Bioinformatics Introduction

David Fazekas

fazekas@netbiol.elte.hu

Department of Genetics, ELTE

Earlham Institute (UK)

Course Information

Lectures and practices

Hungarian lecture: **-**
English lecture: **Monday 10am**
Hungarian practices: **Thursday 2pm**
English practices: **Thursday 8am**, **Friday 10am**, **Friday 2pm**

Course Information

Lecturers

- Eszter Ari - chief trainer and administration
- Balazs Egyed - phylogenetics
- Dániel Gerber - phylogenetics
- Zsuzsa Dosztányi - structural bioinformatics
- Márton Doleschall - population bioinformatics
- Balázs Bohár - network biology
- Dávid Fazekas - network biology

Course Information

Course material

https://genetics.elte.hu/

```
username: genetika2019
password: genetika2019
```

Evaluation

Evaluation of the theoretical part (lecture grade):
The average grade of the 2 mid term tests (written, 45 min.).
You have to take an oral exam during the examination period if one (or both) mid term tests resulted with mark 1 or you missed the mid term exam. Oral exam will be about that part in which you failed or missed. You also can improve your mark (what you got as an average) if you wish in an oral exam – but be aware that you can also decrease the mark.
x.5 averages will be rounded upwards.

Evaluation

Evaluation of the practical part (term grade):
Criterion: Being active on the practicals. You can miss maximum 3 practicals.
Write a project work essay in groups of 3 people (using 3 different proteins) at the end of the semester using the results and the knowledge you have learned during the practicals.

SYLLABUS

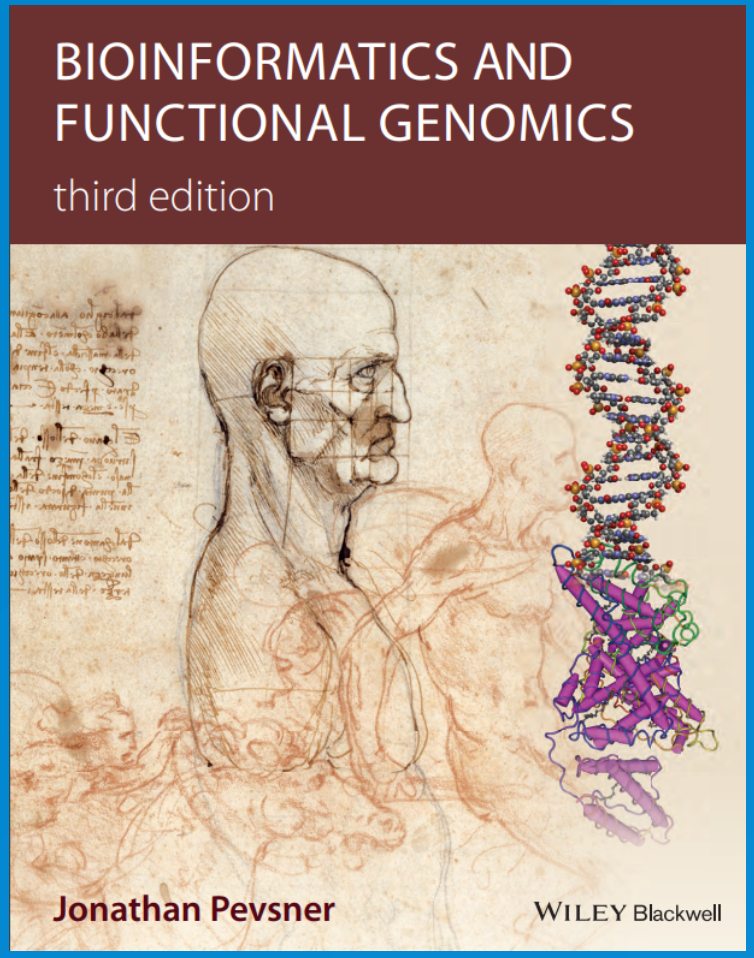
1. What does bioinformatics mean?
2. Databases in molecular biology
3. Sequence comparison and alignment
4. Sequence similarity searching
5. Structural bioinformatics I
6. Structural bioinformatics II
7. Molecular phylogenetics
8. Genomics and transcriptomics I + Mid term test I (Lecture: 1-6)

SYLLABUS

8. Genomics and transcriptomics I + Mid term test I (Lecture: 1-6)
9. Genomics and transcriptomics II
10. Network and systemsbiology I
11. Network and systemsbiology II
12. Genetic background of haplotype reconstruction
13. Mid term test 2 (Lecture 7-12)

Suggested reading

Jonathan Pevsner: BIOINFORMATICS AND FUNCTIONAL GENOMICS



Coose one

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

Contact

Eszter Ari

- arieszter@ttk.elte.hu
- Department of Genetics D5.604
- Office hour: Monday 13:00-15:00 ()

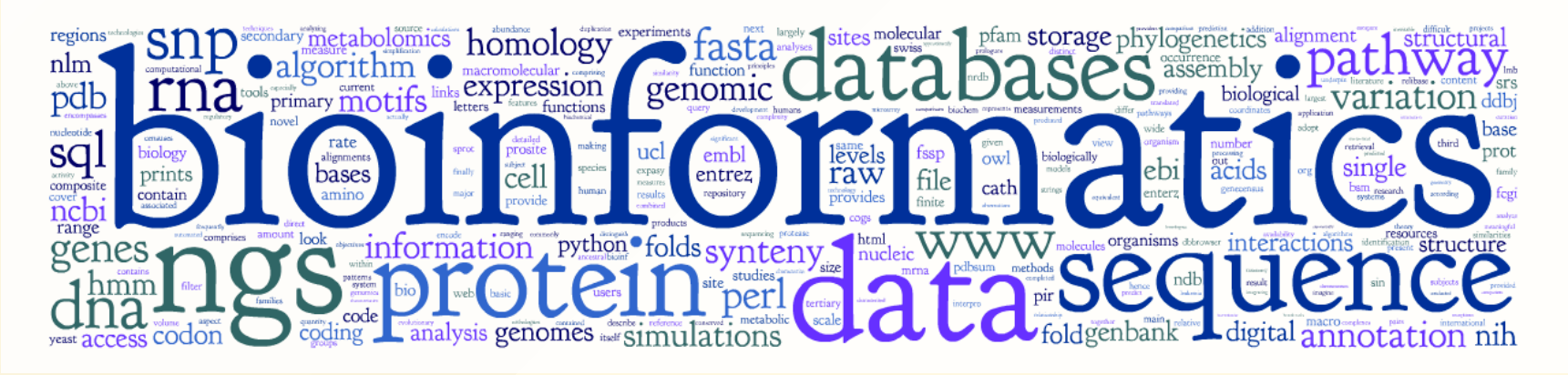
Bring your own device

You can bring your own laptop to the practice. There are wifi and power outlet in the classroom.

BUT

We will not help to make your computer work or connect to wifi.
We will not help to make all required software up and running.

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

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 Structure Biologist !
I am Evolutionary Biologist !
I am System 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?

