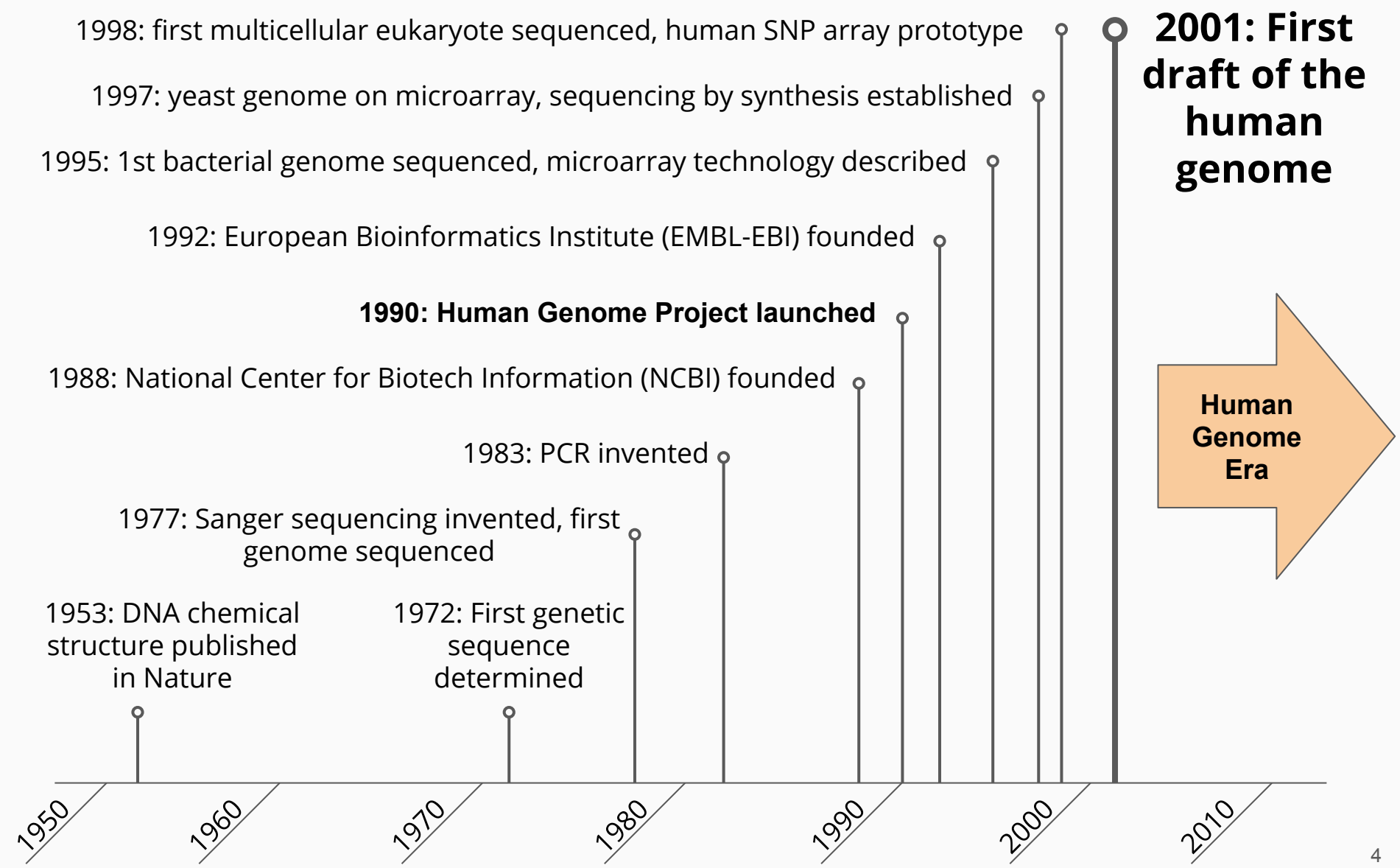
BF528 - Applications in Translational Bioinformatics

