 2000

http://www.bisti.nih.gov/docs/CompuBioDef.pdf

Definition: Computational Biology

“The development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological, behavioral, and social systems.“

“Computational biology uses mathematical and computational approaches to address theoretical and experimental questions in biology.”

Working definition by the NIH Biomedical Information Science and Technology Initiative Consortium, 2000

http://www.bisti.nih.gov/docs/CompuBioDef.pdf

I am Programmer !
I am Geneticist !
I am Biostatician !
I am GeneChip Experts !
I am Omics Experts !
I am Pharmacologist !
I am System Biologist !
I am Structure Biologist !
I am Evolutionary Biologist !
HU HU HU HA HA HA HA HA HA HA HA A A AAA.......
I am one & only the real Bioinformatist and Computational Biologist.

WHAT IS BIOINFORMATICS?

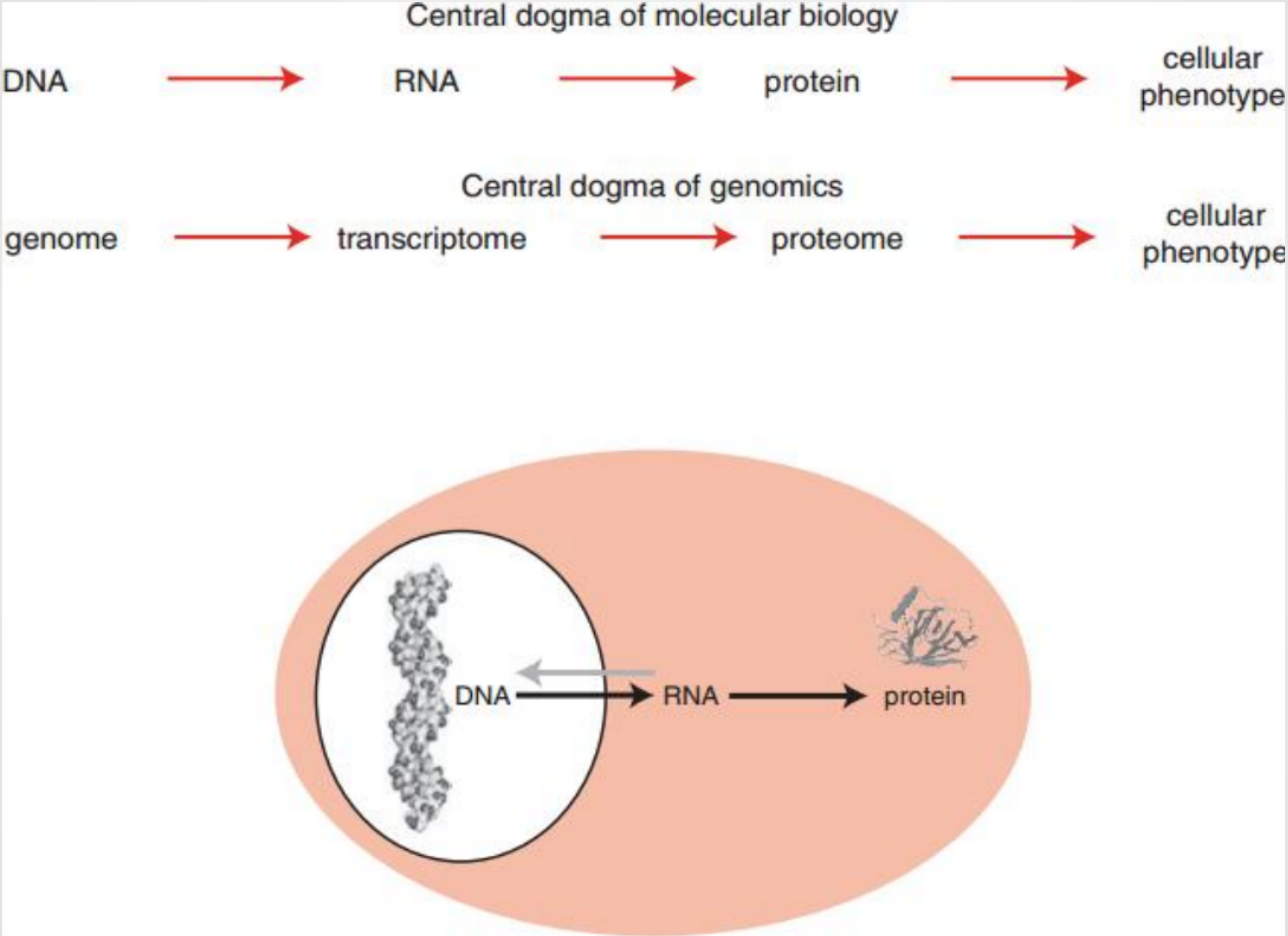
IN A NARROWER SENSE

- Working with data in life sciences

IN THE BROADER SENSE

- Molecular bioinformatics
- Sequence and structure of macro molecules
- Annotations
- Network biology

MOLECULAR BIOINFORMATICS



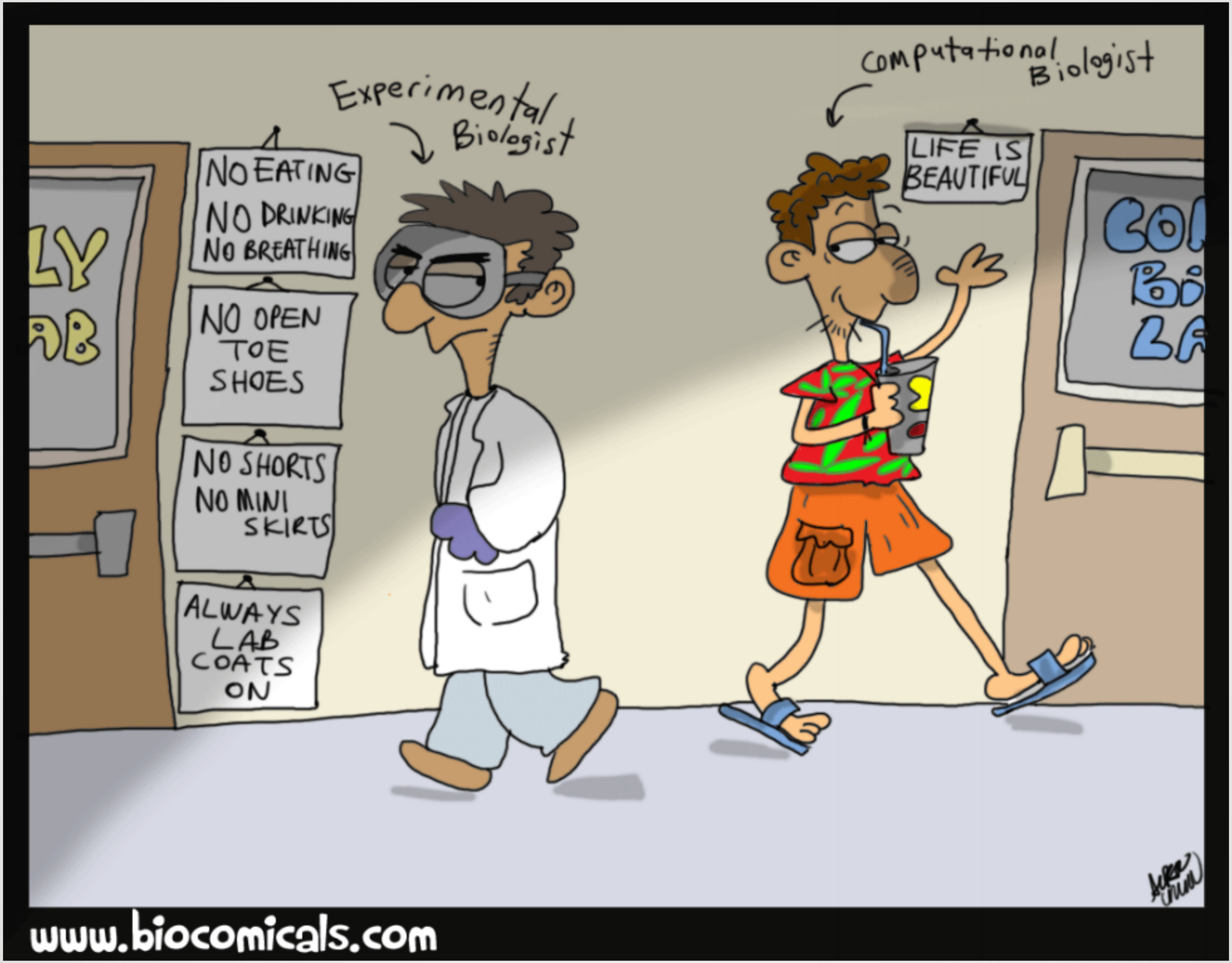
Central dogma of molecular biology

DNA → RNA → protein → cellular phenotype

Central dogma of genomics

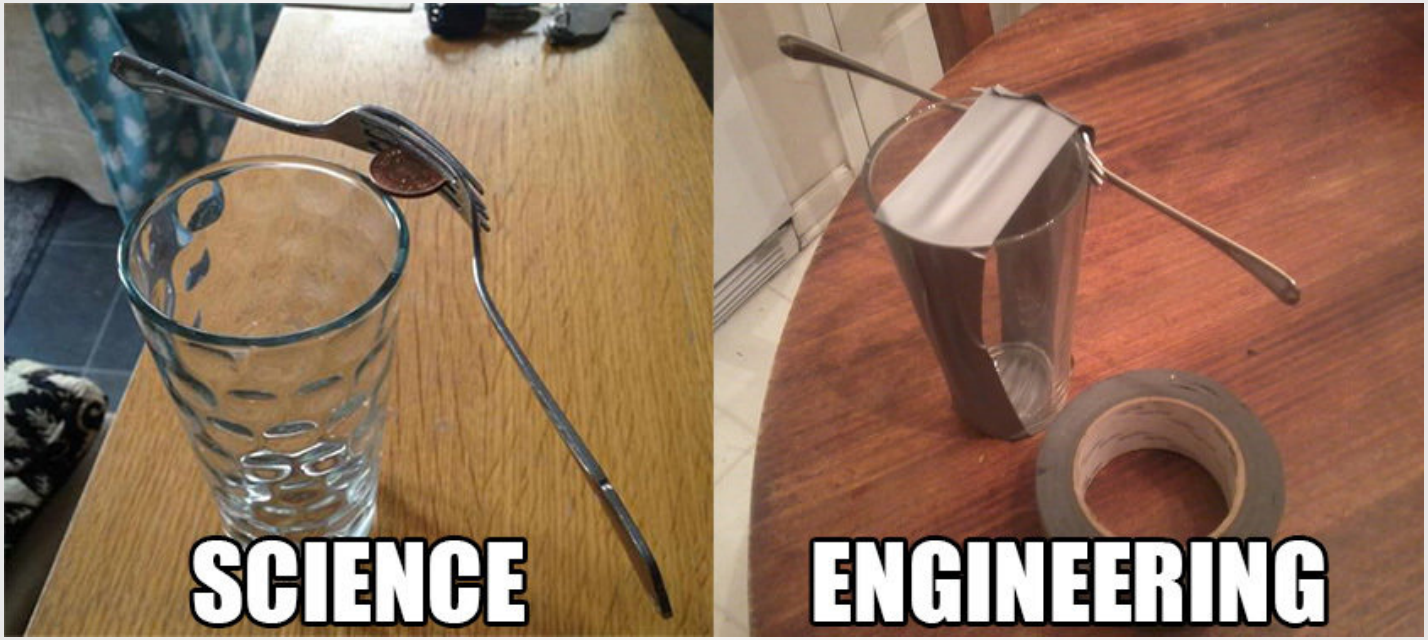
genome → transcriptome → proteome → cellular phenotype

