Introduction to Bioinformatics for Computer Scientists

Lecture 1

Preliminaries

- Lectures will be in English
  - It is the language of science
  - Language of a job interview at Google
- Please send me an email such that:
  - I can set up a course mailing list
- Email: Alexandros.Stamatakis@kit.edu
- I usually reply within a day
- Slides and videos: https://cme.h-its.org/exelixis/web/teaching/slides.html

Preliminaries

- Lab web-site: www.exelixis-lab.org
- Course web-site:

  http://www.exelixis-lab.org/web/teaching/BioinformaticsModule.html
- Exelixis is the Greek word for evolution
- Slides & Videos
  - Are available at the hidden link
  - Live lectures may deviate a bit from the pre-recorded videos
- Help us improve the course :-)

Schedule

- *Lectures until Christmas:* Live on Campus, SARS-CoV-2 permitting

- *Lectures after Christmas:*

  Live via zoom

  or

  Recorded via youtube and then Q & A session via Zoom
- Any preferences?

Schedule

- Why no live lectures after Christmas?
- As of 01.01.2023 I will start setting up a 2nd research group in Crete – The Biodiversity Computing Group and permanently move to Crete while continuing to teach at KIT
- Funded by EU ERA chair project
  - Strengthen so-called widening countries
  - Prevent brain-drain & foster brain gain
  - Propose institutional reforms
- Opportunities for Master theses and PhD theses with competitive salaries in Crete
- First Master thesis student will visit Crete in early 2023

Etiquette

- Address me as Alexis in English if you like
- Please address me by name when writing me an email, don't start emails with *“Hi, “* or *“Hello,”*
- Office hours

  → send me an email to arrange for a virtual meeting
- Live lectures: Laptop. smartphones, tablets
  CLOSED policy
- Feel free to ask as many questions as you like!
- Science needs controversial discussions!

Exam

- Oral exams to take place in March/April, dates to be determined
- I do not know yet if we will do them on-line at KIT or off-line
- I will send around a doodle to assign the exam slots towards the end of November (if this is not too late for you)

Bioinformatics Courses Overview

- Winter
  - *Introduction to Bioinformatics*
    - 2 hours per week lecture
    - Oral exam at the end of the semester
    - 3 ECTS
- Summer
  - *Hot Topics in Bioinformatics - Seminar*
    - You/we select interesting Bioinformatics papers and present them
    - 45 Minute presentation of paper
    - Submit a report of 8 pages at the end of the semester
    - #places restricted to 10 students
    - 3 ECTS
  - *Bioinformatics Practical*
    - Will probably not take place in summer 2023

Knowledge Check

- Please complete the knowledge check
- There are three parts:
  I. HPC background
  II. Algorithms Background
  III. Biology Background
- Just to see where we are

Knowledge Check Courses

- If you are interested to learn more, we recommend the following courses:
- Biology
  - Either "Ergänzungsfach Genetik" or single lectures thereof:
    - Grundlagen der Biologie (WS): introduction to molecular biology, no background required!
    - Molekularbiologie (WS): widely used molecular wet lab methods: cloning, DNA sequencing, PCR etc
    - Genetik (WS): in-depth discussion of replication/transcription/translation machinery
  - These courses are taught to biology bachelors (1st-2nd year), but can also be attend by CS students without problem

- Computer Science
  - Algorithm Engineering: how to make algorithms run fast **in practice**
  - All courses associated with parallel programming, GPU programming, and hardware

Teaching plan

- The current plan is to teach this course together with my PhD students & PostDocs
- This may lead to inconsistencies in language quality & presentation style, that is, a lack of continuity and consistency
- If it does not work, please let me know and I will take over → students rather enjoyed this in the past though

The Lab at Heidelberg

- 5 Phd students: Julia, Dimitri, Ben, Anastasis, Lukas
- 1 PostDoc: Benoit
- 1 senior scientist: Alexey
- 1 master student: Luise
- 1 Bachelor student: Jan

Your Instructors in chronological order

- Alexis

Prof. at KIT & research group leader at Heidelberg Institute for Theoretical Studies

Some Biographical Bullets

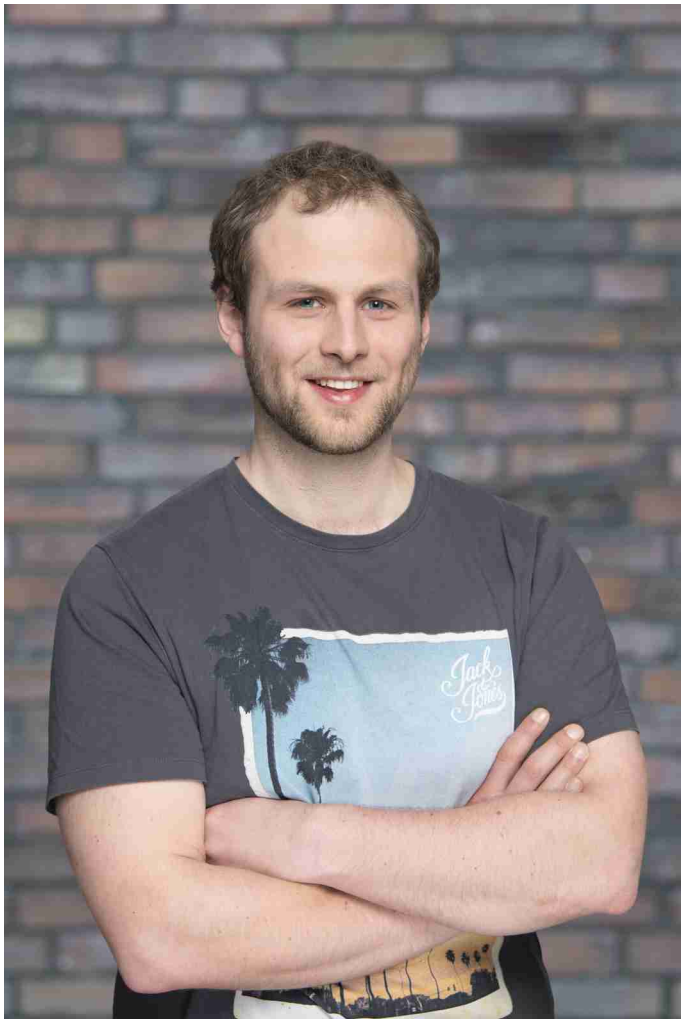
- until 1995: grown up in Athens, Greece
- 1995-2004: Diploma & PhD in CS at TU Munich
- 2005-2006: PostDoc in Crete
- 2006-2008: PostDoc at ETH Lausanne
- 2008-2010: Emmy-Noether group leader at LMU and then TU Munich
- Since 2010: Research group leader at HITS Heidelberg
- Since 2012: Full professor at KIT

The effects of SARS-CoV-2

- I was stuck in our house in Crete, Greece during the pandemic
  … and activated some old contacts on the island
- I am now an affiliated scientist at a research center in Crete with a focus on ancient DNA
- Opportunities for Master thesis projects with a focus on ancient DNA data analysis in Crete
- Lab web-site: https://ancient-dna.gr/index.php/en/
- This is different from the group I will set up in Crete through EU funding
- Nobel prize just went to ancient DNA researcher Svante Pääbo