Introduction

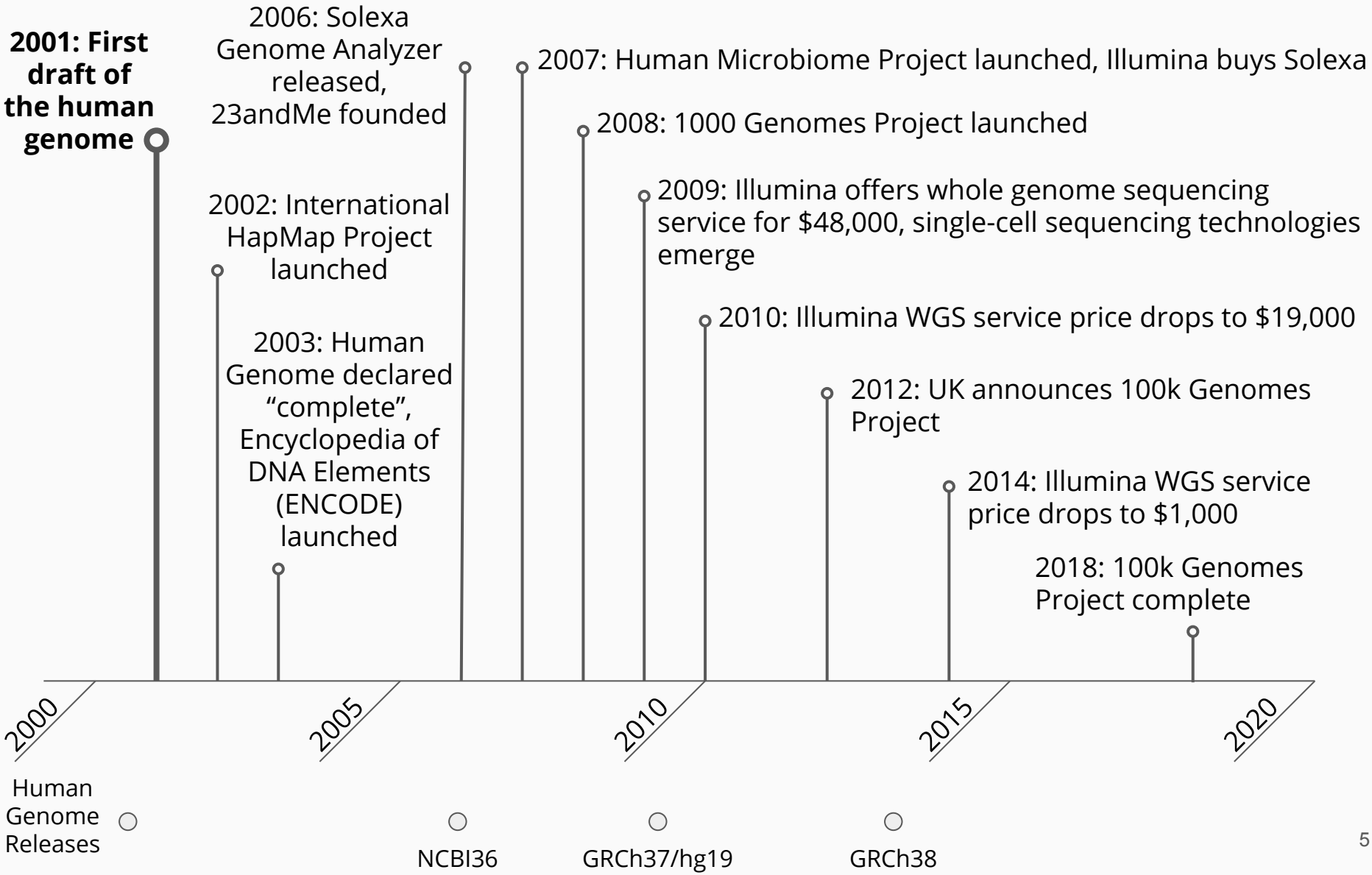
All Lectures Are Being Recorded
but
Only my screen!

