

Bioinformatics

Announcements

- Homework 8 has been released and will be due **Wed 8/9 11:59pm**
 - There are **two** components --> surveys and written response
 - The surveys are NOT eligible for any extensions
 - The written response is eligible for extensions as usual
- All students who receive full credit on this homework are eligible to receive **1 additional extra credit point** if at least 80% of the course gets full credit on this homework (submits all surveys and completes the written response)
- Exam alterations form priority deadline was yesterday
 - If you need an exam alteration please request ASAP
- HW Recovery 1-4 has been processed
 - No HW Recovery for HW 7 & 8
- Lab 13 is optional
- Topical Review Sessions today

Biomedical Engineering vs. Bioinformatics vs. Computational Biology vs. Biotechnology

Biomedical Engineering

using engineering to treat disease

knee replacements

fitness trackers

deep brain stimulation

Bioinformatics

using tech to analyze DNA, RNA, proteins, and Big Data in biology

AATCGTACGAAGATC

Leu Ala Cys Phe

supercomputing and genomics

Computational Biology

using computer science, math, and statistics to understand biology

comparing brain cells to computer networks

Proto

science of sound and hearing

Biotechnology

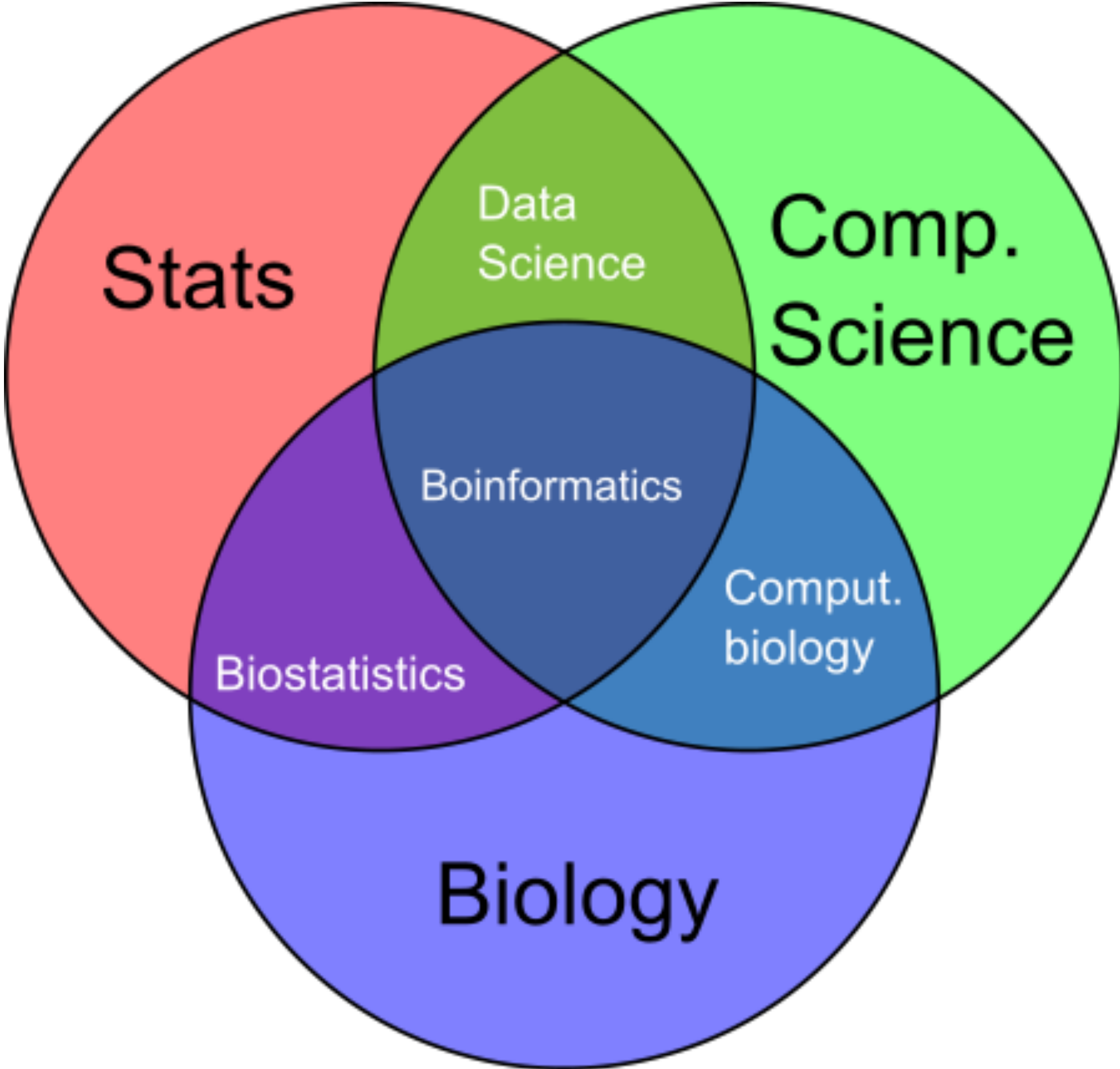
using biology as technology to improve research and industry

