

BIOINFORMATICS

(FOR COMPUTER SCIENTISTS)

MPCS56420
AUTUMN 2020
SESSION 1



THE UNIVERSITY OF
CHICAGO



WELCOME TO BIOINFORMATICS

COURSE LOGISTICS

- Prerequisites
 - Algorithms
 - Core Programming
- Consent of instructor
 - Requires departmental approval



COURSE LOGISTICS

- Flipped Classroom
 - Videos of lectures
 - Meet for discussions and office hours



COURSE LOGISTICS

- Regular Class Meeting
 - Thursday 5:30-6:30pm
- Lecture Video Recap/Office Hours
 - Monday 1:00 (See slack poll)
- Office hours
 - Tuesday 5:30 (Neal Conrad)
 - Monday 1:30pm (Andrew) (See slack poll)

Introduce upcoming weeks material
Introduce/update assignment
Open Q and A

Student presentations



COURSE LOGISTICS

- Regular Class Meeting
 - Thursday 5:30-6:30pm
- Lecture Video Recap/Demos
 - Monday 1:00 (See slack poll)
- Office hours
 - Tuesday 5:30 (Neal Conrad)
 - Monday 1:30pm (Andrew) (See slack poll)

Ask questions about the video (session/slides#)

Questions from Slack to be addressed

Demos and walkthroughs

COURSE LOGISTICS

- Regular Class Meeting
 - Thursday 5:30-6:30pm
- Lecture Video Recap/Demos
 - Monday 1:00 (See slack poll)
- Office hours
 - Tuesday 5:30 (Neal Conrad)
 - Monday 1:30pm (Andrew) (See slack poll)

Please try and schedule a specific time to meet for office hours.

#office-hours

COURSE LOGISTICS

- No biology required
 - Weekly “need to know” basis
 - Class will adjust for student backgrounds
 - Online resources for learning biology
 - Khan Academy, EdX, Coursera, Udacity, YouTube

COURSE LOGISTICS

- Biology is a learn-as-you-go
- Researchers study one gene for a lifetime

6 hours later



COURSE DESIGN

INSTRUCTORS

- Andrew
 - The University of Chicago
 - Center for Structural Genomics of Infectious Diseases
 - 10 years teaching MPICS Program
 - ANL
 - Midwest Center for Structural Genomics, Leadership Computing Facility
 - Computation Institute

COURSE DESIGN

INSTRUCTORS

- Research programs
 - Biomolecular modeling; Drug Discovery
 - Protein Crystallography
 - HPC for bioinformatics pipelines
 - STEM Education (K-12)

COURSE DESIGN

INSTRUCTORS

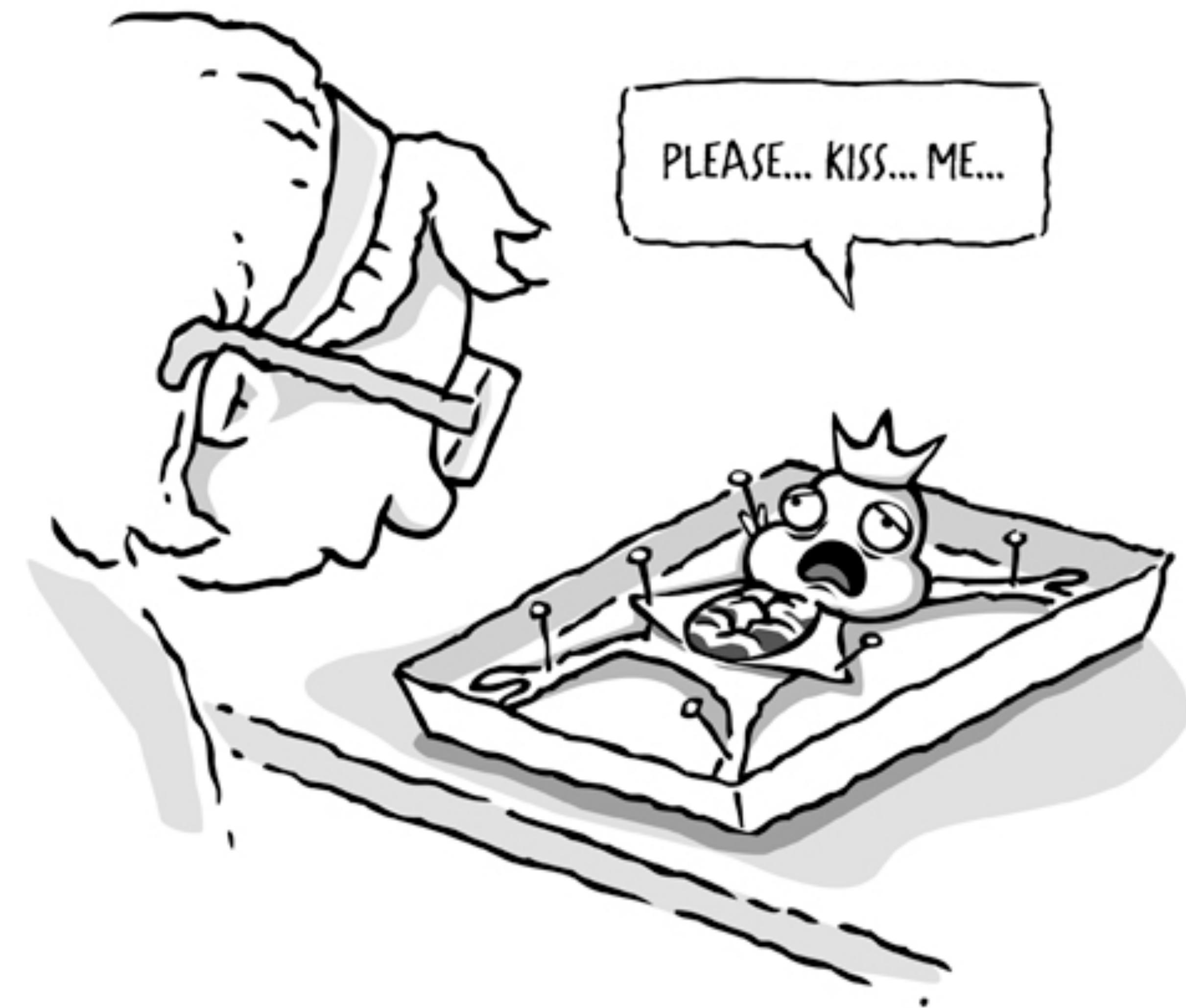
- Neal Conrad
 - Software Engineering Associate 2 at Argonne National Laboratory's Data Science and Learning Division
 - Graduated from MPSCS in March, 2019
 - His work includes designing and building large-scale web applications, data visualization tools, and open-source libraries to support scientific research in bioinformatics.

COURSE DESIGN

COURSE DESIGN

CHALLENGES

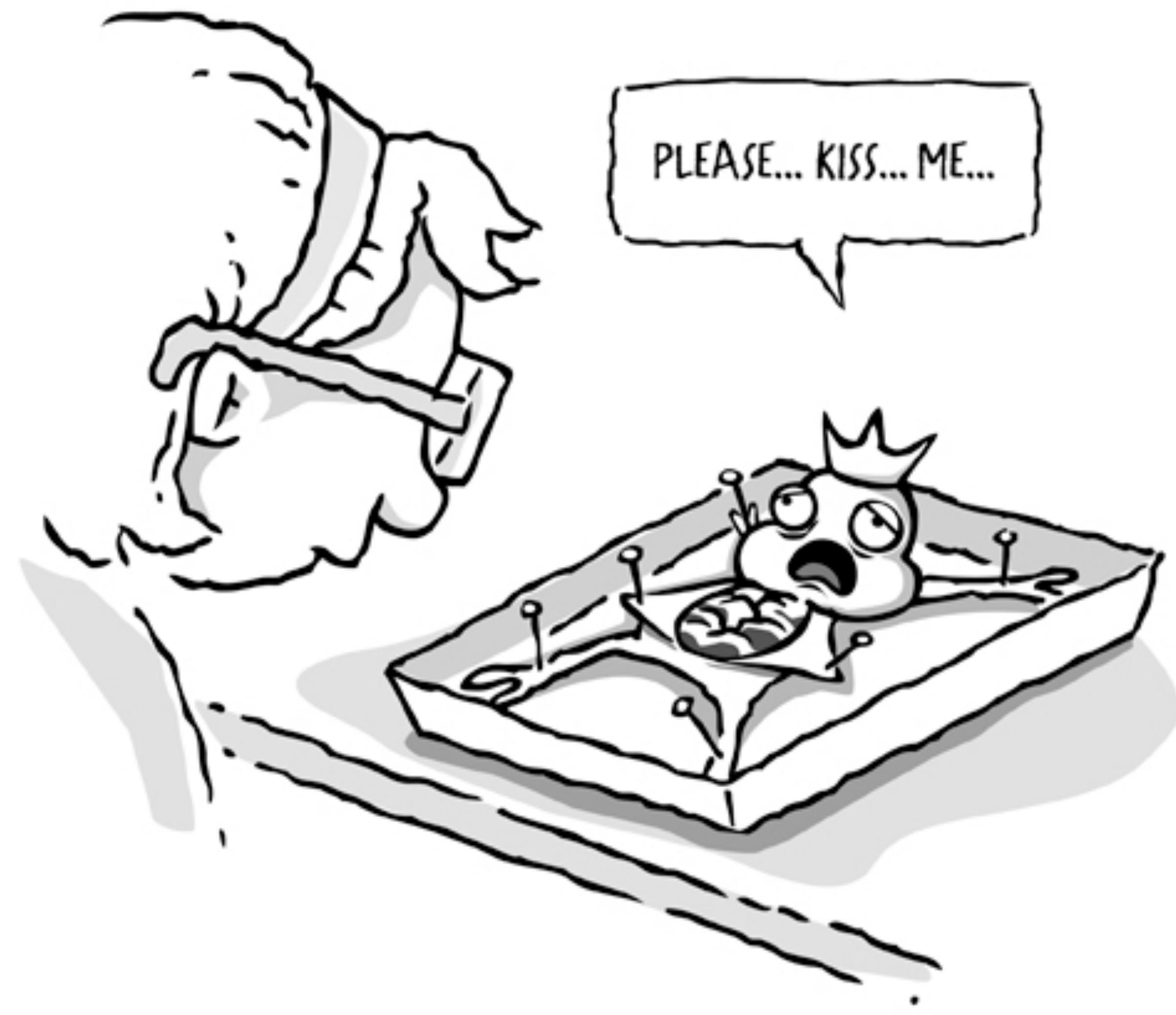
- People from different backgrounds
 - No biology since high school



COURSE DESIGN

CHALLENGES

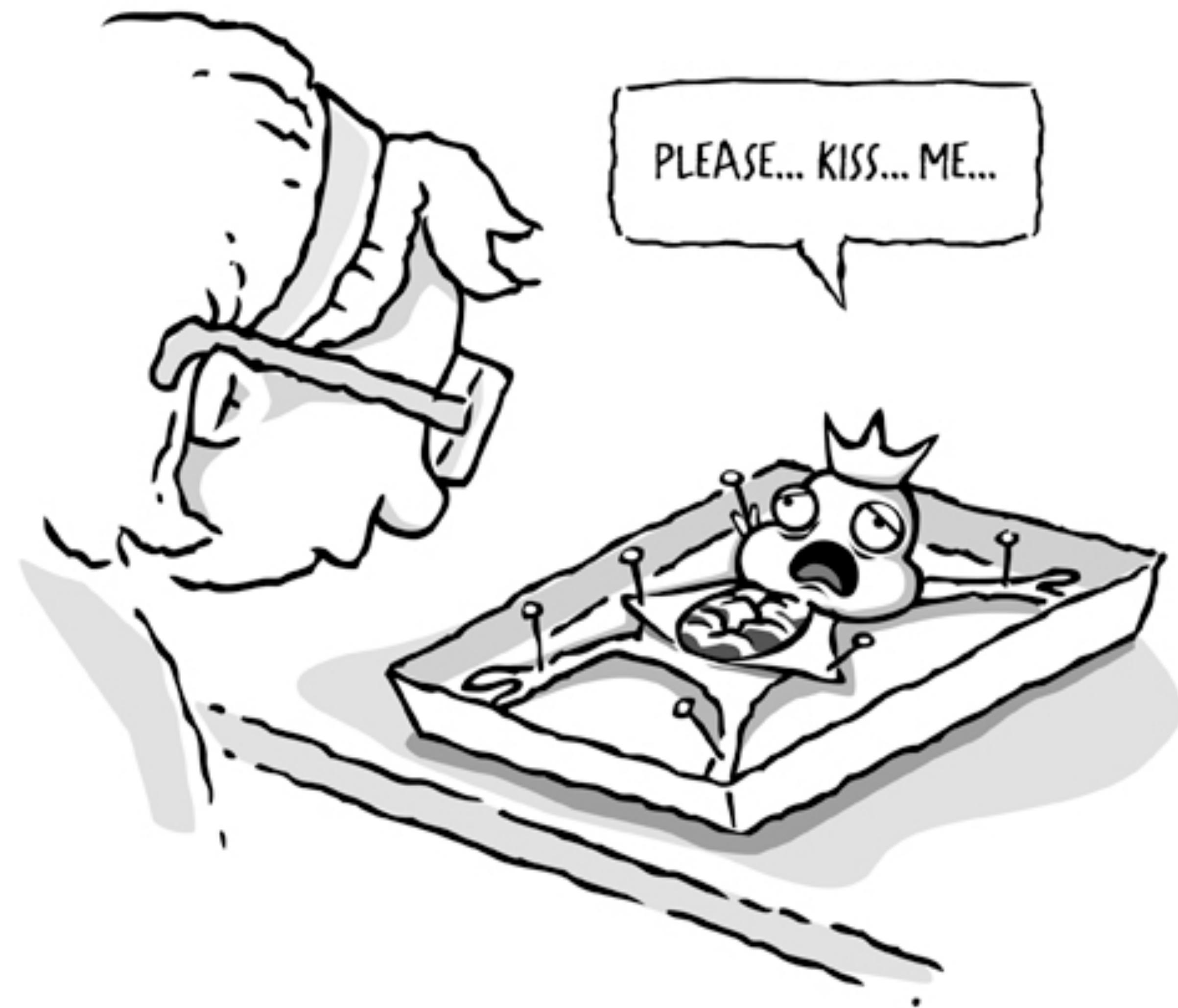
- Wide topic that touches many fields and sub-fields
 - Molecular biology, structural biology, genomics, medicine



COURSE DESIGN

CHALLENGES

- Different programming backgrounds and skill sets
 - HPC, scripting, compiled languages, unix, web, etc.



COURSE DESIGN

CHALLENGES

- Different interests and goals for the course
 - Casual interest in biology
 - Elective credit



COURSE DESIGN

PROVIDE KNOWLEDGE AND TOOLS TO
UNDERSTAND AND SOLVE A
RESEARCH PROBLEM
(IN BIOLOGICAL SPACE) WITH THE
AID OF COMPUTATION



THINGS YOU
LEARN HERE CAN
BE UNIVERSALLY
APPLIED

COURSE DESIGN

COURSE GOALS

- Introduce major techniques and bioinformatics algorithms
 - Implement some of them
- Develop applications for bioinformatic analysis
 - Make them accessible to other researchers
- Become proficient in most important bioinformatics databases
 - NCBI, EBI, PDB, etc.
- Provide practical experience in research

COURSE DESIGN

SECONDARY GOAL:
BECOME "COCKTAIL
PARTY" CONVERSANT
IN BIOLOGY



SYLLABUS

COURSE DESIGN

- Week 1,2 - Biology and Biological Databases
- Week 3,4 - Sequence Alignment
- Week 5,6 - Protein Bioinformatics
- Week 7,8,9 - Discovery, Genomes and Human Variation

COURSE DESIGN

- Week 1: Genomics, Bioinformatics and Molecular Biology
 - A high-level view of increasingly important role of computing in the biological sciences will be presented.
- Week 2: Genomes, Sequences and Databases
 - A survey of the current state of the art in storing, organizing and analyzing large data sets will be discussed
 - The advantages and disadvantages of these methods will be explored in the context of academic and commercial research initiatives.