Biology became a data science in 1953

Biology as a Data Science



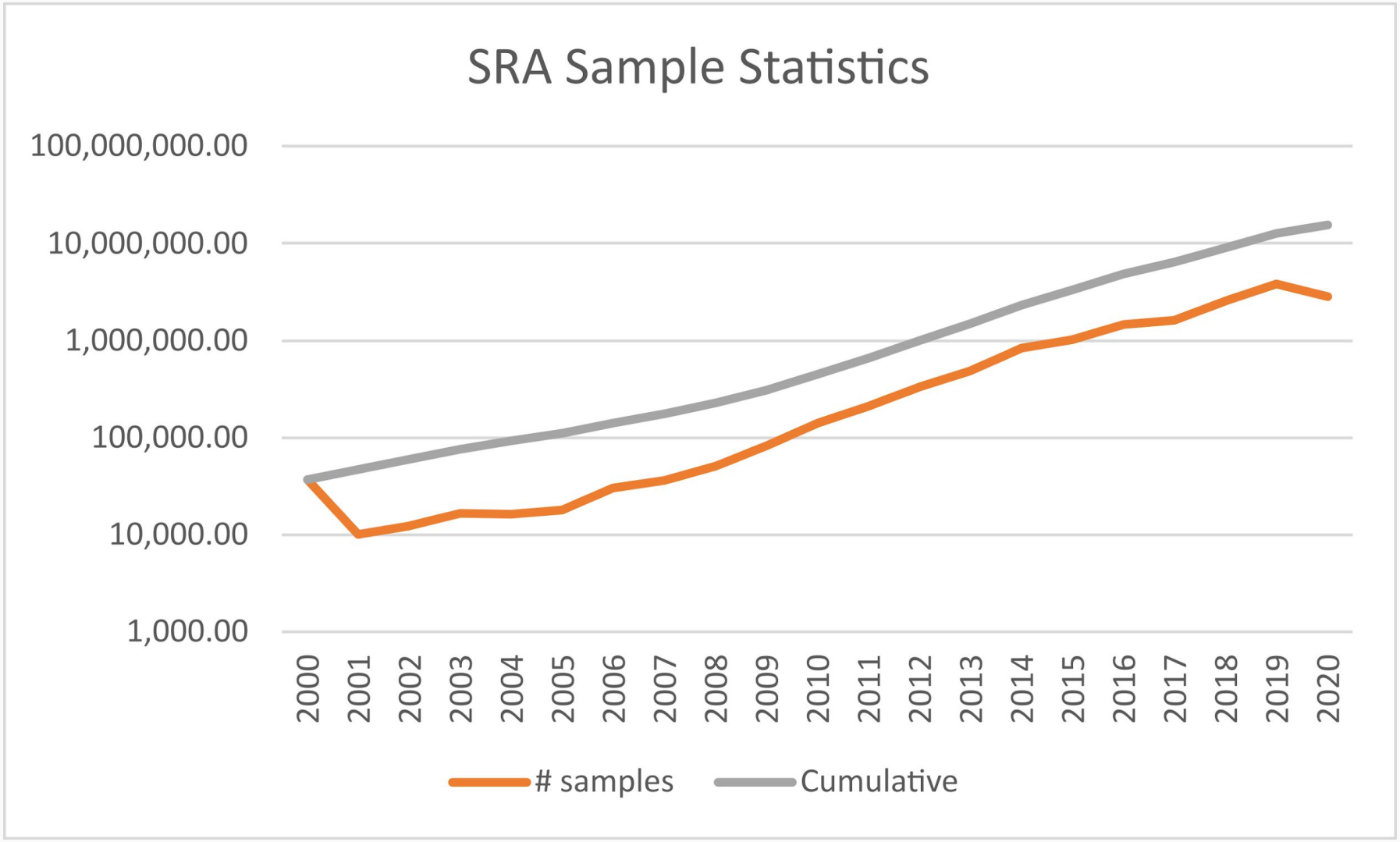
The Human Genome Era



“Big” Data

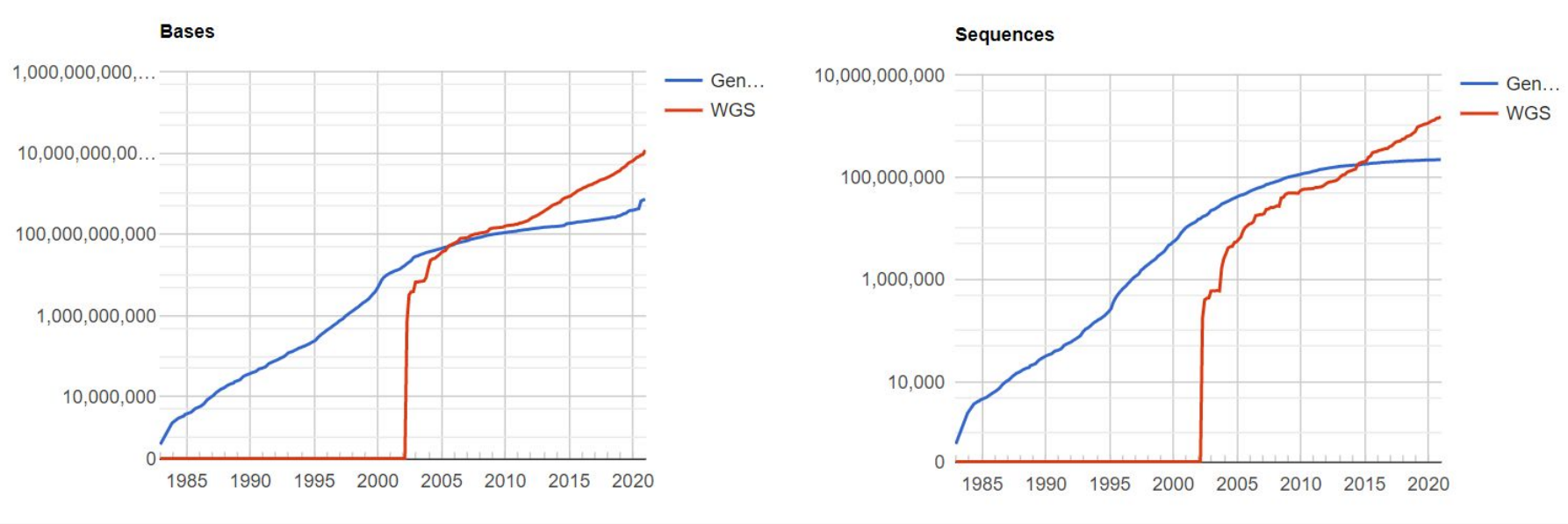
- Single Microarray dataset: ~500Mb
- Single short read dataset: ~2Gb-300Gb
- Human genome reference sequence: ~2Gb
- One run of Illumina instruments:
  - HiSeq 2500: ~1Tb
  - NovaSeq 6000: ~6Tb
- NCBI Short Read Archive (SRA) samples:
  - 2015: 3,351,430: ~37 Tb
  - 2020: 15,671,845: 17,500 Tb (136,718 iPhones)

Short Read Archive Sample Repository



NCBI Genbank Sequences

NCBI GenBank - https://www.ncbi.nlm.nih.gov/genbank/



What is Bioinformatics?

