Intro to bioinformatics

What is bioinformatics?
Who does bioinformatics?
What domains of bioinformatics exist?
Do life scientists benefit from learning bioinformatics?
What do you need to learn to analyze genomic data?

What is bioinformatics?

- Application of computations to answer biological questions

- Bioinformatics is an umbrella term that covers various loosely connected or even disconnected fields with computations being the only common denominator

- This term is frequently applied to the analysis of high-throughput datasets, although it is only a fraction of it
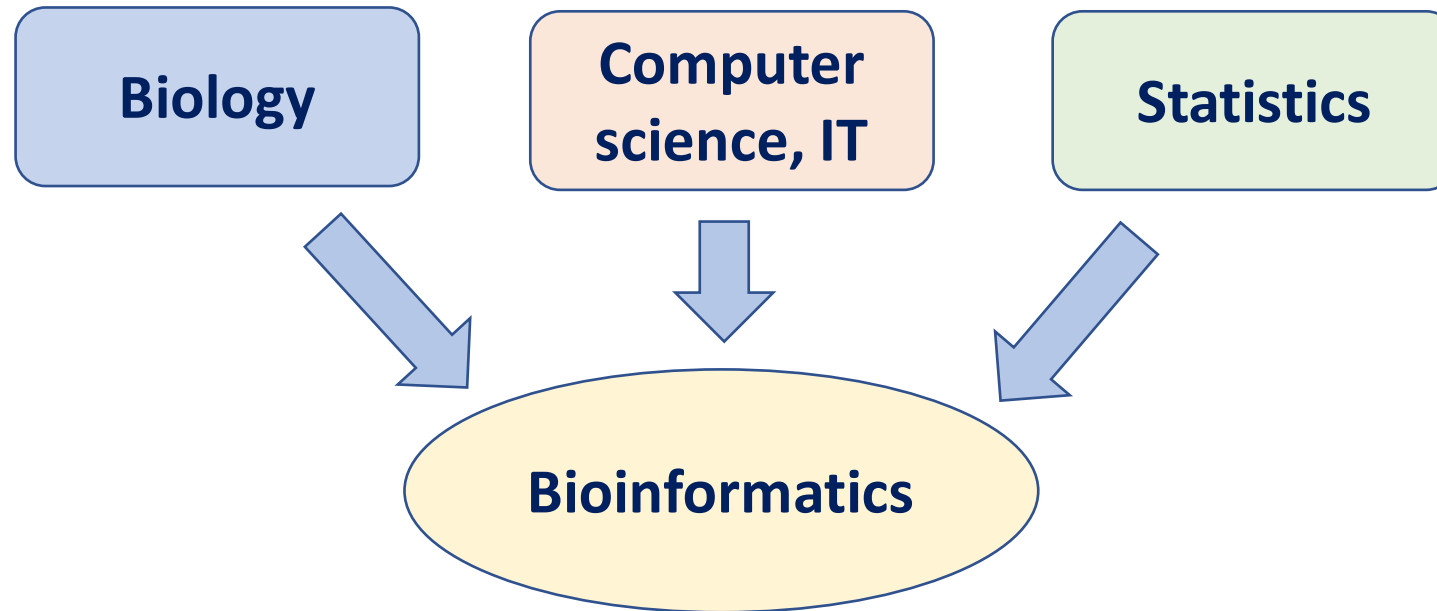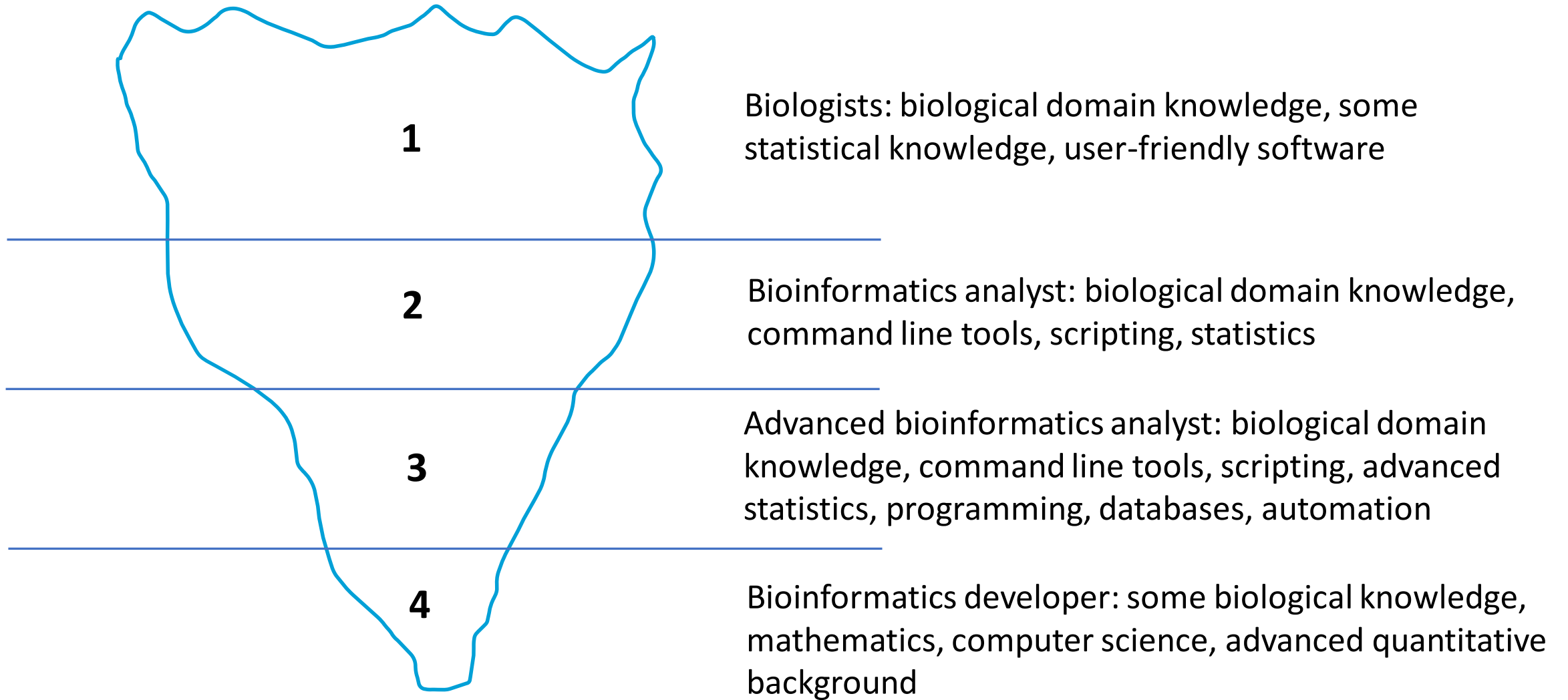
Who does bioinformatics?