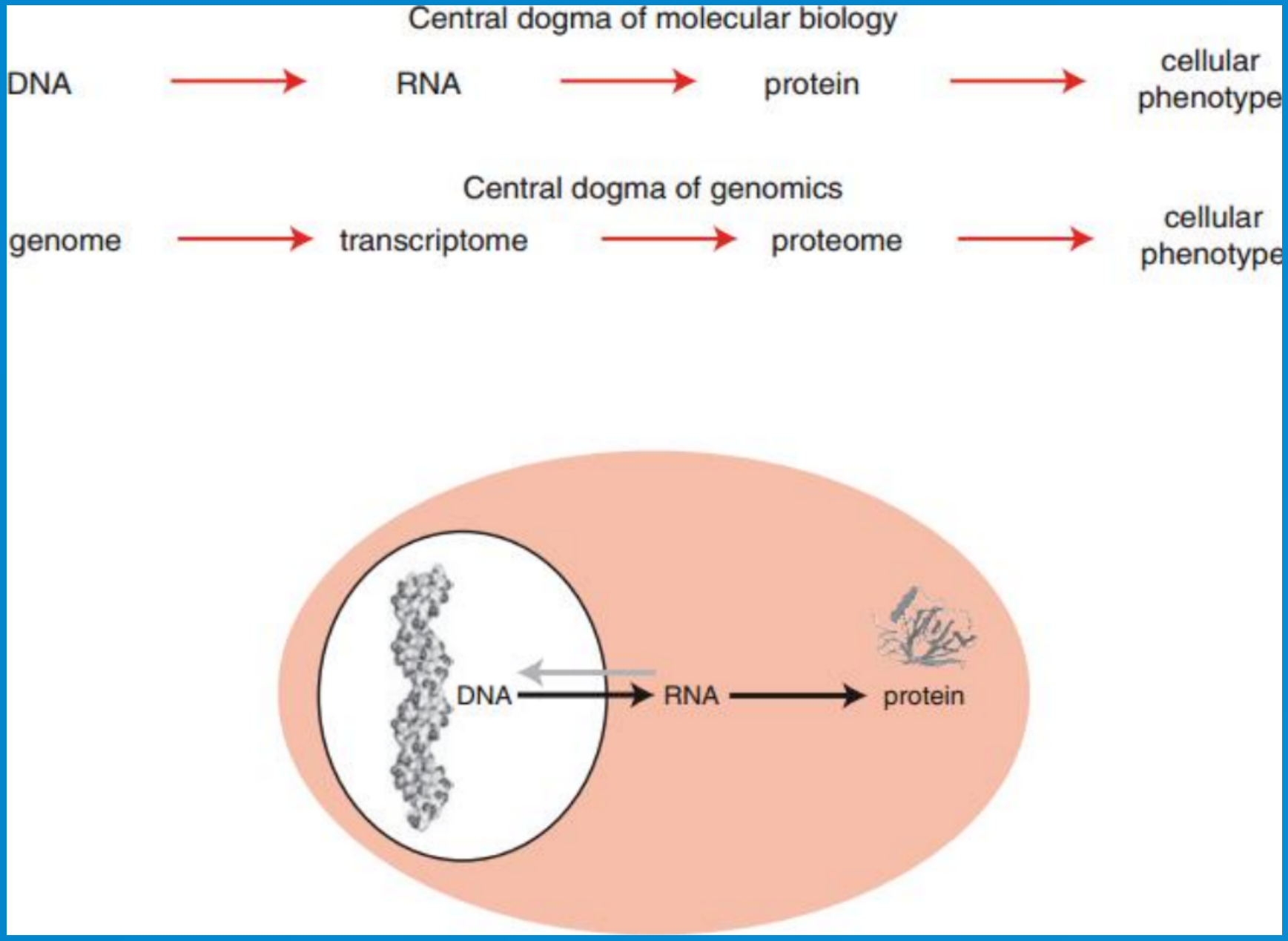
In a narrower sense

- Working with data in life sciences

In the broader sence

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

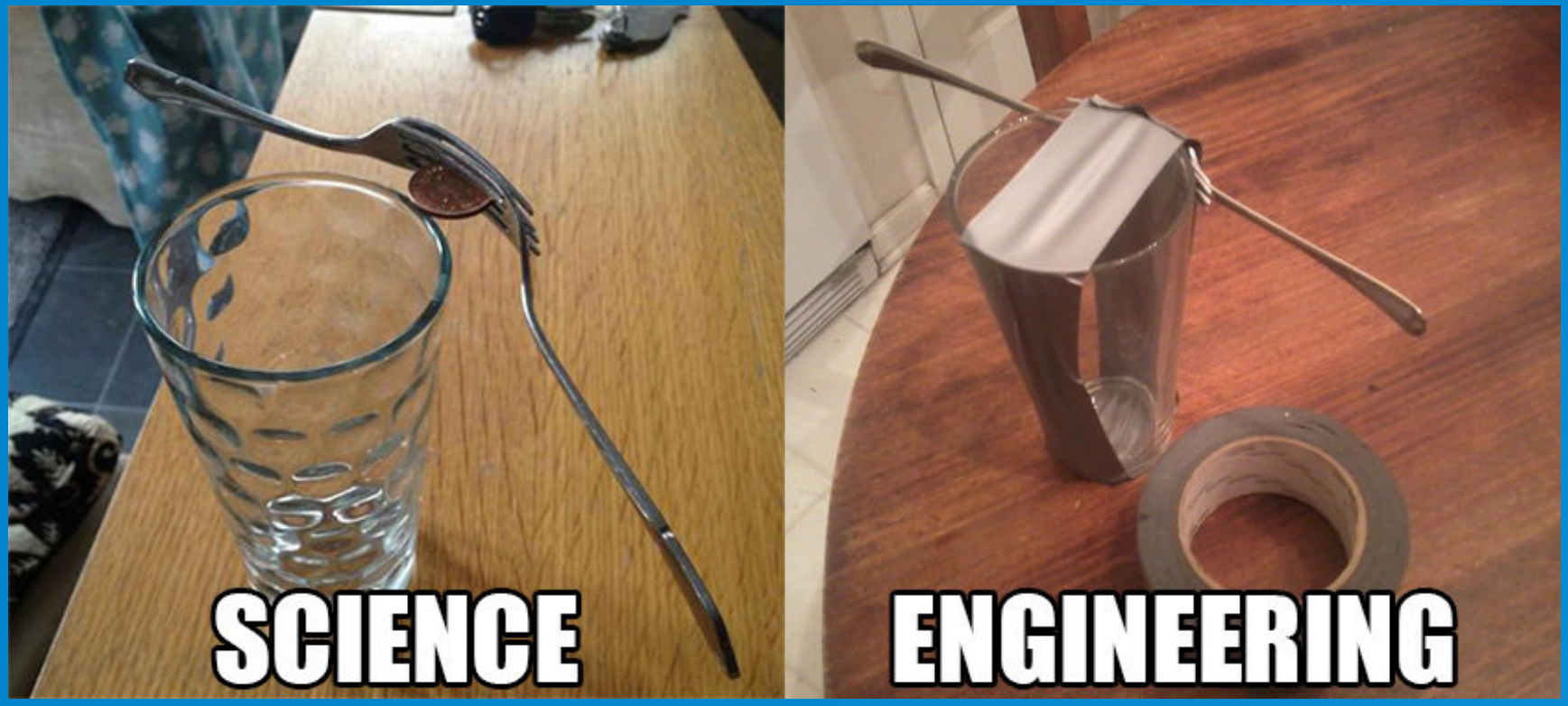
Molecular Bioinformatics

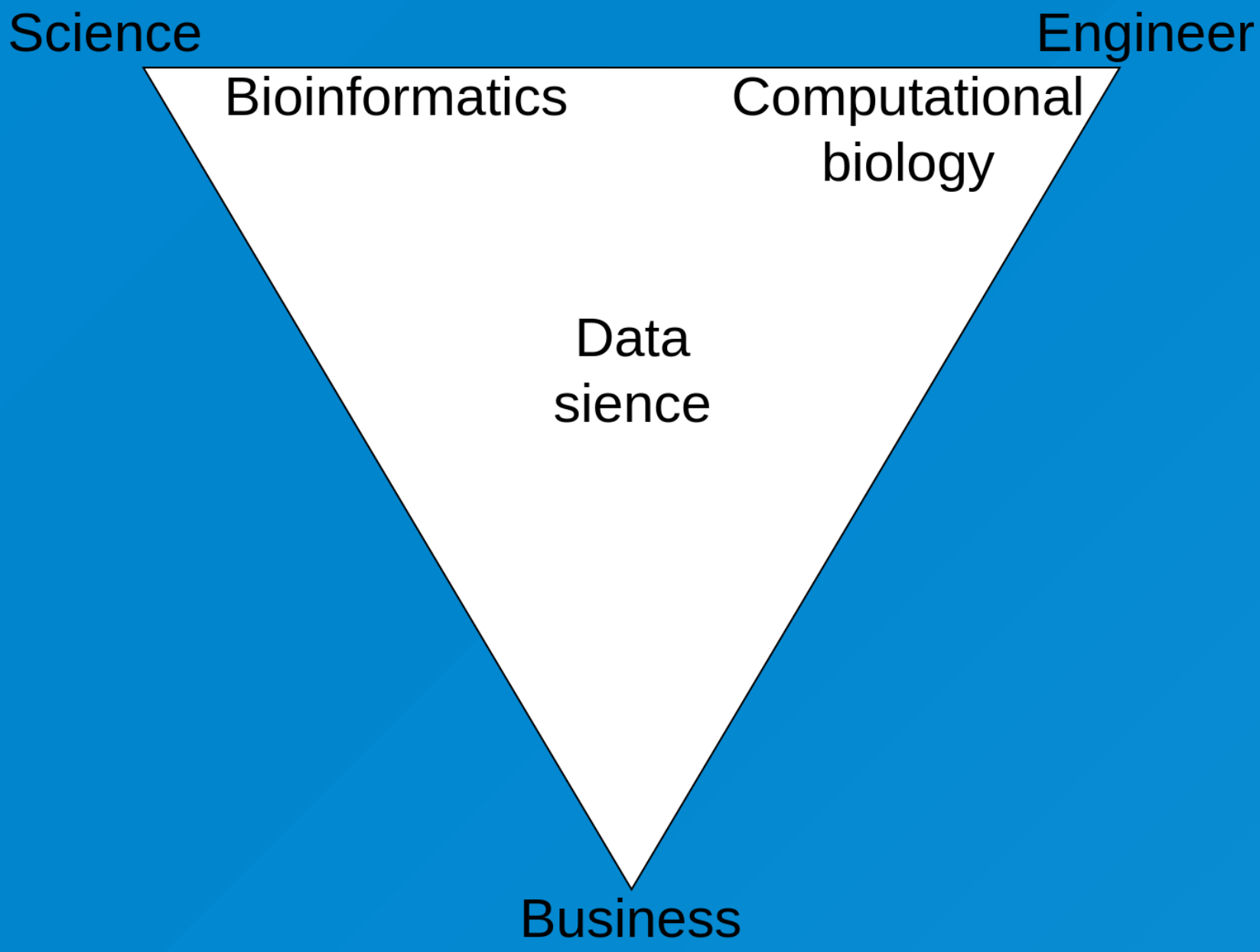