COURSE DESIGN

- Week 3: Sequence Alignment
 - Fast, reliable alignment of text strings started the bioinformatics revolution. This lecture will show how these seemingly simple strings form the basis of almost all bioinformatics research.
- Week 4: Analyzing DNA and Protein Sequences
 - Techniques for sequence analysis will be discussed

COURSE DESIGN

- Week 5: Protein Structure and Function
 - Proteins are central building blocks of all organisms
 - Take bioinformatics to the third-dimension, showcasing how the spatial assembly and interactions of proteins support life and cause of disease
- Week 6: Structure Analysis and Molecular Modeling
 - Understanding protein function holds the promise developing therapeutics and curing diseases, but the computational complexity of analyzing three-dimensional models presents obstacles that have been difficult to overcome. This lecture will discuss approaches to shape analysis and comparison that can be scaled to large data sets.

COURSE DESIGN

- Week 7: In-Silico Drug Discovery
 - Approaches to using computer models to develop new drugs will be presented. We will discuss how years of playing Tetris might be more useful than you thought in combating antibiotic resistant pathogens.
- Week 8: Student Presentations / The Human Genome and Disease
 - The cause of diseases can be as simple as a single misplaced letter in a DNA sequence. From gene to disease, we will trace the genetic origin of disease. We will explore different approaches to cataloging and analyzing these changes.

COURSE DESIGN

- Week 9: Personal Genomics and Drug Discovery
 - Personalized genomic analysis is being used by consumers to better understand their health and their ancestry. The technologies used to power these services will be introduced as well as the different approaches used to provide web services to analyze the data.
- Final Exam Week 10: Final Project Presentations
 - Students will present a research topic in bioinformatics of their own choosing

COURSE WORK

COURSE WORK

- 6 assignments (10% each)
 - Split between practical problems and implementation problems
- Gene Study Presentations (15%)
- Final project (25%):
 - Implement a bioinformatic method of your choosing

WE WILL DISCUSS
THIS IN DETAIL
LATER ON

COURSE WORK

Week	Material	Homework	Final Project
1	Bioinformatics, Genomics, and Molecular Biology	Assignment 1	
2	Biological Databases	Assignment 2	
3	Sequence Alignment	Assignment 3	
4	Analyzing Sequences	Assignment 4	
5	Protein Structure and Function	Assignment 5	
6	Molecular Modelling	Assignment 6	
7	in-silico drug discovery	Gene Study	Proposals Due before class
8	Student Gene Presentations		
9	The Human Genome and Disease		
Final	Student Video Presentations		Projects Due

COURSE WORK

- Identify a gene of interest
 - Related to disease, biofuels, manufacturing, etc.
 - Family of genes, what pathway, what does it do?
 - Find a gene of unknown function, predict its function...
- Become an expert on that gene
 - Publications
 - Sequence, structure and function analysis on it
- Present to classs

K-RAS AND CANCER

AIN'T NUTHIN' BUT A G PROTEIN, BABY.

ASSIGNMENTS & PROJECTS

FINAL PROJECT

- Project of your choosing
- Examples
 - Implement a published bioinformatic method
 - Invent your own
 - Research a topic
 - Anything related to anything we've talked about...
- Present it to the class

ASSIGNMENTS & PROJECTS

- Honor Code
 - All the assignments should be your own work
 - Department policies are strictly enforced
- Citing Resources
 - Cite any resources you use on homework, presentation and projects
 - Includes online resources
 - StackOverflow, BioStars, blogs, GitHub, etc.
- Third-party libraries and software
 - There are many great libraries for bioinformatics
 - Unless specifically stated, you should not use them

COURSE TECHNOLOGIES

COURSE TECHNOLOGIES

- Bioinformatics is focused on developing solutions to biological problems
 - Not on mastery of any particular language
 - Many bioinformatics resources violate all CS good design and implementation rules
 - Utility trumps all
 - Flat files are the lifeblood of many bioinformatic pipelines
 - The rise and fall of DoubleTwist

```
proteinworks — abinkows@miraclac1:~ more — 127x77
abinkows@miraclac1:~ bash

import os
import sys
import math

def sort_by_atom_number(txt):
    lines = {}
    new_lines = []
    for line in txt.splitlines():
        if line.startswith("ATOM"):
            atom_number = int(line[7:11])
            lines[atom_number] = line

    for line in sorted(lines):
        new_lines.append(lines[line])
    return '\n'.join(new_lines)

#-
def strip_lines(pdb_txt, tag_func):
    new_lines = []
    for line in pdb_txt.splitlines():
        if tag_func(line):
            continue
        new_lines.append(line)
    return '\n'.join(new_lines)

#-
def strip_pdb_extension(filename):
    return os.path.splitext(os.path.basename(filename))[0]

#-
def atomic_distance(atom1_xyz, atom2_xyz):
    """ atom1_xyz is a list [x,y,z] coordinate """
    return math.sqrt((atom1_xyz[0]-atom2_xyz[0])**2-
                    (atom1_xyz[1]-atom2_xyz[1])**2-
                    (atom1_xyz[2]-atom2_xyz[2])**2)

#-
def xyz_from_pdbleline(line):
    x = float(line[30:38])
    y = float(line[38:46])
    z = float(line[46:54])
    return [x,y,z]

#-
def extract_atom_neighbors(pdb_txt, ligand_xyz, cutoff):
    new_lines = []
    for line in pdb_txt.splitlines():
        if line.startswith("ATOM"):
            protein = xyz_from_pdbleline(line)
            for ligand_atom in ligand_xyz:
                dist = atomic_distance(protein, ligand_atom)
                if dist < cutoff:
                    #print "%s - %f" % (line, dist)
                    new_lines.append(line)
    return new_lines

#-
def extract_ligand_coordinates(pdb_txt, ligand_key):
    """ Return an [x,y,z] list of the coords of a given ligand
    TODO: Optionally print to file?
    """
    print "## Extracting coordinates for ligand key "
    print ligand_key

    coords = []
    for line in pdb_txt.splitlines():
        if line.startswith("HETATM"):
            res_type = (line[17:20]).strip()
            chain_id = line[21]
            res_num = int(line[22:26])
            #print ligand_key
            coords.append([res_type, chain_id, res_num])

    return coords
```

COURSE TECHNOLOGIES

- Bioinformatics requires aptitude in a variety of programming languages
 - Scripting (Python, Perl)
 - Command line (Bash, wget, curl, etc.)
 - Compiled languages (Fortran, C, C++)
 - Specialized languages (R, Matlab)
 - Web programming (Javascript, PHP, HTML)

```
import os
import sys
import math

def sort_by_atom_number(txt):
    lines = {}
    new_lines = []
    for line in txt.splitlines():
        if line.startswith("ATOM"):
            atom_number = int(line[7:11])
            lines[atom_number] = line

    for line in sorted(lines):
        new_lines.append(lines[line])
    return '\n'.join(new_lines)

#-----
def strip_lines(pdb_txt, tag_func):
    new_lines = []
    for line in pdb_txt.splitlines():
        if tag_func(line):
            continue
        new_lines.append(line)
    return '\n'.join(new_lines)

#-----
def strip_pdb_extension(filename):
    return os.path.splitext(os.path.basename(filename))[0]

#-----
def atomic_distance(atom1_xyz, atom2_xyz):
    """ atom1_xyz is a list [x,y,z] coordinate """
    return math.sqrt((atom1_xyz[0]-atom2_xyz[0])**2+
                     (atom1_xyz[1]-atom2_xyz[1])**2+
                     (atom1_xyz[2]-atom2_xyz[2])**2)

#-----
def xyz_from_pdbleline(line):
    x = float(line[30:38])
    y = float(line[38:46])
    z = float(line[46:54])
    return [x,y,z]

#-----
def extract_atom_neighbors(pdb_txt,ligand_xyz,cutoff):
    new_lines = []
    for line in pdb_txt.splitlines():
        if line.startswith("ATOM"):
            protein_xyz = xyz_from_pdbleline(line)
            for ligand_atom in ligand_xyz:
                dist = atomic_distance(protein_xyz,ligand_atom)
                if dist < cutoff:
                    new_lines.append(line)
    return new_lines

#-----
def extract_ligand_coordinates(pdb_txt,ligand_key):
    """ Return an [x,y,z] list of the coords of a given ligand
        TODO: Optionally print to file?
    """
    print "### Extracting coordinates for ligand key "
    print ligand_key

    coords = []
    for line in pdb_txt.splitlines():
        if line.startswith("HETATM"):
            res_type = (line[17:20]).strip()
            chain_id = line[21]
```

MORE IMPORTANT
THAN YOU MIGHT
THINK

COURSE TECHNOLOGIES

- Command line tools
 - “Dirty little secret” of bioinformatics
 - You have to do something with your data before/after your big cluster runs

Command-line tools can be 235x faster than your Hadoop cluster

Sat 25 January 2014 by Adam Drake

Introduction

As I was browsing the web and catching up on some sites I visit periodically, I found a cool article from [Tom Hayden](#) about using [Amazon Elastic Map Reduce](#) (EMR) and [mrjob](#) in order to compute some statistics on win/loss ratios for chess games he downloaded from the [millionbase archive](#), and generally have fun with EMR. Since the data volume was only about 1.75GB containing around 2 million chess games, I was skeptical of using Hadoop for the task, but I can understand his goal of learning and having fun with mrjob and EMR. Since the problem is basically just to look at the result lines of each file and aggregate the different results, it seems ideally suited to stream processing with shell commands. I tried this out, and for the same amount of data I was able to use my laptop to get the results in about 12 seconds (processing speed of about 270MB/sec), while the Hadoop processing took about 26 minutes (processing speed of about 1.14MB/sec).

After reporting that the time required to process the data with 7 c1.medium machine in the cluster took 26 minutes, Tom remarks

"This is probably better than it would take to run serially on my machine but probably not as good as if I did some kind of clever multi-threaded application locally."

This is absolutely correct, although even serial processing may beat 26 minutes. Although Tom was doing the project for fun, often people use Hadoop and other so-called *Big Data* (*tm*) tools for real-world processing and analysis

COURSE TECHNOLOGIES

CONSIDERATIONS FOR BIOINFORMATICS PROGRAMMING

- Portability - Will it run on my desktop and a legacy SGI machine?
- Scalability - Will it run on a cluster?
- Development speed - How long will it take to write
- Longevity - Is this a one time script or a full fledged application?
- Deliverability - Will this eventually be a web app or in the App Store?
- Target Audience - Who (besides me) may be running this? What is their background?

COURSE TECHNOLOGIES

- Programming languages for this course
 - Lectures and demonstrations will be conducted mostly in Python
 - Historical language of bioinformatics is Perl
- Why Python?
 - Great online resources and support
 - Support for SciPy scientific programming
 - Extensibility for C modules
 - Optimize when you need it
 - Native web application language

DRUDGE REPORT 2014® Hacker News Google News Screen Time Analytics Journals GTasks GrabLinks

PLOS Collect... Untitled 1 Introduction t... uchicago-link... PLOS Collect... Online Resou... MPCS 56420

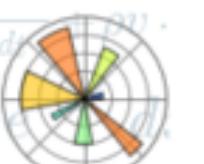
SciPy.org

Sponsored By ENTHOUGHT

Install Getting Started Documentation Report Bugs Blogs

SciPy (pronounced "Sigh Pie") is a Python-based ecosystem of open-source software for mathematics, science, and engineering. In particular, these are some of the core packages:

 NumPy Base N-dimensional array package	 SciPy library Fundamental library for scientific computing
 Matplotlib Comprehensive 2D Plotting	 IP[y]: IPython Enhanced Interactive Console
 Sympy Symbolic mathematics	 pandas Data structures & analysis

[More information...](#)

CORE PACKAGES:

Numpy [SciPy library](#) Matplotlib IPython Sympy Pandas

News

NumPy 1.9.0 released See [Obtaining NumPy & SciPy libraries](#).
(2014-09-07)

NumPy 1.8.2 released See [Obtaining NumPy & SciPy libraries](#).
(2014-08-09)

SciPy 0.14.0 released See [Obtaining NumPy & SciPy libraries](#).
(2014-05-03)

NumPy 1.8.1 released See [Obtaining NumPy & SciPy libraries](#).
(2014-03-26)

Search Go

COURSE TECHNOLOGIES

- Jupyter notebooks
 - Support for reproducible workflows
 - Excellent support for SciPy
 - Consistent environment across platforms
 - Quickly becoming a “industry” standard
 - Publishing reproducible results

ipython.org

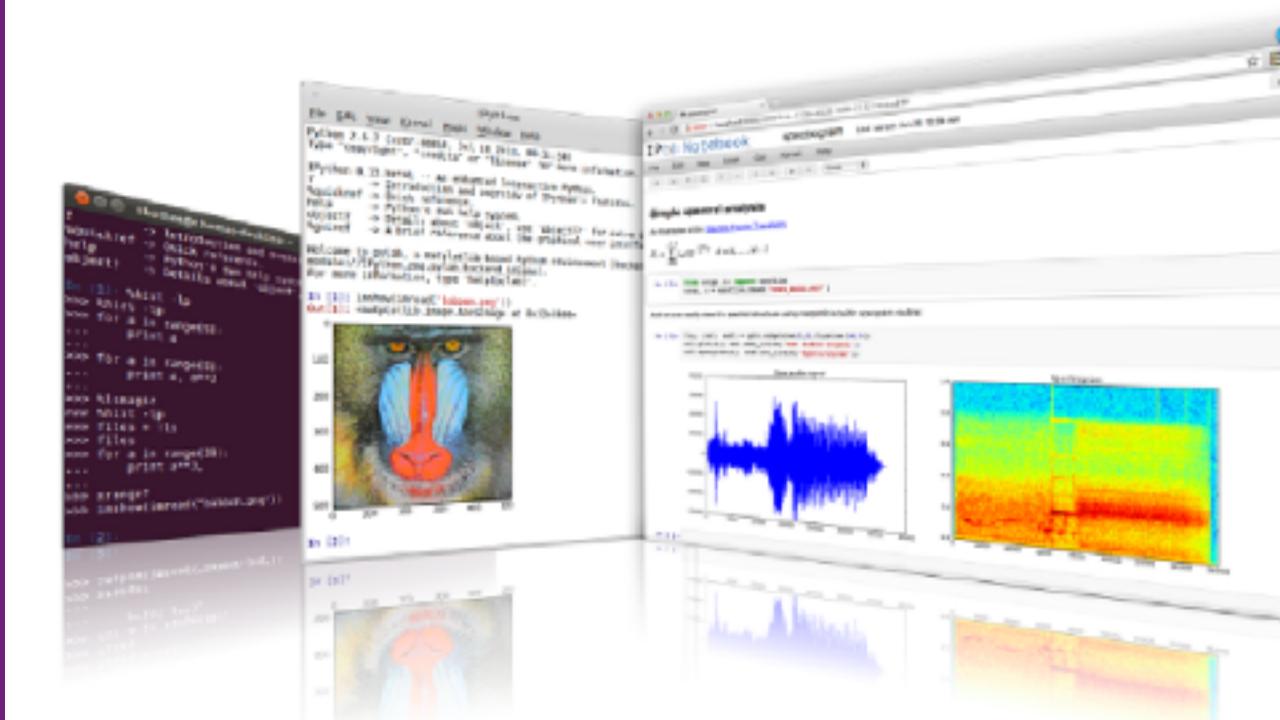
IP[y]: IPython

Interactive Computing

[Install](#) · [Documentation](#) · [Project](#) · [Jupyter](#) · [News](#) · [Cite](#) · [Donate](#)

IPython provides a rich architecture for interactive computing with:

- A powerful interactive shell.
- A kernel for [Jupyter](#).
- Support for interactive data visualization and use of [GUI toolkits](#).
- Flexible, [embeddable](#) interpreters to load into your own projects.
- Easy to use, high performance tools for [parallel computing](#).