"**Bioinformatics** is an interdisciplinary field that develops methods and software tools for understanding biological data, in particular when the data sets are large and complex. As an interdisciplinary field of science, bioinformatics combines biology, computer science, information engineering, mathematics and statistics to analyze and interpret biological data."

Wikipedia - (Accessed 1/18/2021)

What is Bioinformatics?!?

"an interdisciplinary field that develops methods and software tools"?

"an interdisciplinary field of science"?

DEPENDS ON WHO YOU ASK

NB: "**Science** (from the Latin word scientia, meaning "knowledge") is a systematic enterprise that builds and organizes knowledge in the form of testable explanations and predictions about the universe." - Wikipedia

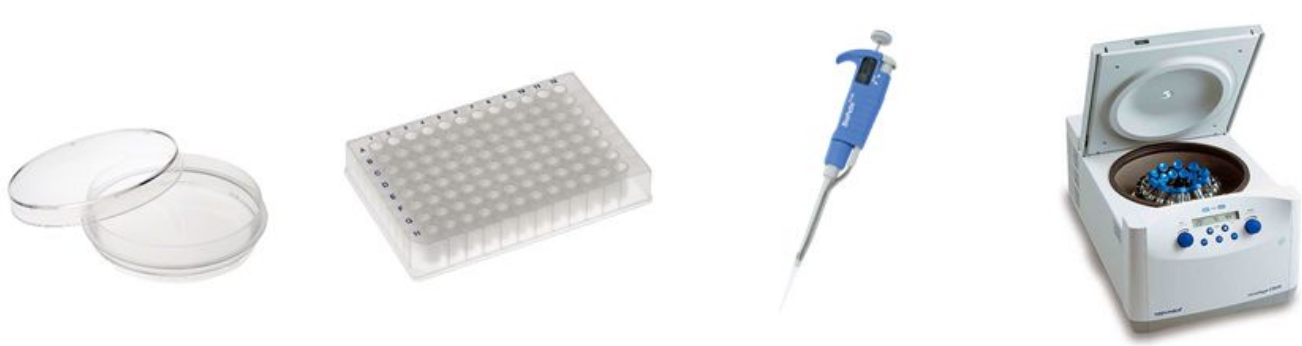
What is Bioinformatics?

"**Bioinformatics** is an interdisciplinary field that develops methods and software tools for understanding biological data, in particular when the data sets are large and complex. As an interdisciplinary field of science, bioinformatics combines biology, computer science, information engineering, mathematics and statistics to analyze and interpret biological data."

Wikipedia - (Accessed 1/18/2021)

The Biologist's Tools

Wet lab biologists:

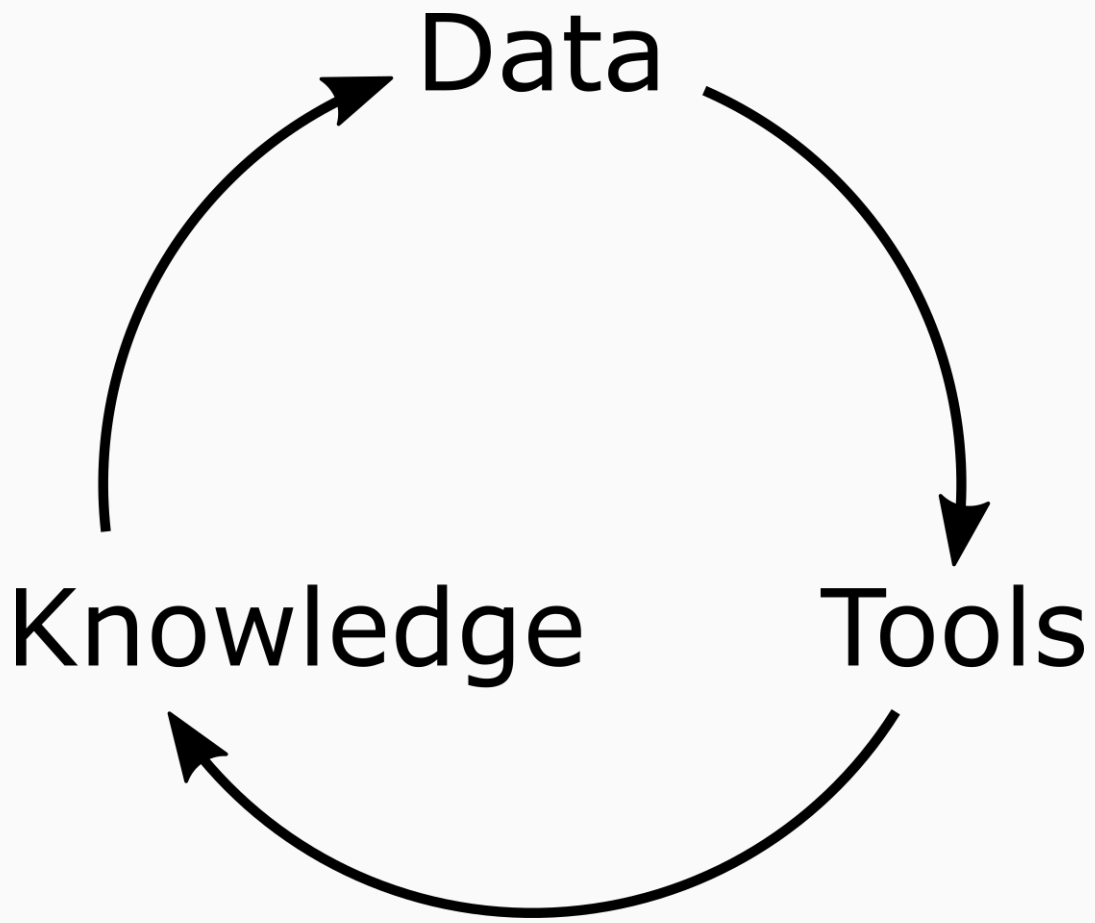


Bioinformaticians:

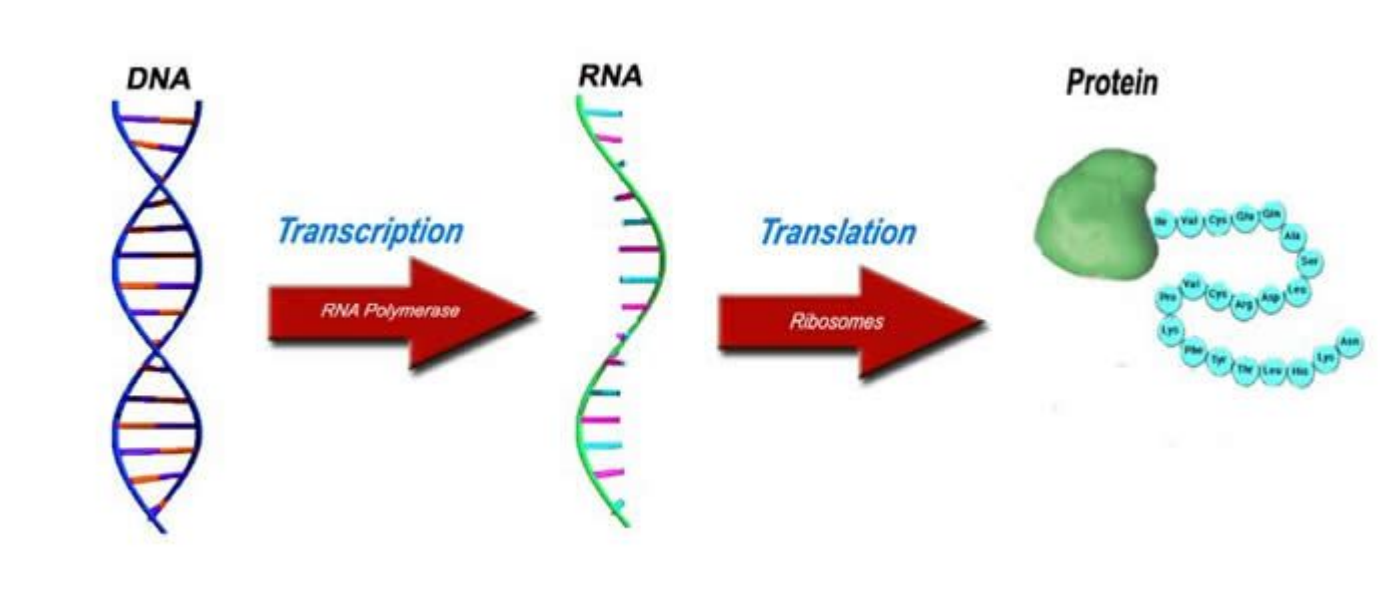
Conceptual History of Bioinformatics

1. Biological sequences digitized
2. Biological databases needed to store sequences
3. Search tools needed for databases
4. Tools for analyzing data from searches
5. New tools required to analyze human genome
6. Those new tools enable analysis of large amounts of sequencing data
7. Sequencing data volume explodes, requiring new tools
8. ...