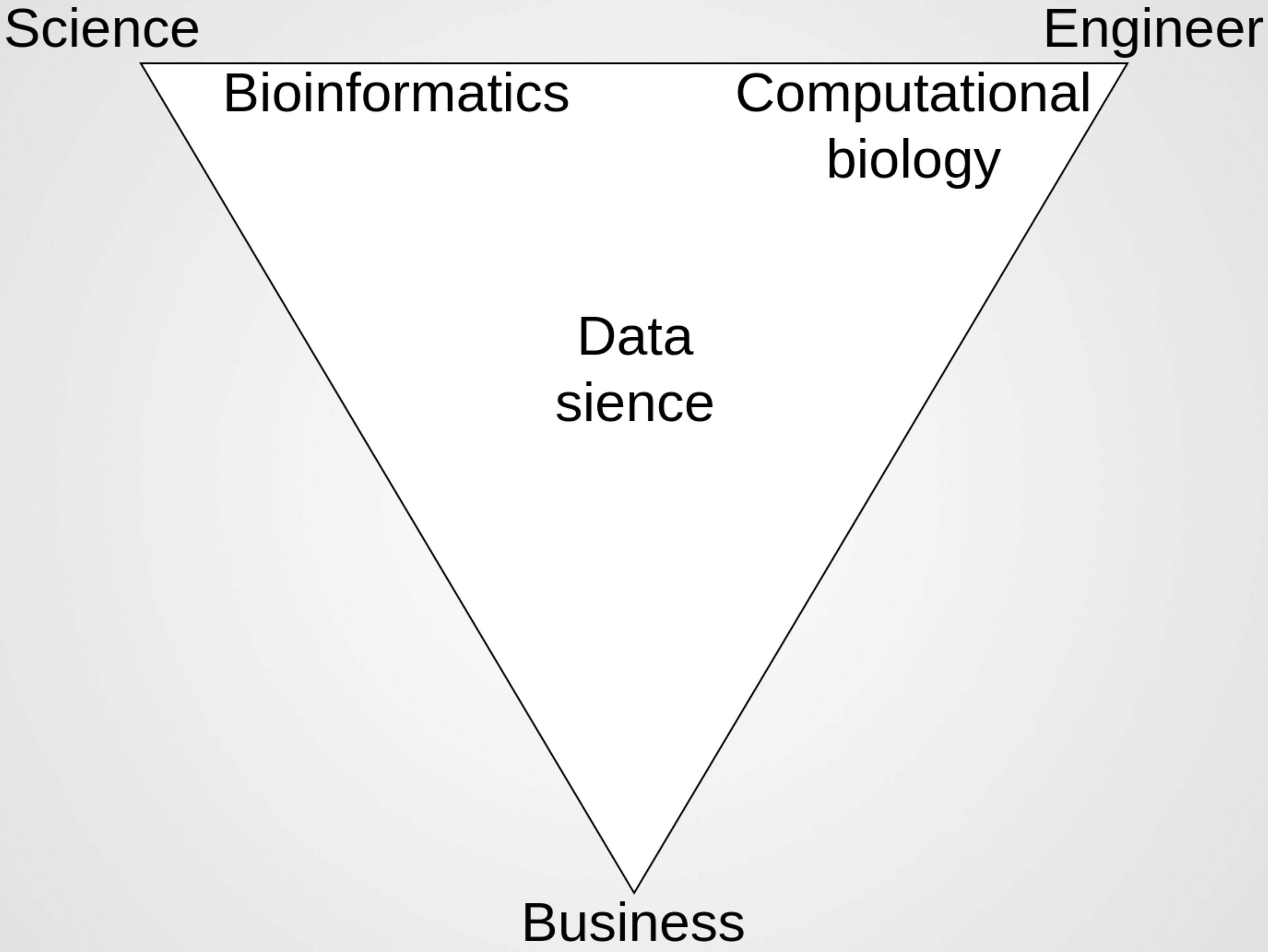
WET LAB - DRY LAB



DATA SCIENCE?

SCIENCE VS ENGINEERING





Science

Engineer

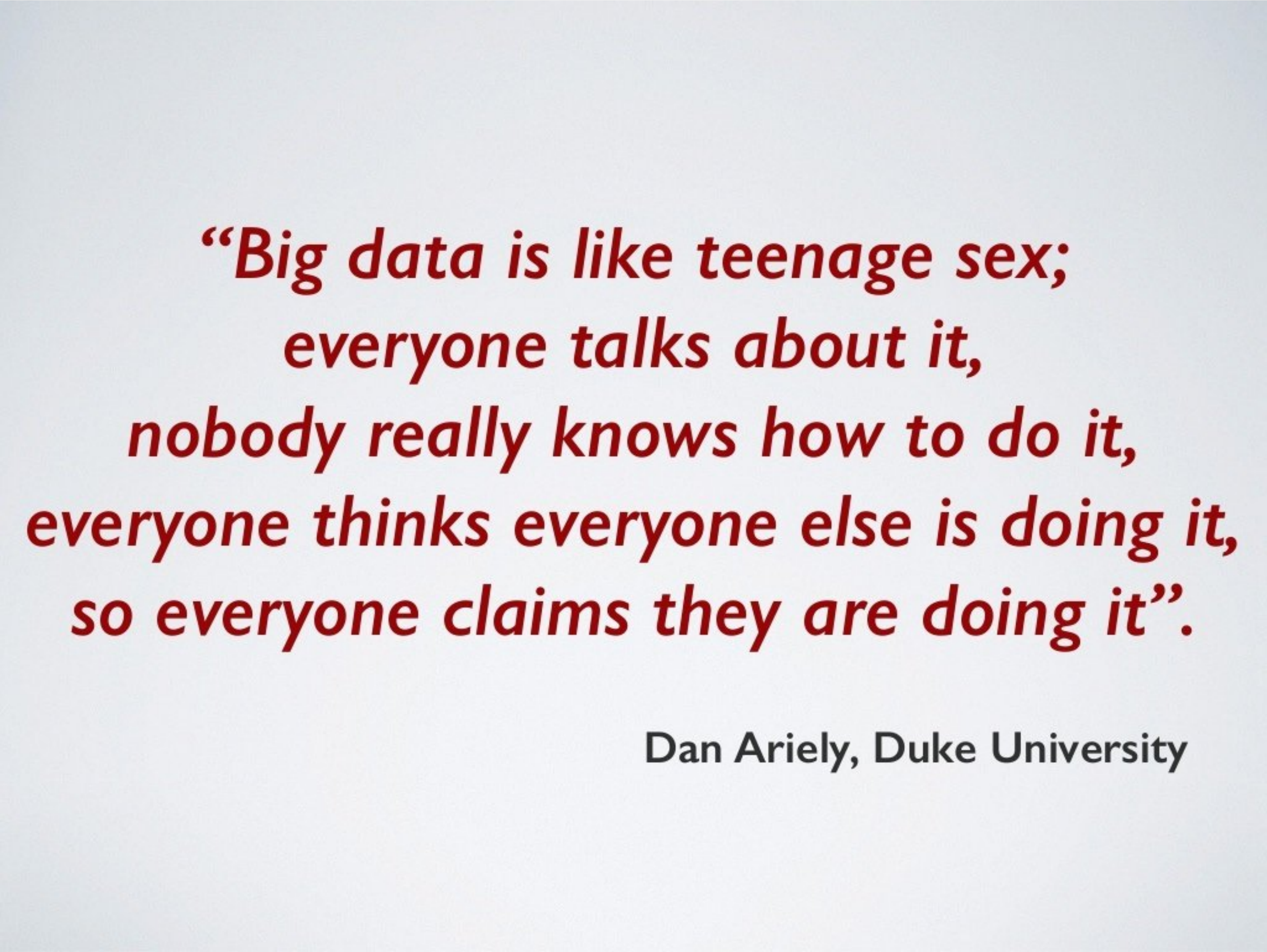
Bioinformatics

Computational biology

Data sience

Business

BIG DATA

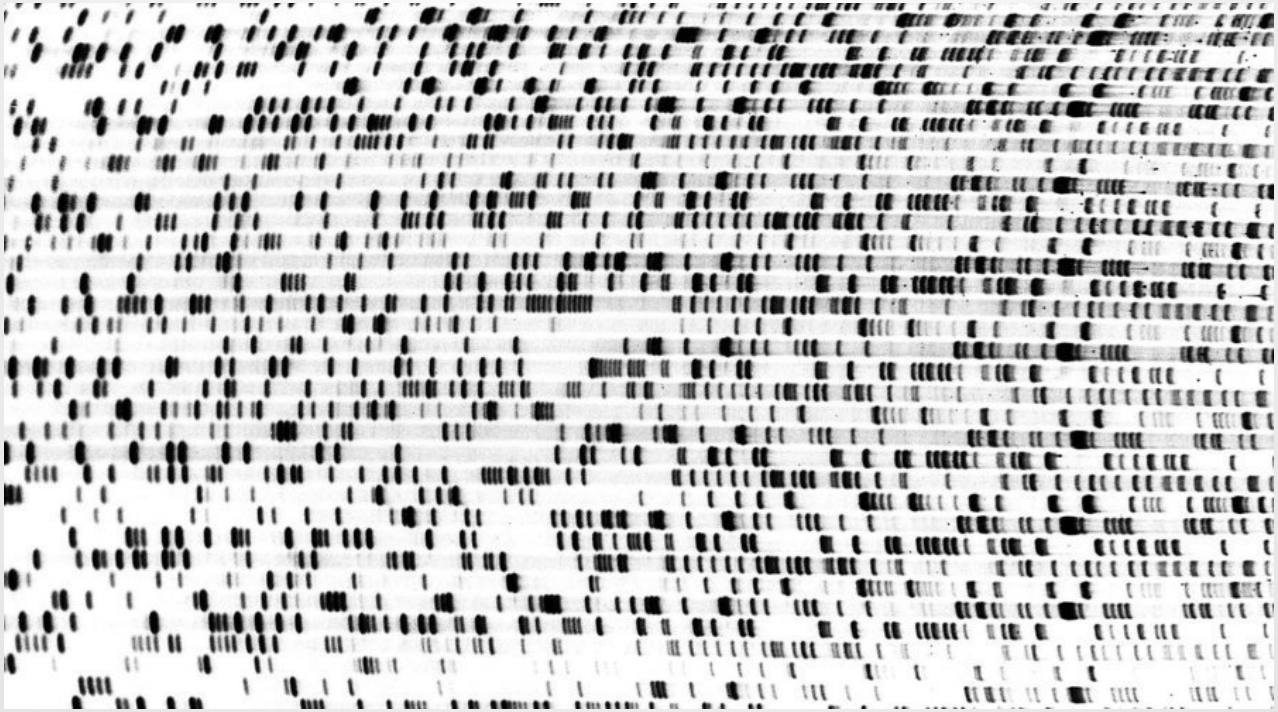


"Big data is like teenage sex;
everyone talks about it,
nobody really knows how to do it,
everyone thinks everyone else is doing it,
so everyone claims they are doing it".