- Anyone who uses bioinformatics application with graphical use interface (GUI) or web interface to run computations, for example align sequences with BLAST, search for protein domains, detect enriched biological terms, etc.

- Bioinformatics analysts who apply **command line** tools and programming to analyze large scale data sets, where GUI tools and web servers are to restrictive and no longer practical

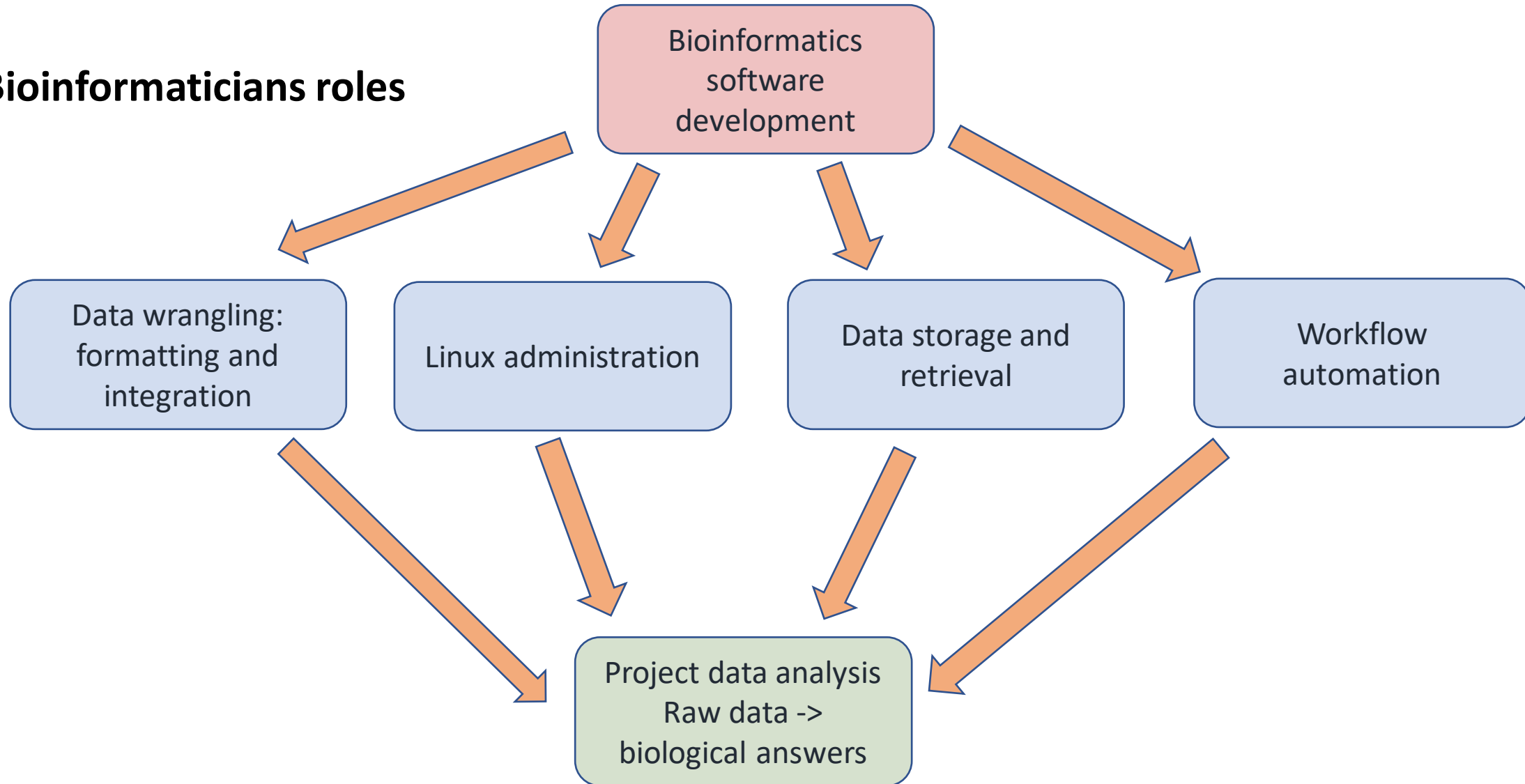- Computer scientists and mathematicians that develop bioinformatics algorithms and software