Data → Tools → Knowledge Cycle



Central Dogma of Biology



Each of these molecules is a **sequence**

http://www.lhsc.on.ca/Patients_Families_Visitors/Genetics/Inherited_Metabolic/Mitochondria/DiseasesattheMolecularLevel.htm

The Sequence: The Fundamental Datatype

Molecular Sequence

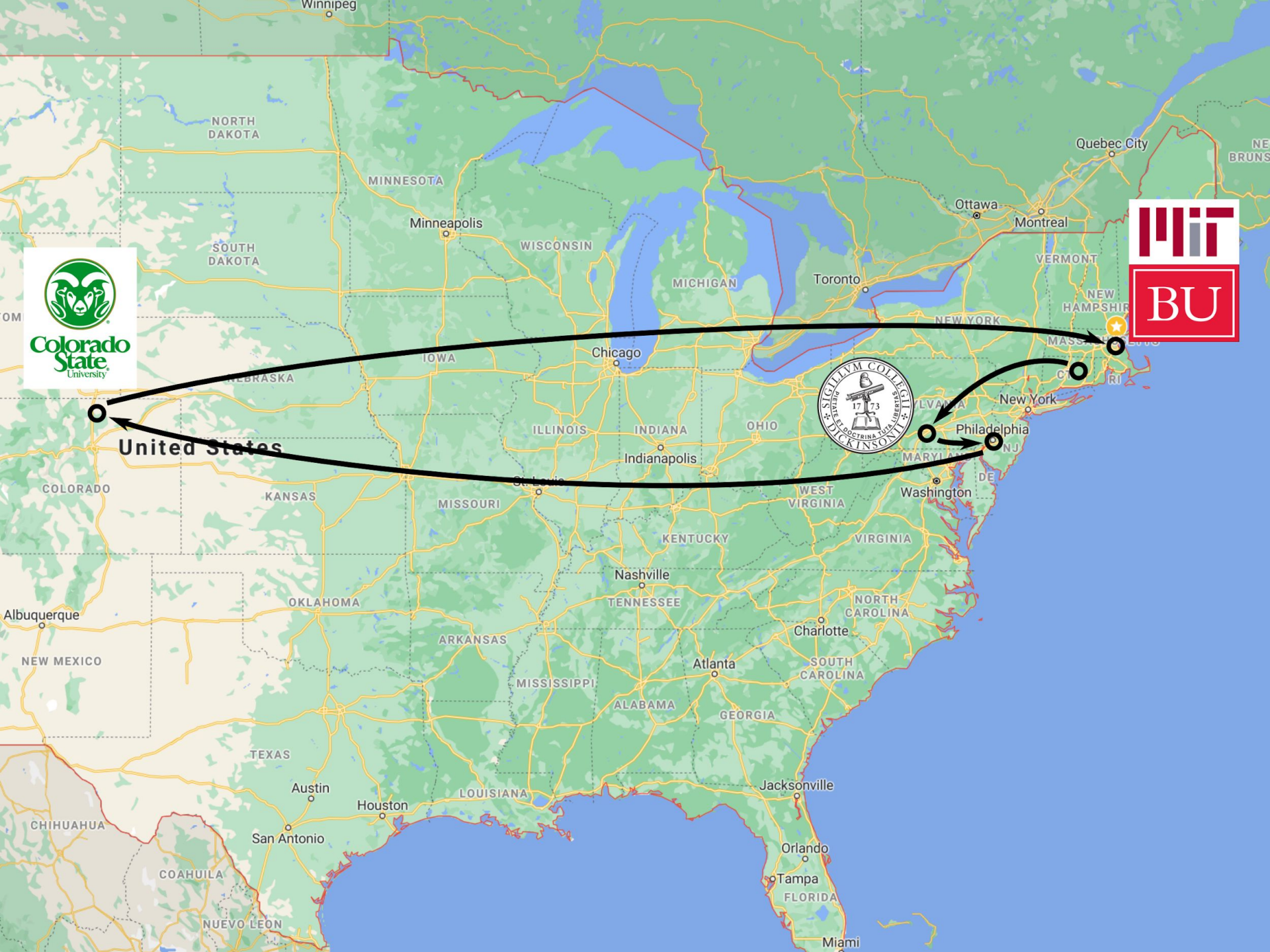
+ Computer Science
  = genome assembly, homology, phylogeny
+ Physics
  = DNA/RNA/protein structure, drug prediction
+ Statistics
  = gene expression, population genetics, biomarkers
+ Mathematics
  = metabolic modeling, synthetic biology, systems biology

Translational Bioinformatics

"**Translational Bioinformatics (TBI)** is an emerging field in the study of health informatics, focused on the convergence of molecular bioinformatics, biostatistics, statistical genetics, and clinical informatics."