To get started with IPython in the Jupyter Notebook, see our [official example collection](#). Our [notebook gallery](#) is an excellent way to see the many things you can do with IPython while learning about a variety of topics, from basic programming to advanced statistics or quantum mechanics.

To learn more about IPython, you can watch our videos and screencasts, download our [talks and presentations](#), or read our [extensive documentation](#). IPython is open source (BSD license), and used by a range of [other projects](#); add your project to that list if it uses IPython as a library, and please don't forget to [cite the project](#).

IPython supports Python 2.7 and 3.3 or newer. Our older 1.x series supports Python 2.6 and 3.1.

Jupyter and the future of IPython

Display a menu

COURSE TECHNOLOGIES

The screenshot shows the Google Collaboratory interface. At the top, there's a purple header bar with the title "COURSE TECHNOLOGIES". Below it is a dark grey toolbar with the "CO" logo, a file icon, and the text "mpcs56420-2018-spring-assignment-1.ipynb". The main workspace is a dark grey area with a sidebar on the left containing icons for code, text, and files. A yellow speech bubble in the top right corner contains the text: "GOOGLE COLLABORATORY CAN IMPORT DIRECTLY FROM GITHUB". The main content area displays an assignment notebook titled "MPCS 56420 - 2020 - Spring - Assignment 1". Inside, there's a text block about assignment submission and a problem section labeled "Problem 1." with a detailed question. A toolbar with various icons is visible at the bottom right.

mpcs56420-2018-spring-assignment-1.ipynb

File Edit View Insert Runtime Tools Help

+ Code + Text

MPCS 56420 - 2020 - Spring - Assignment 1

This assignment is due Tuesday, March 10, 2018 at 5:29 PM. The instructors will clone a copy of your assignment repository and use the last version checked in before the due date. Please answer immediately below each question. If you have not done so, please email instructor your GitHub account user name as soon as possible.

Problem 1.

This course is composed of students from a variety of backgrounds and experiences. Compose a brief introduction providing us with your educational background and work experience. Specifically, let us know about your background (if any) in biology. This will allow us to tailor the class materials at a level that is appropriate for everyone. Next, please share your interests and motivation for taking this course and what you hope to get out of it. Finally, if there is any particular topic that you would like us to cover in class, please make a note of it.

Type your answer here.

2.

COURSE TECHNOLOGIES

- Google Could Platform
 - App Engine, Compute Engine, Storage, Genomics
 - Python, Java, PHP and Go
 - NumPy and Matplotlib are available on Google App Engine
 - Support for “big data” workflows

The screenshot shows the Google Cloud Platform Compute Engine landing page. At the top, there's a navigation bar with icons for back, forward, search, and account. The URL 'cloud.google.com' is visible. Below the header, there's a large image of server racks with a blue hexagonal icon containing a white circuit board symbol overlaid. The text 'Compute Engine' is prominently displayed in large white letters. A subtitle explains: 'Run large-scale workloads on virtual machines hosted on Google's infrastructure. Choose a VM that fits your needs and gain the performance of Google's worldwide fiber network.' A 'Get Started' button is located in the bottom-left corner of the main image area. At the very bottom of the page, there's a navigation bar with links for 'Features', 'Case Studies', 'Pricing Calculator', 'Pricing', and 'Documentation'.

Features



High-performance virtual machines

Compute Engine's Linux VMs are consistently performant, scalable, highly secure and reliable. Supported distros include Debian and CentOS. You can choose from

COURSE TECHNOLOGIES

- Research Computing Center
 - High-performance computing and visualization resources
 - High-capacity storage and backup
- Software
 - High-speed networking
- Hosted data sets

The screenshot shows the homepage of the Research Computing Center (RCC) at the University of Chicago. The header features the university's crest and the text "THE UNIVERSITY OF CHICAGO" next to "Research Computing Center". The main navigation menu includes "Getting Started", "Resources", "Research", and "Support & Services". Below the header is a large, close-up photograph of green cereal plants against a blue sky. A blue banner across the image contains the text "Computing Cereals in Parallel". To the right of the banner, a text box says: "When trying to simulate the many facets of Earth's climate, a 'mashup' of cereals can come in handy." At the bottom of the page, there are three columns with icons and text: "Enabling Research" (book icon), "Data Visualization" (bar chart icon), and "Training & Education" (two people icon). Each column also has a brief description.

THE UNIVERSITY OF
CHICAGO | Research Computing Center

Getting Started Resources Research Support & Services

Computing Cereals in Parallel

When trying to simulate the many facets of Earth's climate, a "mashup" of cereals can come in handy.

Enabling Research

Our department has helped enable the advancement of critical inquiry from the physical sciences to the social sciences to the humanities.

Data Visualization

RCC maintains data visualization resources including high-end graphics processing hardware, visualization software, and custom remote visualization tools.

Training & Education

The Research Computing Center offers training and workshops on a variety of topics relevant to research computing.

Display a menu

COURSE TECHNOLOGIES

- You can use any other languages and technologies for your final projects
- Limited support from instructors



COURSE RESOURCES

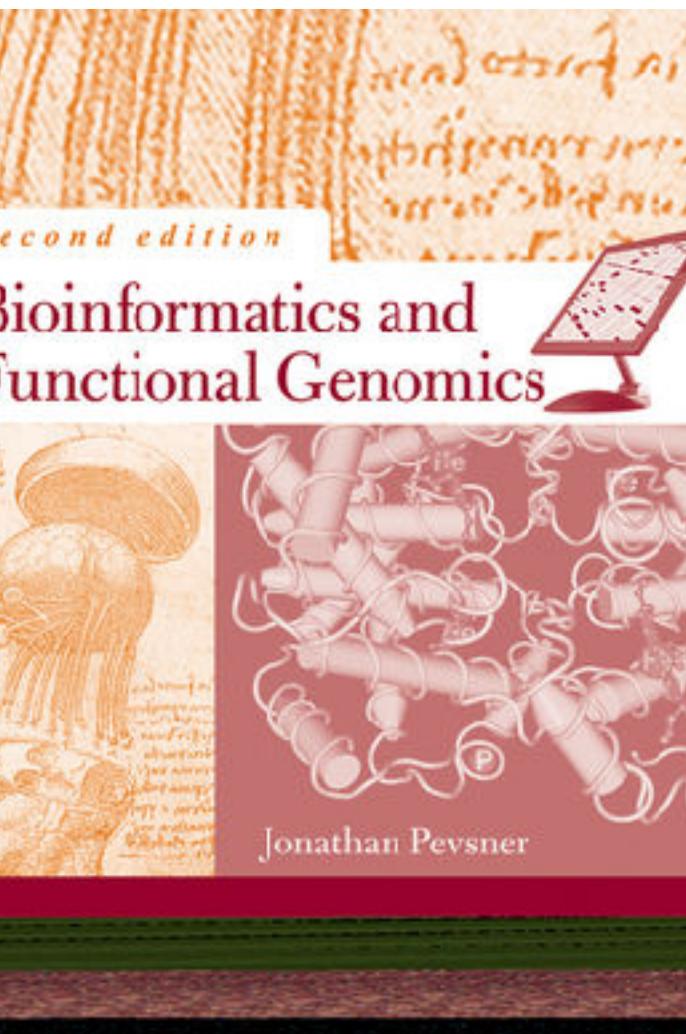
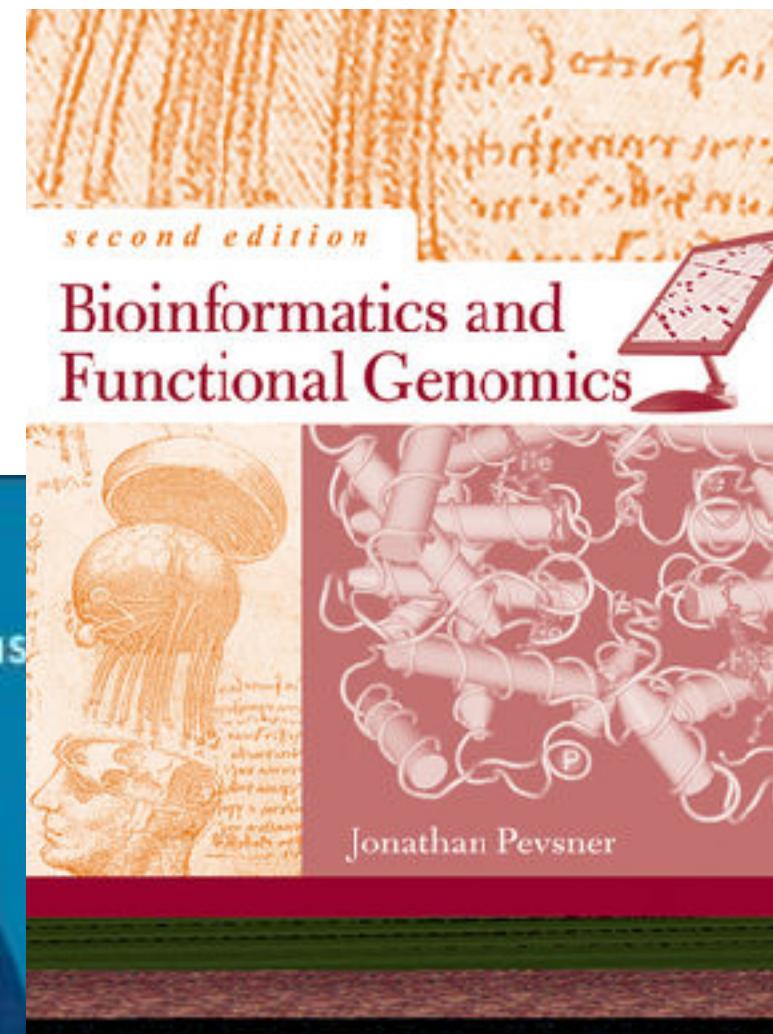
COURSE RESOURCES

- Text Books
 - No required text books
 - Books become outdated quickly
 - Publications (scientific journals) have latest information
 - Quality online references from reputable sources
 - NCBI Handbook (<http://www.ncbi.nlm.nih.gov/books/NBK143764/?term=handbook>)

The screenshot shows the NCBI Bookshelf website. At the top, there's a navigation bar with links for "Home - Books - NCBI", "How To", "Bookshelf", and a dropdown menu set to "Books". Below the navigation is a search bar with options "Browse Titles", "Limits", and "Advanced". A large green image of microorganisms serves as the header for the "Bookshelf" section. To the right, a text box describes Bookshelf as providing free access to books and documents in life science and healthcare, enabling users to browse, retrieve, and read content. Below this are sections for "Using Bookshelf" (with links to "Quick Start Guide", "FAQ", "Tutorials", "Bookshelf News", and "Copyright and Permissions"), "New & Updated" (listing several recent publications like "The Influence of Global Environmental Change on Infectious Disease Dynamics" and "Fitness Measures and Health Outcomes in Youth"), and "Featured Titles" (listing titles like "Multigene Panels in Prostate Cancer Risk Assessment" and "Negotiating Bioethics"). On the far right, there are "Participate" links for "Authors and Publishers", "How to Apply", and "Participation Agreement", and a "More Information" section for "NLM Literature Archive", "Open Access Subset", and "Librarians".

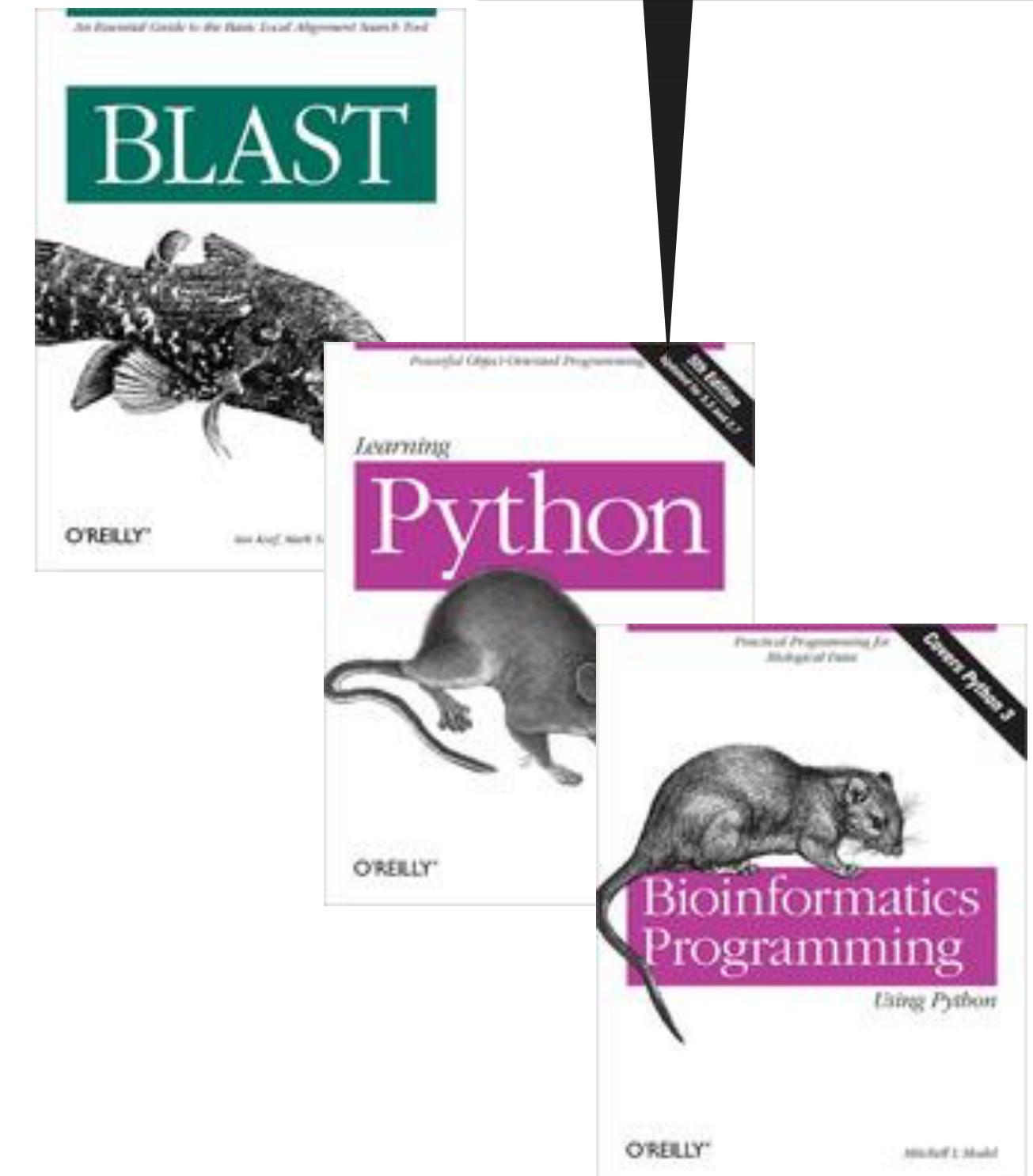
COURSE RESOURCES

TEXT BOOKS



RECOMMENDED

FREE DIGITAL COPIES
AVAILABLE THROUGH
UCHICAGO LIBRARY



COURSE RESOURCES

- Biopython
 - Open sources framework for bioinformatics
 - Great resources and place to get started for algorithms and data structures
 - Unless specified, you should not use it



Page Discussion

Biopython

(Redirected from [Main Page](#))

Navigation

[Main Page](#)
[Downloads](#)
[Mailing lists](#)
[Documentation](#)
[Cookbook](#)
[News](#)
[Source Code](#)
[New issue tracker](#)
[Old issue tracker](#)
[Buildbot Tests](#)
[Participants](#)
[Script Central](#)
[Recent changes](#)
[Random page](#)

Toolbox

[What links here](#)
[Related changes](#)
[Special pages](#)
[Printable version](#)
[Permanent link](#)

Introduction

Biopython is a set of freely available tools for biological computation written in [Python](#) by an international team of developers.