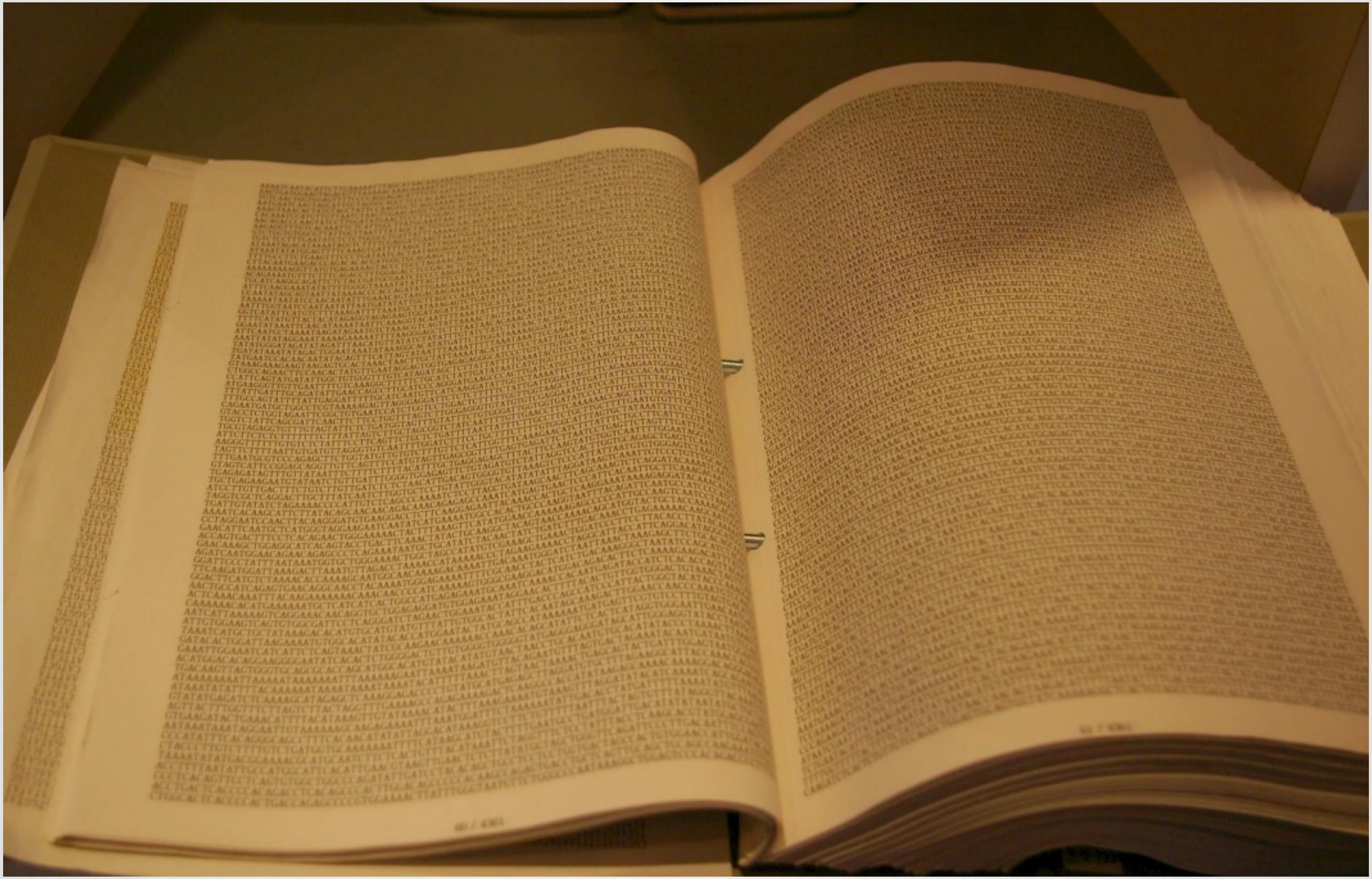
Dan Ariely, Duke University

THE SUBJECT OF BIOINFORMATICS

Classic bioinformatics

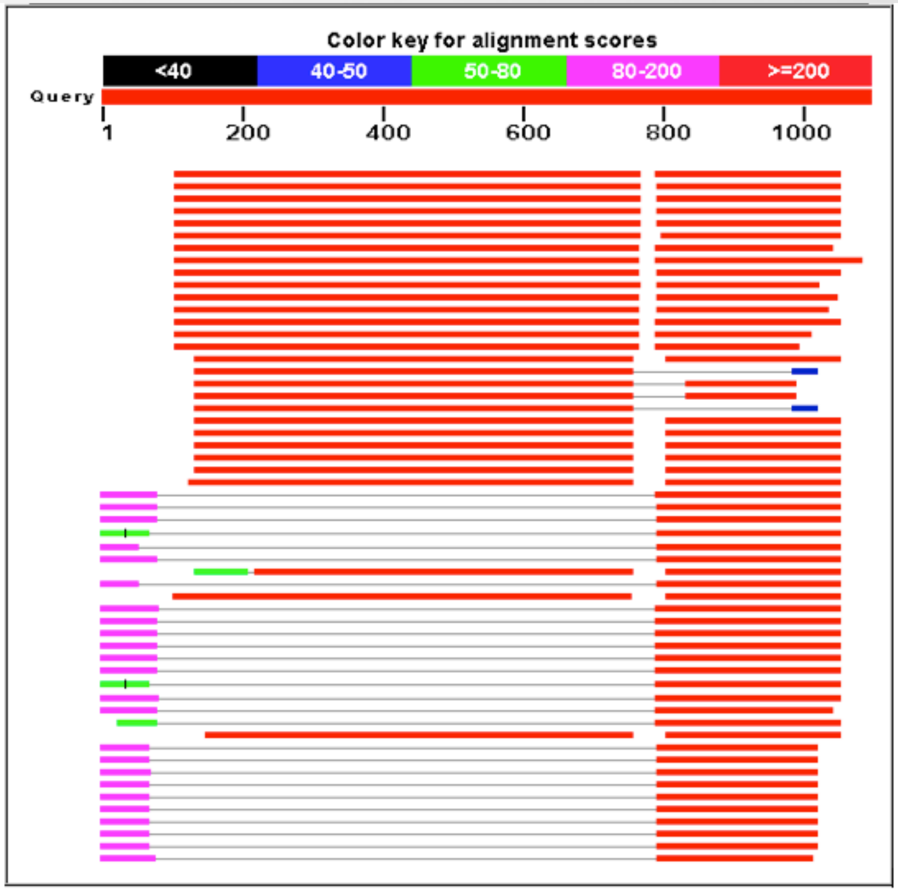


WITHOUT COMPUTER



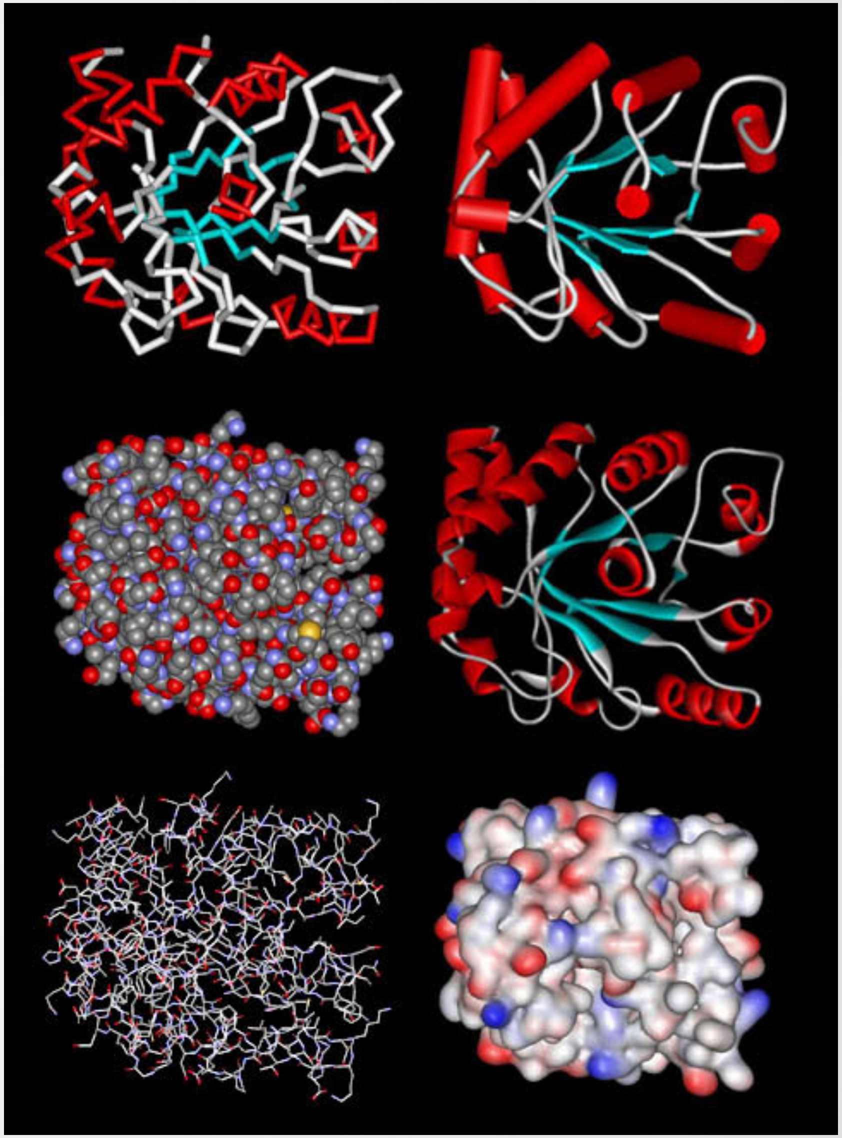
CLASSIC BIOINFORMATICS

- Sequence alignment
- Statistical analysis (e.g. CG ratio, gene length)
- Genome annotation:
  - ORF, gene prediction
  - promoter analysis
- Sequence database
- Sequence searching



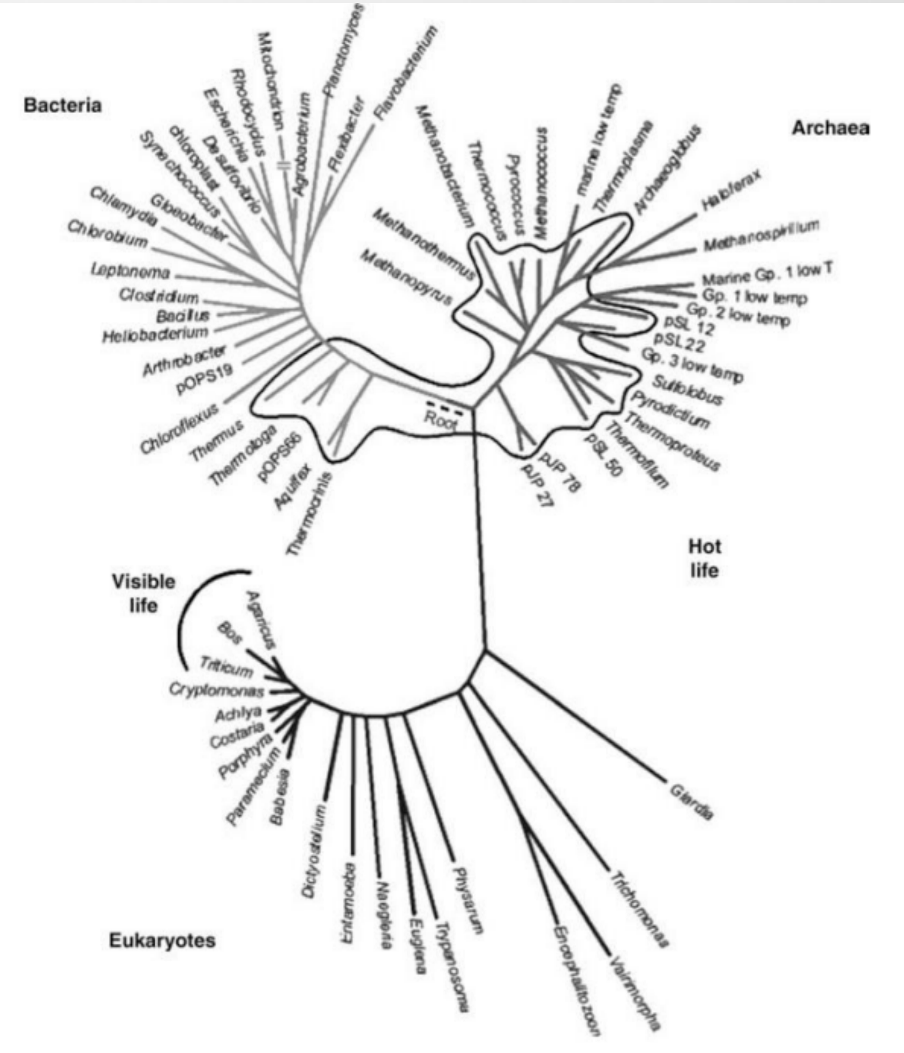
CLASSIC BIOINFORMATICS

- 3D structure of macro molecules
- Protein docking



CLASSIC BIOINFORMATICS

- Molecular phylogenetics



"OLD" AND "NEW" BIOLOGY

In the (near) past:

For researchers the greatest the challenge was to produce good quality data.

Today:

The biggest challenge for researchers to interpret a massive set of data is because biological data collection is done in bulk, has become industry-standard.