The levels of knowledge

Tiers in bioinformatics, inverted iceberg

1

Biologists: biological domain knowledge, some statistical knowledge, user-friendly software

2

Bioinformatics analyst: biological domain knowledge, command line tools, scripting, statistics

3

Advanced bioinformatics analyst: biological domain knowledge, command line tools, scripting, advanced statistics, programming, databases, automation

4

Bioinformatics developer: some biological knowledge, mathematics, computer science, advanced quantitative background

Who does bioinformatics?

Different domains in bioinformatics
(as applied to next generation sequencing).

Metagenomics: using sequencing to determine species or/and strain composition in the population of organisms

Assembly: constructing genomes and transcriptomes from shorter nucleotide sequences

Resequencing: finding mutations and variations in genomic sequences

Quantification: diverse field focused on using sequencing to determine functional characteristics of the cell

Different domains in bioinformatics (as applied to next generation sequencing).

Ever-growing quantification domain (Something-seq):

✓ RNA-seq: quantification and comparison of gene and transcript expression

✓ Small RNA-seq: sequencing effort focused on small RNAs

✓ ChIP-seq, ATAC-seq, DNAse-seq, CLIP-seq: quantification of genomic regions bound by protein factors

✓ Bisulfite-seq: sequencing of bisulfite converted DNA to detect DNA methylation

✓ Ribo-seq: sequencing-based profiling of active ribosomes

✓ Hi-C: determining linkage between distant chromosomal regions

Etc.

Different domains in bioinformatics
(as applied to next generation sequencing).

Short read sequencing (<= 300 bp)
Illumina, Ion torrent

vs

Long read sequencing (kb and mb)
PacificBio and Nanopore

Bulk sequencing (average across the
cellular population)

vs

Single-cell sequencing (sequencing
profile obtained from individual cell
or small cluster of cells)

Do life scientists benefit from knowing bioinformatics?

Yes, they do!

✓ Bioinformatics is now relevant in most areas of biological studies

✓ Learning it helps to understand publications

✓ Helps to collaborate with bioinformaticians, ask the right questions, and incorporate the results of bioinformatics analyzes

✓ Submit the requests and understand data reports from bioinformatics service providers

✓ Analyze and publish bioinformatics data on their own

✓ Understand principles and limitations behind bioinformatics methods used in the analysis

What do you need to learn to analyze genomic data?

Must have

- Linux command line
- Scripting language (bash)
- Statistics and R language
- Command line utilities specific to your work

Nice to have

- Interpreted programming languages (perl, python, etc)
- SQL language (databases)
- Pipeline management (CWL, nextflow, bpipe)
- Git version control system
- Docker
- Compiled languages (Java, C++, C#, etc)