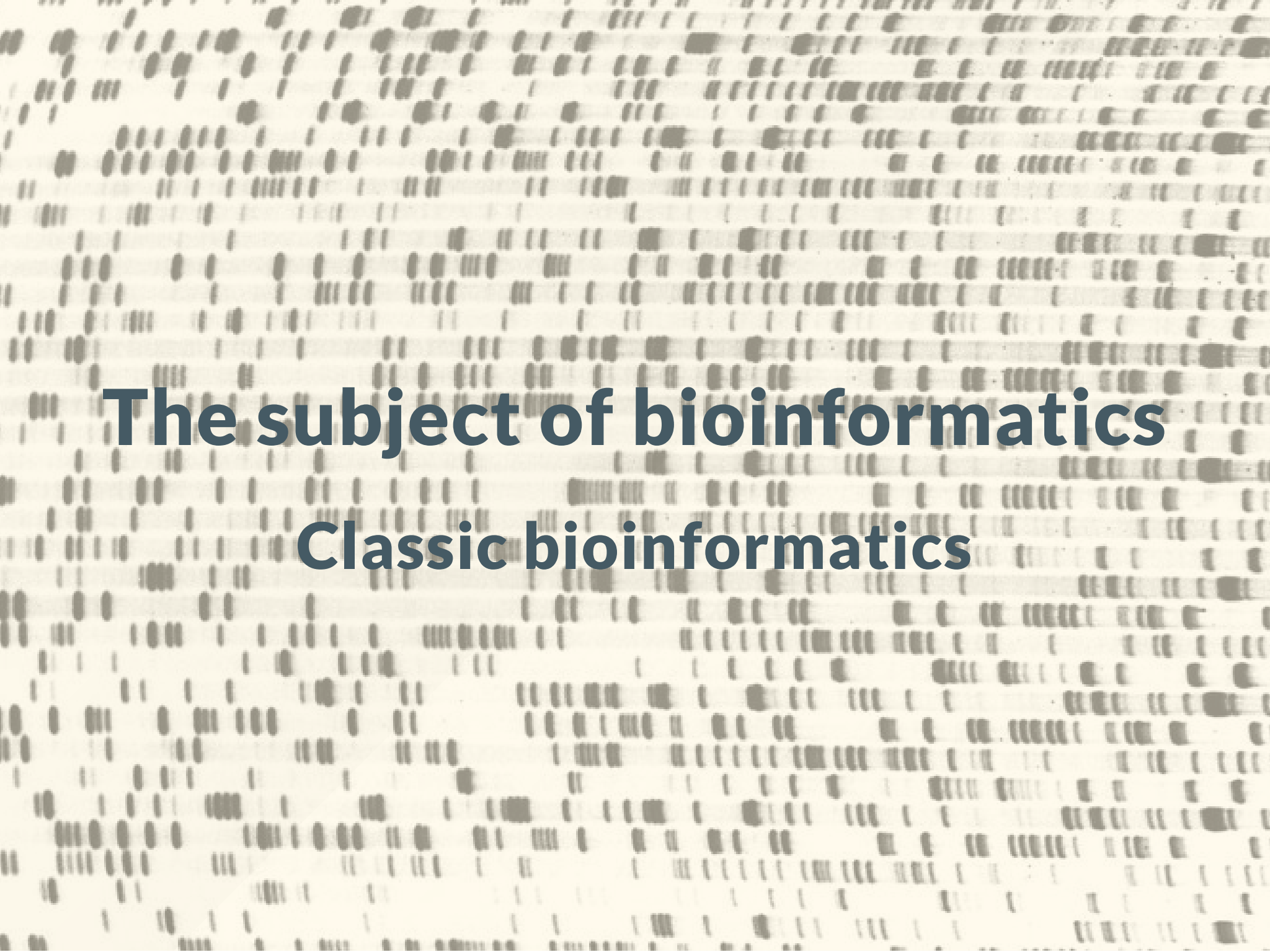
Wet lab - Dry lab

Bioinformatics and Data Science

Science vs Engineering



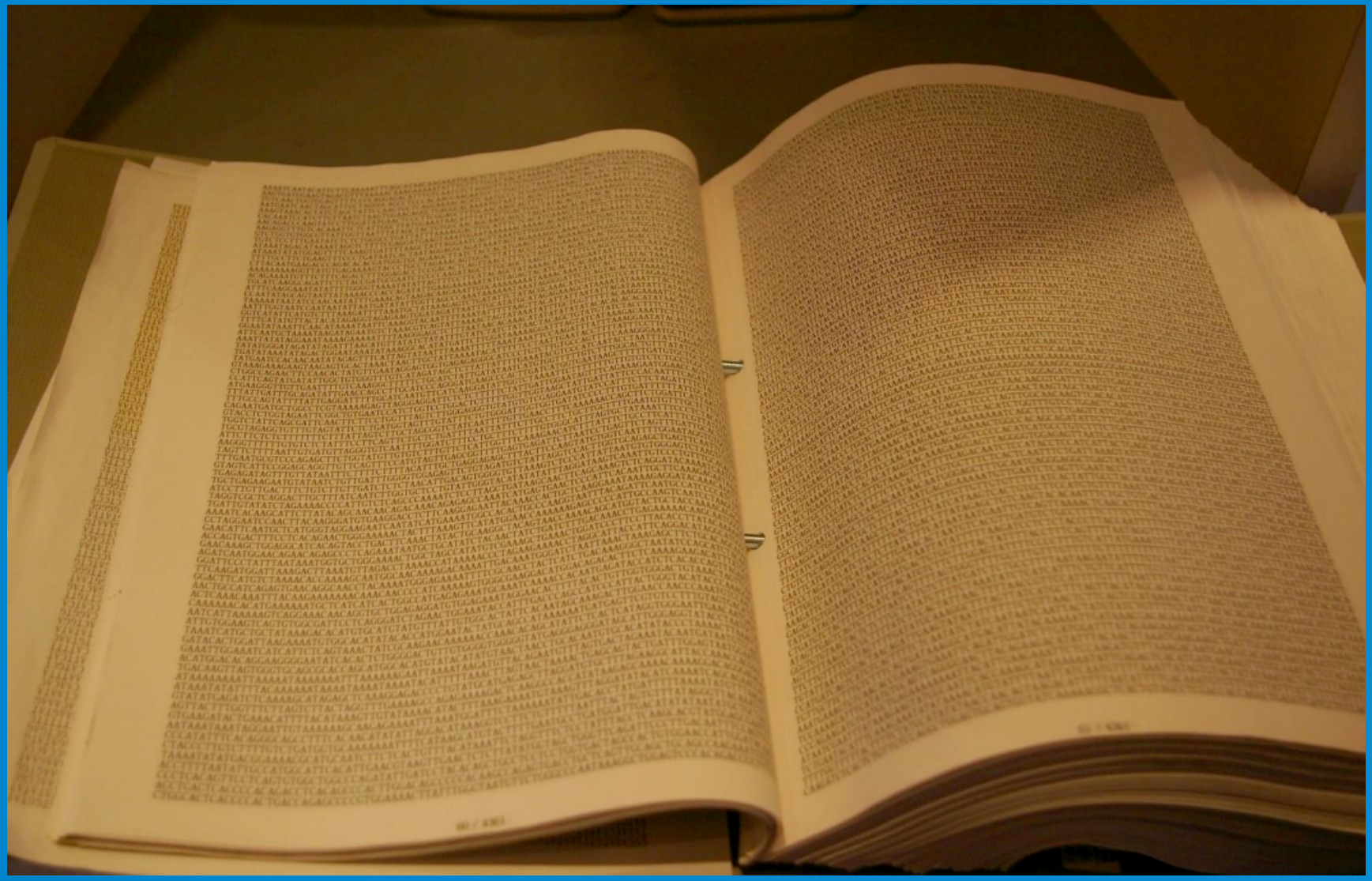

Science
Engineer
Bioinformatics
Computational biology
Data sience
Business



The subject of bioinformatics

Classic bioinformatics