Your Instructors in chronological order

- Lukas Hübner



Shared PhD student with Peter Sanders & former Master's student at KIT

Your Instructors in chronological order

- Alexey Kozlov

Former PhD student & former Master's student at KIT, the supplementary lecture tips are from Alexey, now staff scientist at HITS

Your Instructors in chronological order

- Benoit Morel

PostDoc, worked in industry prior to his PhD

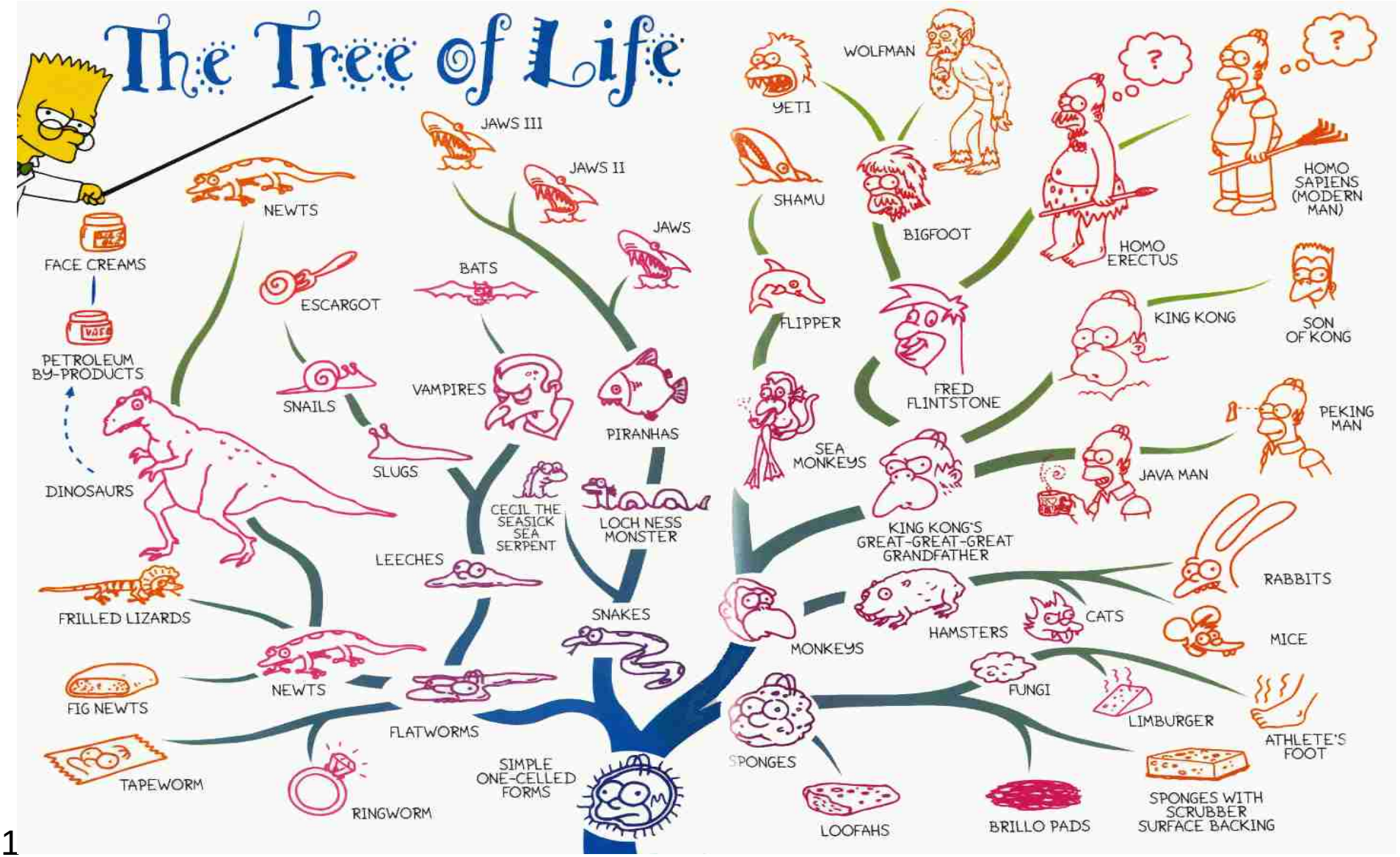
Goals of this Course

- introduce *some* biological terminology
- present *some* areas of Bioinformatics
- provide an overview
- show that there are interesting algorithmic & computational problems
- provide you the knowledge you need to work with us on research projects

Course Structure

- Introduction & biological Terminology (2 lectures → Alexis)
- Sequence Analysis (3 lectures → Lukas, Alexey, Alexis)
  - Pair-wise Sequence alignment (Lukas)
  - Searches on strings (Alexey)
  - Multiple Sequence Alignment (Alexis)
- Phylogenetics (6 lectures → Alexis, Benoit)
  - Intro to Phylogenetics (Alexis)
  - Phylogenetic search algorithms (Alexis)
  - Markov Chains (Alexis)
  - Discrete operations on trees (Benoit)
  - Likelihood of trees I (Alexis)
  - Likelihood of trees II (Alexis)
- Bayesian Phylogenetic Inference & MCMC (2 lectures → Alexis)
  - Introduction
  - Advanced Topics
- Population Genetics (1 lecture → Alexis)
  - Introduction
- Course revision (1 lecture → Alexis)

Main Research Focus of my Lab



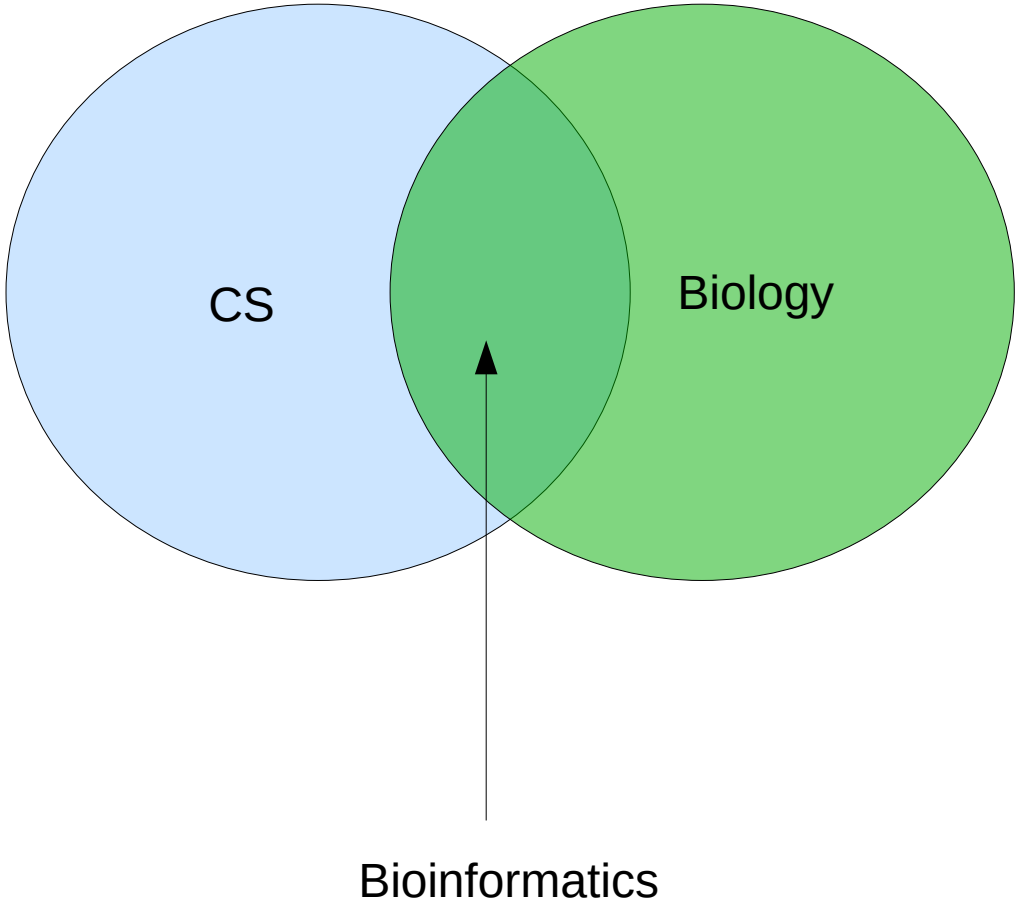
What is Bioinformatics?