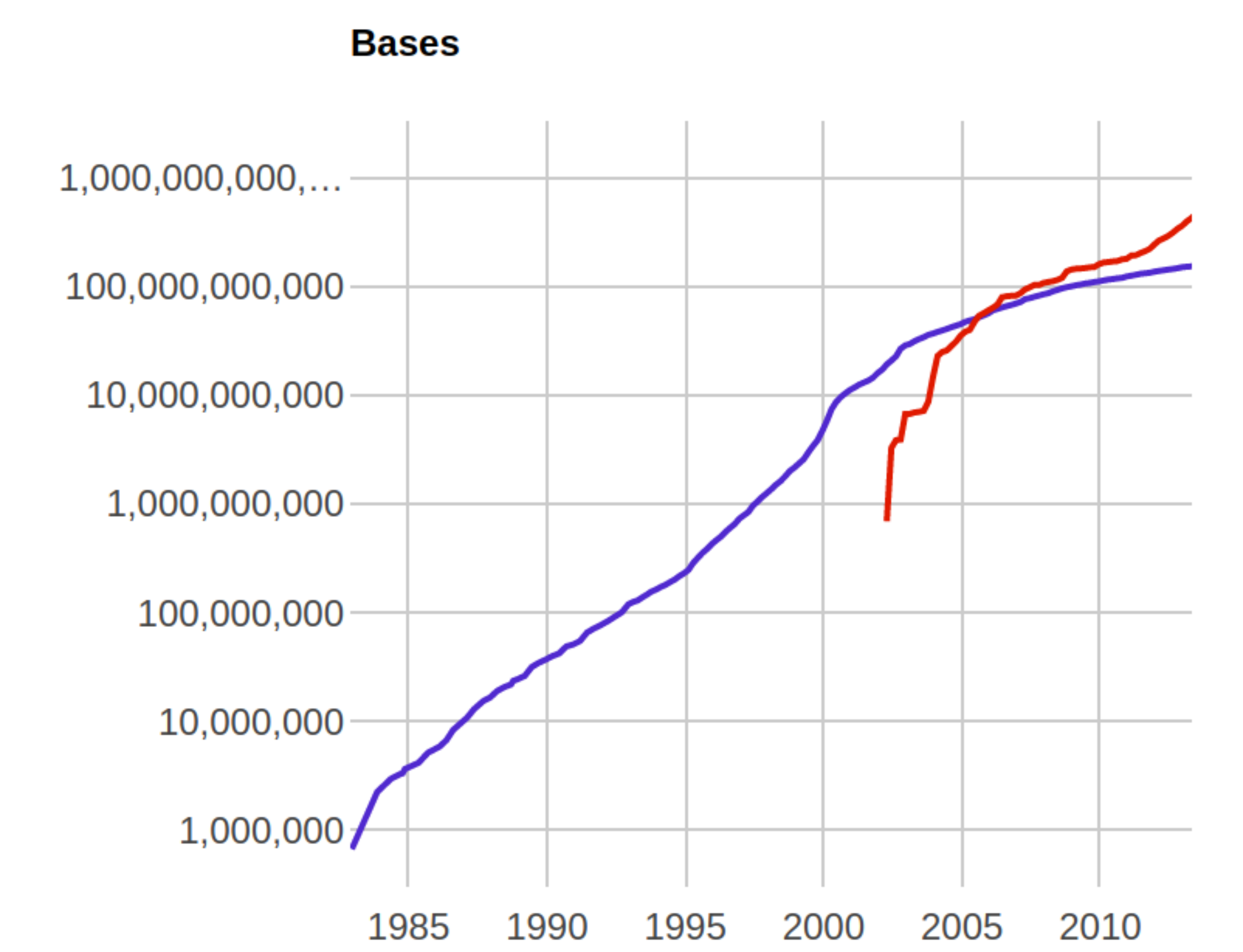
MODERN BIOINFORMATICS

- Post-genome era
- Next gen sequencing
- Comparative genomics
- Transcriptomics
- Proteomics
- Systems biology