Without computer



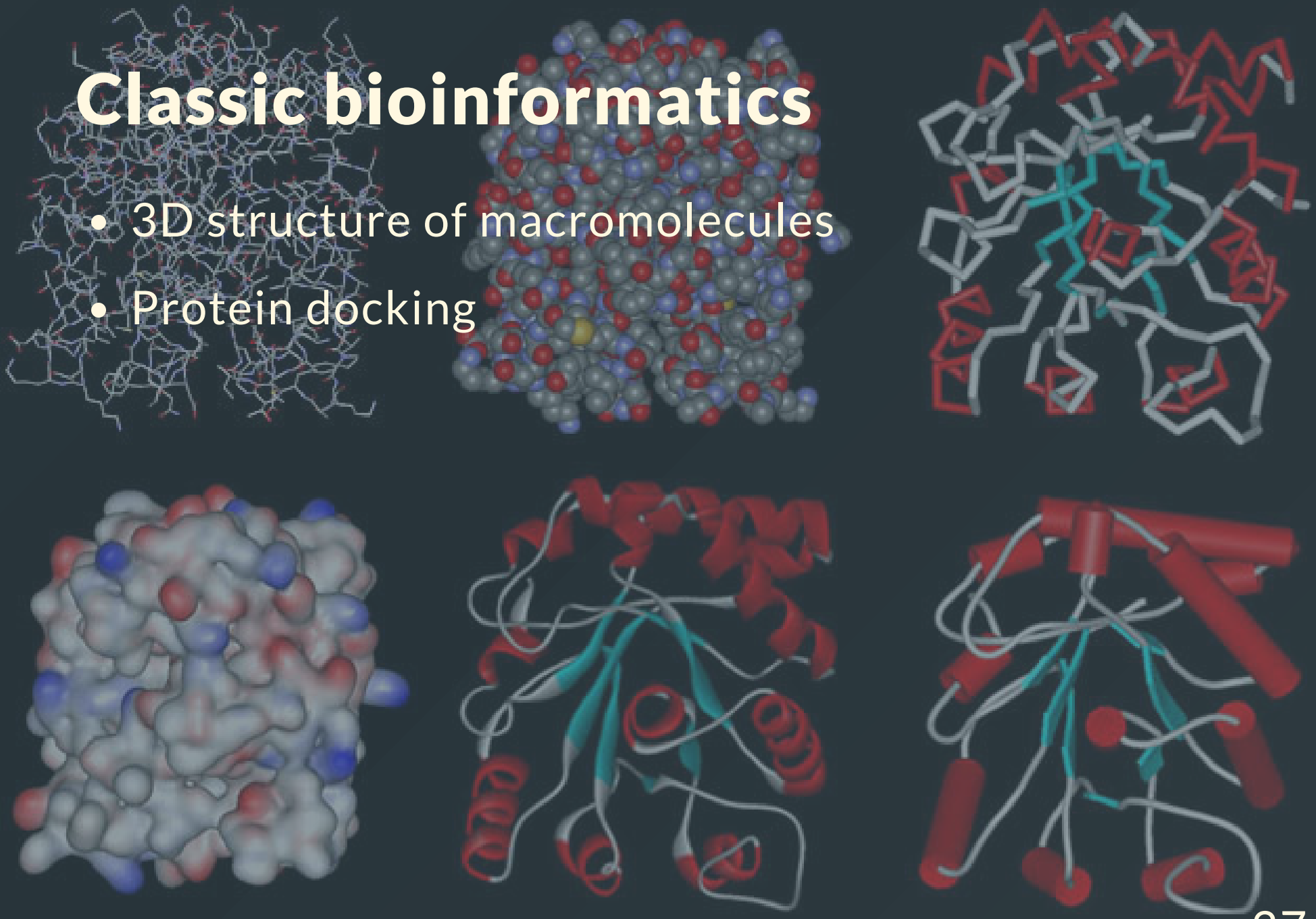
Without computer

Classic bioinformatics

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

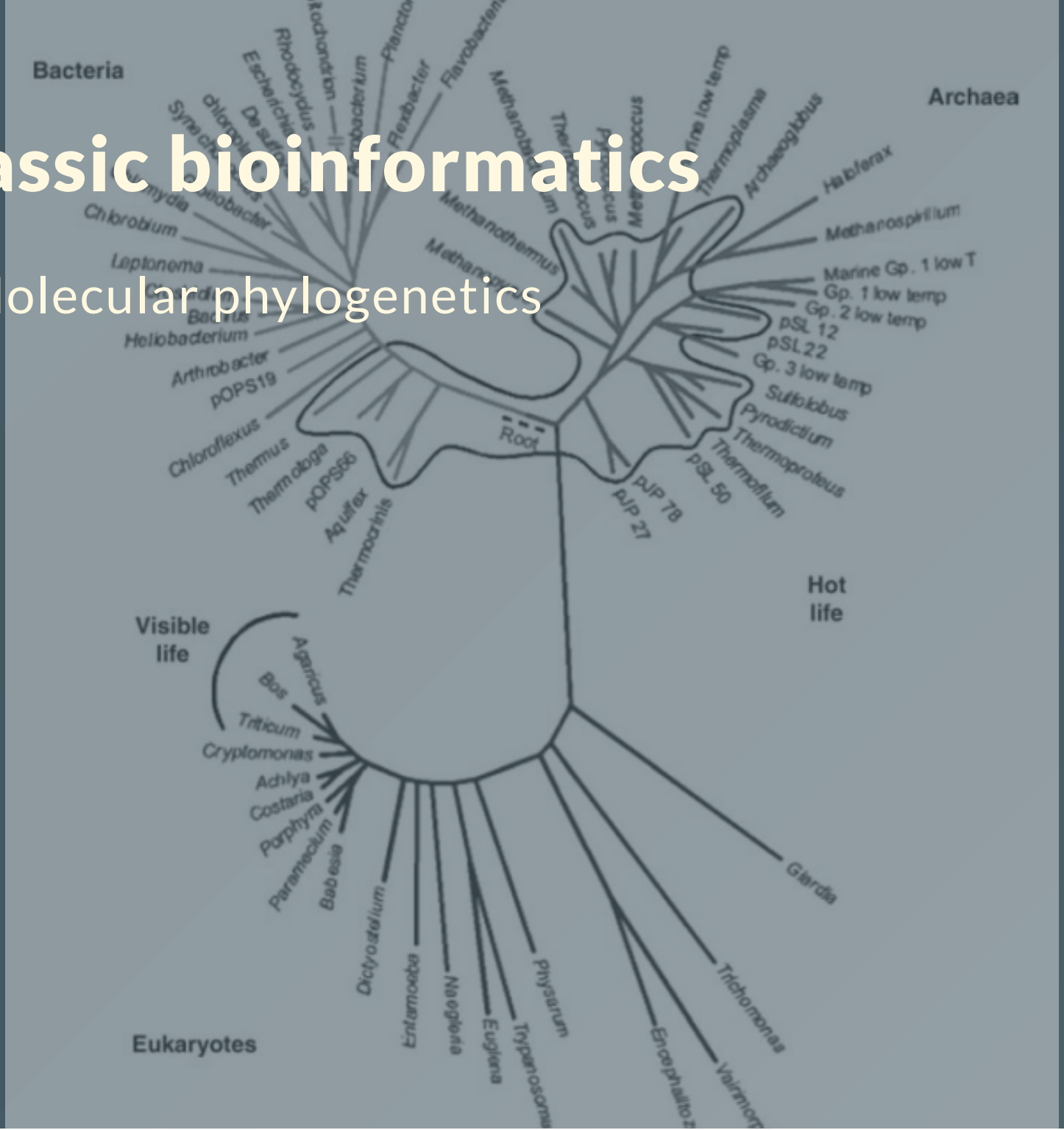
Classic bioinformatics

- 3D structure of macromolecules
- Protein docking



Classic bioinformatics