- Term introduced by P. Hogeweg & B. Hesper in 1970 http://en.wikipedia.org/wiki/Paulien_Hogeweg
- There are many definitions
- I will provide my own:
  - In bioinformatics we intend to develop, optimize, and parallelize algorithms, models, and production-level software for analyzing, storing, and extracting knowledge from, biological raw data.
  - Key differences to CS
    - proof-of-concept implementations are not sufficient
    - we need to produce code that can be used by biologists
    - we need to provide support for the code
    - have a look at http://groups.google.com/group/raxml
    - Most famous Bioinformaticians are known for one or more widely-used and highly cited algorithms & tools they have developed
- “Biology easily has 500 years of exciting problems to work on” – Donald Knuth

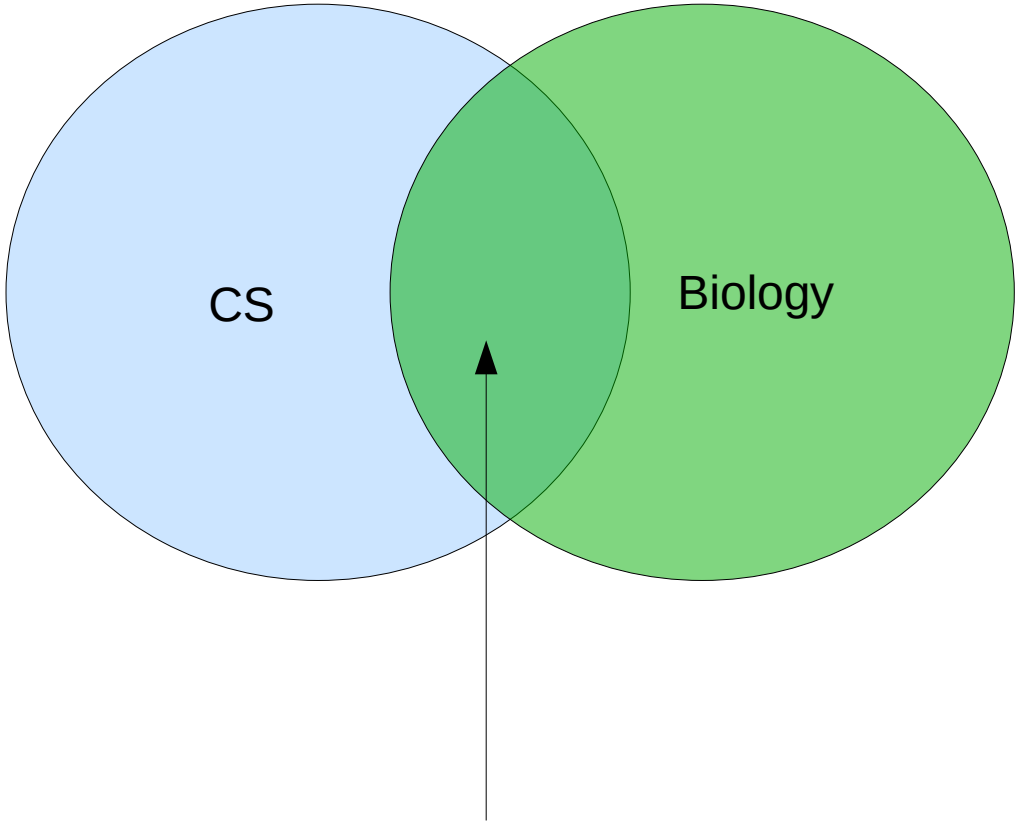
The ideal Bioinformatics tool

What is my hypothesis?

What is Bioinformatics?

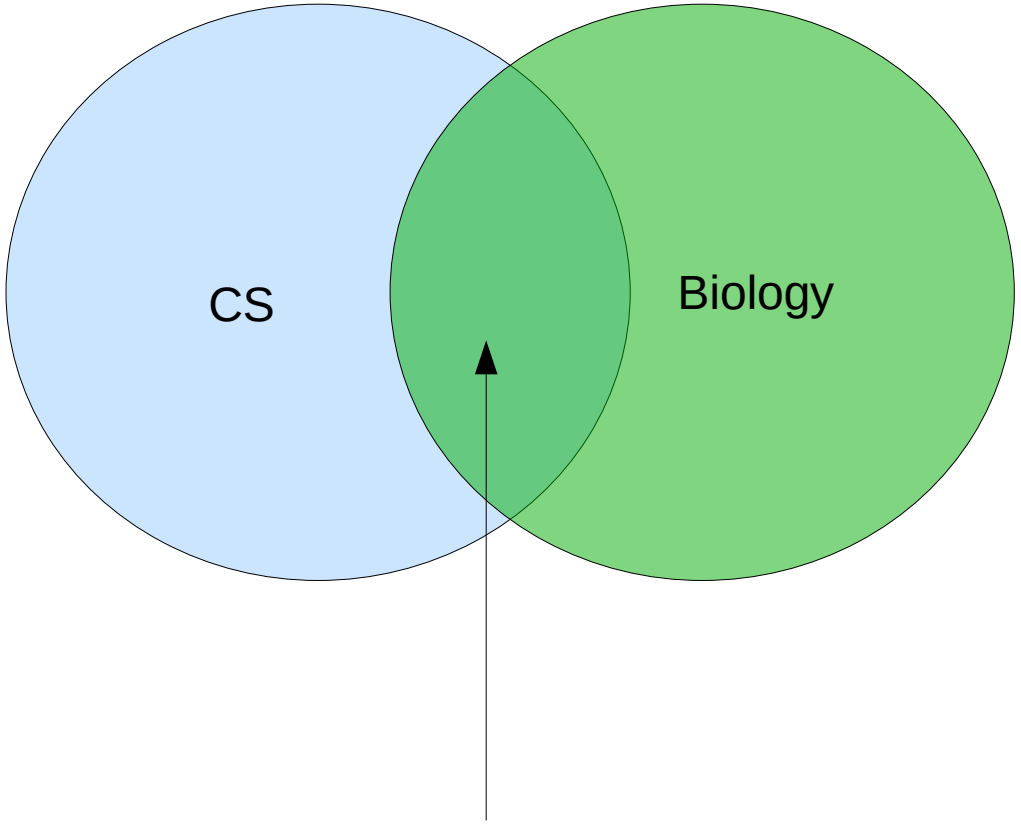


Why is this exciting?