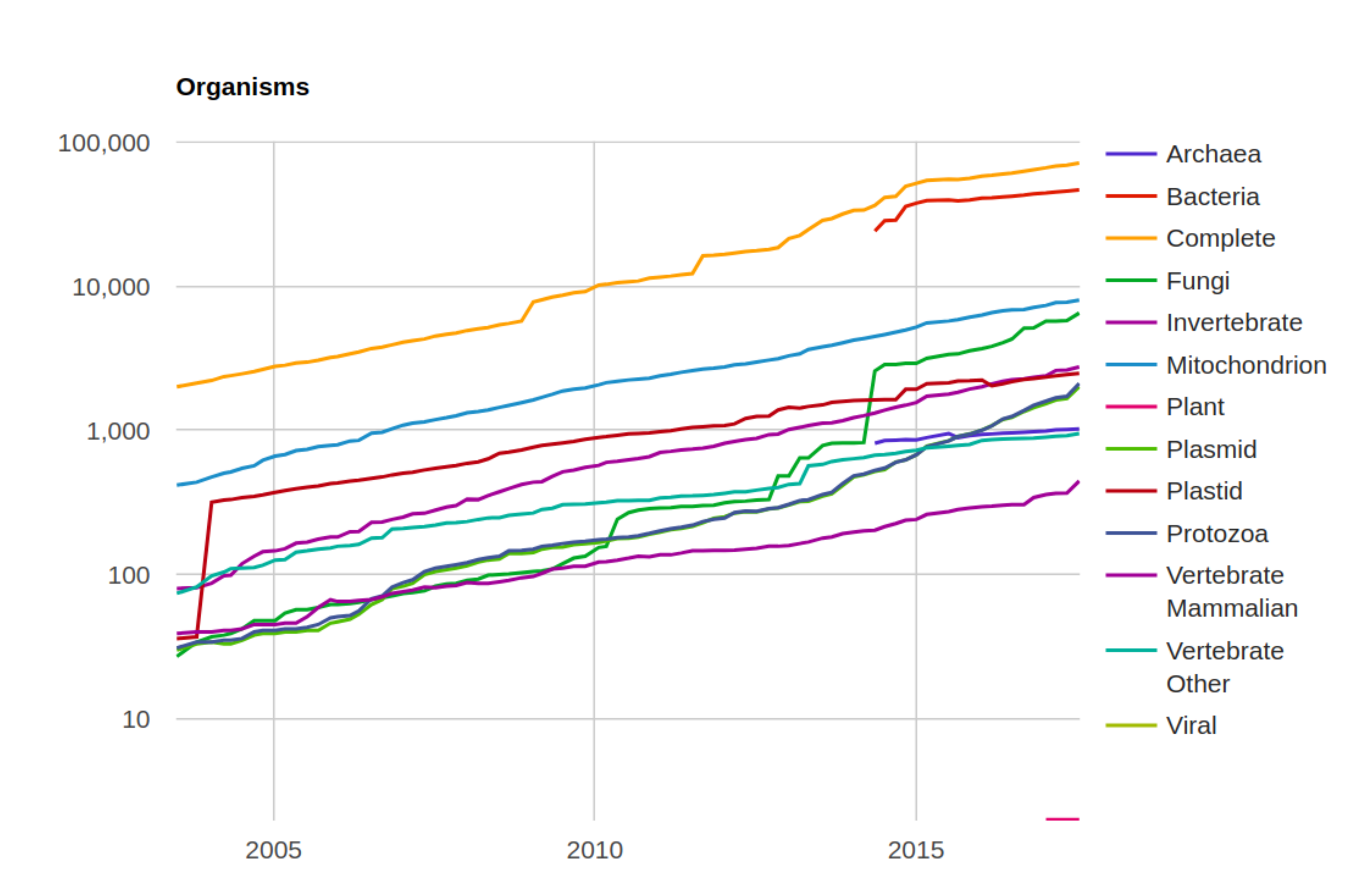
THE SUBJECT OF BIOINFORMATICS

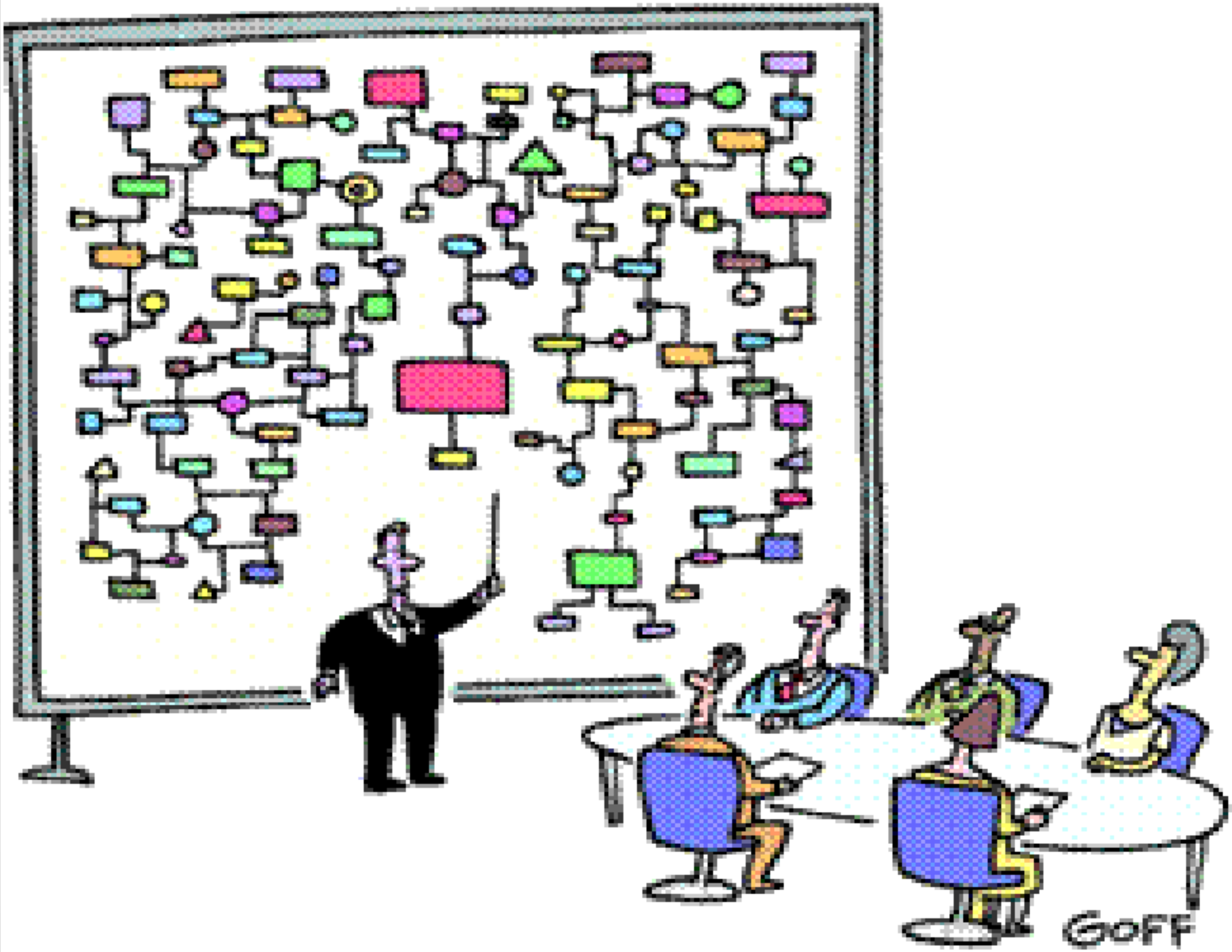
Data

NCBI GENEBANK: 2.6*10^{12} BASE



SEQUENCED GENOMES: 72000 ORGANISM

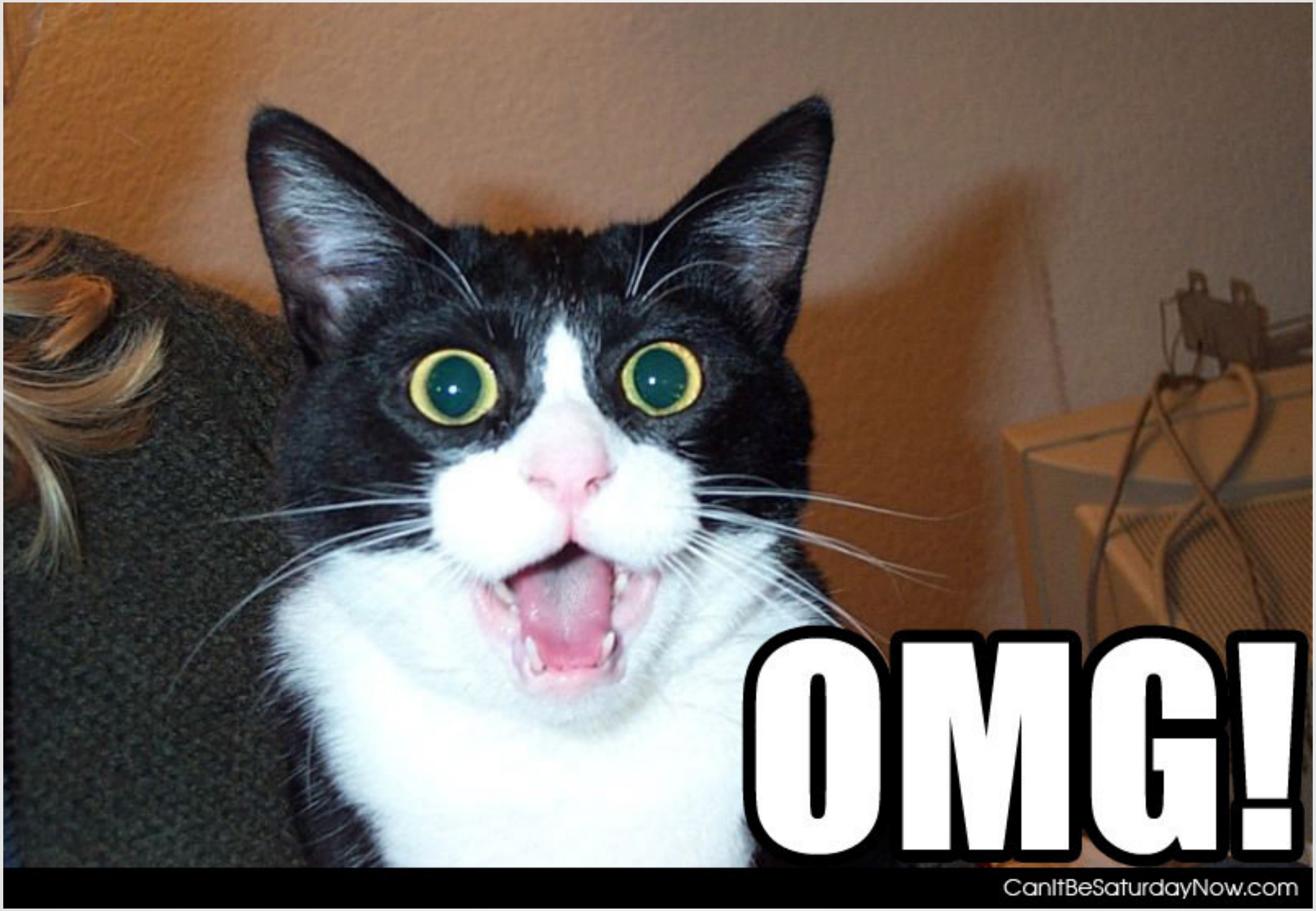




"And that's why we need a computer."

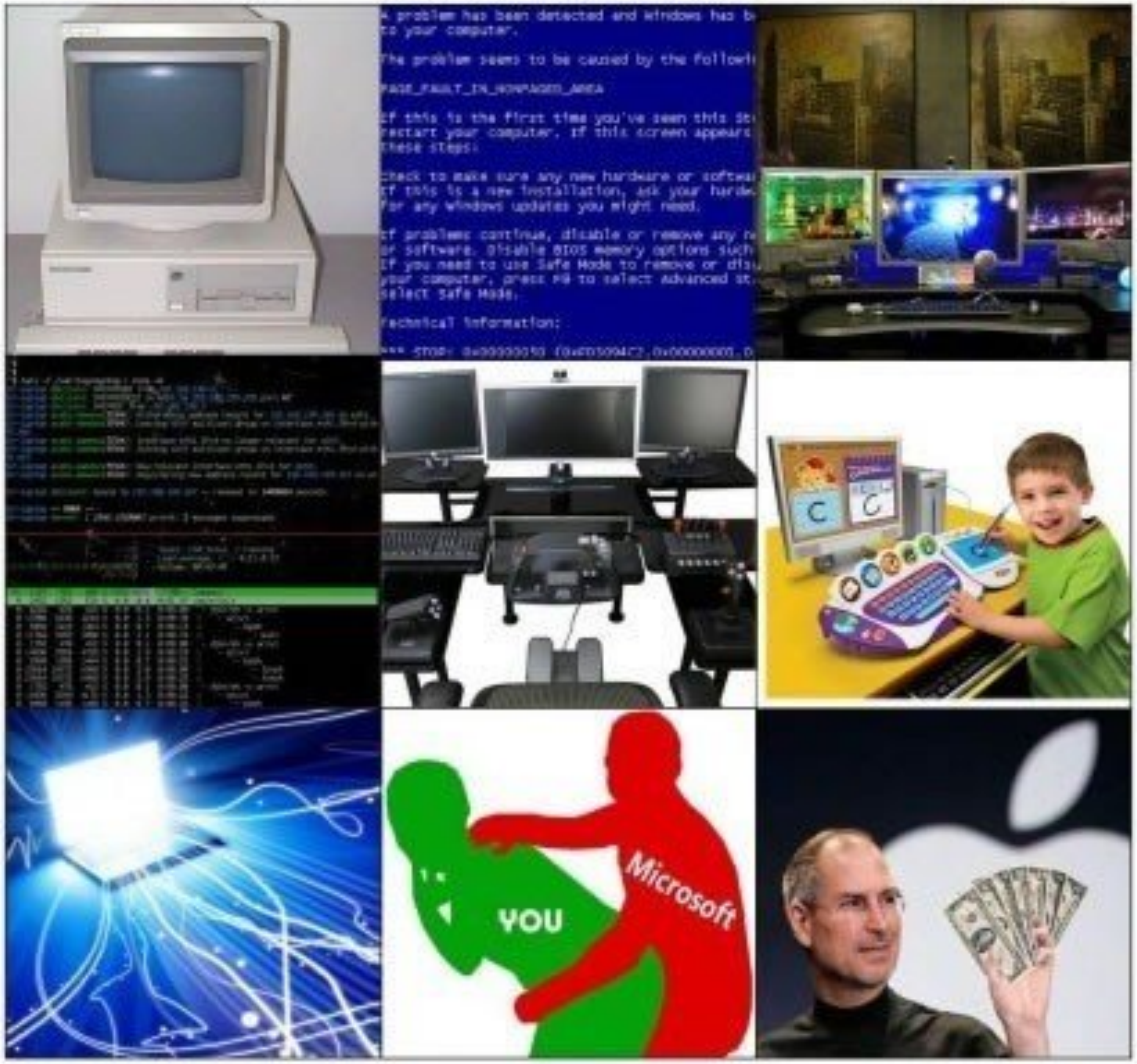
TOOLS WE'LL USE DURING THIS COURSE

- Database querying, searching
- Data managing in text files and table
- Web services
- Graphical programs
- 3D structure modelling
- Network analysis and visualization tools
- ...
- LINUX


OMG!
CanItBeSaturdayNow.com

Linux Windows Mac

as seen by...



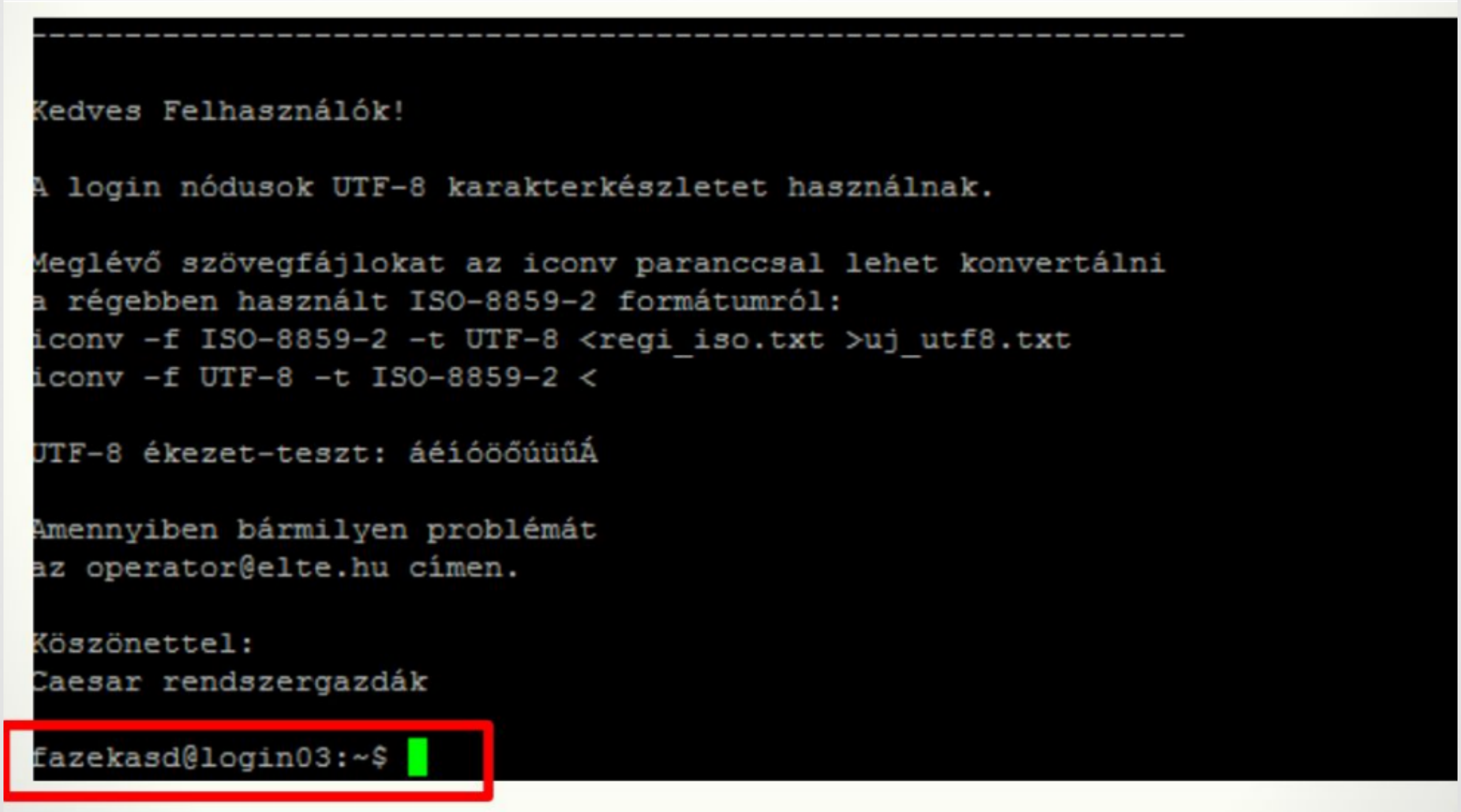
```
A problem has been detected and Windows has b
to your computer.

The problem seems to be caused by the followi

PAGE_FAULT_IN_NONPAGED_AREA

If this is the first time you've seen this St
restart your computer. If this screen appears
these steps:

Check to make sure any new hardware or softwa
If this is a new installation, ask your hardw
for any Windows updates you might need.

If problems continue, disable or remove any n
or software. Disable BIOS memory options such
If you need to use Safe Mode to remove or dis
your computer, press F8 to select Advanced St
select Safe Mode.

Technical information:

*** STOP: 0x00000050 (0xFD3094C2,0x00000001,0
```