- Molecular phylogenetics

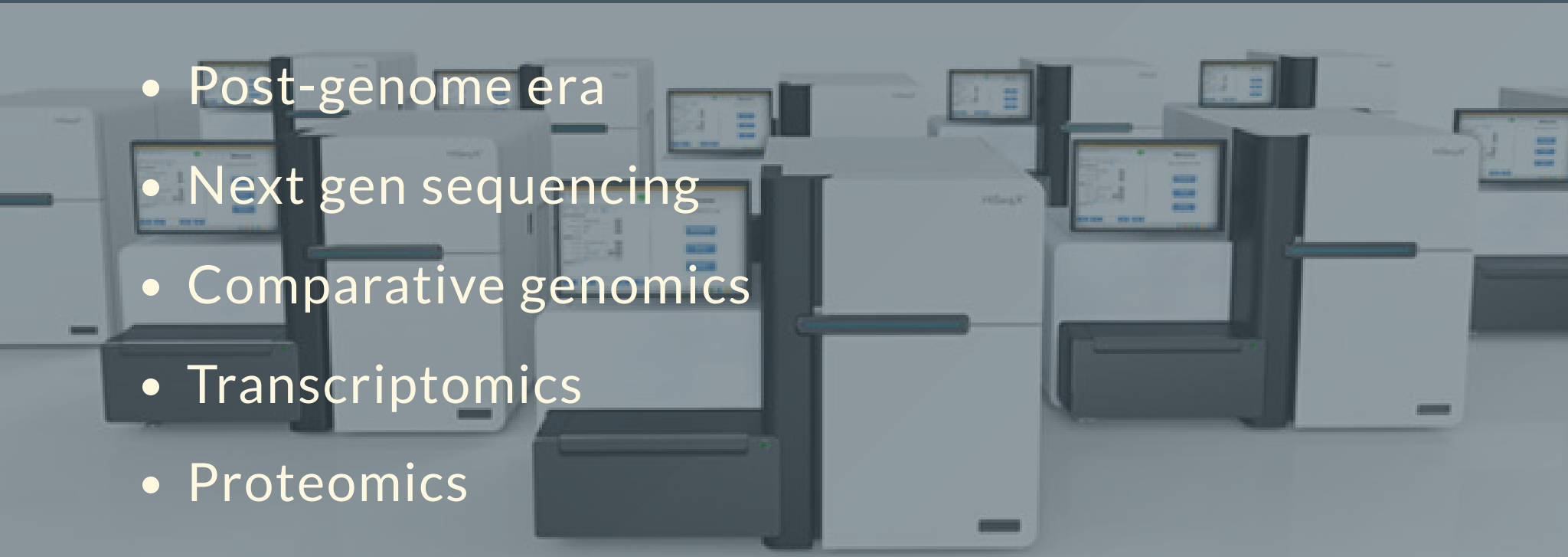


"old" and "new" biology

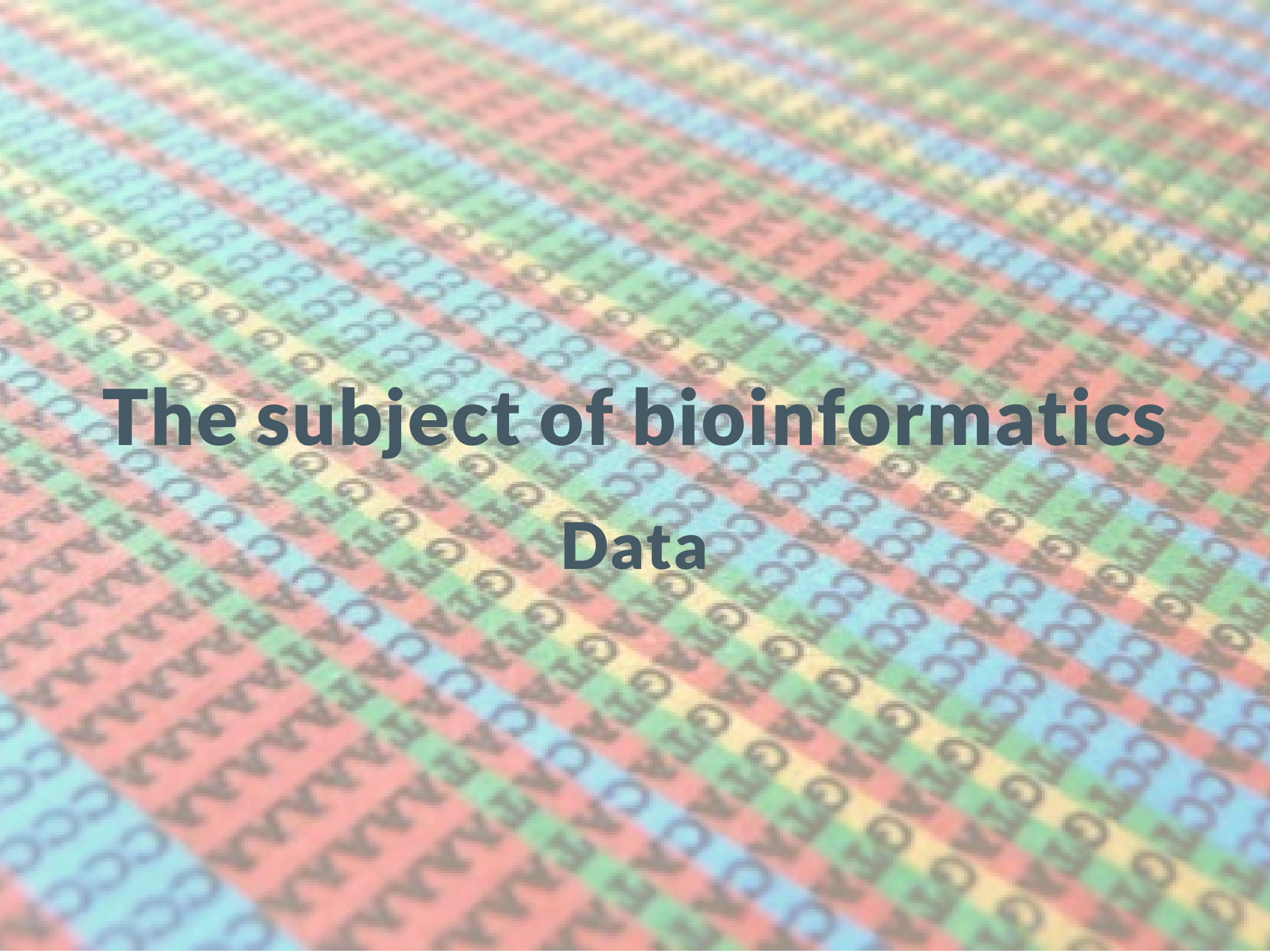
In the (near) past:
For researchers the greatest the challenge was to produce good quality data.

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.