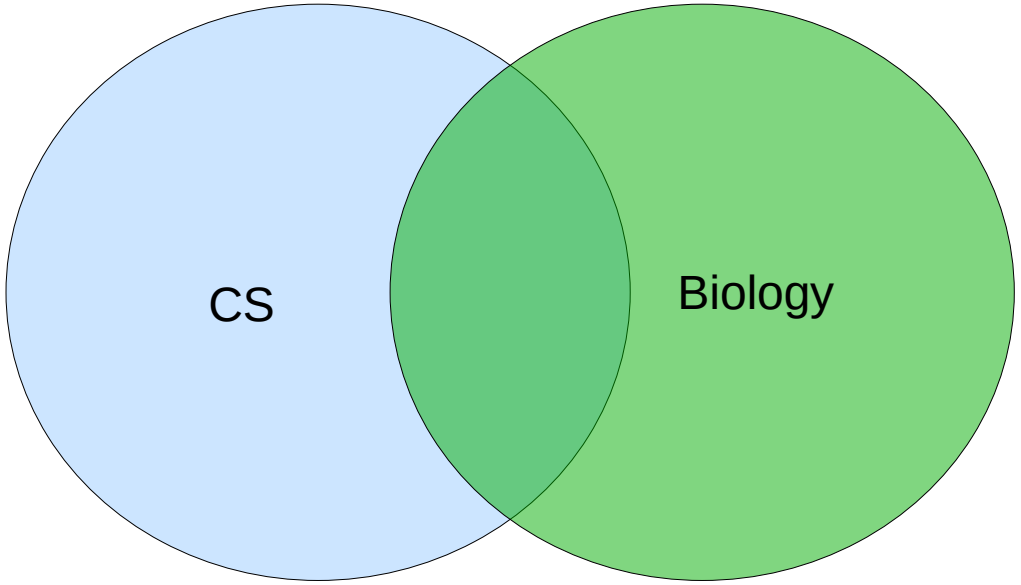
Important problems → medical applications,
Infectious diseases, genetic defects etc.
Masses of data → storage and analysis challenges
HPC → increased need for parallel codes

What are the challenges?



We can't be experts in everything → interdisciplinary collaboration
We need a culture of asking questions when we don't understand a term/concept!

Disciplines involved



Numerics
Statistics
Discrete Algorithms
Algorithm Engineering
Parallel Computing
Supercomputing
Software Engineering (in practice)

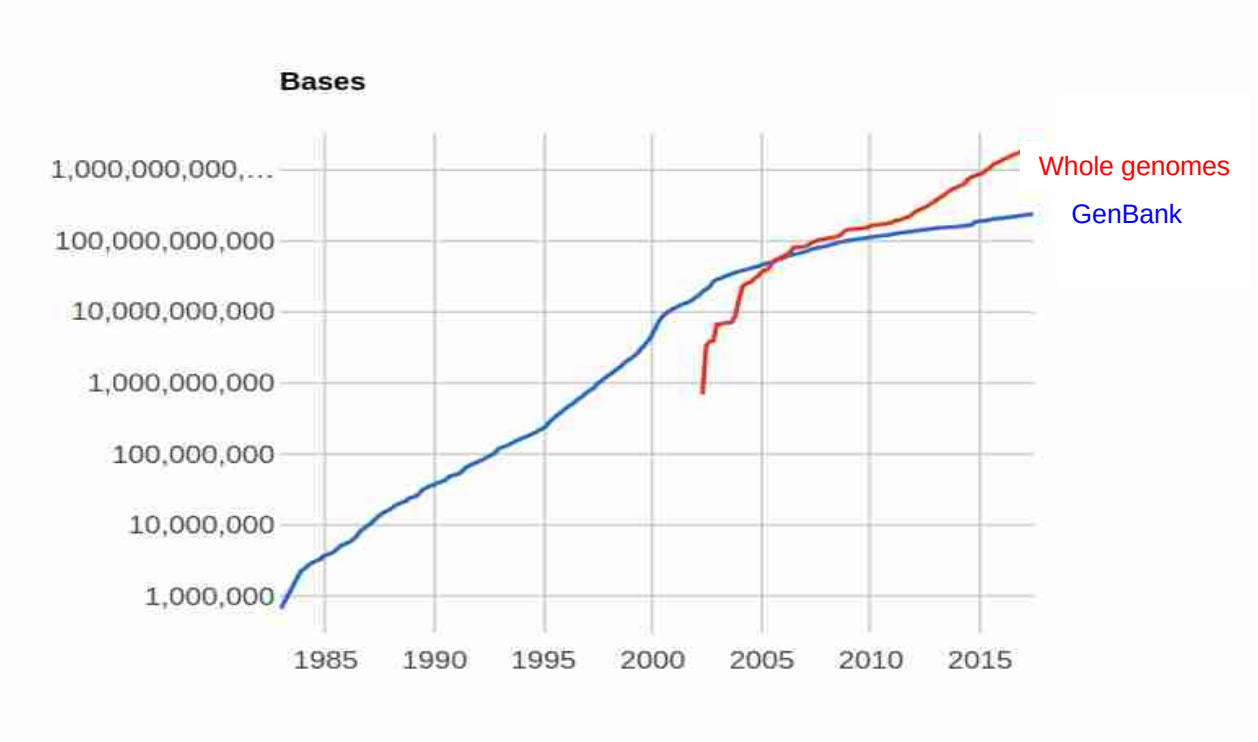
What is Biological Raw Data?

- There are many types of biological raw data
  - Images from microscopes
  - Microarray data
  - Protein structure data
  - Morphological data
  - Ecological data
  - Biogeographical data
  - …
- In this course we will mainly focus on *classic* Bioinformatics, that is, the analysis of molecular sequence data (DNA, protein data)

DNA data

- DNA data is available in public databases
- The most well-known one is GenBank
- Maintained by NCBI: National Center for Biotechnology Information, US
- Other databases for DNA data: EMBL (EU), DDBJ (Japan)

# of nucleotides/
base pairs
log-scale!



DNA data

- Genetic sequence
- Alphabet of 4 basic characters (nucleotides):
  - **A**denine
  - **C**ytosine
  - **G**uanine
  - **T**hymine
- A DNA sequence: *AACGTTTGA*
  - This sequence has 9 base pairs/nucleotides
- In RNA data: **T** is replaced by **U**racil
- A RNA sequence: *AACGUUUGA*
- We will see what RNA is later
- If we use **T** or **U** does usually not matter, computationally