Mac Fanboys

Windows Fanboys

Linux Fanboys

LINUX COMMAND LINE

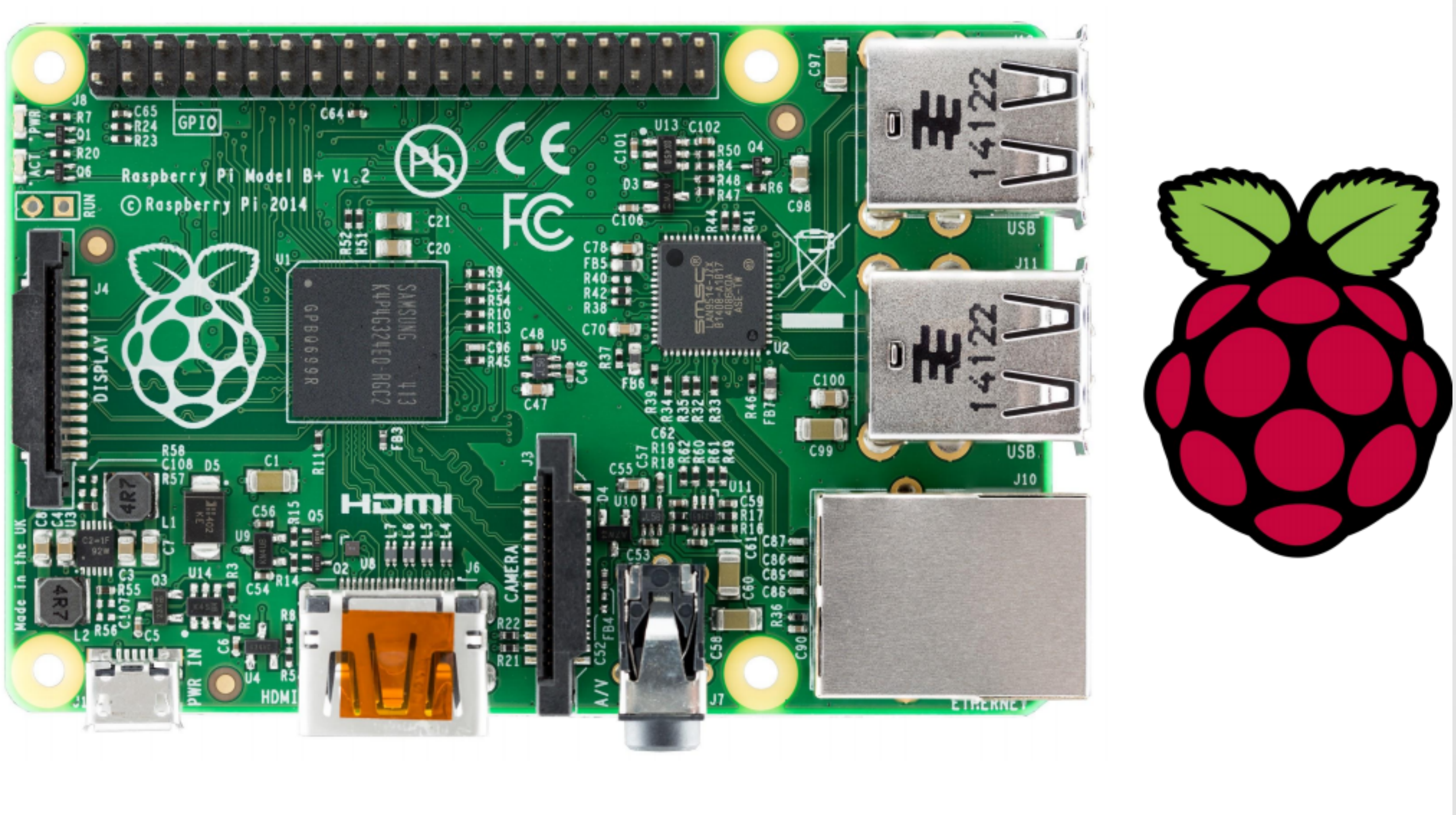


```
-------------------------------------------------------------

Kedves Felhasználók!

A login nódusok UTF-8 karakterkészletet használnak.

Meglévő szövegfájlokat az iconv paranccsal lehet konvertálni
a régebben használt ISO-8859-2 formátumról:
iconv -f ISO-8859-2 -t UTF-8 <regi_iso.txt >uj_utf8.txt
iconv -f UTF-8 -t ISO-8859-2 <

UTF-8 ékezet-teszt: áéíóöőúüűÁ

Amennyiben bármilyen problémát
az operator@elte.hu címen.

Köszönettel:
Caesar rendszergazdák

fazekasd@login03:~$
```

LINUX

- Distributions
- Package managers
- Kernel
- Shell
- File system
- Graphical user interface
- Remote access

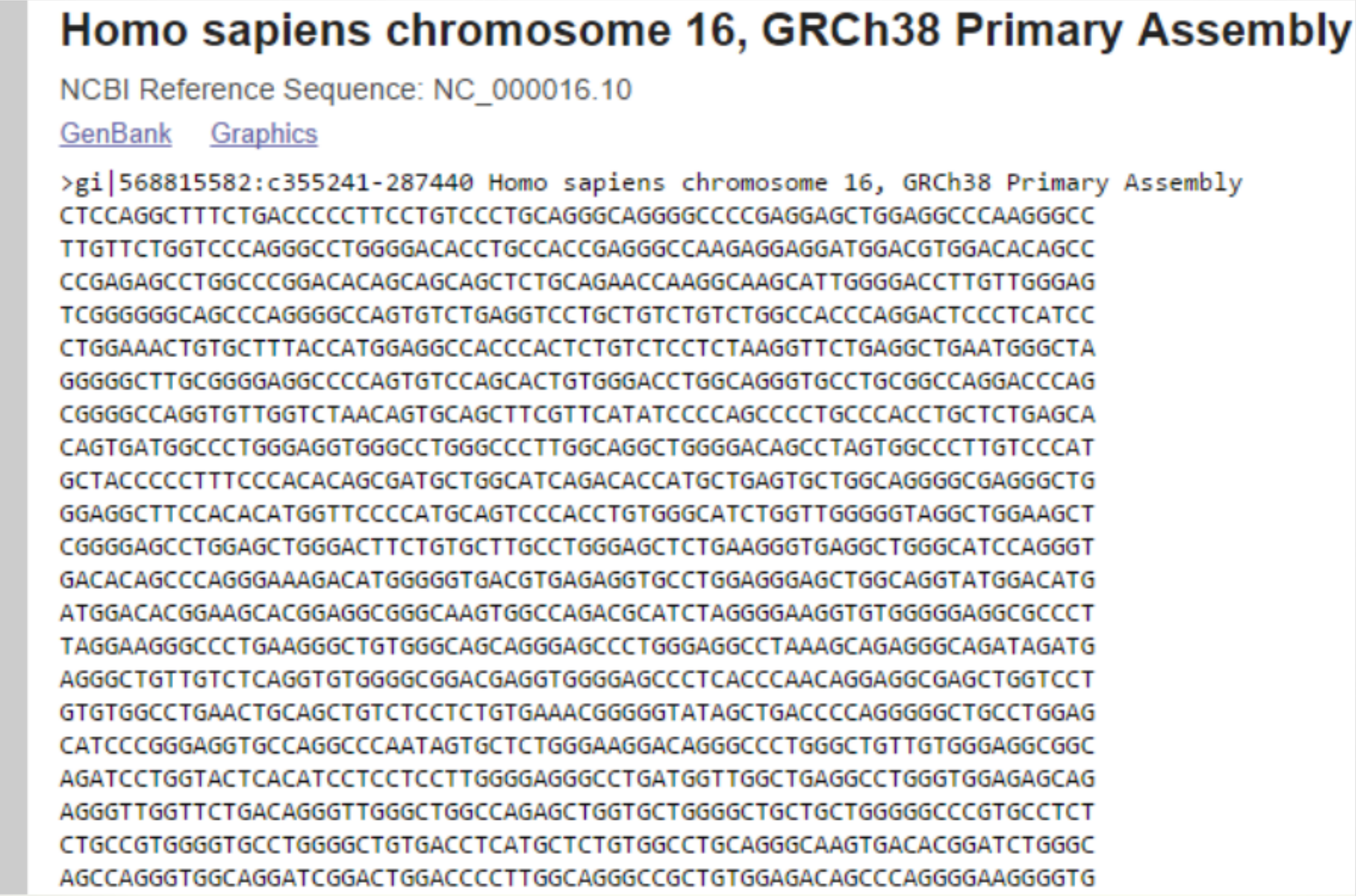
PCLAB


GPIO
Raspberry Pi Model B+ V1.2
©Raspberry Pi 2014
DISPLAY
HDMI
CAMERA
A/V
PWR IN
USB
USB
ETHERNET
Made in the UK

DATA STORING

SEQUENCE - FLAT FILE

SEQUENCE - FASTA FILE



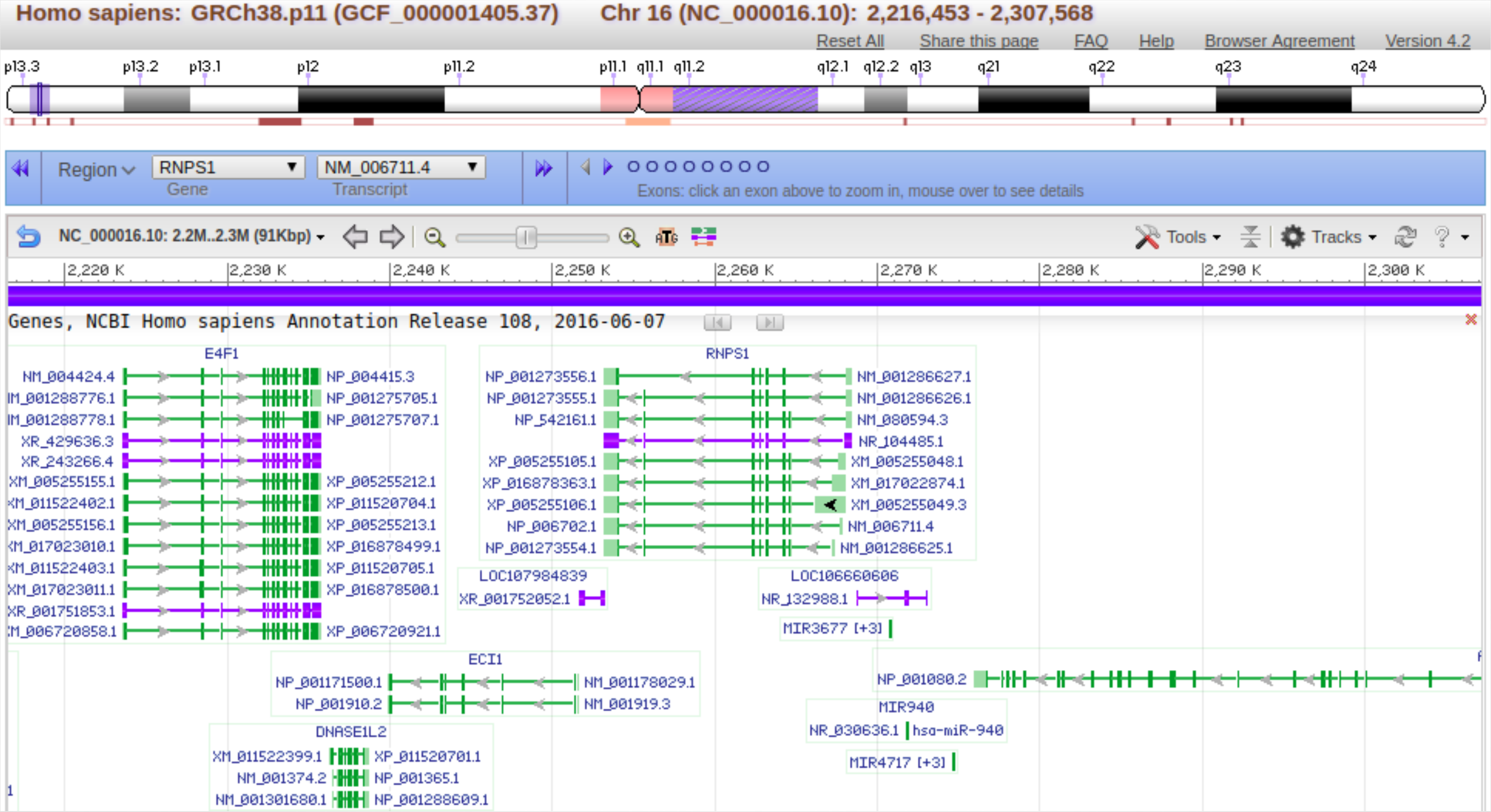
Homo sapiens chromosome 16, GRCh38 Primary Assembly