It is a distributed collaborative effort to develop Python libraries and applications which address the needs of bioinformatics. The source code is made available under the [Biopython License](#), which is extremely permissive and has been adopted by many other projects around the world. We work along with the [Open Bioinformatics Foundation](#), who generously host this wiki.

This wiki will help you download and install Biopython, and start using the libraries and tools.

Get Started

- [Download Biopython](#)
- [Installation help](#) (PDF)

Get help

- [Tutorial](#) (PDF)
- [Documentation on this wiki](#)
- [Cookbook \(working examples\)](#)
- [Discuss and ask questions](#)

- [W](#)
- [D](#)
- [G](#)
- [R](#)

The latest release is [Biopython 1.65](#), released on 17 December 2014.

This page was last modified on 17 December 2014, at 21:09.

This page has been accessed 1,875,589 times.

Content is available under [GNU Free Documentation License 1.2](#).

[Privacy policy](#) [About Biopython](#) [Disclaimers](#)

COURSE RESOURCES

- Canvas "Lite"
- Links to materials
 - Github
 - Panopto

2020.02

[Home](#)

[Announcements](#) 

[Syllabus](#)

[Modules](#) 

[Assignments](#) 

[Discussions](#)

[Library Reserves](#)

[People](#)

[Grades](#)

[Panopto Video](#)

[Purchase UChicago Bookstore Course Materials](#)

[Collaborations](#) 

[Conferences](#) 

[Files](#) 

[Syllabus](#) 

Recent Announcements

MPCS 56420 1 (Spring 2020) Bioinformatics for Computer Scientists

This is a test.

Course Summary:

Date

Details

Course Status

 Unpublished

 Publish

 Import from Commons

 Choose Home Page

 View Course Stream

COURSE RESOURCES

2020.02

[Home](#)

[Announcements](#) 

[Syllabus](#)

[Modules](#) 

[Assignments](#) 

[Discussions](#)

[Library Reserves](#)

[People](#)

[Grades](#)

[Panopto Video](#)

[Purchase UChicago Bookstore Course Materials](#)

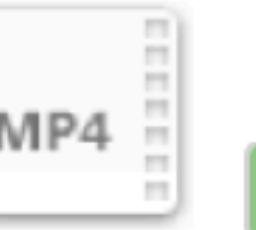
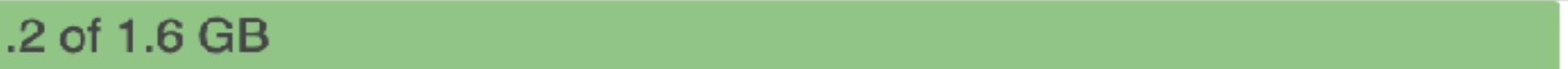
Search in folder  

Add files to session1  

Session 1
Sort by: Name 

+ Add folder

Choose video or audio files

1  mcps56420-2020-spring-session-1a
1.2 of 1.6 GB  13:06

COURSE RESOURCES

- Github Repository for all course materials (no website 😢)
- <https://github.com/uchicago-mobi/mpcs51032-2020-spring>

uchicago-bio / [mpcs56420-2020-spring](#)

Code Issues 0

No description, website, or topics.

Manage topics

12 commits 1 branch 0 packages

Branch: master New pull request Create new file

tabinks Update mpcs56420-2020-spring-assignment-1.ipynb

assignment-1 Update mpcs56420-2020-spring-assignment-1.ipynb

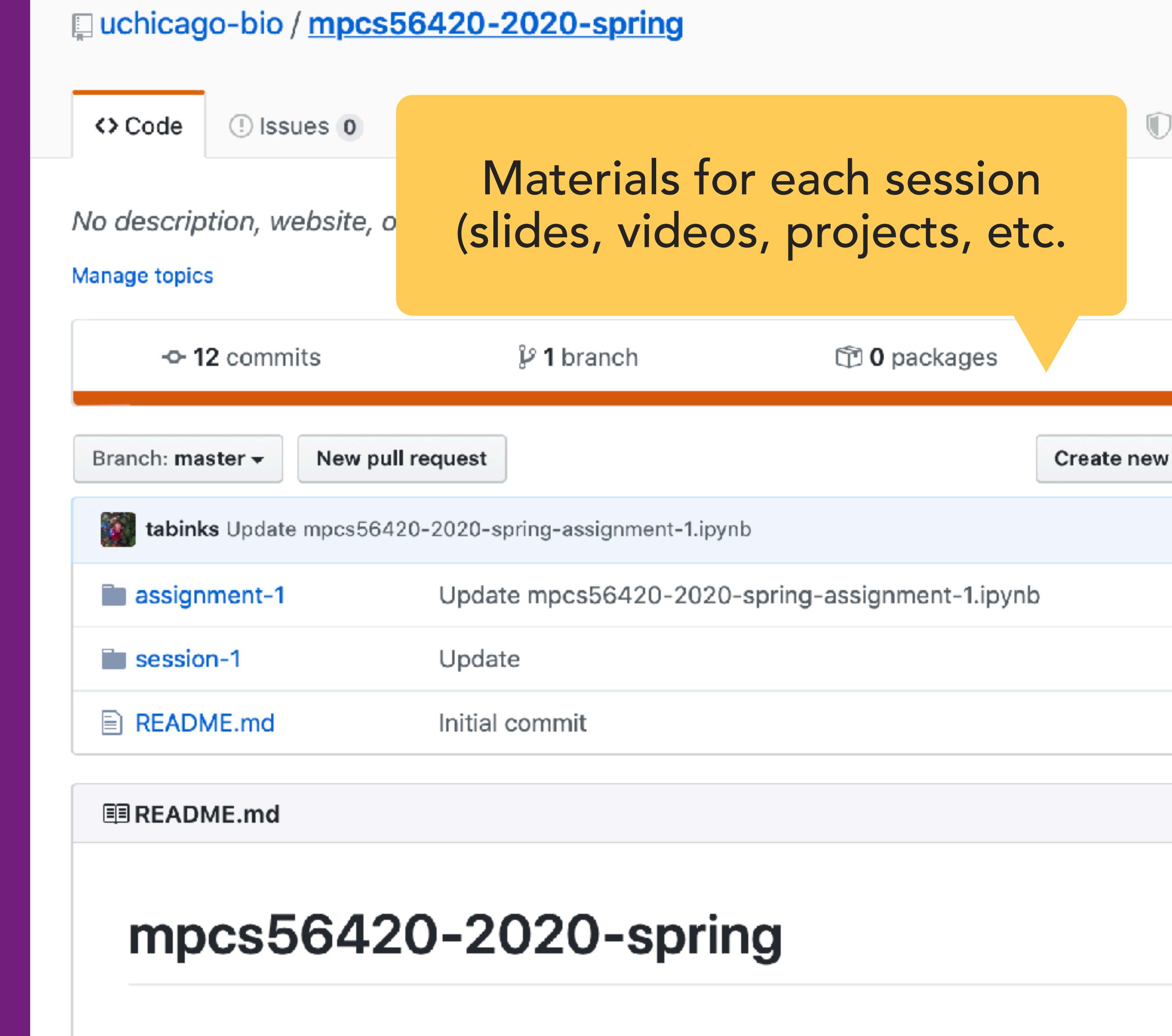
session-1 Update

README.md Initial commit

README.md

mpcs56420-2020-spring

Materials for each session (slides, videos, projects, etc.)



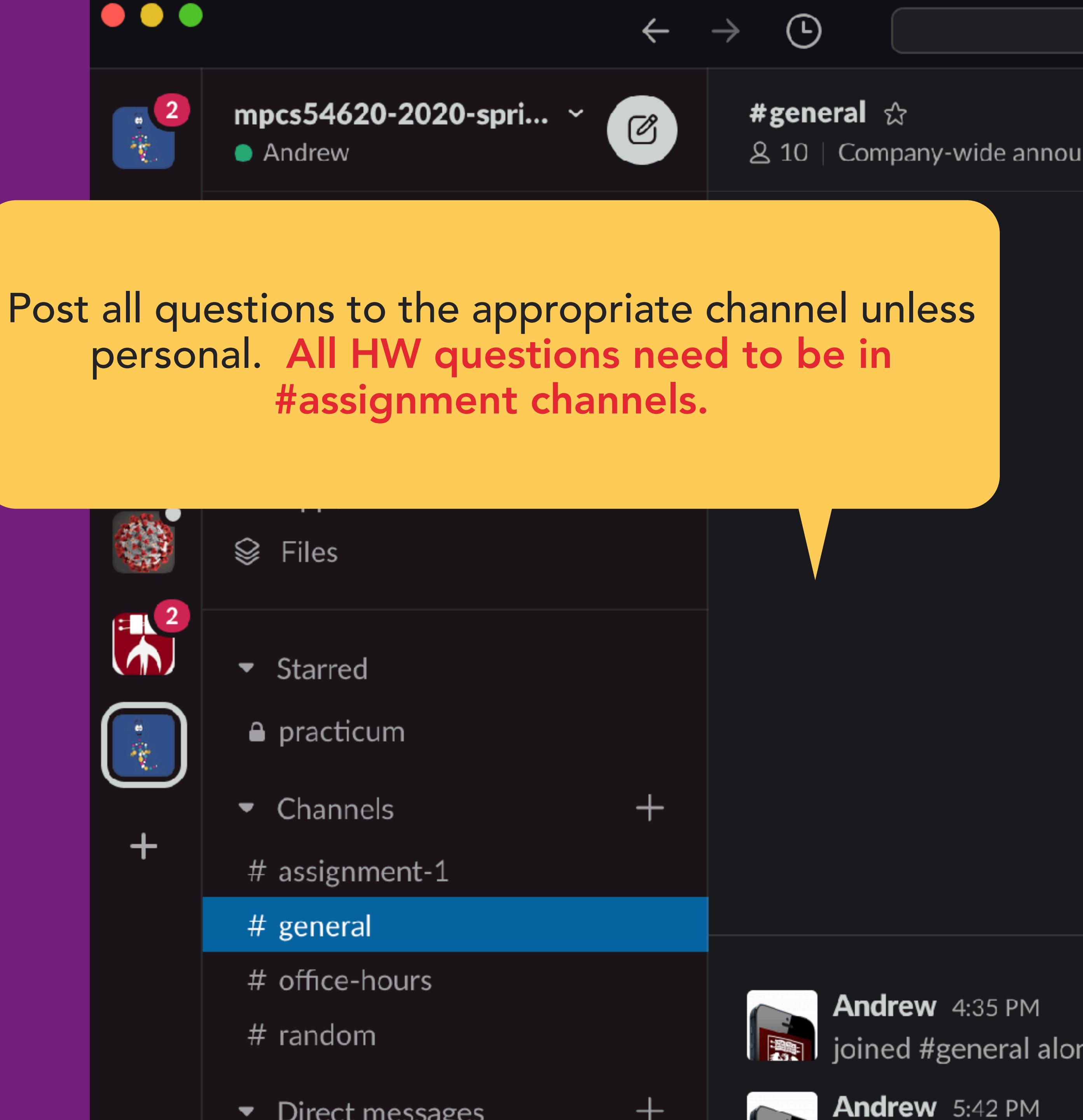
COURSE RESOURCES

- Github for coursework
 - Commit your repo to “turn in” assignments
 - Private repositories
 - Link on each assignment

The screenshot shows the GitHub Classroom interface for the course MPCS56420 - Bioinformatics (for computer scientists). The repository is named "mpcs56420-2018-spring" and is associated with the user "uchicago-bio". A yellow callout box contains the text "UNIQUE REPO FOR EACH ASSIGNMENT". Below the repository list, there is a section titled "Give this to your students" with a link: <https://classroom.github.com/a/vZ-nWORD>. A large callout box at the bottom right contains the text "**"mpcs56420-2018-spring-assignment-1"** does not exist" and "Share the invitation link with your students to get started".