Wikipedia - (Accessed 1/18/2021)

Instructor Introduction


Colorado State University
MIT
BU
United States
Winnipeg
NORTH DAKOTA
MINNESOTA
SOUTH DAKOTA
Minneapolis
WISCONSIN
MICHIGAN
Quebec City
Ottawa
Montreal
Toronto
VERMONT
NEW HAMPSHIRE
NEW YORK
IOWA
Chicago
NEBRASKA
ILLINOIS
INDIANA
OHIO
Indianapolis
New York
Philadelphia
Washington
COLORADO
KANSAS
MISSOURI
WEST VIRGINIA
KENTUCKY
VIRGINIA
Nashville
OKLAHOMA
TENNESSEE
NORTH CAROLINA
Albuquerque
Charlotte
ARKANSAS
NEW MEXICO
Atlanta
SOUTH CAROLINA
MISSISSIPPI
ALABAMA
GEORGIA
TEXAS
Austin
LOUISIANA
Houston
Jacksonville
CHIHUAHUA
San Antonio
Orlando
COAHUILA
Tampa
FLORIDA
NUEVO LEON
Miami

Course Structure

Instructors

- **Instructor:** Adam Labadorf
- **TAs:**
  - Gargi Dayama
  - Jacquelyn (Jackie) Turcinovic
  - Shruthi Bandyadka
  - Jing Zhang

Course Organization

- http://bf528.readthedocs.io
- Lectures are on Zoom WF starting at 2:30
- Lectures will be recorded and posted online
  - Recordings do not capture Zoom video
- 5 projects over the course of the semester
  - 4 group, 1 individual
- Some online content early in semester

Course Overview

- Survey course in bioinformatics
- Focus on high-throughput sequencing data, tools, and techniques
- Focus on practical skills
- Group work simulates real-world collaborative environment

Course Goals

- Survey current bioinformatics techniques in translational studies
- Give you hands-on experience working with high-throughput biological data and tools
- Read and understand papers that use bioinformatics in translational studies
- Develop shared vocabulary between biology and computation

Schedule of Topics

Lecture	Day	Date	Topic	Secondary	Project
1	**W**	**Jan 27**	Introduction	Command Line Interface	
2	**F**	**Jan 29**	Genomics, Genes, and Genomes	Cluster Usage	
3	**W**	**Feb 3**	Array Technologies	git	1
4	**F**	**Feb 5**	Gene sets and enrichment	R+RStudio Primer	
5	**W**	**Feb 10**	2nd Gen Sequencing		
6	**F**	**Feb 12**	Sequence Analysis Fundamentals		
7	**W**	**Feb 17**	Sequence Analysis - RNA-Seq 1		
8	**F**	**Feb 19**	Sequence Analysis - RNA-Seq 2		
9	**W**	**Feb 24**	Genomic Variation and SNP Analysis		1/2
10	**F**	**Feb 26**	Sequence Analysis - ChIP-Seq		
11	**W**	**Mar 3**	Biological Data Formats	Something Fancy	
12	**F**	**Mar 5**	Sequence Visualization	Genome Browsers	
13	**W**	**Mar 10**	Biological Databases		
14	**F**	**Mar 12**	Replicability vs Reproducibility Strategies		

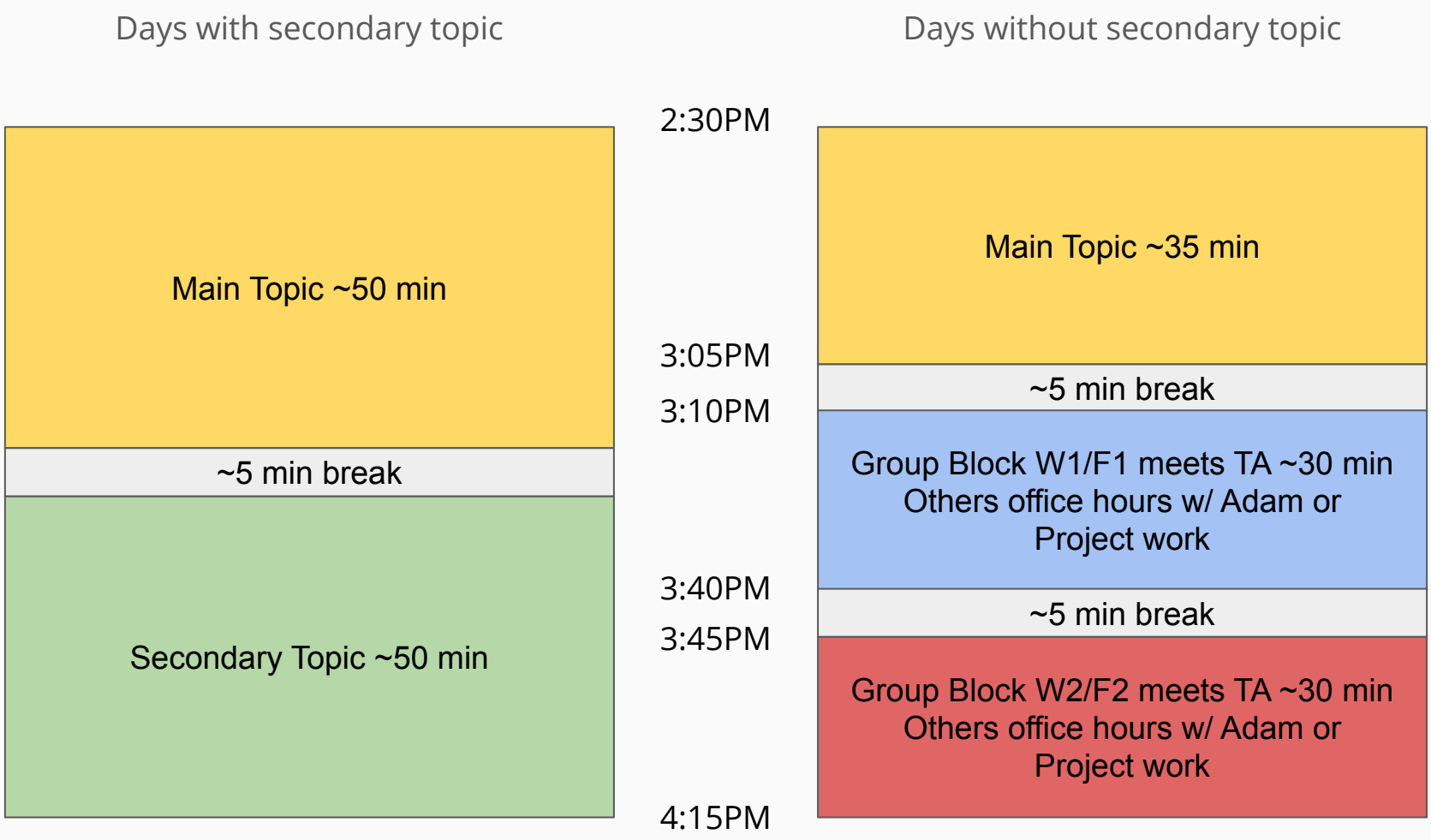
NB: during weeks without secondary topic, you will meet with your TA during the second half of class