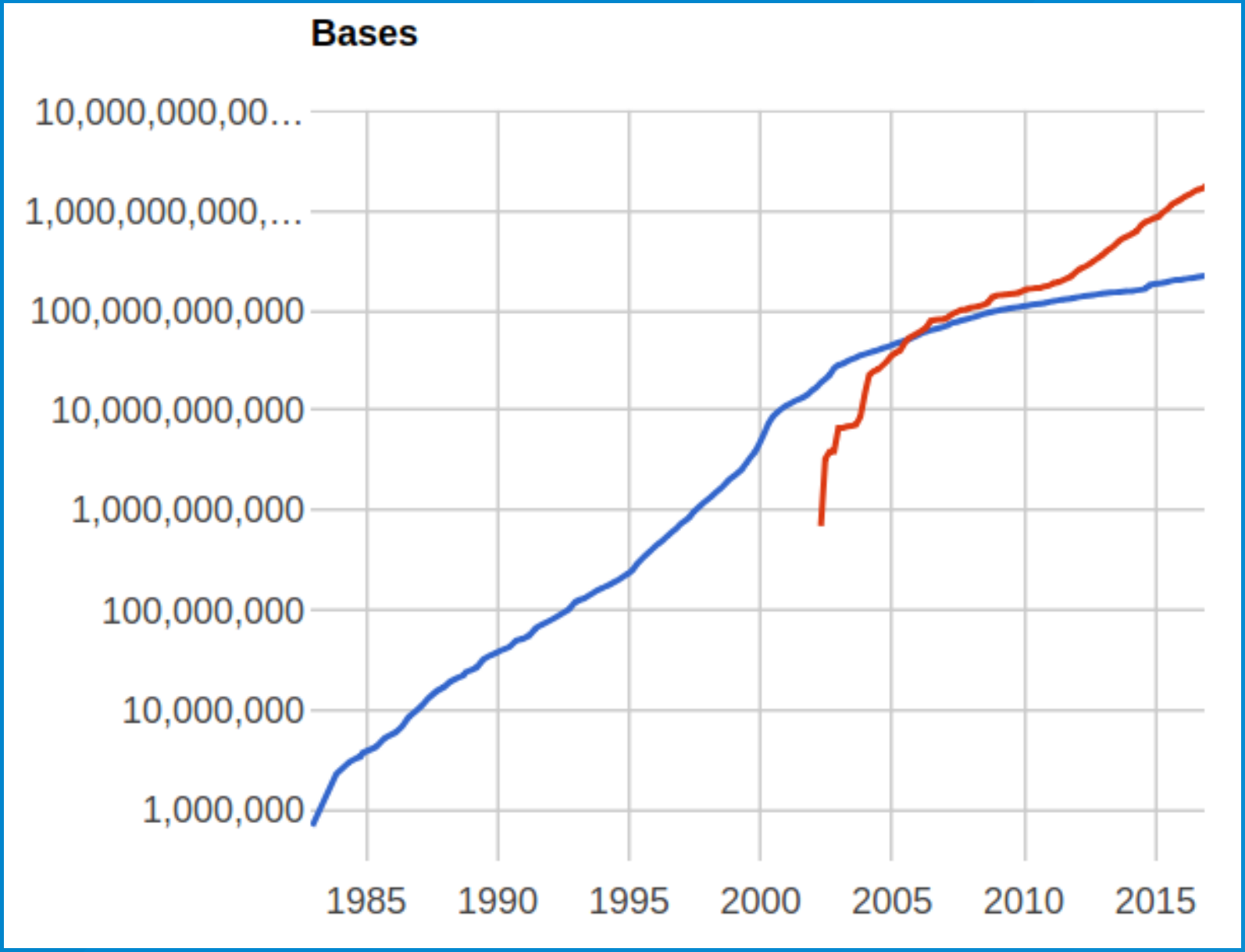
Modern bioinformatics



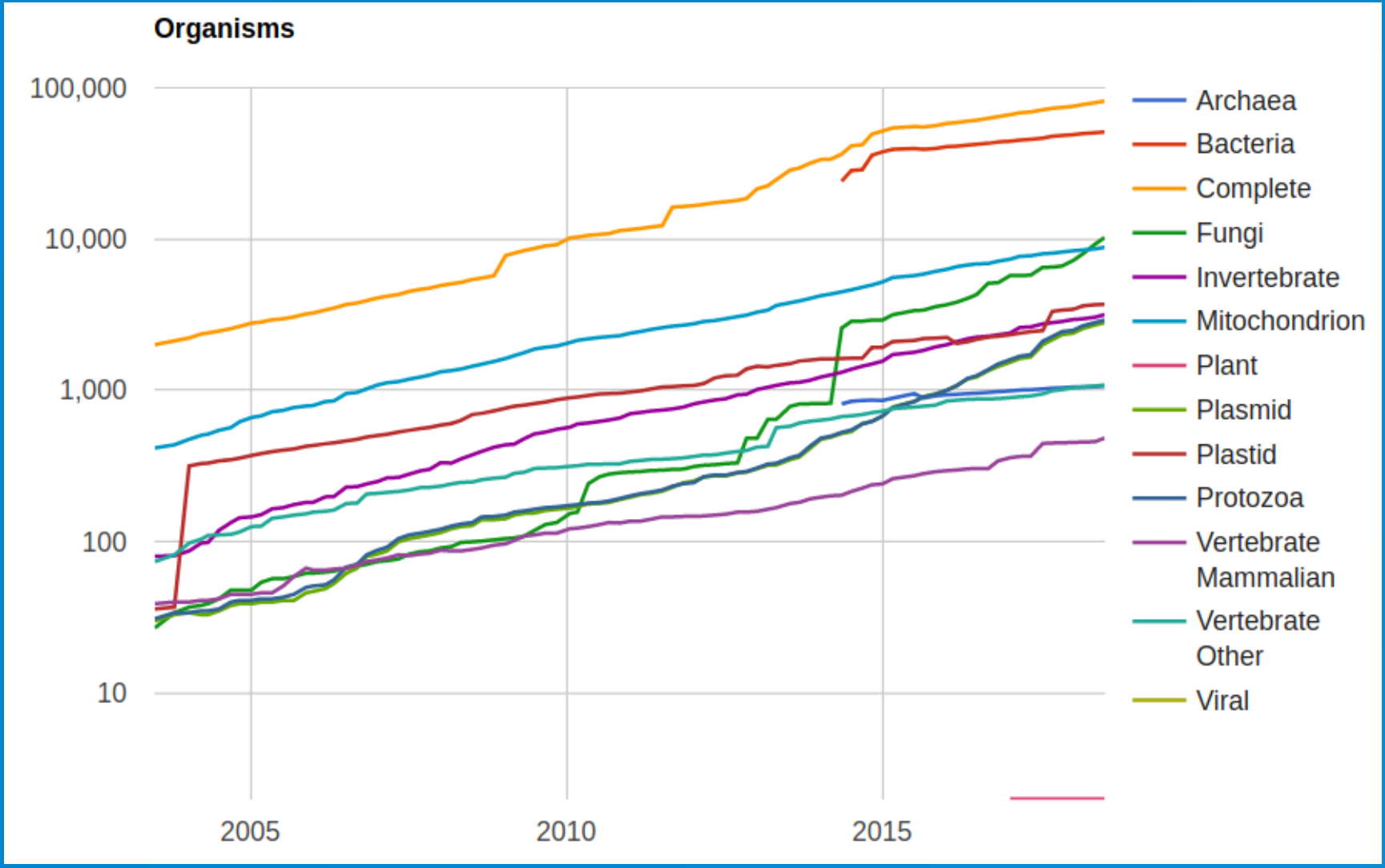
- Post-genome era
- Next gen sequencing
- Comparative genomics
- Transcriptomics
- Proteomics
- Systems biology