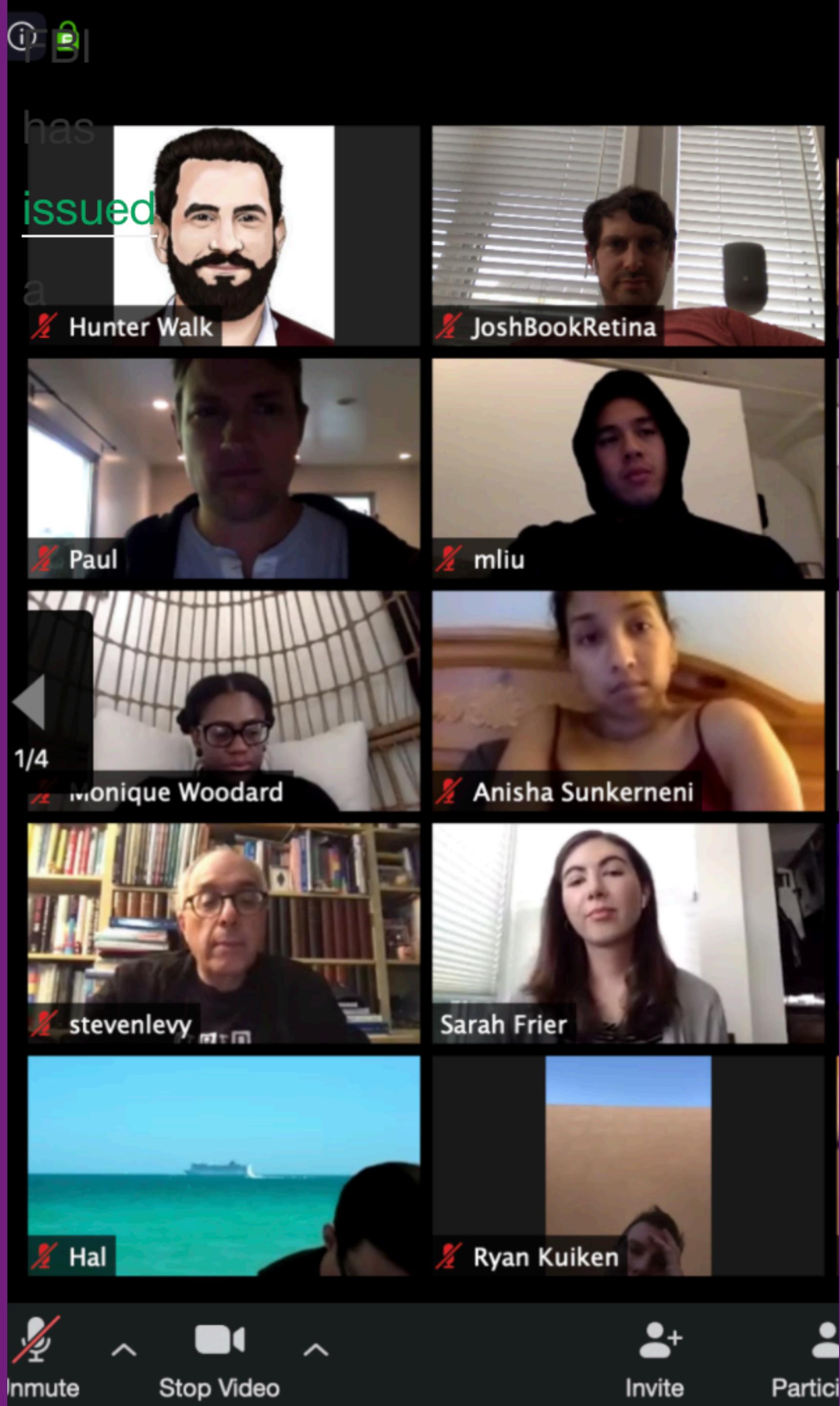
COURSE LOGISTICS

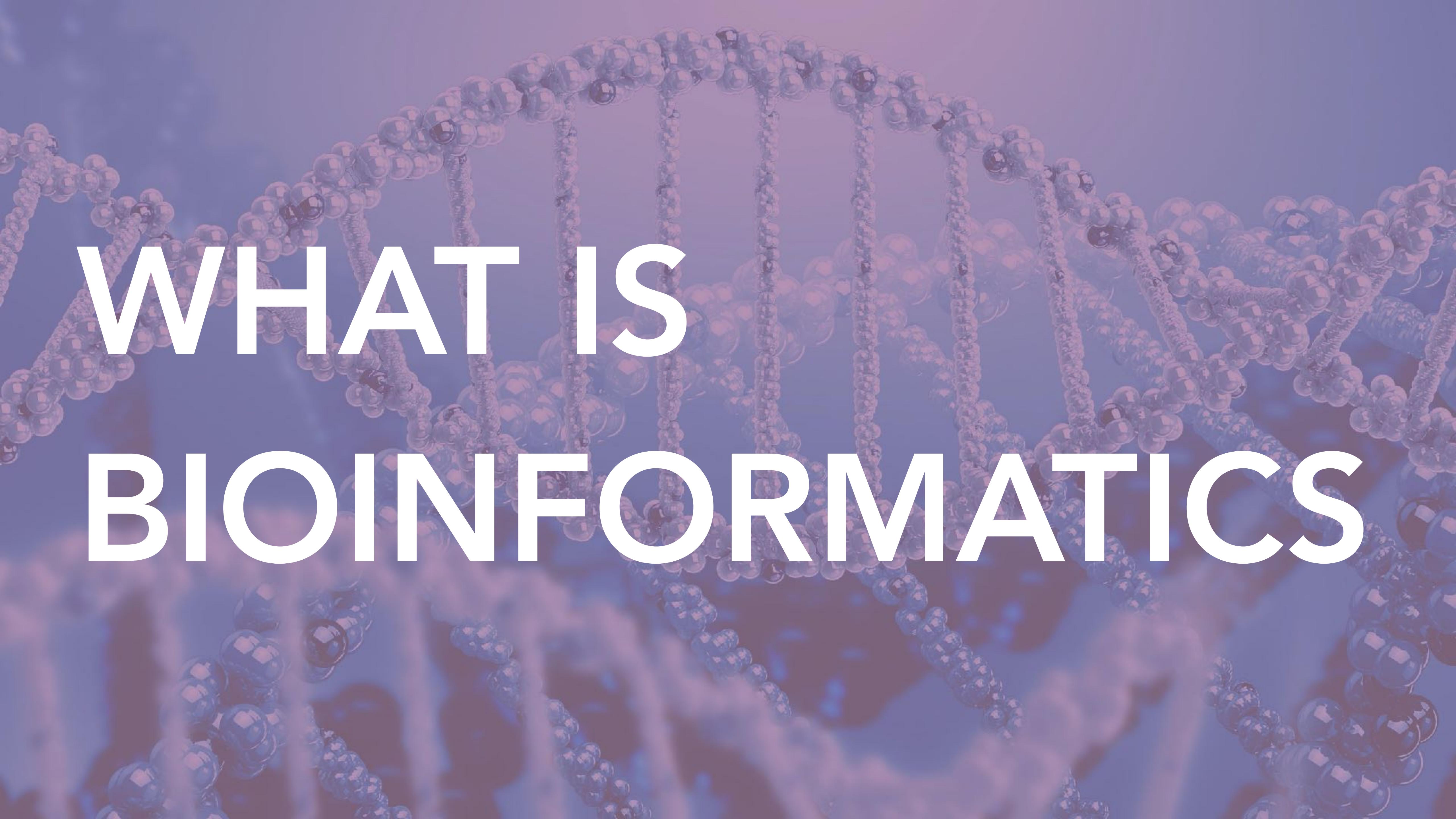
- Discussion forum rules
 - Anyone can answer a question
 - Post sources (if available) when answering questions



COURSE RESOURCES

- Please make sure you are muted if you are not talking
- Questions
 - "Raise Hand" if you have a question
 - Type it in the chat
 - Post in slack
- All sessions will be recorded
 - You can opt-out by turning camera and microphone off
 - Will not be distributed outside of class





WHAT IS BIOINFORMATICS

WHAT IS BIOINFORMATICS

- "Bioinformatics is the field of science in which biology, computer science, and information technology merge into a single discipline"
 - Official definition from the National Center for Biotechnology Information (NCBI)

WHAT IS BIOINFORMATICS

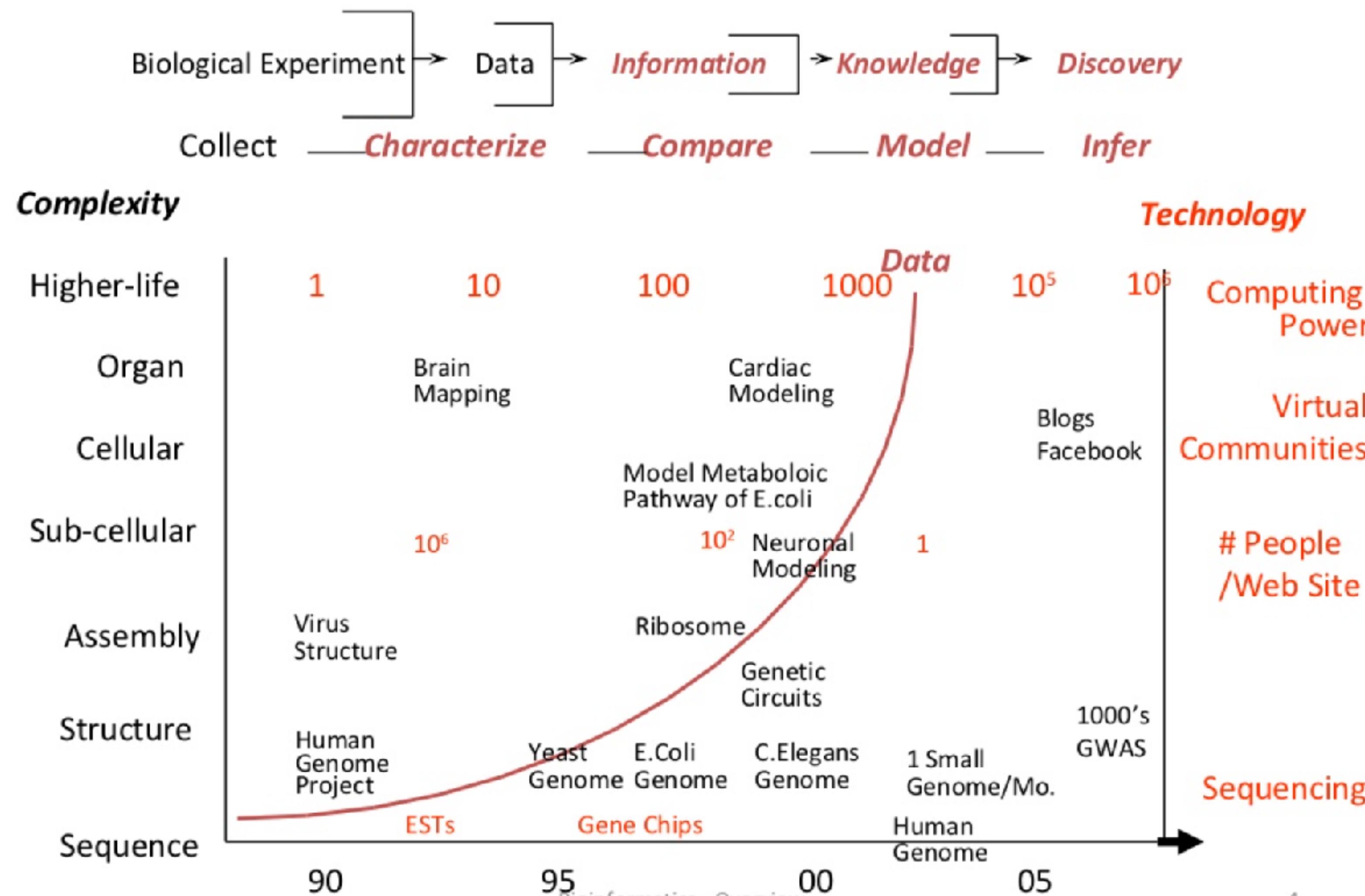
THREE IMPORTANT SUB-DISCIPLINES

- Development of new algorithms and statistics with which to assess relationships among members of large data sets
- Analysis and interpretation of various types of data
 - Nucleotide and amino acid sequences, proteins
- Development and implementation of tools that enable efficient access and management of different types of information

WHAT IS BIOINFORMATICS?

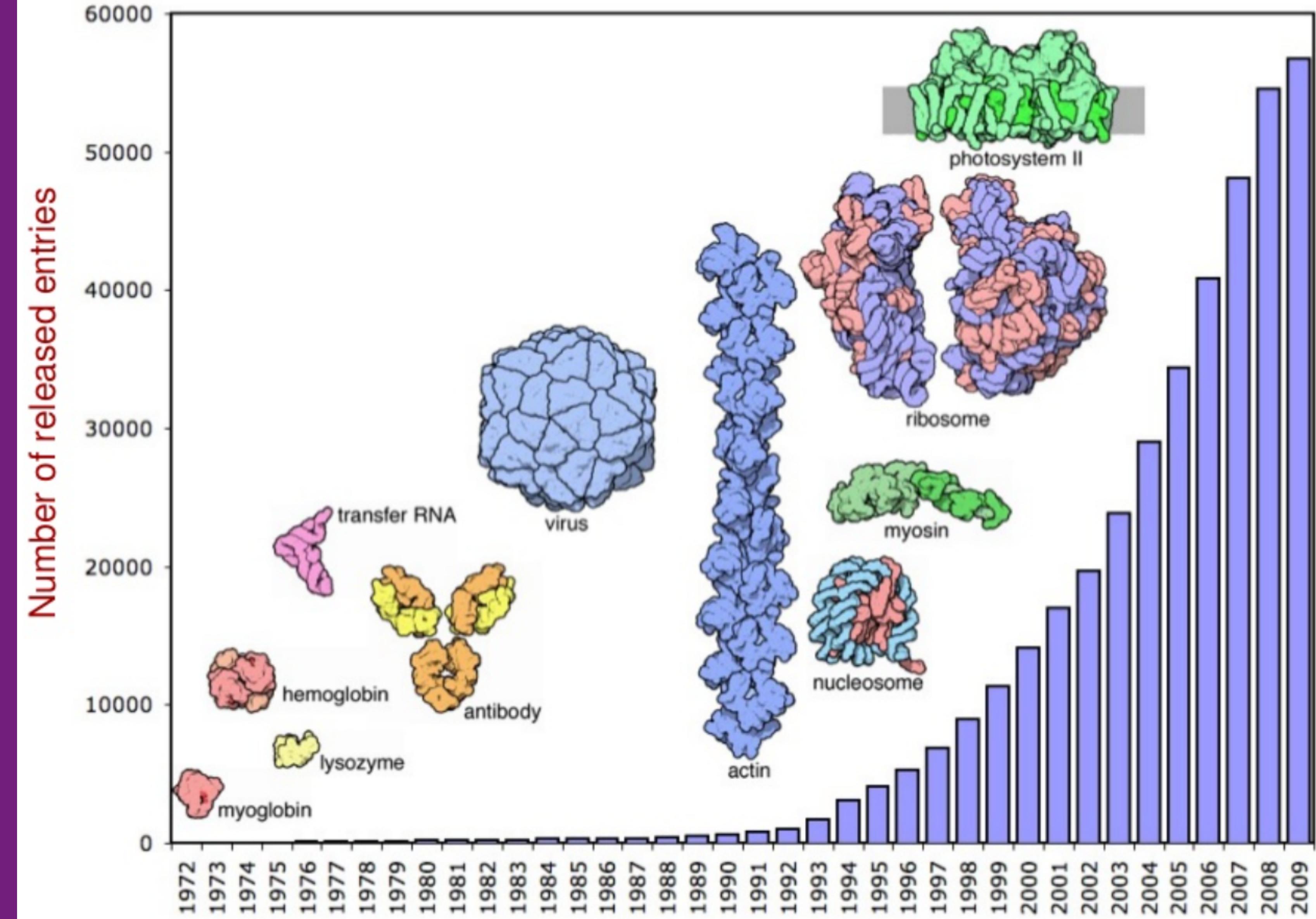
PHILIP BOURNE,
UCSD

Bioinformatics In One Slide



WHAT IS BIOINFORMATICS

- Growth of data
 - Protein Data Bank (PDB)



WHAT IS BIOINFORMATICS

- Using computers to answer biological questions
 - Including management and use of biological information
- Not LIMS (laboratory information management systems)
 - Collect data from experiments
 - Organize lab notebooks

WHAT IS BIOINFORMATICS

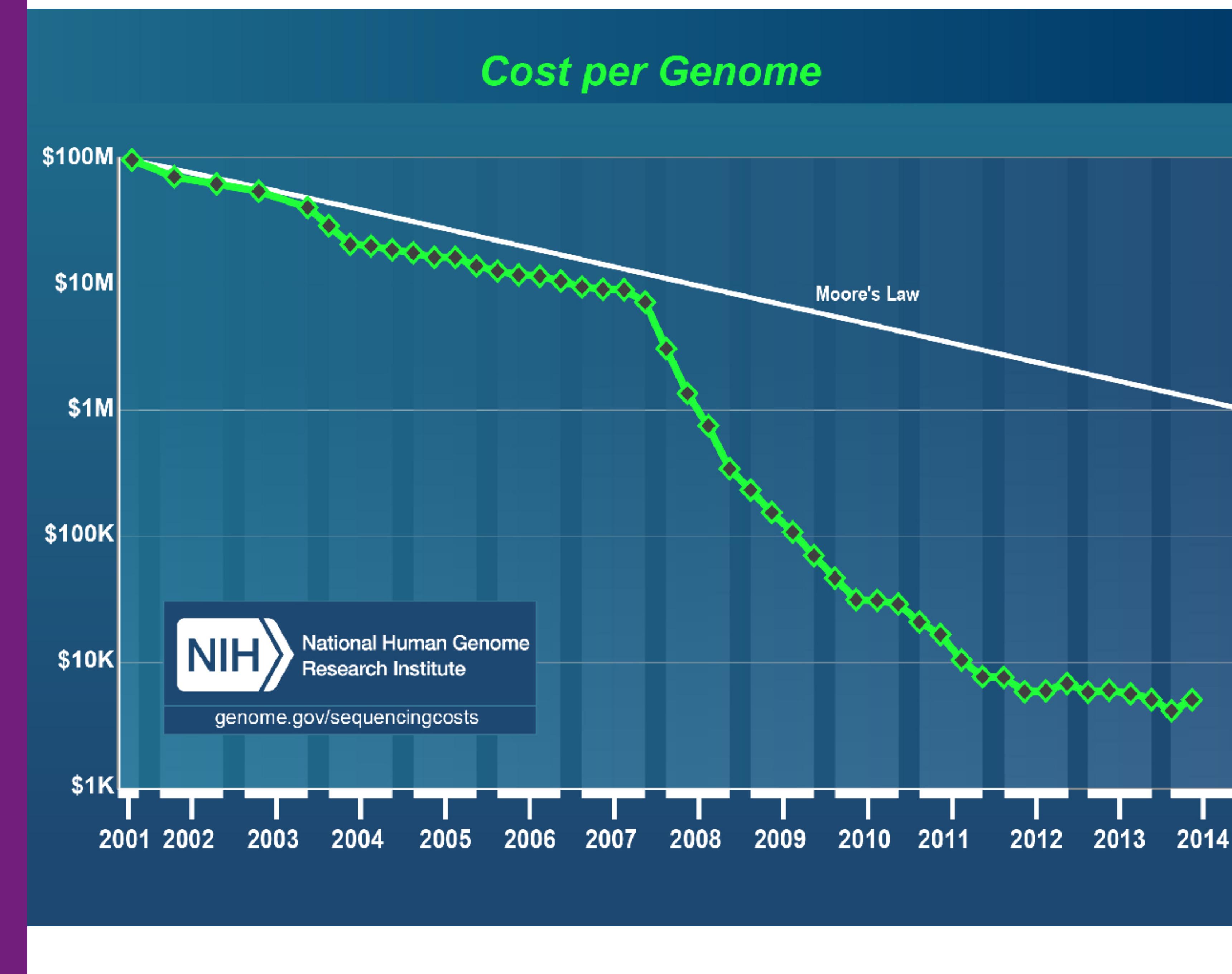
BANKING USES
COMPUTERS BUT ITS
NOT CALLED
BANKFORMATICS



- My answer: Its the next step in biology and biological research

WHAT IS BIOINFORMATICS

- The drive for the next-new-new revolution in bioinformatics
 - Genome sequencing costs
 - Do things that were literally not possible years ago



WHAT IS BIOINFORMATICS

- What can you do with bioinformatics?
- Really important scientific research that contributes to a better humanity

Trending News Bill Gates Charleston tornado Hajj Stampede Sleep Apnea

II 0:42 NOW WATCHING UP NEXT JOHN BOEHNER STEPPING DOWN

Police confirm waiter spat in customer's soda using DNA testing

by Reuters Videos 0:47 mins

A disgruntled customer who found spit in his drink after visiting a Chili's restaurant in Clay, New York, is suing the company that owns Chili's, the franchise owner and a former waiter.