NCBI Reference Sequence: NC_000016.10

GenBank Graphics

```
>gi|568815582:c355241-287440 Homo sapiens chromosome 16, GRCh38 Primary Assembly
CTCCAGGCTTTCTGACCCCCTTCCTGTCCCTGCAGGGCAGGGGCCCCGAGGAGCTGGAGGCCCAAGGGCC
TTGTTCTGGTCCCAGGGCCTGGGGACACCTGCCACCGAGGGCCAAGAGGAGGATGGACGTGGACACAGCC
CCGAGAGCCTGGCCCGGACACAGCAGCAGCTCTGCAGAACCAAGGCAAGCATTGGGGACCTTGTTGGGAG
TCGGGGGGCAGCCCAGGGGCCAGTGTCTGAGGTCCTGCTGTCTGTCTGGCCACCCAGGACTCCCTCATCC
CTGGAAACTGTGCTTTACCATGGAGGCCACCCACTCTGTCTCCTCTAAGGTTCTGAGGCTGAATGGGCTA
GGGGGCTTGCGGGGAGGCCCCAGTGTCCAGCACTGTGGGACCTGGCAGGGTGCCTGCGGCCAGGACCCAG
CGGGGCCAGGTGTTGGTCTAACAGTGCAGCTTCGTTCATATCCCCAGCCCCTGCCCACCTGCTCTGAGCA
CAGTGATGGCCCTGGGAGGTGGGCCTGGGCCCTTGGCAGGCTGGGGACAGCCTAGTGGCCCTTGTCCCAT
GCTACCCCCTTTCCCACACAGCGATGCTGGCATCAGACACCATGCTGAGTGCTGGCAGGGGCGAGGGCTG
GGAGGCTTCCACACATGGTTCCCCATGCAGTCCCACCTGTGGGCATCTGGTTGGGGGTAGGCTGGAAGCT
CGGGGAGCCTGGAGCTGGGACTTCTGTGCTTGCCTGGGAGCTCTGAAGGGTGAGGCTGGGCATCCAGGGT
GACACAGCCCAGGGAAAGACATGGGGGTGACGTGAGAGGTGCCTGGAGGGAGCTGGCAGGTATGGACATG
ATGGACACGGAAGCACGGAGGCGGGCAAGTGGCCAGACGCATCTAGGGGAAGGTGTGGGGGAGGCGCCCT
TAGGAAGGGCCCTGAAGGGCTGTGGGCAGCAGGGAGCCCTGGGAGGCCTAAAGCAGAGGGCAGATAGATG
AGGGCTGTTGTCTCAGGTGTGGGGCGGACGAGGTGGGGAGCCCTCACCCAACAGGAGGCGAGCTGGTCCT
GTGTGGCCTGAACTGCAGCTGTCTCCTCTGTGAAACGGGGGTATAGCTGACCCCAGGGGGCTGCCTGGAG
CATCCCGGGAGGTGCCAGGCCCAATAGTGCTCTGGGAAGGACAGGGCCCTGGGCTGTTGTGGGAGGCGGC
AGATCCTGGTACTCACATCCTCCTCCTTGGGGAGGGCCTGATGGTTGGCTGAGGCCTGGGTGGAGAGCAG
AGGGTTGGTTCTGACAGGGTTGGGCTGGCCAGAGCTGGTGCTGGGGCTGCTGCTGGGGGCCCGTGCCTCT
CTGCCGTGGGGTGCCTGGGGCTGTGACCTCATGCTCTGTGGCCTGCAGGGCAAGTGACACGGATCTGGGC
AGCCAGGGTGGCAGGATCGGACTGGACCCCTTGGCAGGGCCGCTGTGGAGACAGCCCAGGGGAAGGGGTG
```

SEQUENCE - WITH ANNOTATIONS

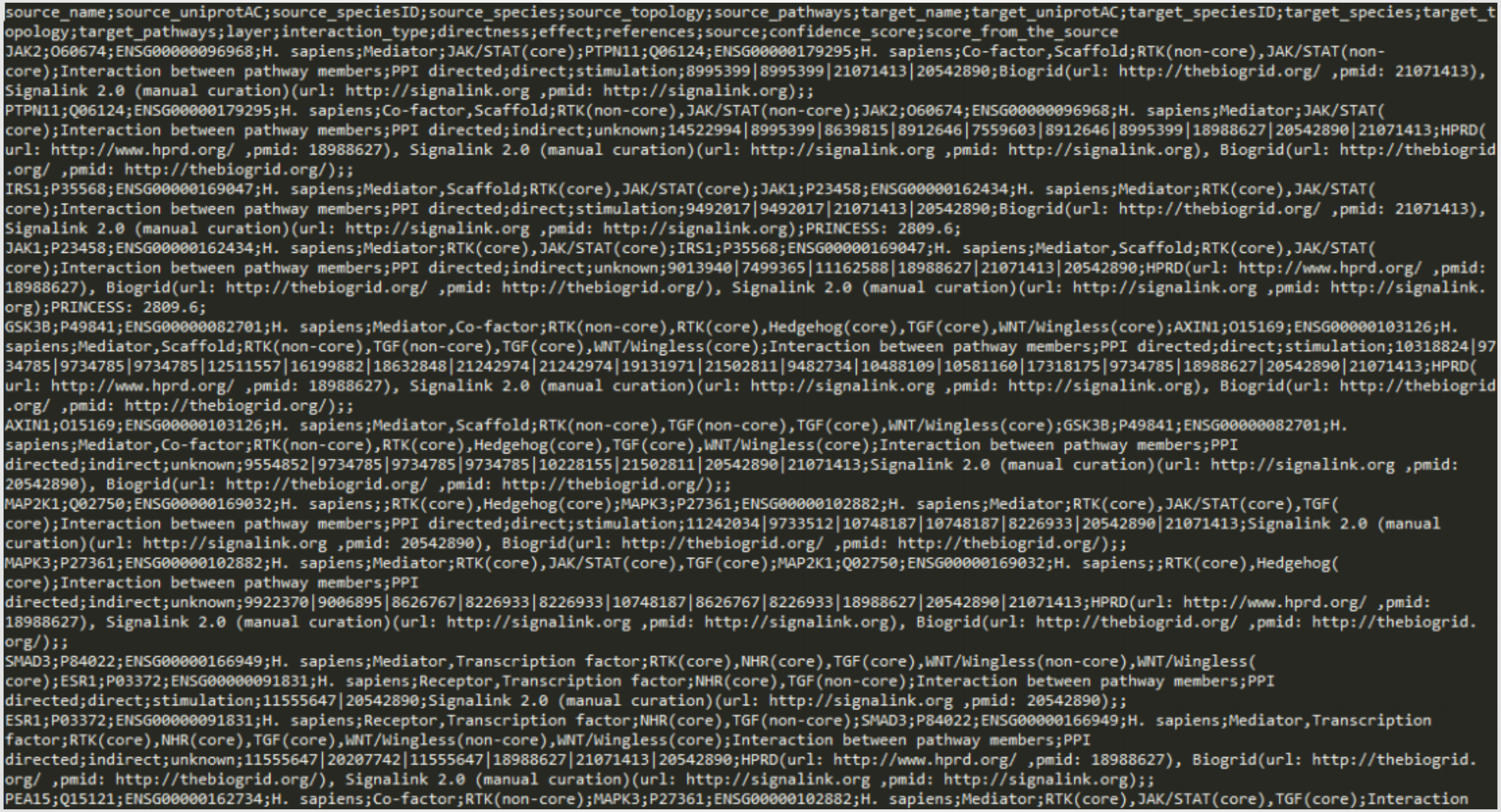


Homo sapiens: GRCh38.p11 (GCF_000001405.37) Chr 16 (NC_000016.10): 2,216,453 - 2,307,568

Exons: click an exon above to zoom in, mouse over to see details

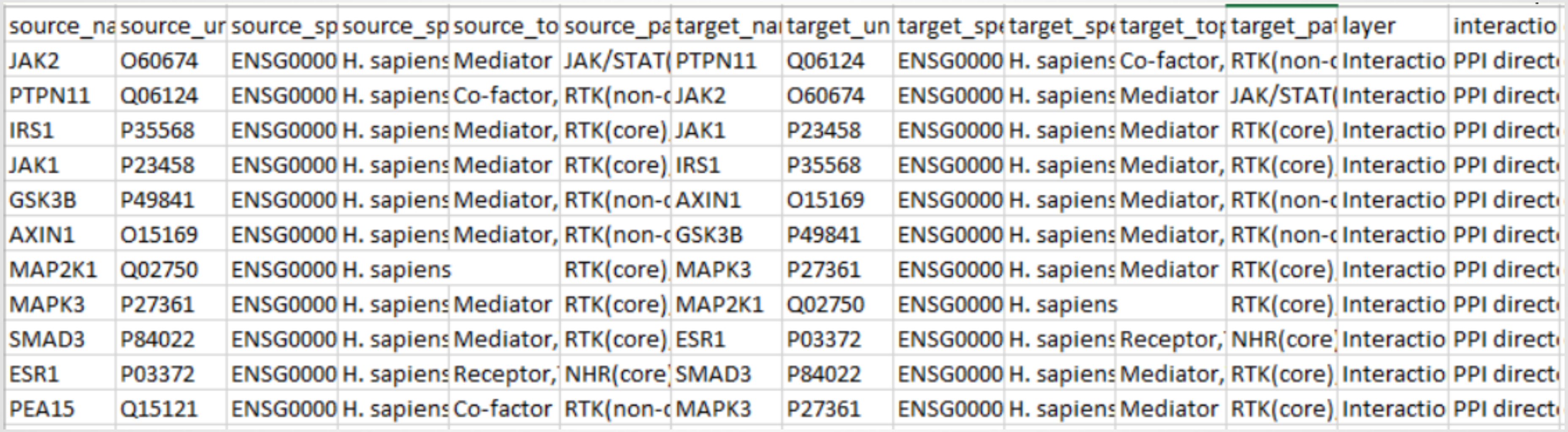
Genes, NCBI Homo sapiens Annotation Release 108, 2016-06-07