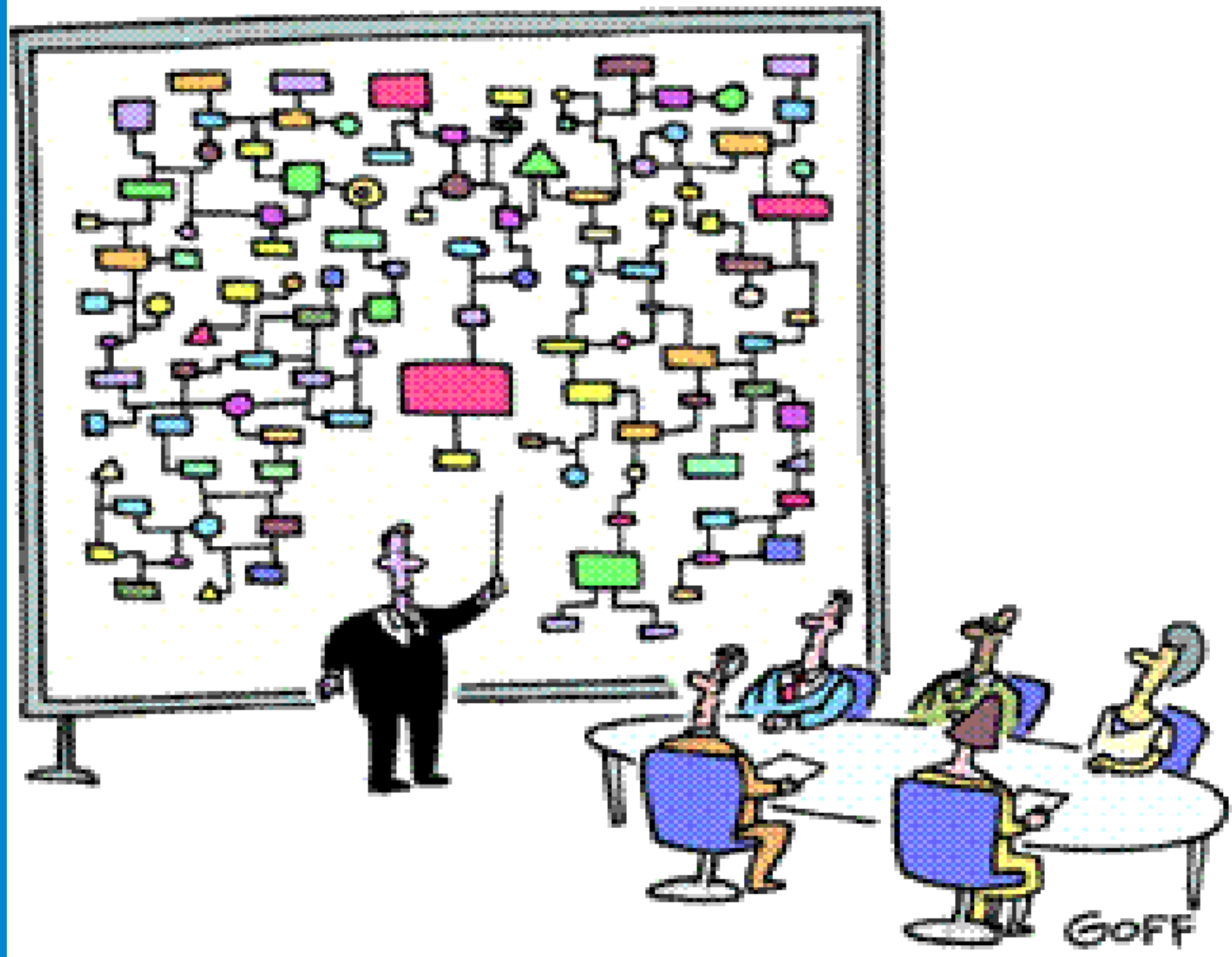

The subject of bioinformatics
Data

NCBI GeneBank: $3.2*10^{12}$ bases



Sequenced genomes: 82000 organisms





"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

Linux Windows Mac

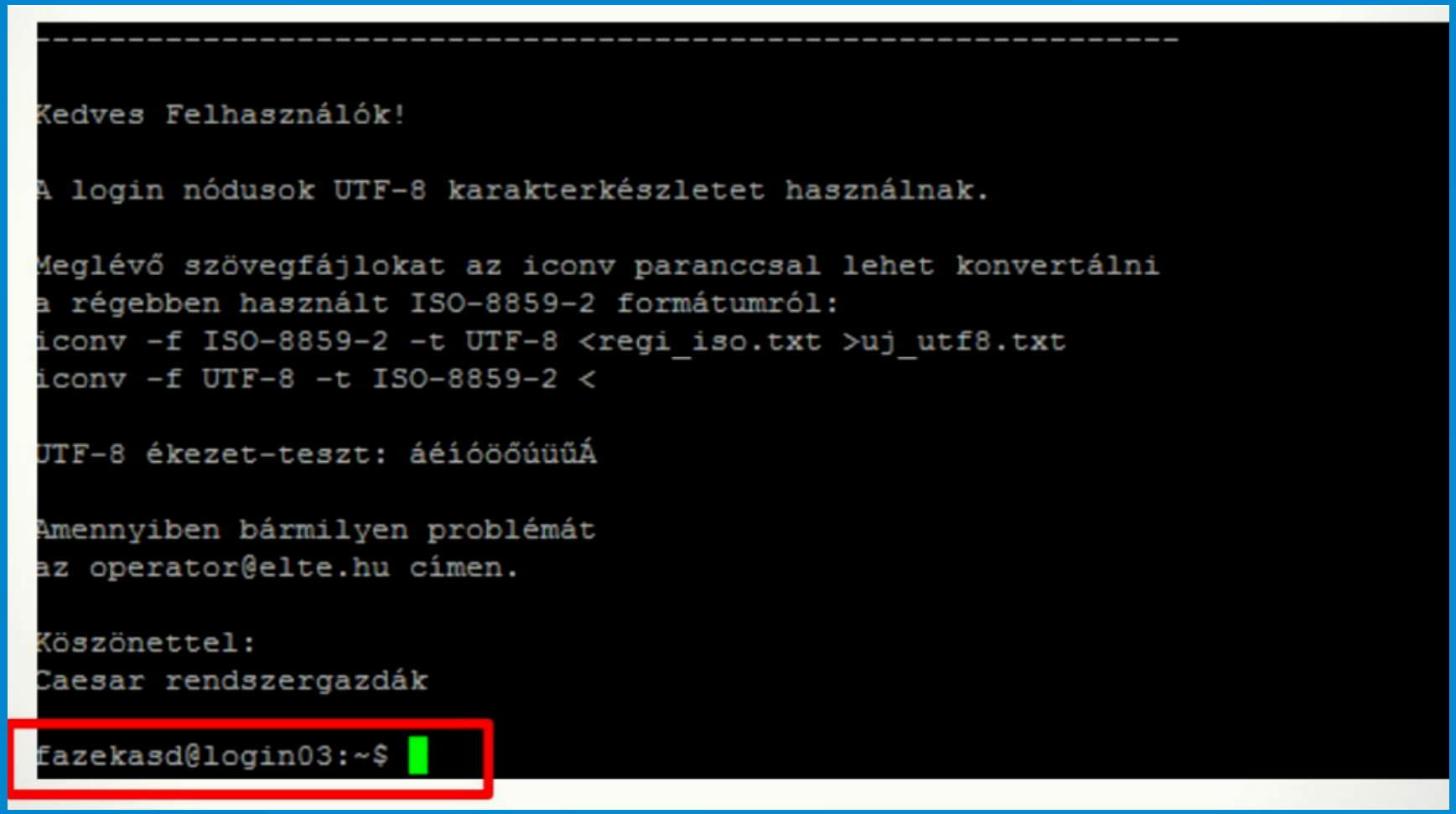
as seen by...

Mac Fanboys

Windows Fanboys

Linux Fanboys

Linux comand line




Data storing

Sequence - flat file

```
ID   AXIN1_HUMAN             Reviewed;         862 AA.
AC   O15169; Q4TT26; Q4TT27; Q86YA7; Q8WVW6; Q96S28;
DT   01-DEC-2000, integrated into UniProtKB/Swiss-Prot.
DT   10-MAY-2002, sequence version 2.
DT   03-SEP-2014, entry version 163.
DE   RecName: Full=Axin-1;
DE   AltName: Full=Axis inhibition protein 1;
DE            Short=hAxin;
GN   Name=AXIN1; Synonyms=AXIN;
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
OC   Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   NUCLEOTIDE SEQUENCE [MRNA] (ISOFORM 1).
RX   PubMed=9230313; DOI=10.1016/S0092-8674(00)80324-4;
RA   Zeng L., Fagotto F., Zhang T., Hsu W., Vasicek T.J., Perry W.L. III,
RA   Lee J.J., Tilghman S.M., Gumbiner B.M., Costantini F.;
RT   "The mouse Fused locus encodes Axin, an inhibitor of the Wnt signaling
RT   pathway that regulates embryonic axis formation.";
RL   Cell 90:181-192(1997).
RN   [2]
RP   NUCLEOTIDE SEQUENCE [LARGE SCALE GENOMIC DNA].
RX   PubMed=11157797; DOI=10.1093/hmg/10.4.339;
RA   Daniels R.J., Peden J.F., Lloyd C., Horsley S.W., Clark K.,
RA   Tufarelli C., Kearney L., Buckle V.J., Doggett N.A., Flint J.,
RA   Higgs D.R.;
RT   "Sequence, structure and pathology of the fully annotated terminal 2
RT   Mb of the short arm of human chromosome 16.";
RL   Hum. Mol. Genet. 10:339-352(2001).
RN   [3]
RP   NUCLEOTIDE SEQUENCE [LARGE SCALE GENOMIC DNA].
```

Sequence - FASTA file

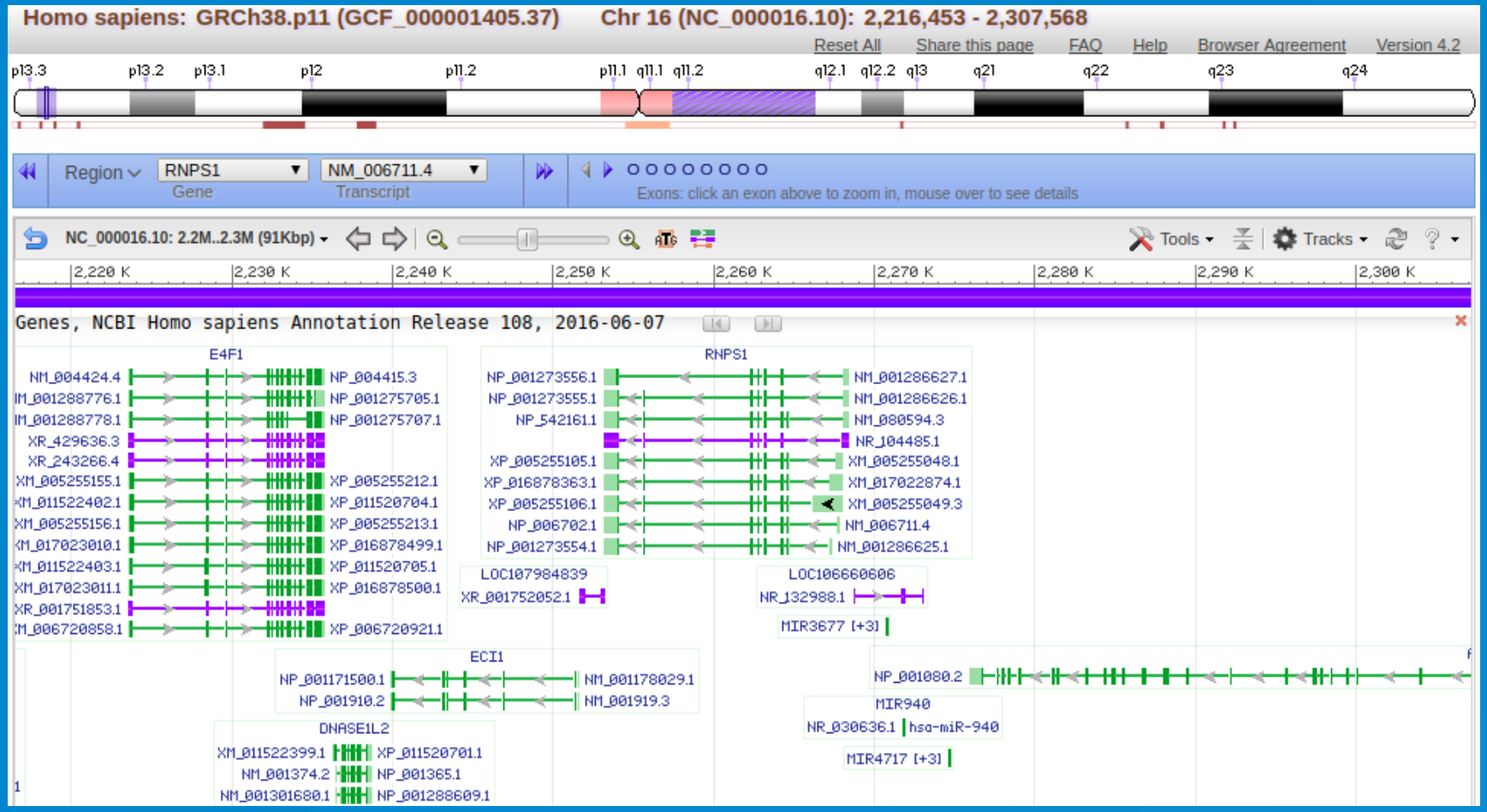
Homo sapiens chromosome 16, GRCh38 Primary Assembly