OR THIS

<https://news.yahoo.com/video/police-confirm-waiter-spat-cust-1000000000000000000.html>

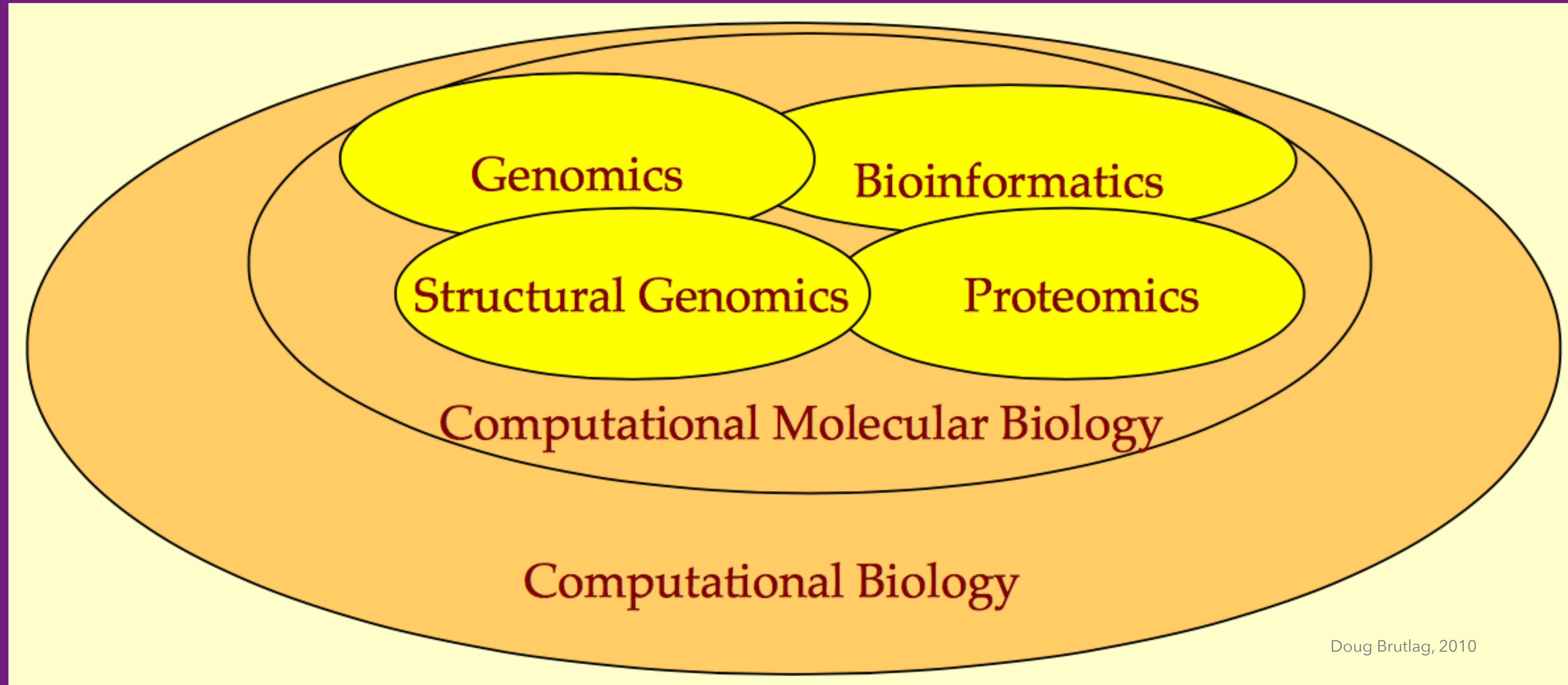
WHAT IS BIOINFORMATICS

WHAT ABOUT "COMPUTATIONAL BIOLOGY"?

INCLUDES
BIOINFORMATICS

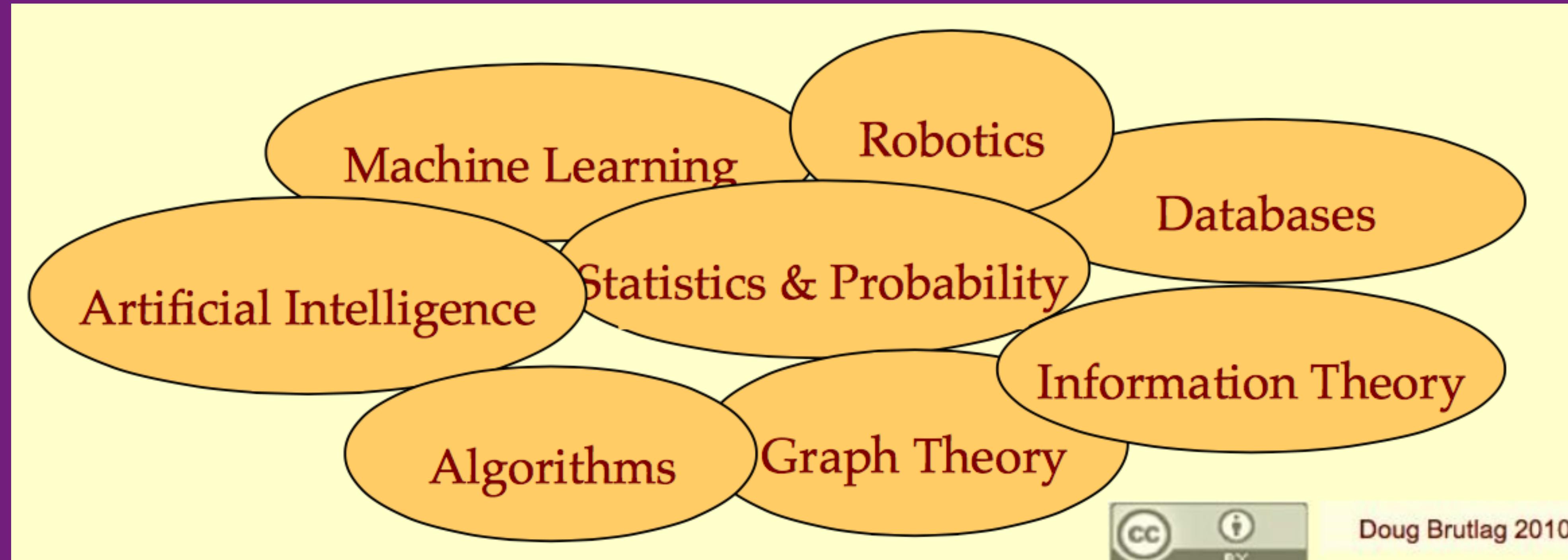
- Computational biology is an interdisciplinary field that applies the techniques to address biological problems
 - Computer science
 - Applied mathematics
 - Statistics
- A broader term (Can be used interchangeably, in my opinion)

WHAT IS BIOINFORMATICS



- Using computers to answer biological question

WHAT IS BIOINFORMATICS?



- Computational biology pulls from many subjects for ideas
 - Unconventional ways (mosquito robot, docking game)

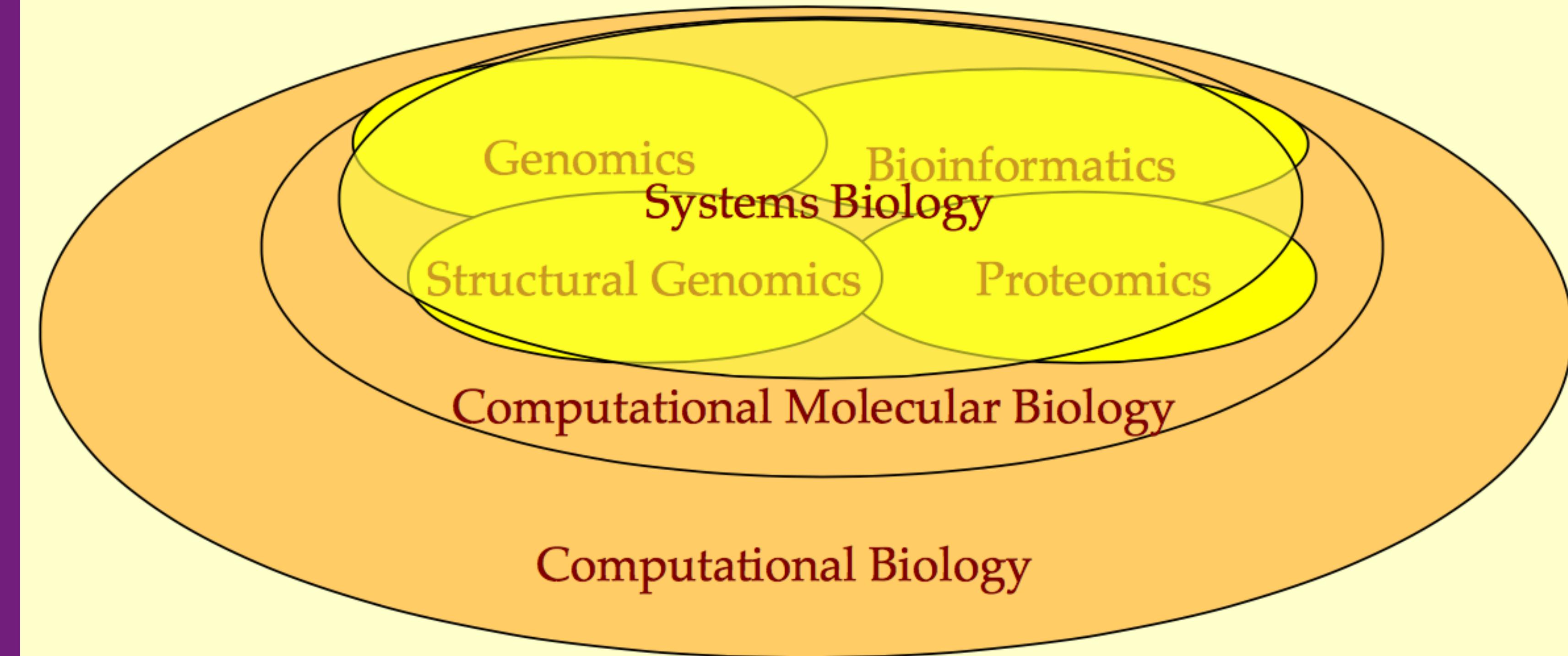
WHAT IS BIOINFORMATICS?

THERE ARE MANY
MORE SUB
DISCIPLINES

- What about all the “-omics”?
 - Genomics - Discovery, arrangement, expression of genes
 - Proteomics - Study the proteins in a system; mass spec analysis; markers
 - Structural Genomics - Study structural representatives of all proteins
 - Bioinformatics - Studying biological information

WHAT IS BIOINFORMATICS

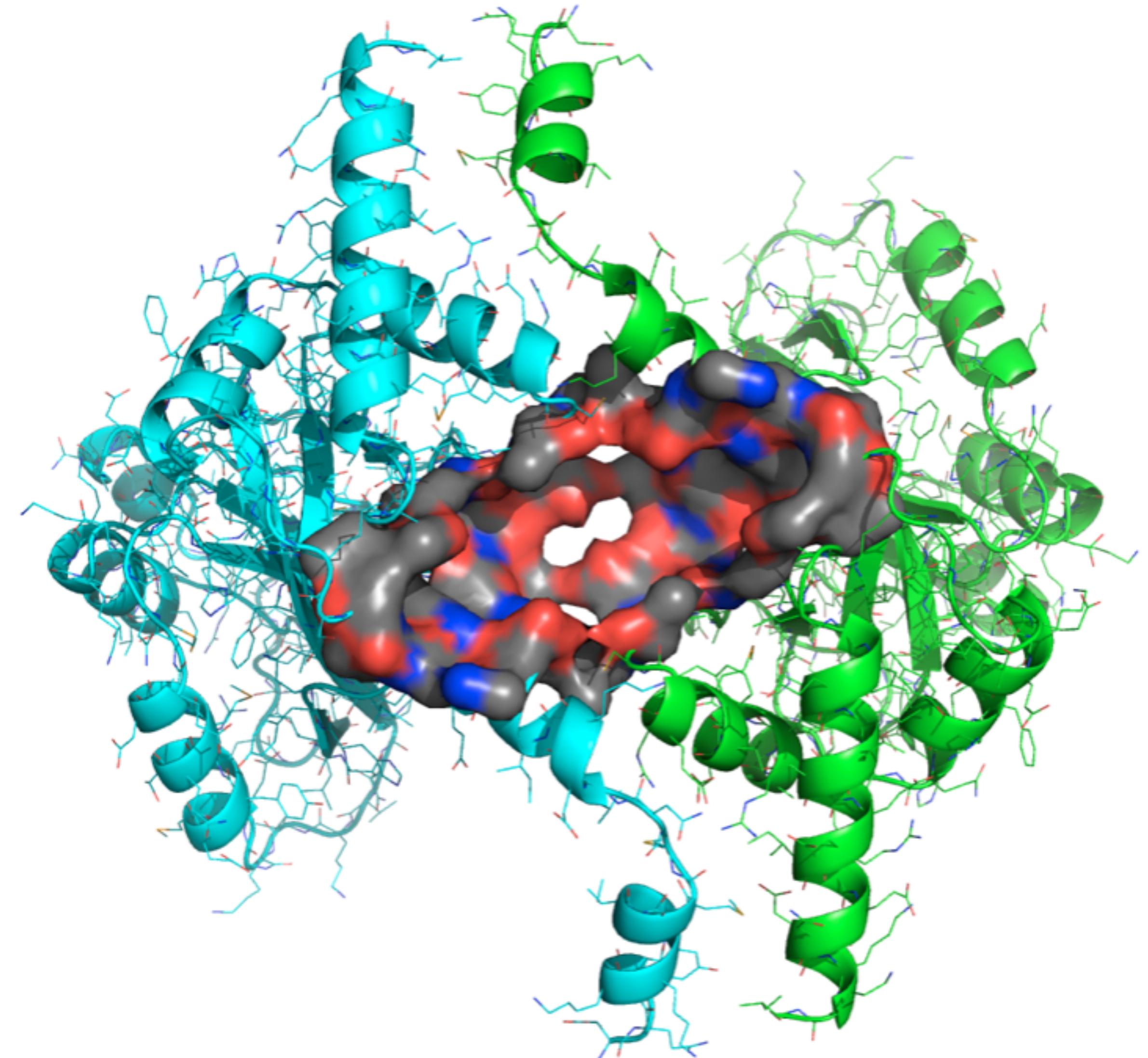
- Systems biology
 - New(er) field uses information to understand how systems work together
 - Enabled by the other fields



Doug Brutlag, 2010

WHAT IS BIOINFORMATICS

- Means different things to different people
 - Population modeling, disease modeling, molecular modeling, etc.
- Spend a career in one small subset
 - Protein surface analysis
- We will try to cover a little bit of everything



GOALS OF BIOINFORMATICS

GOALS OF BIOINFORMATICS

- Discover
- Predict
- Infer
- Organize
- Integrate
- Simulate
- Engineer

GOALS OF BIOINFORMATICS

- Analysis of genes and proteins
 - Homolog detection
 - Alignment (the residual-level mapping among homologous genes/proteins)
 - Application of the alignments
 - Detect the conserved residues, Functional sites, Prediction of protein structures, Motif finding (cis-elements)
 - Phylogeny
 - Engineer

CRISPR-altered plants are not going to be regulated (for now)



[Photo: [dimitrisvetsikas1969/Pixabay](#)]

Good news for people who like genetically altered tomatoes and other plants. The U.S. Department of Agriculture announced it will no longer regulate them.