Extended DNA alphabet

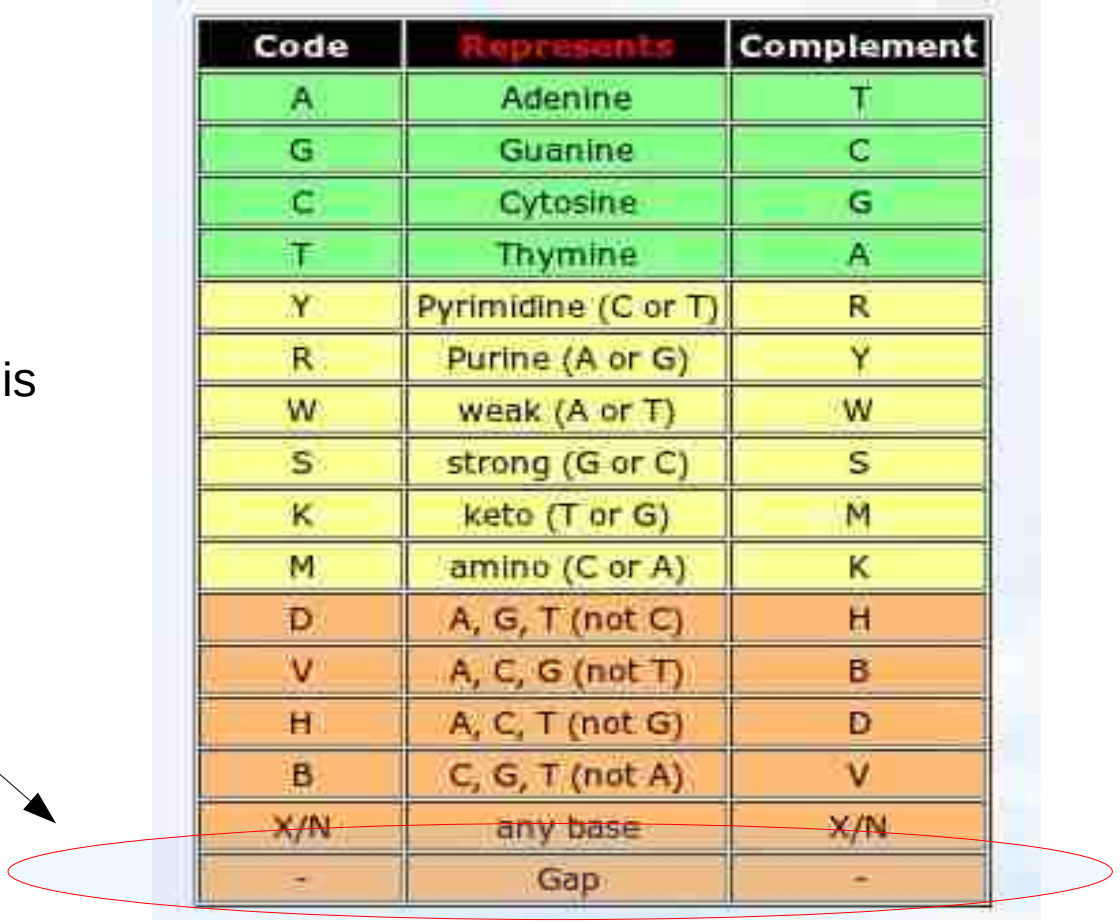
- DNA sequencing techniques are not exact
- Need to extend character set to denote:
  - could be an A *or* C
  - could be an A *or* C *or* G
  - ...
- International Union for Pure and Applied Chemistry (IUPAC) encoding

Ambiguity Code

Code	Represents	Complement
A	Adenine	T
G	Guanine	C
C	Cytosine	G
T	Thymine	A
Y	Pyrimidine (C or T)	R
R	Purine (A or G)	Y
W	weak (A or T)	W
S	strong (G or C)	S
K	keto (T or G)	M
M	amino (C or A)	K
D	A, G, T (not C)	H
V	A, C, G (not T)	B
H	A, C, T (not G)	D
B	C, G, T (not A)	V
X/N	any base	X/N

Ambiguity Code

We will talk about this later!



Code	Represents	Complement
A	Adenine	T
G	Guanine	C
C	Cytosine	G
T	Thymine	A
Y	Pyrimidine (C or T)	R
R	Purine (A or G)	Y
W	weak (A or T)	W
S	strong (G or C)	S
K	keto (T or G)	M
M	amino (C or A)	K
D	A, G, T (not C)	H
V	A, C, G (not T)	B
H	A, C, T (not G)	D
B	C, G, T (not A)	V
X/N	any base	X/N
-	Gap	-

DNA Sequencing

- The process of reading the nucleotide bases in a DNA molecule
- There exist various sequencing technologies
- Properties
  - Cost
  - Speed
  - Amount of data/Number of Sequences
  - Sequence length
  - Error rate

DNA Sequencing

- Sanger sequencing (*since 1977)*
  - High accuracy: 99.9%
  - Long sequences: 300-900 nucleotides
  - Expensive: $2400 per 1,000,000 nucleotides
  - Few sequences: up to ≈ 100
- Next-generation sequencing (since *2007*)
  - Lower accuracy 98-99.9%
  - Short sequences (100-400 nucleotides)
  - Inexpensive $1 - $10 per 1,000,000 nucleotides
  - Many sequences: 500 – 3,000,000,000 per sequencer run