NETWORK - PLAIN TEXT



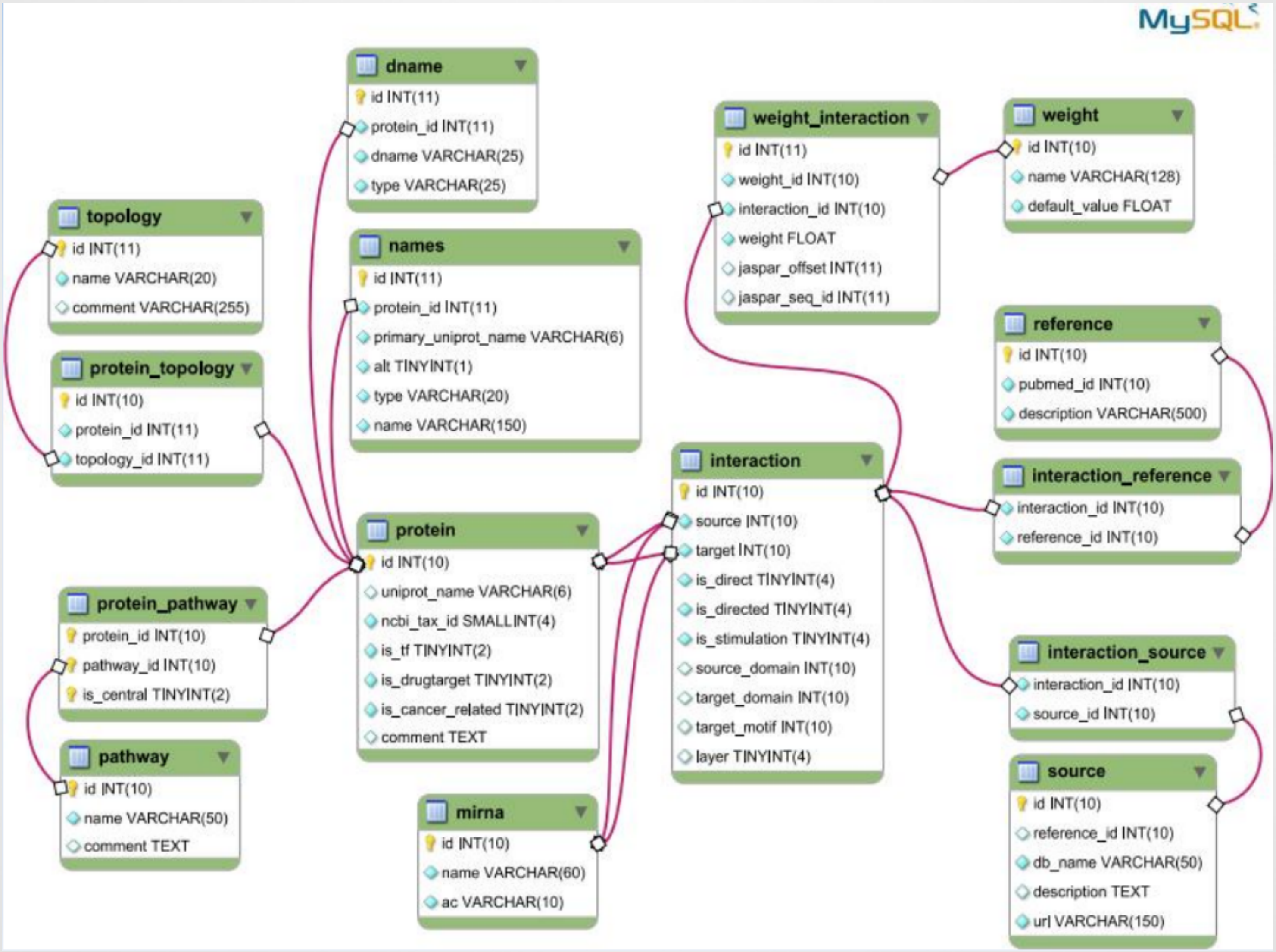
```
source_name;source_uniprotAC;source_speciesID;source_species;source_topology;source_pathways;target_name;target_uniprotAC;target_speciesID;target_species;target_t
opology;target_pathways;layer;interaction_type;directness;effect;references;source;confidence_score;score_from_the_source
JAK2;O60674;ENSG00000096968;H. sapiens;Mediator;JAK/STAT(core);PTPN11;Q06124;ENSG00000179295;H. sapiens;Co-factor,Scaffold;RTK(non-core),JAK/STAT(non-
core);Interaction between pathway members;PPI directed;direct;stimulation;8995399|8995399|21071413|20542890;Biogrid(url: http://thebiogrid.org/ ,pmid: 21071413),
Signalink 2.0 (manual curation)(url: http://signalink.org ,pmid: http://signalink.org);;
PTPN11;Q06124;ENSG00000179295;H. sapiens;Co-factor,Scaffold;RTK(non-core),JAK/STAT(non-core);JAK2;O60674;ENSG00000096968;H. sapiens;Mediator;JAK/STAT(
core);Interaction between pathway members;PPI directed;indirect;unknown;14522994|8995399|8639815|8912646|7559603|8912646|8995399|18988627|20542890|21071413;HPRD(
url: http://www.hprd.org/ ,pmid: 18988627), Signalink 2.0 (manual curation)(url: http://signalink.org ,pmid: http://signalink.org), Biogrid(url: http://thebiogrid
.org/ ,pmid: http://thebiogrid.org/);;
IRS1;P35568;ENSG00000169047;H. sapiens;Mediator,Scaffold;RTK(core),JAK/STAT(core);JAK1;P23458;ENSG00000162434;H. sapiens;Mediator;RTK(core),JAK/STAT(
core);Interaction between pathway members;PPI directed;direct;stimulation;9492017|9492017|21071413|20542890;Biogrid(url: http://thebiogrid.org/ ,pmid: 21071413),
Signalink 2.0 (manual curation)(url: http://signalink.org ,pmid: http://signalink.org);PRINCESS: 2809.6;
JAK1;P23458;ENSG00000162434;H. sapiens;Mediator;RTK(core),JAK/STAT(core);IRS1;P35568;ENSG00000169047;H. sapiens;Mediator,Scaffold;RTK(core),JAK/STAT(
core);Interaction between pathway members;PPI directed;indirect;unknown;9013940|7499365|11162588|18988627|21071413|20542890;HPRD(url: http://www.hprd.org/ ,pmid:
18988627), Biogrid(url: http://thebiogrid.org/ ,pmid: http://thebiogrid.org/), Signalink 2.0 (manual curation)(url: http://signalink.org ,pmid: http://signalink.
org);PRINCESS: 2809.6;
GSK3B;P49841;ENSG00000082701;H. sapiens;Mediator,Co-factor;RTK(non-core),RTK(core),Hedgehog(core),TGF(core),WNT/Wingless(core);AXIN1;O15169;ENSG00000103126;H.
sapiens;Mediator,Scaffold;RTK(non-core),TGF(non-core),TGF(core),WNT/Wingless(core);Interaction between pathway members;PPI directed;direct;stimulation;10318824|97
34785|9734785|9734785|12511557|16199882|18632848|21242974|21242974|19131971|21502811|9482734|10488109|10581160|17318175|9734785|18988627|20542890|21071413;HPRD(
url: http://www.hprd.org/ ,pmid: 18988627), Signalink 2.0 (manual curation)(url: http://signalink.org ,pmid: http://signalink.org), Biogrid(url: http://thebiogrid
.org/ ,pmid: http://thebiogrid.org/);;
AXIN1;O15169;ENSG00000103126;H. sapiens;Mediator,Scaffold;RTK(non-core),TGF(non-core),TGF(core),WNT/Wingless(core);GSK3B;P49841;ENSG00000082701;H.
sapiens;Mediator,Co-factor;RTK(non-core),RTK(core),Hedgehog(core),TGF(core),WNT/Wingless(core);Interaction between pathway members;PPI
directed;indirect;unknown;9554852|9734785|9734785|9734785|10228155|21502811|20542890|21071413;Signalink 2.0 (manual curation)(url: http://signalink.org ,pmid:
20542890), Biogrid(url: http://thebiogrid.org/ ,pmid: http://thebiogrid.org/);;
MAP2K1;Q02750;ENSG00000169032;H. sapiens;;RTK(core),Hedgehog(core);MAPK3;P27361;ENSG00000102882;H. sapiens;Mediator;RTK(core),JAK/STAT(core),TGF(
core);Interaction between pathway members;PPI directed;direct;stimulation;11242034|9733512|10748187|10748187|8226933|20542890|21071413;Signalink 2.0 (manual
curation)(url: http://signalink.org ,pmid: 20542890), Biogrid(url: http://thebiogrid.org/ ,pmid: http://thebiogrid.org/);;
MAPK3;P27361;ENSG00000102882;H. sapiens;Mediator;RTK(core),JAK/STAT(core),TGF(core);MAP2K1;Q02750;ENSG00000169032;H. sapiens;;RTK(core),Hedgehog(
core);Interaction between pathway members;PPI
directed;indirect;unknown;9922370|9006895|8626767|8226933|8226933|10748187|8626767|8226933|18988627|20542890|21071413;HPRD(url: http://www.hprd.org/ ,pmid:
18988627), Signalink 2.0 (manual curation)(url: http://signalink.org ,pmid: http://signalink.org), Biogrid(url: http://thebiogrid.org/ ,pmid: http://thebiogrid.
org/);;
SMAD3;P84022;ENSG00000166949;H. sapiens;Mediator,Transcription factor;RTK(core),NHR(core),TGF(core),WNT/Wingless(non-core),WNT/Wingless(
core);ESR1;P03372;ENSG00000091831;H. sapiens;Receptor,Transcription factor;NHR(core),TGF(non-core);Interaction between pathway members;PPI
directed;direct;stimulation;11555647|20542890;Signalink 2.0 (manual curation)(url: http://signalink.org ,pmid: 20542890);;
ESR1;P03372;ENSG00000091831;H. sapiens;Receptor,Transcription factor;NHR(core),TGF(non-core);SMAD3;P84022;ENSG00000166949;H. sapiens;Mediator,Transcription
factor;RTK(core),NHR(core),TGF(core),WNT/Wingless(non-core),WNT/Wingless(core);Interaction between pathway members;PPI
directed;indirect;unknown;11555647|20207742|11555647|18988627|21071413|20542890;HPRD(url: http://www.hprd.org/ ,pmid: 18988627), Biogrid(url: http://thebiogrid.
org/ ,pmid: http://thebiogrid.org/), Signalink 2.0 (manual curation)(url: http://signalink.org ,pmid: http://signalink.org);;
PEA15;Q15121;ENSG00000162734;H. sapiens;Co-factor;RTK(non-core);MAPK3;P27361;ENSG00000102882;H. sapiens;Mediator;RTK(core),JAK/STAT(core),TGF(core);Interaction
```

NETWORK - TABLE



source_na	source_ur	source_sp	source_sp	source_to	source_pa	target_nai	target_un	target_spe	target_spe	target_top	target_pat	layer	interactio
JAK2	O60674	ENSG0000	H. sapiens	Mediator	JAK/STAT(	PTPN11	Q06124	ENSG0000	H. sapiens	Co-factor,	RTK(non-c	Interactio	PPI direct
PTPN11	Q06124	ENSG0000	H. sapiens	Co-factor,	RTK(non-c	JAK2	O60674	ENSG0000	H. sapiens	Mediator	JAK/STAT(	Interactio	PPI direct
IRS1	P35568	ENSG0000	H. sapiens	Mediator,	RTK(core)	JAK1	P23458	ENSG0000	H. sapiens	Mediator	RTK(core)	Interactio	PPI direct
JAK1	P23458	ENSG0000	H. sapiens	Mediator	RTK(core)	IRS1	P35568	ENSG0000	H. sapiens	Mediator,	RTK(core)	Interactio	PPI direct
GSK3B	P49841	ENSG0000	H. sapiens	Mediator,	RTK(non-c	AXIN1	O15169	ENSG0000	H. sapiens	Mediator,	RTK(non-c	Interactio	PPI direct
AXIN1	O15169	ENSG0000	H. sapiens	Mediator,	RTK(non-c	GSK3B	P49841	ENSG0000	H. sapiens	Mediator,	RTK(non-c	Interactio	PPI direct
MAP2K1	Q02750	ENSG0000	H. sapiens		RTK(core)	MAPK3	P27361	ENSG0000	H. sapiens	Mediator	RTK(core)	Interactio	PPI direct
MAPK3	P27361	ENSG0000	H. sapiens	Mediator	RTK(core)	MAP2K1	Q02750	ENSG0000	H. sapiens		RTK(core)	Interactio	PPI direct
SMAD3	P84022	ENSG0000	H. sapiens	Mediator,	RTK(core)	ESR1	P03372	ENSG0000	H. sapiens	Receptor,	NHR(core)	Interactio	PPI direct
ESR1	P03372	ENSG0000	H. sapiens	Receptor,	NHR(core)	SMAD3	P84022	ENSG0000	H. sapiens	Mediator,	RTK(core)	Interactio	PPI direct
PEA15	Q15121	ENSG0000	H. sapiens	Co-factor	RTK(non-c	MAPK3	P27361	ENSG0000	H. sapiens	Mediator	RTK(core)	Interactio	PPI direct

NETWORK - RELATIONAL DATABASE



dname
- id INT(11)
- protein_id INT(11)
- dname VARCHAR(25)
- type VARCHAR(25)

topology
- id INT(11)
- name VARCHAR(20)
- comment VARCHAR(255)

protein_topology
- id INT(10)
- protein_id INT(11)
- topology_id INT(11)

names
- id INT(11)
- protein_id INT(11)
- primary_uniprot_name VARCHAR(6)
- alt TINYINT(1)
- type VARCHAR(20)
- name VARCHAR(150)

weight_interaction
- id INT(11)
- weight_id INT(10)
- interaction_id INT(10)
- weight FLOAT
- jaspar_offset INT(11)
- jaspar_seq_id INT(11)

weight
- id INT(10)
- name VARCHAR(128)
- default_value FLOAT

reference
- id INT(10)
- pubmed_id INT(10)
- description VARCHAR(500)

interaction_reference
- interaction_id INT(10)
- reference_id INT(10)

interaction
- id INT(10)
- source INT(10)
- target INT(10)
- is_direct TINYINT(4)
- is_directed TINYINT(4)
- is_stimulation TINYINT(4)
- source_domain INT(10)
- target_domain INT(10)
- target_motif INT(10)
- layer TINYINT(4)

protein
- id INT(10)
- uniprot_name VARCHAR(6)
- ncbi_tax_id SMALLINT(4)
- is_tf TINYINT(2)
- is_drugtarget TINYINT(2)
- is_cancer_related TINYINT(2)
- comment TEXT

protein_pathway
- protein_id INT(10)
- pathway_id INT(10)
- is_central TINYINT(2)

pathway
- id INT(10)
- name VARCHAR(50)
- comment TEXT

mirna
- id INT(10)
- name VARCHAR(60)
- ac VARCHAR(10)

interaction_source
- interaction_id INT(10)
- source_id INT(10)

source
- id INT(10)
- reference_id INT(10)
- db_name VARCHAR(50)
- description TEXT
- url VARCHAR(150)

NETWORK - VISUALIZATION