Schedule of Topics cont'd

15	**W**	**Mar 17**	Computational Environment Management	Conda	2/3
16	**F**	**Mar 19**	Computational Pipeline Strategies	Snakemake	
17	**W**	**Mar 24**	Single Cell Techniques		
18	**F**	**Mar 26**	Single Cell Analysis		
	W	**Mar 31**	Wellness Day, no class		
19	**F**	**Apr 2**	Microbiome: 16S		
20	**W**	**Apr 7**	Microbiome: Metagenomics		3/4
21	**F**	**Apr 9**	Proteomics		
22	**W**	**Apr 14**	Metabolomics		
23	**F**	**Apr 16**	Integrative Genomics		
	W	**Apr 21**	Schedule adjustment, no class		
	F	**Apr 23**	Project work		
24	**W**	**Apr 28**	The Future + retrospective, end of classes		4/5
	F	**May 8**	Final projects due		5

NB: during weeks without secondary topic, you will meet with your TA during the second half of class

Lecture Structure



Group Block W1/W2 meet on Wednesdays, Group Blocks F1/F2 meet on Fridays

Group Blocks Illustration

Wednesday Groups

	Group Block	Gargi	Jackie	Jing	Shruthi
3:10	W1	1	2	3	
3:45	W2		4	5	6

Friday Groups

	Group Block	Gargi	Jackie	Jing	Shruthi
3:10	F1	7	8	9	
3:45	F2		10	11	12

Projects

- Reproduce findings from published studies
- 5 projects over the course of the semester
  - Last project is an individual project
- Each project has a full writeup
- Your writing must be your own, but you may share everything else!
- No homeworks or exams

Project Groups

- Students assigned into groups of 3 or 4
- Each group has a primary TA
- You will receive knowledge survey soon
- Assigned into groups based on experience
- Groups are for the entire semester

Project Groups cont'd

- Group members will play one of four roles:
  - *Data Curator* - find, download, and organize data
  - *Programmer* - process data into analyzable form
  - *Analyst* - transform processed data into interpretable form
  - *Biologist* - understand paper and biological context, help interpret results
- Roles rotate for each project
- Structured class time to help facilitate group work and help each other!

Project Group Meeting

- Your group will meet once most weeks with your TA over Zoom
- Feel free to bring questions about the lecture material
- Discuss project progress:
  - “What did I work on since our last meeting?”
  - “What challenges did I encounter?”
  - “Are there any obstacles to completing my work?”
  - “What will I be working on for next meeting?”

Project Report

- Organized like a published study
- Sections (primary role):
  - Intro - background and motivation (Biologist)
  - Data - data description (Data Curator)
  - Methods - processing and tools (Programmer)
  - Results - findings (Analyst)
  - Discussion - interpret findings (Biologist)
  - Conclusion (all)

Assessment

- Each project is 20% of your total grade
- Evaluation
  - You will receive feedback for each project
  - Your work will be assessed in the following areas:
    - Accuracy
    - Clarity of language
    - Figure/table quality
    - Report formatting
    - Thoroughness
    - Cohesion
  - Detailed guidance found on course website

Core Course Values

- OPENNESS
  - Anyone can work together - share code!
  - Copy code from Stack Exchange!
- ACCEPTANCE
  - All backgrounds and beliefs are welcome.
  - This class is a safe place for dissent.
- PATIENCE
  - We are all learning. We all bring value. Be kind!