A next-Generation Sequencer

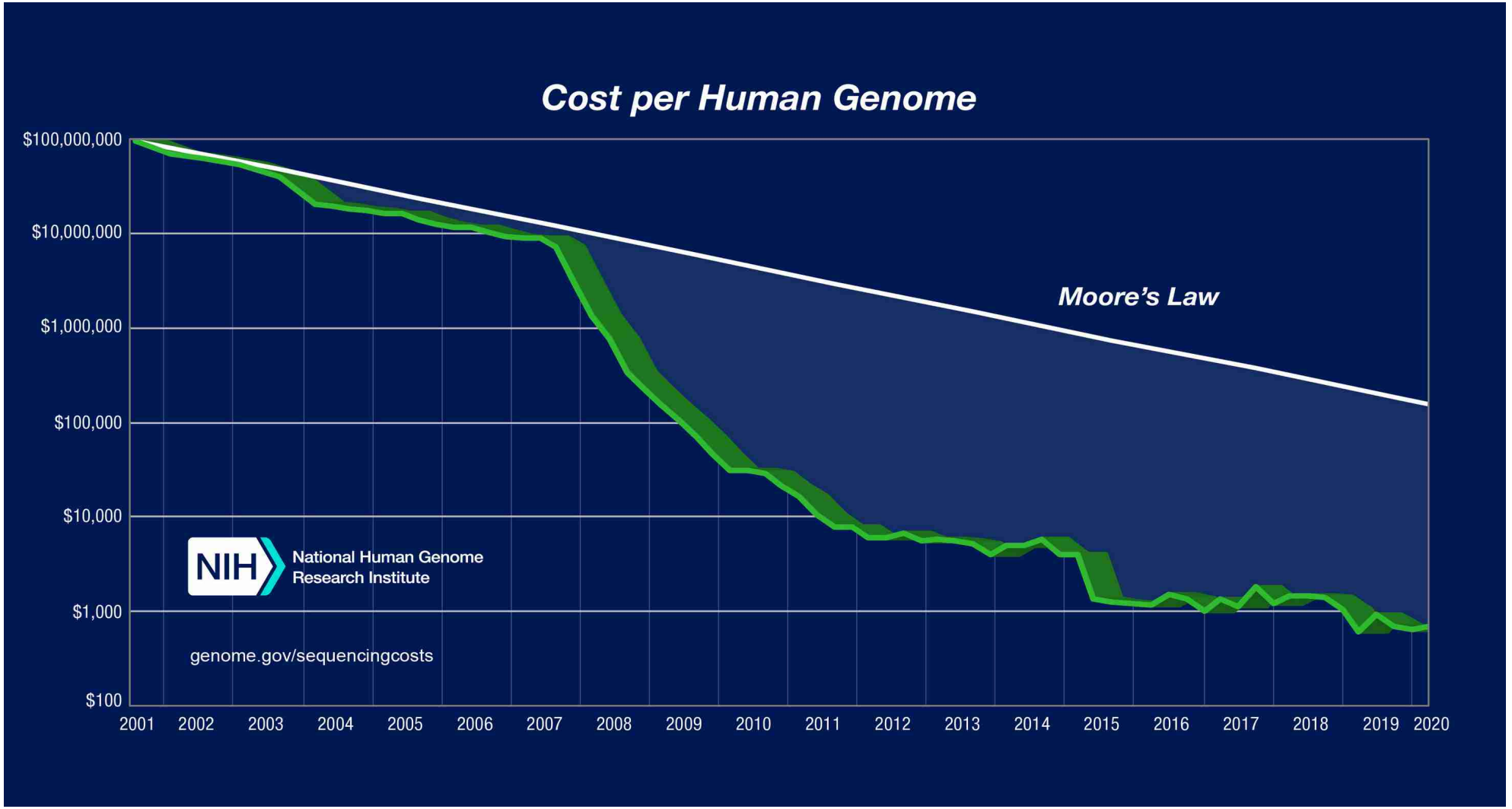
A Next^2 Generation Sequencer

DNA Sequencing

- Sanger sequencing
  - High accuracy: 99
  - Long sequences:
  - Expensive: $2400 per 1 00 nucleotides
  - Few sequences: up to ≈ 00
- Next-generation sequencing (since *2007*)
  - Lower accuracy: 98-99.9%
  - Short sequences (100-400 nucleotides)
  - Inexpensive $1 - $10 per 1,000,000 nucleotides
  - Many sequences: 500 – 3,000,000,000 per sequencer run

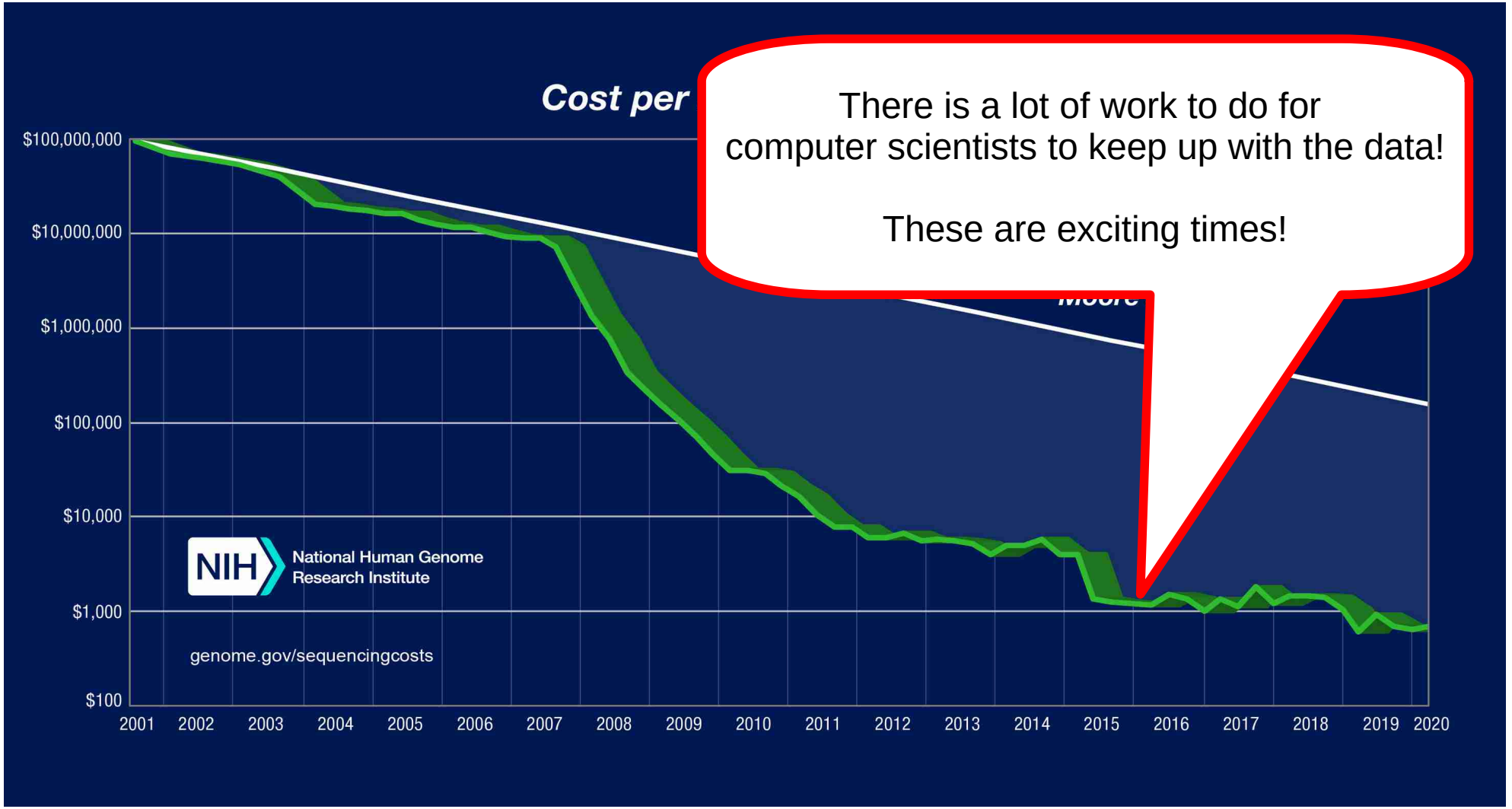
This is a revolution!
We will see how this data can be used and analyzed in this course!

The revolution



Sequencing cost versus processing cost!

The revolution



Sequencing cost versus processing cost!