The USDA not only rolled back Obama-era rules regulating genetically edited plants, but now it claims that plants whose genomes have been altered using gene-editing technology

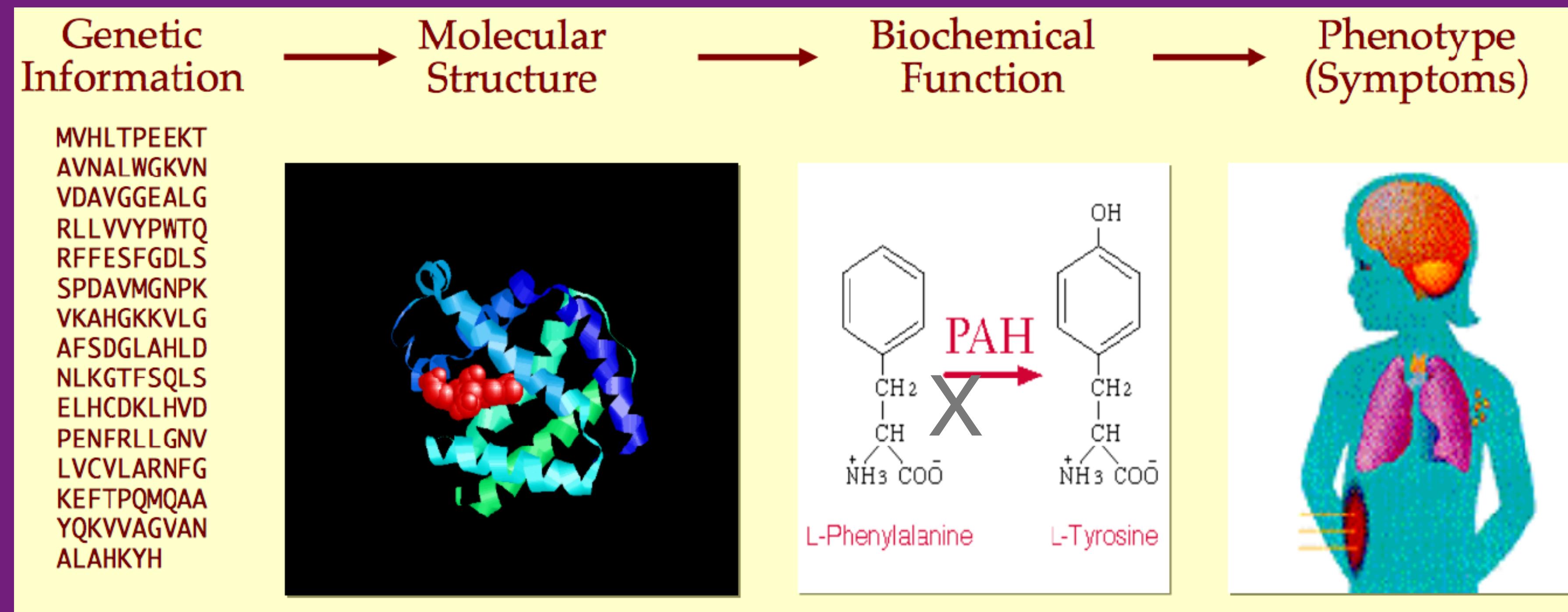
GOALS OF BIOINFORMATICS

LEARN BY
COMPARISON
AND INFERENCE

- Is Protein A similar to Protein B?
 - Sequence similarity (alignment!)
 - Structure similarity (structural comparison)
 - Co-expression (Microarray data analysis)
 - Any types of correlation (operon-structure, etc)

GOALS OF BIOINFORMATICS

CENTRAL PARADIGM OF
BIOINFORMATICS:
PREDICT/INFERENCE



- Phenylalanine hydroxylase (PAH) gene adds a hydroxyl group to create tyrosine
 - Tyrosine import to formation of hormones and neurotransmitters
- Mutation in PAH allows toxic buildup of Phenylalanine
- PKU (Phenylketonuria) - intellectual disability and seizures; low birth weight in babies

GOALS OF BIOINFORMATICS

- Challenges in bioinformatics (molecular biology)
 - Genetic information is redundant
 - Structural information is redundant
 - Genes and proteins aren't stable
 - Flexibility important for function
 - Genes have multiple functions
 - Translate 1 dimensional to 3 dimensional data

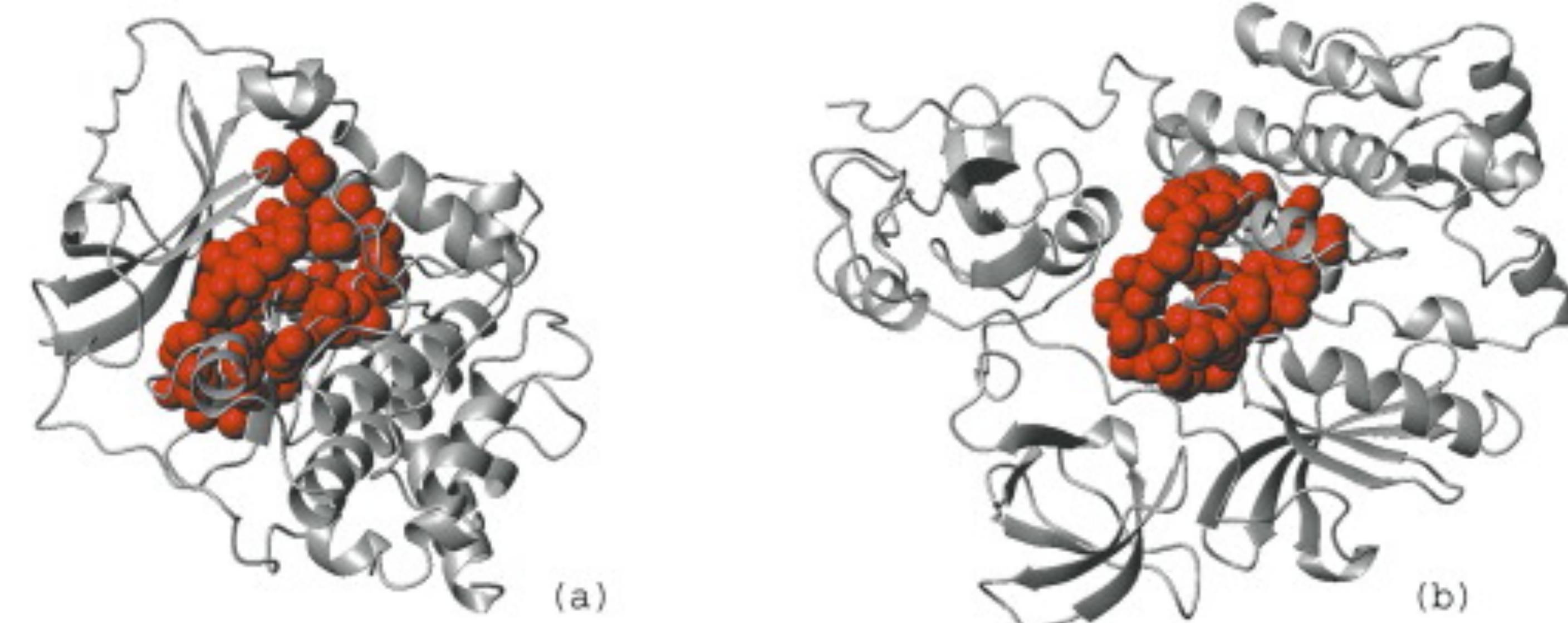
Genetic Information

MVHLTPEEKT
AVNALWGKVN
VDAVGGEALG
RLLVVYPWTQ
RFFESFGDLS
SPDAVMGNPK
VKAHGKKVLG
AFSDGLAHLD
NLKGTFSQLS
ELHCDKLHVD
PENFRLLGKV
LVCVLARNFG
KEFTPQMQAA
YQKVVAGVAN
ALAHKYH

GOALS OF BIOINFORMATICS

- Challenges: 1D to 3D

IMPORTANT RESIDUES
ARE NEIGHBORS IN 3D
SPACE, NOT IN 2D



(c)

>1cdk_A

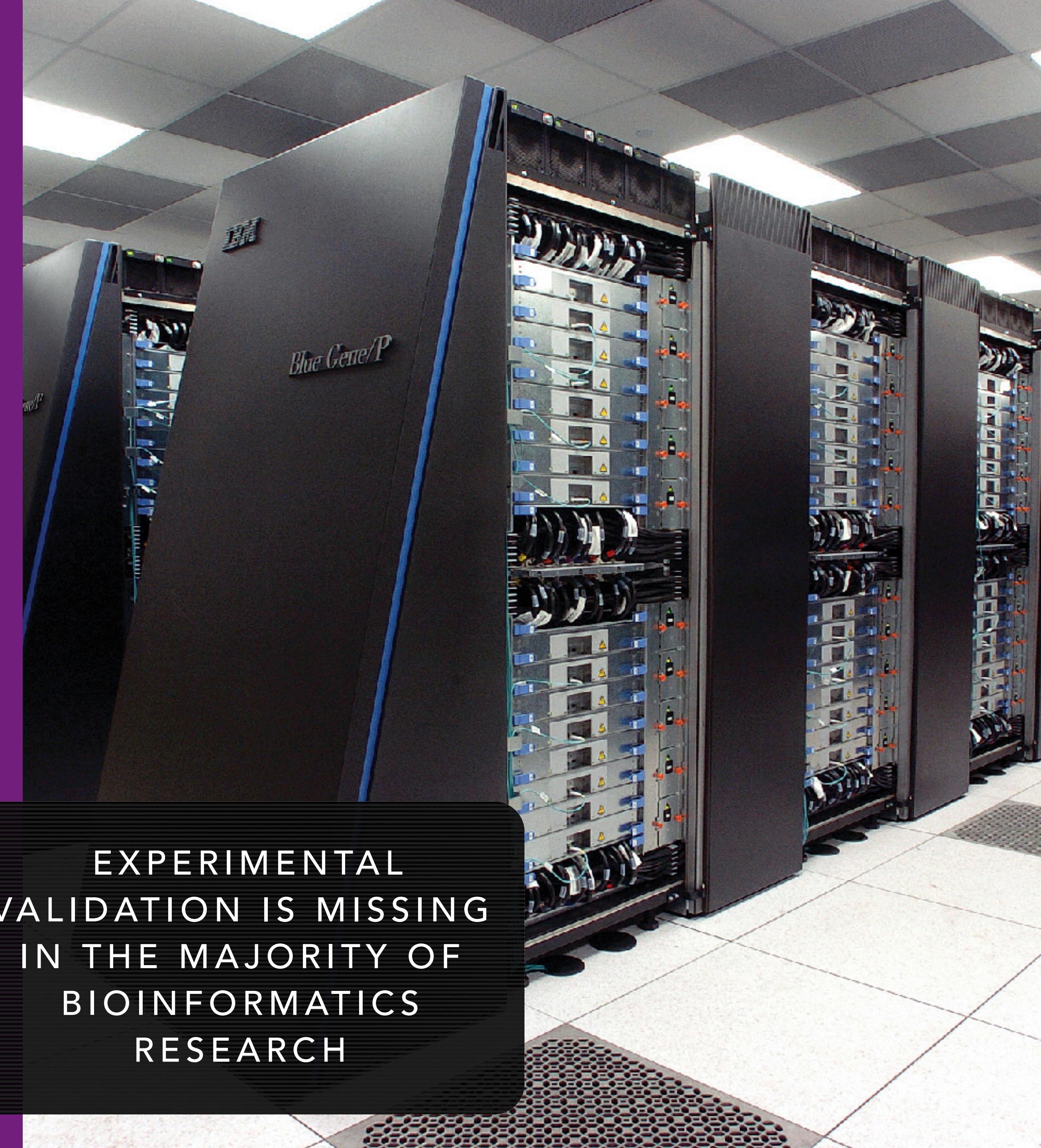
GNAAAAKKGSEQESVKEFLAKAKEDFLKKWENPAQNTAHDQFERIKT**LGTGSFGRV**MLVKHKETGNHF**AMKILD**
KQKVVKLQIEHTLNEKRILQAVNFPFLVKLEYSFKDNSNLYMVMEYVPGGEMFSHLRRIGRFSEPHARFYAAQI
VLTFEYLHSLDLIYRDLKPENLLIDQQGYIQV**TDFGF**AKRVKGRTWTLCGTPEYLAPEIILSKGYNKAVDWALG
VLIYEMAAGYPPFFADQPIQIYEKIVSGKVRFPSHSSDLKDLLRNLLQVDLTKRGNLKDGVNDIKNHKWFA
DWIAIYQRKVEAPFIPFKFGPGDTSNFDDYEEEEIRVSINEKCGKEFSEF

>2src_

MVTTFVALYDYESRTETDLSFKKGERLQIVNNTEGDWWLAHSLSTGQTGYIPSNYVAPSDSIQAEEWYFGKITRR
ESERLLLNAENPRGTFLVRESETTKGAYCLSVSDFDNAKGLNVHYKIRKLDSGGFYITSRTQFNSLQQLVAYYS
KHADGLCHRLTTVCPTSKPQTQGLAKDAWEIPRESLRLEV**KLGQGCFGEV**WMGTWNGTTRV**AIKTL**KPGT**MSPEA**
FLQEAQVMKKLRHEKLVQLYAVVSEEPYIVT**EYMSKGSLDFLKGETGKYLRLPQLVDMAAQIASGMAYVERMN**
YVHRDL**RAANIL**VGENLVCKV**ADFGLARLIEDNEYTARQGAKFPIKWT**PEAALYGRFTIKSDVWSFGILLTEL
TKGRVPYPGMVNREVLDQVERGYRMPCPPECPESLHDLMCQCWRKEPEERPTFEYLQAFLDYFTSTEPQXQPGE
NL