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



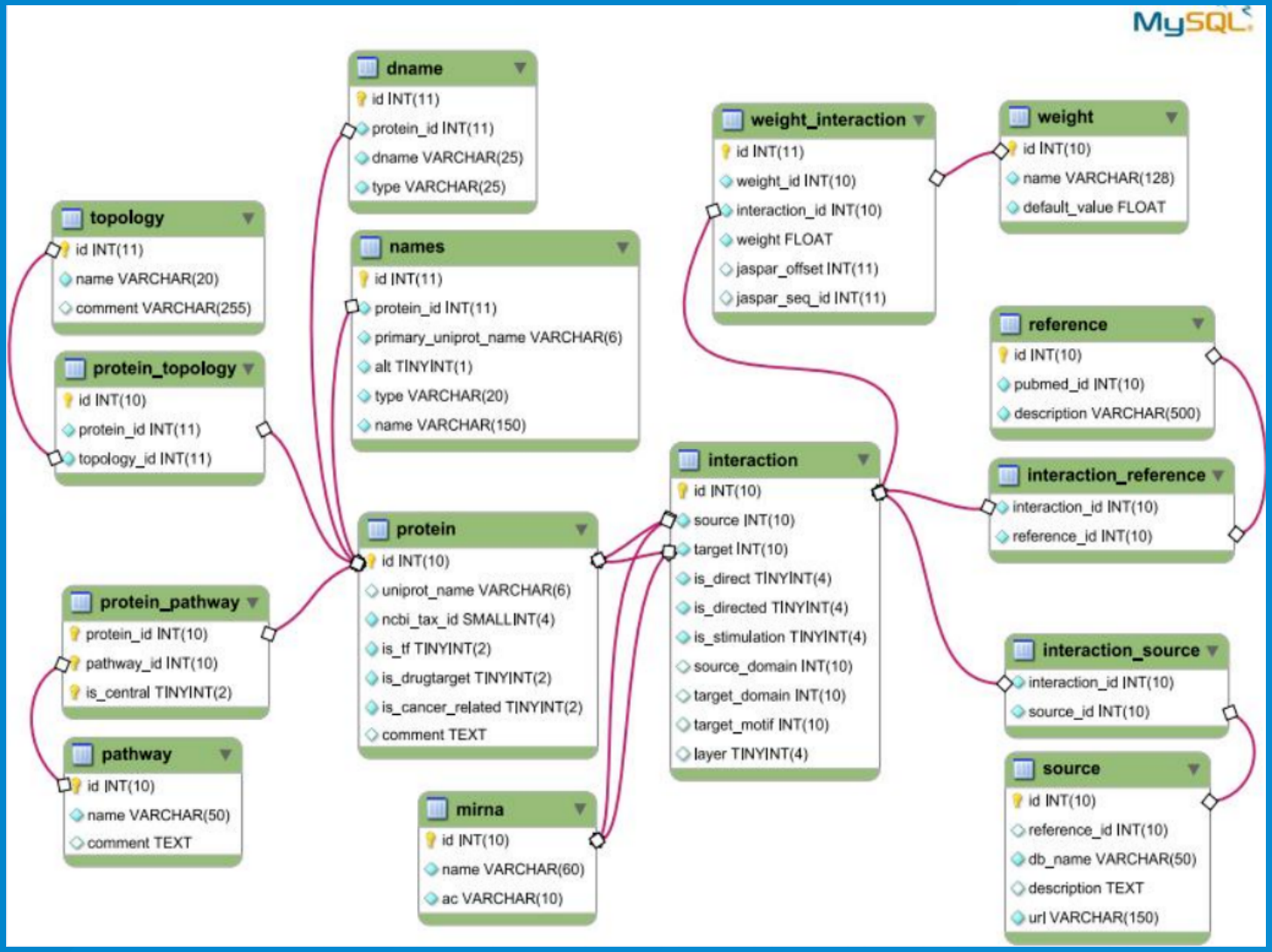
Network - plain text

```
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
```

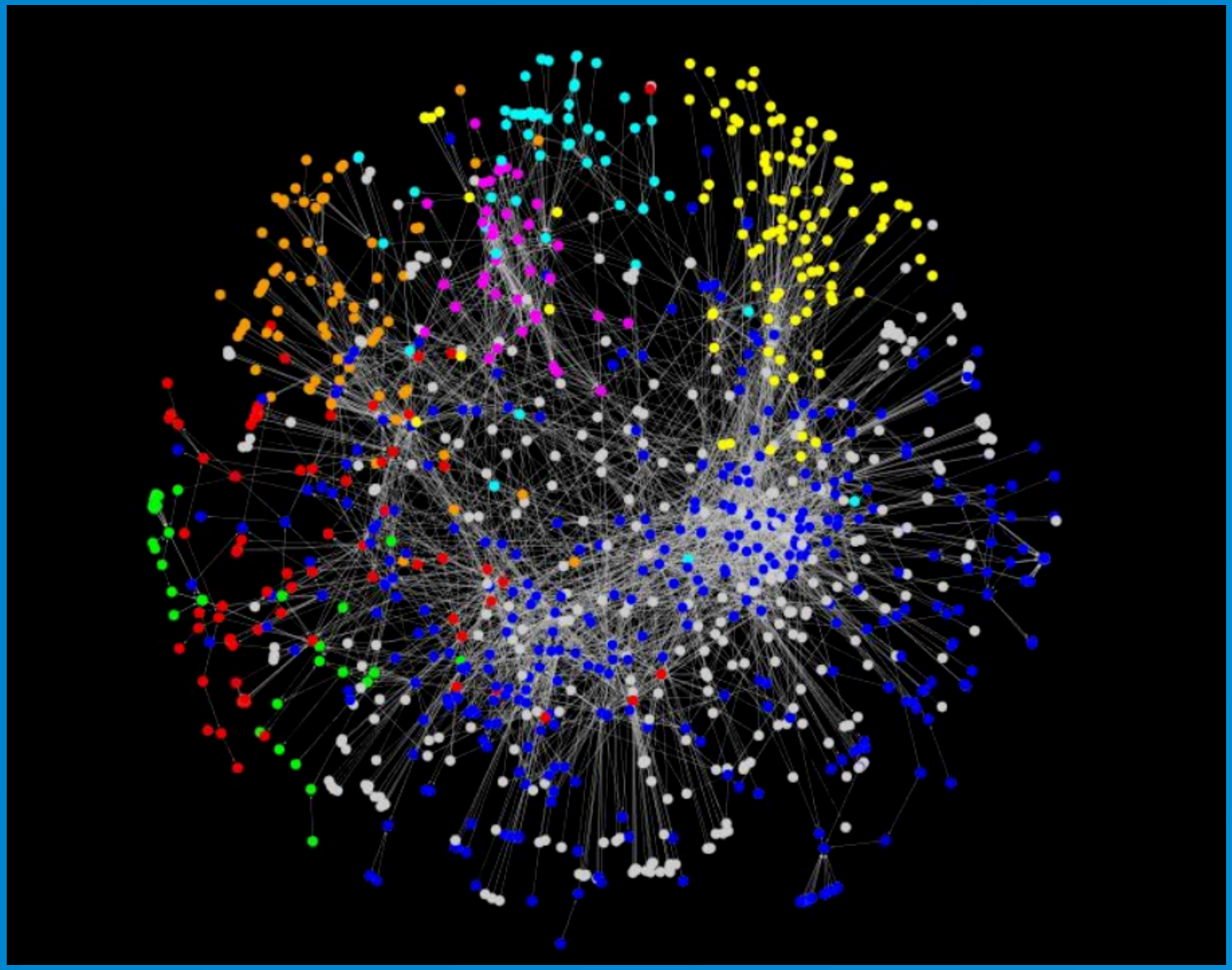
Network - table

source_na	source_ur	source_sp	source_sp	source_to	source_pa	target_na	target_un	target_spe	target_spe	target_to	target_pa	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



Network - visualization



Software requirements for the 1st practice:

Windows:

- PuTTY https://www.putty.org

macOS:

- terminal emulator (built in)

Linux:

- terminal emulator (built in)