Remember

- Back in 2001 the complete sequencing of the human genome made the news!
- Papers appeared in *Science & Nature*
- Now it's almost boring: aha, somebody sequenced yet another genome
- Our lab in 2014
  - Evolutionary analysis of 50 bird *genomes*
  - Evolutionary analysis of 140 insect transcriptomes → we will see what a *transcriptome* is later

Bird & Insect Papers

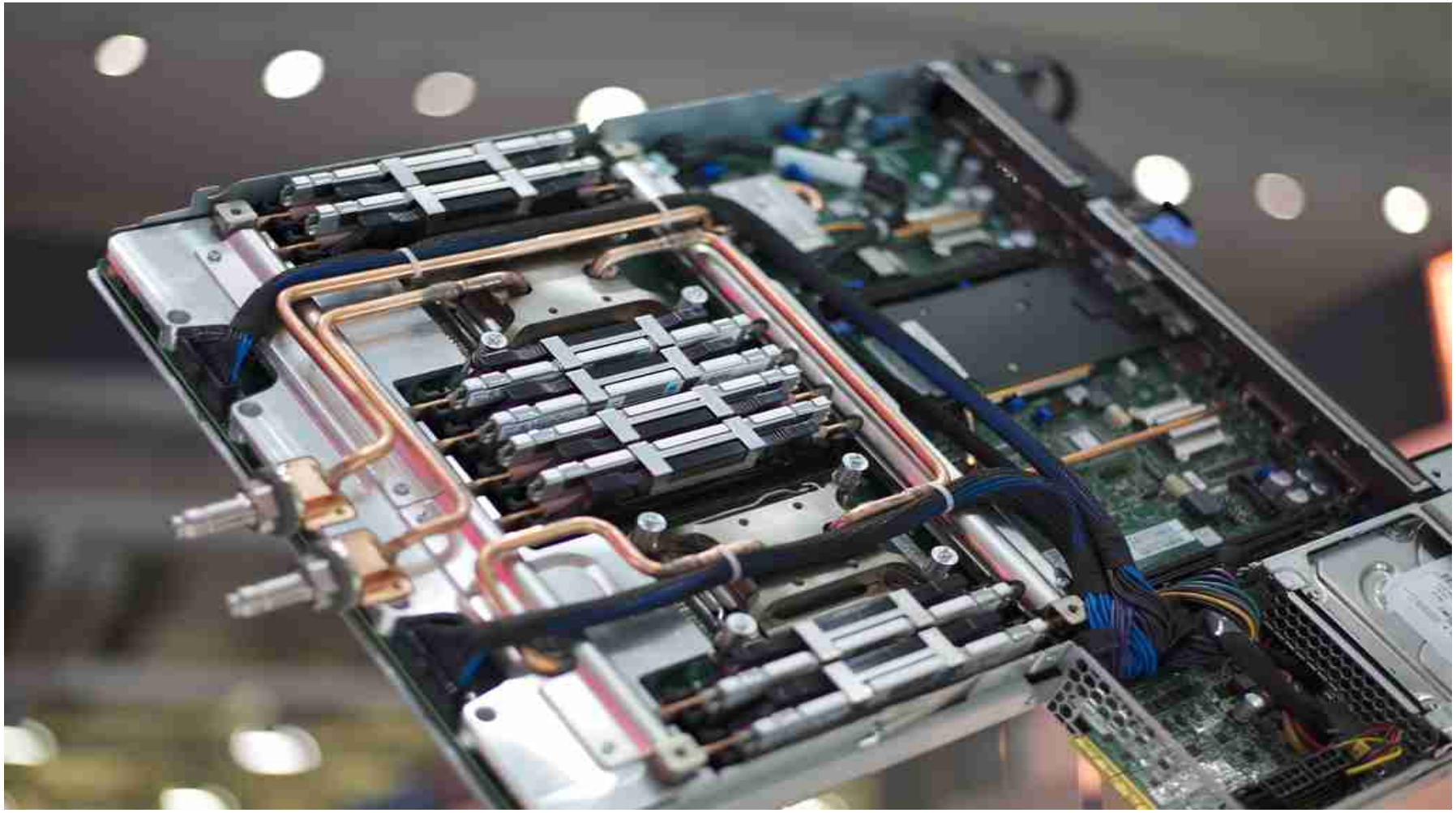
Bird & Insect Papers

In 2018/19 we are analyzing datasets of 1500 insects and 350 birds

Supercomputing

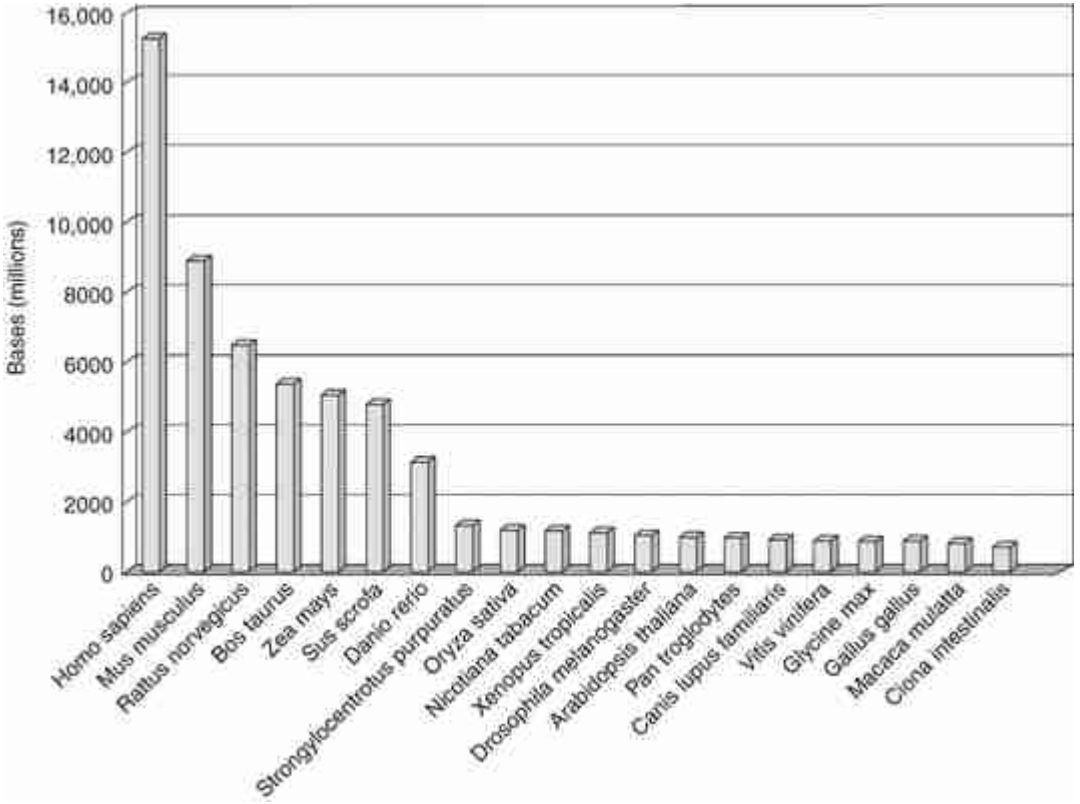
Munich supercomputer: SuperMUC

SuperMUC Cooling



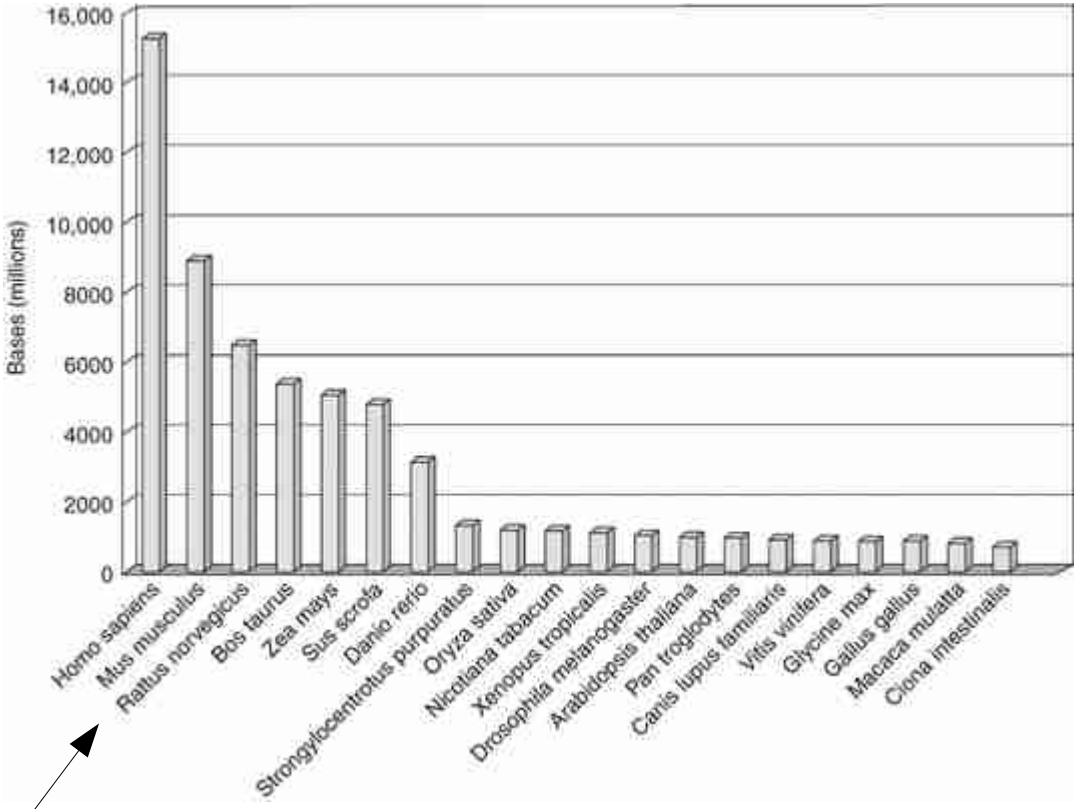
DNA data

- GenBank: most-sequenced species



DNA data

- GenBank: most-sequenced species



 Some of these species are so-called *model organisms*

Model Organism

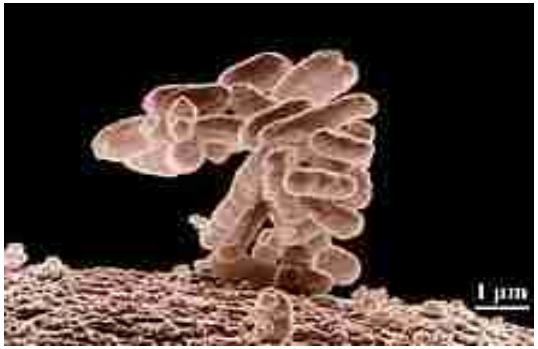
- A species that is extensively studied/sequenced to understand particular biological phenomena, with the expectation that discoveries made for the model organism will provide insight into the workings of other organisms.
- Selection criteria:
  - easy experimental manipulation
  - ease of genetic manipulation
  - easy to grow

    → short life-cycle/generation times
  - easy to extract DNA data
  - Economical importance → rice
- Often researchers reverse-engineer organisms
- Full list of model organisms: http://www.life.umd.edu/labs/mount/Models.html

Some Model Organisms

- *Escheria coli*

  gut bacterium → can cause food poisoning, grows fast, inexpensive to cultivate



- *Drosophila Melanogaster*