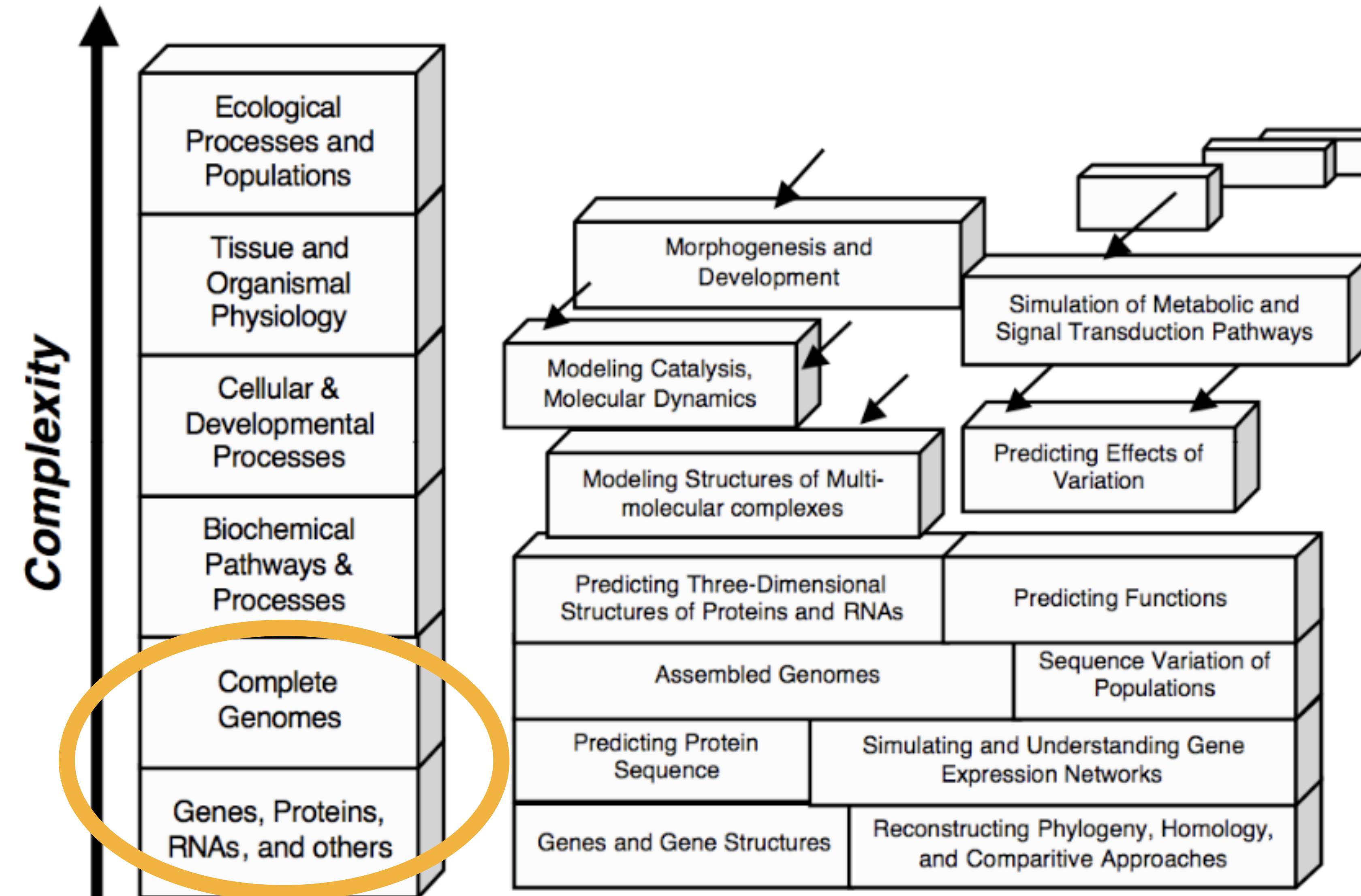
GOALS OF BIOINFORMATICS

- Challenges in bioinformatics (computing)
 - Growth, complexity and volume of biological data
 - Propagation of errors
 - Problem of inference
 - Economics of research
 - Computation is cheap (and doesn't go on vacation)
 - People are expensive
 - Wet lab work is expensive



EXPERIMENTAL
VALIDATION IS MISSING
IN THE MAJORITY OF
BIOINFORMATICS
RESEARCH

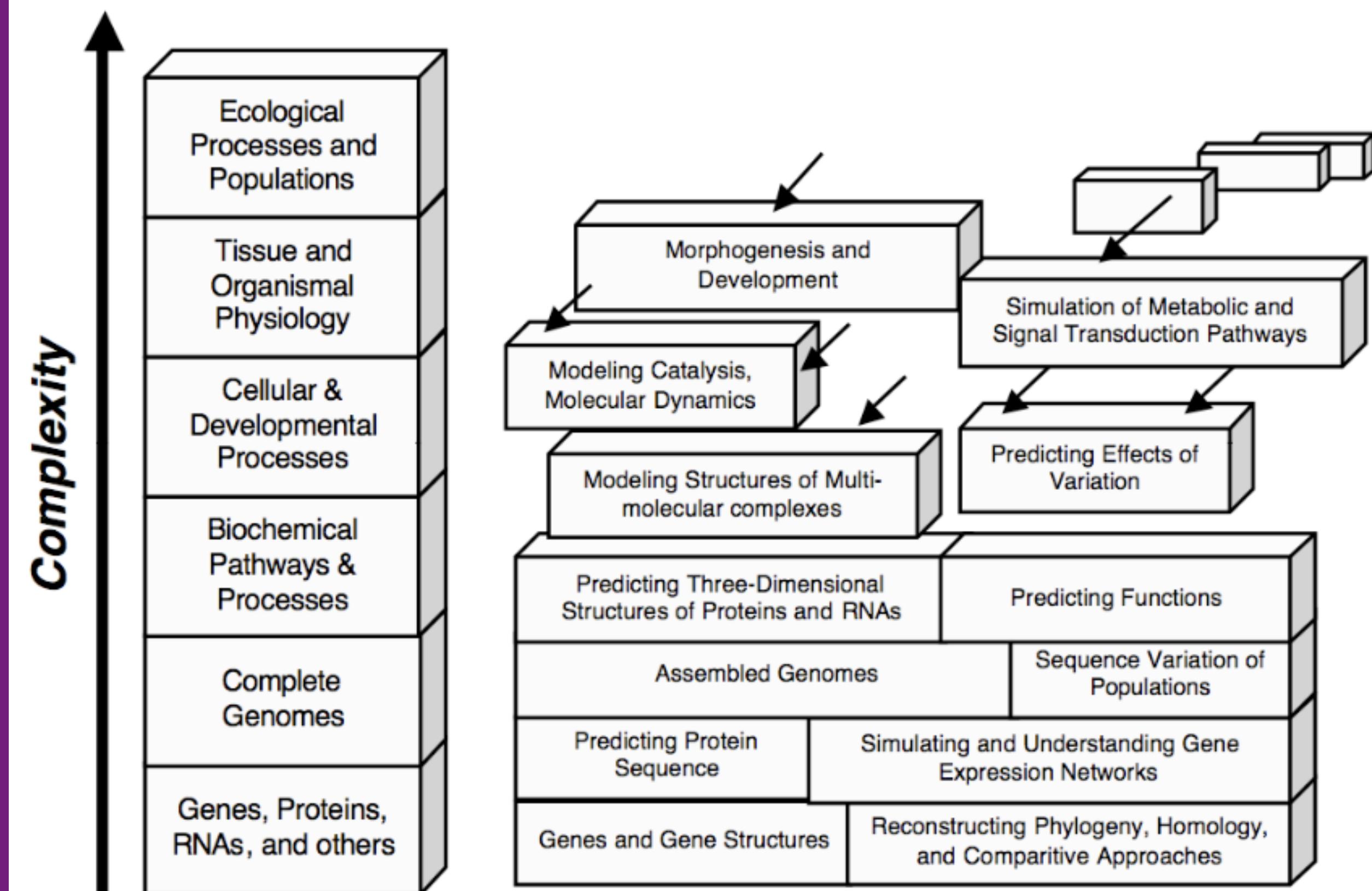
GOALS OF BIOINFORMATICS



IMPACT OF BIOINFORMATICS

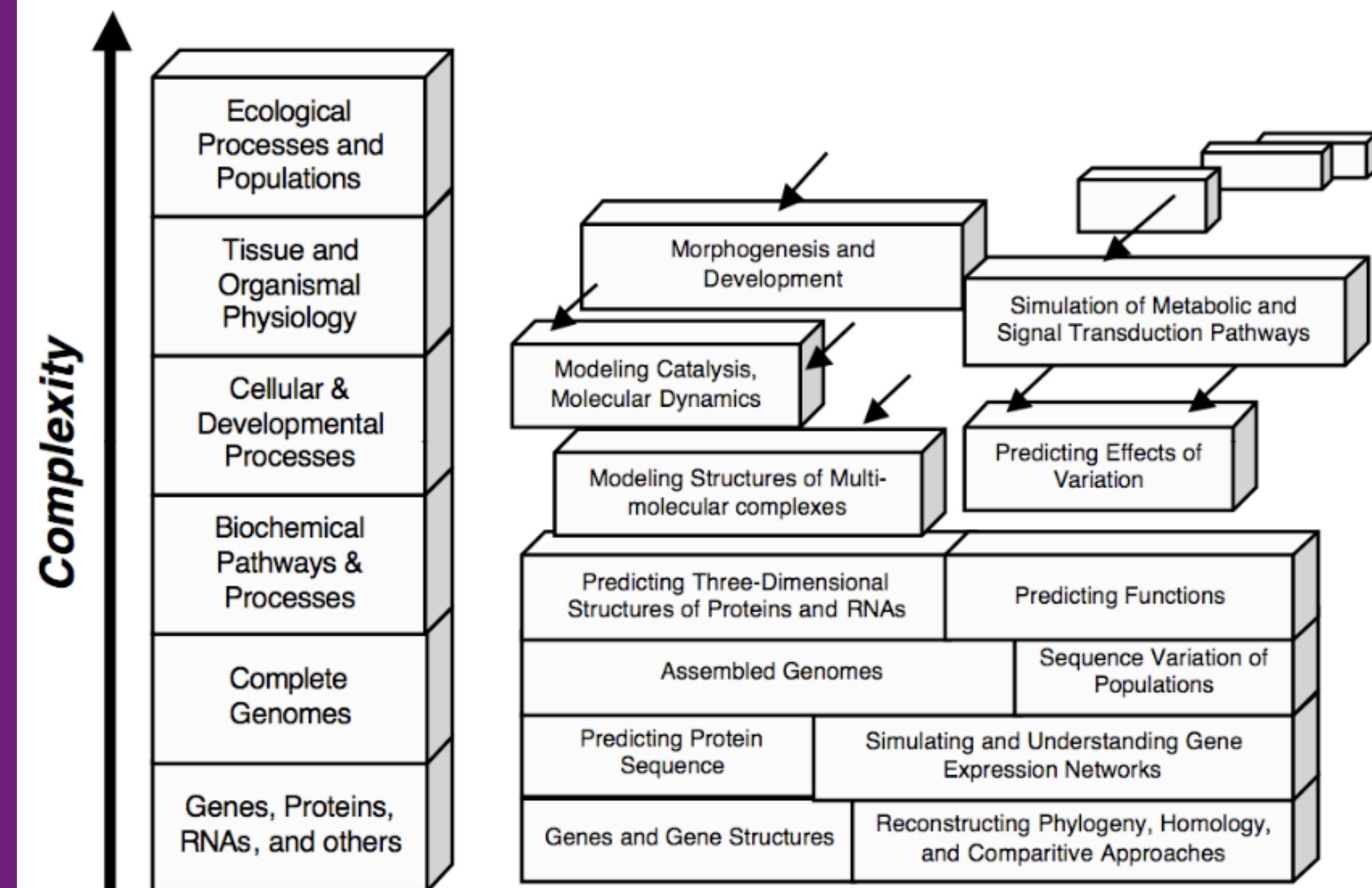
IMPACT OF BIOINFORMATICS

- On Biological sciences
 - Large scale experimental techniques
 - Analyze an entire genome
 - Information growth
 - Integration of data between fields



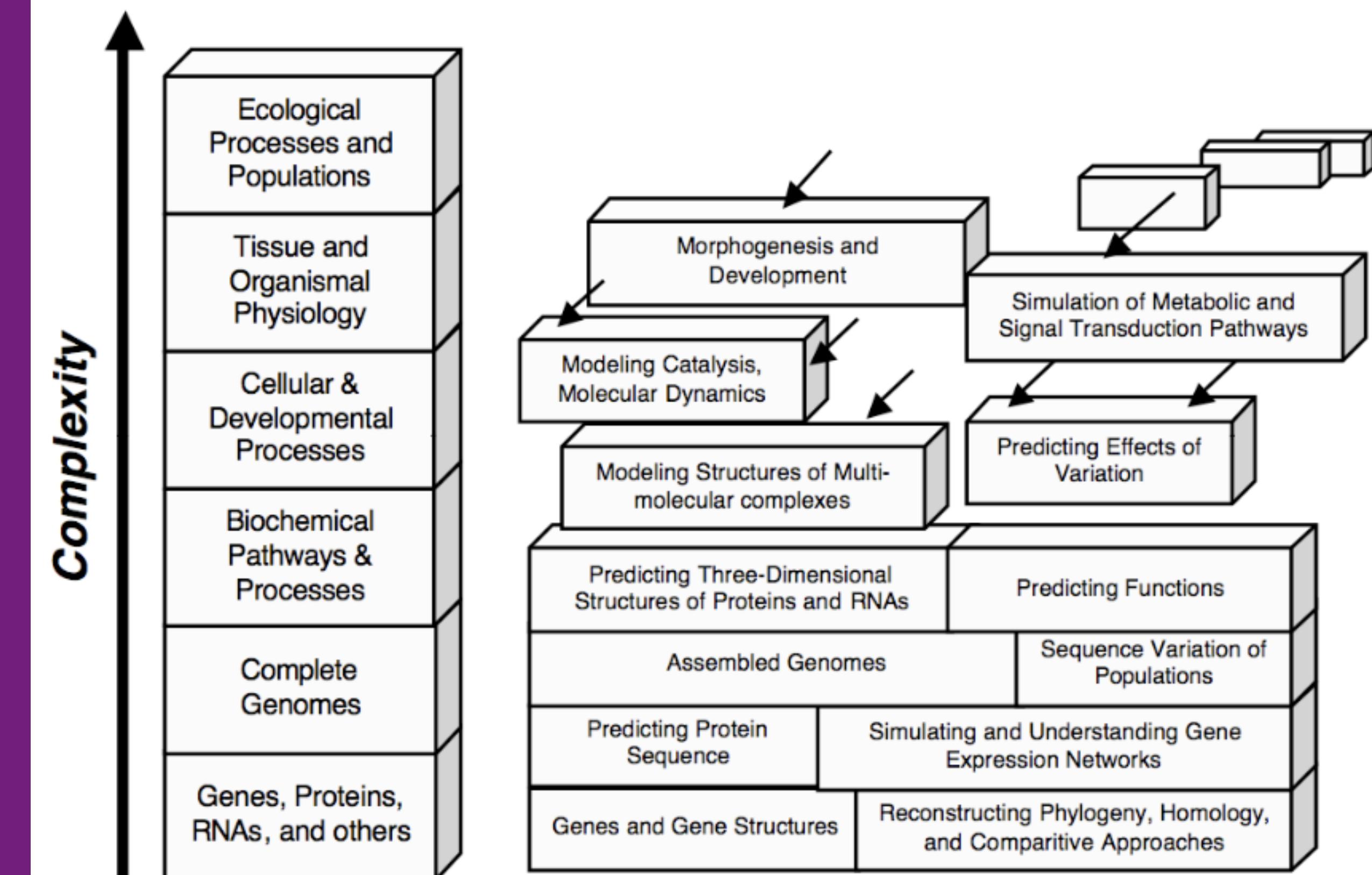
IMPACT OF BIOINFORMATICS

- On computational sciences
 - New algorithmic and statistical problems
 - Big data



IMPACT OF BIOINFORMATICS

- On Medical sciences
 - Translational research
 - Personalized treatment



ASSIGNMENT

ASSIGNMENT 1

uchicago-bio / **mpcs56420-2020-spring**

Unwatch ▾ 2

Star 0

Fork 0

Code

Issues 0

Pull requests 0

Actions

Projects 0

Wiki

Security

Insights

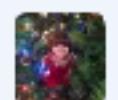
Settings

Branch: master ▾

mpcs56420-2020-spring / assignment-1 / mpcs56420-2020-spring-assignment-1.ipynb

Find file

Copy path



tabinks Update mpcs56420-2020-spring-assignment-1.ipynb

ad0baed now

1 contributor

371 lines (371 sloc) | 15.7 KB



Raw

Blame

History



MPCS 56420 - 2020 - Spring - Assignment 1

This assignment is due Thursday, April 16, 2020 at 5:29 PM. The instructors will clone a copy of your assignment repository and use the last version checked in before the due date. Please answer immediately below each question.

Github Classroom Repo: <https://classroom.github.com/a/ro8lKKH2> If you have not done so, please change your Github account user name as soon as possible.

HTTPS://GITHUB.COM/
UCHICAGO-BIO/

MPCS56420-2020-SPRING/
BLOB/MASTER/ASSIGNMENT-1/
MPCS56420-2020-SPRING-
ASSIGNMENT-1.IPYNB