  fruit fly → breeds quickly

- *Arabidopsis Thaliana*

  flowering plant → small genome

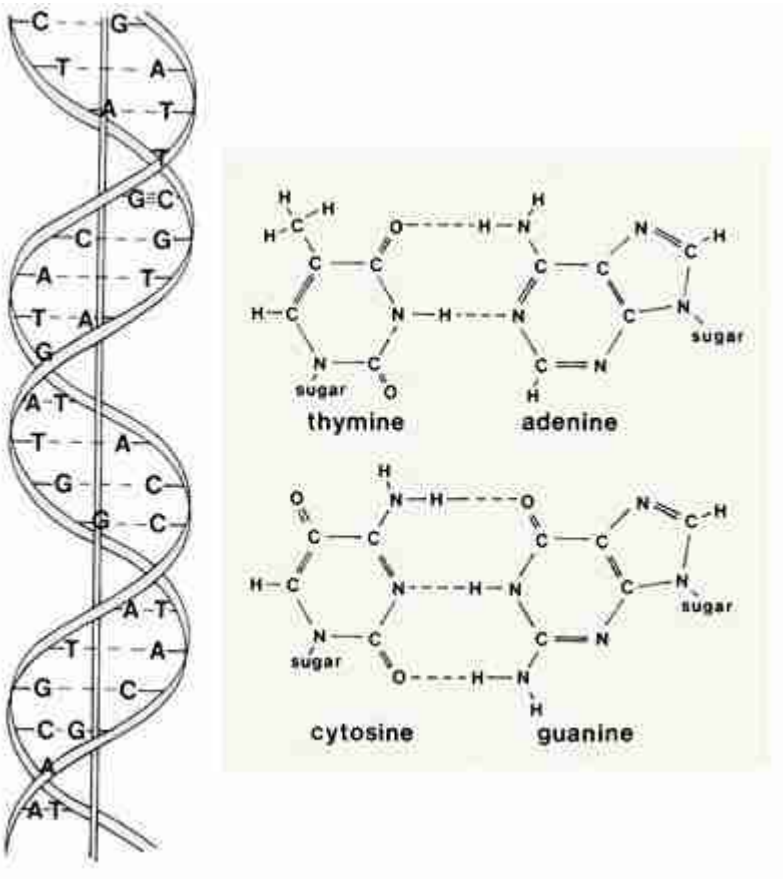
Back to DNA

- What's a base pair?
- Pairing of A with T *or* C with G in double-stranded DNA

AATTGGC

TTAACCG

complement



Sloppy terminology

- The # of base pairs is frequently used as synonym for the # of nucleotides in a single-strand sequence
- This sequence has 5 nucleotides: *ACGGT*
- We can also say that it has 5 base pairs
- As in CS we use kilo, giga, etc for sequence lengths
  - kb → kilo-bases
  - Mb → Mega-bases
  - Gb → Giga-bases

Genome

- The full genetic information of an organism
  - Contains all chromosomes
  - Comprises the coding & non-coding sequence data of the organism
  - Coding sequence data → part of the genome that encodes proteins
  - Non-coding (in earlier days: junk) DNA → part of the genome that does not encode proteins but still has a function
    - The function of non-coding DNA is only partially known
    - Non-coding DNA regulates protein processes

Genome Size

- Not necessarily correlated with organism complexity
- Homo Sapiens: 3.2 Gb (Giga-bases)
- Marbled lungfish: 130 Gb (Giga-bases)
- Plants often have very large genomes → partially due to redundant information caused by hybridization

Terminology introduced

- Sequence data/sequence
- Nucleotide/base-pair
- DNA/RNA
- Ambiguity coding
- Sequencing
  - Sanger Sequencing
  - Next (Next) Generation Sequencing
- Genome
- Model Organism
- Double-stranded DNA
- Coding versus non-coding DNA

Drop me an Email!

- Alexandros.Stamatakis@kit.edu