making cells fluorescent under the microscope

CRISPR gene editing

seedless fruit

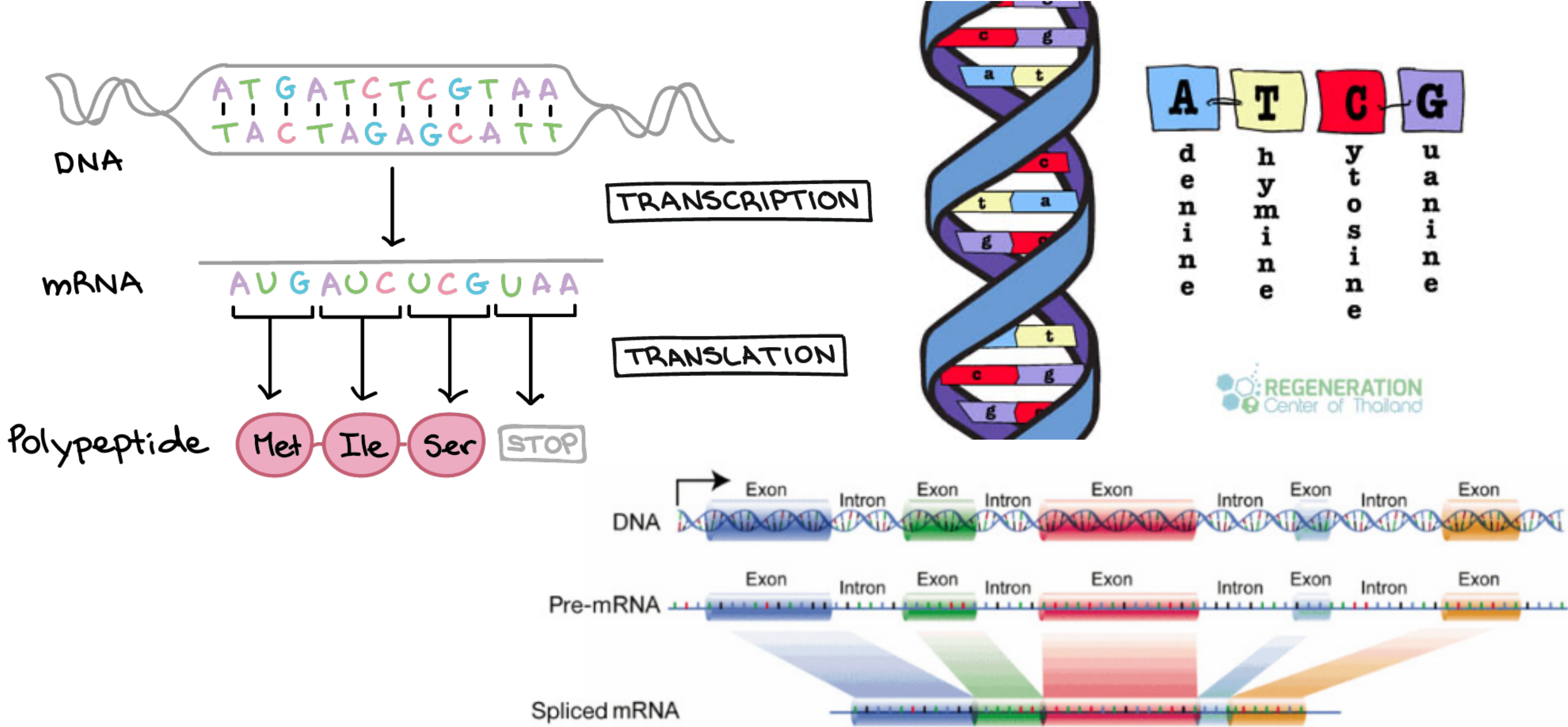
Focus for Today: Bioinformatics



Fields in Bioinformatics

- ***Translational Bioinformatics***– Development of techniques for transforming voluminous biomedical (especially genomic) data to support proactive, predictive, preventive, and participatory health
- ***Clinical Research Informatics***– Development of approaches for enabling the discovery, management, and evaluation of new health knowledge
- ***Clinical Informatics***– Development and application of techniques to improve health care delivery services; clinical informatics is a subspecialty of the American Board of Medical Specialties
- ***Consumer Health Informatics***– Development of information structures and approaches for supporting patient-centric health care needs
- ***Public Health Informatics***– Development of methodologies for supporting public health needs, including surveillance, prevention, preparedness, and health promotion

Central Dogma of Biology

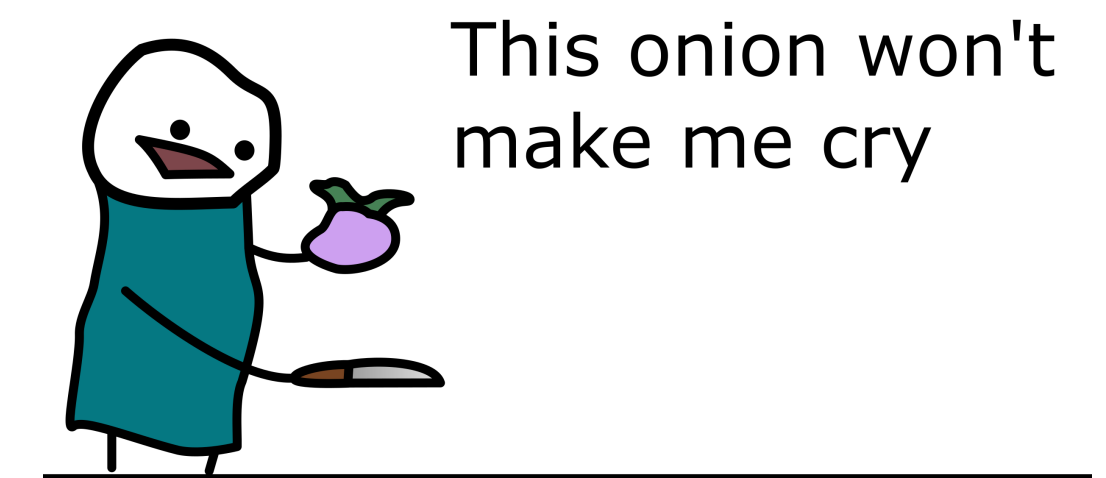
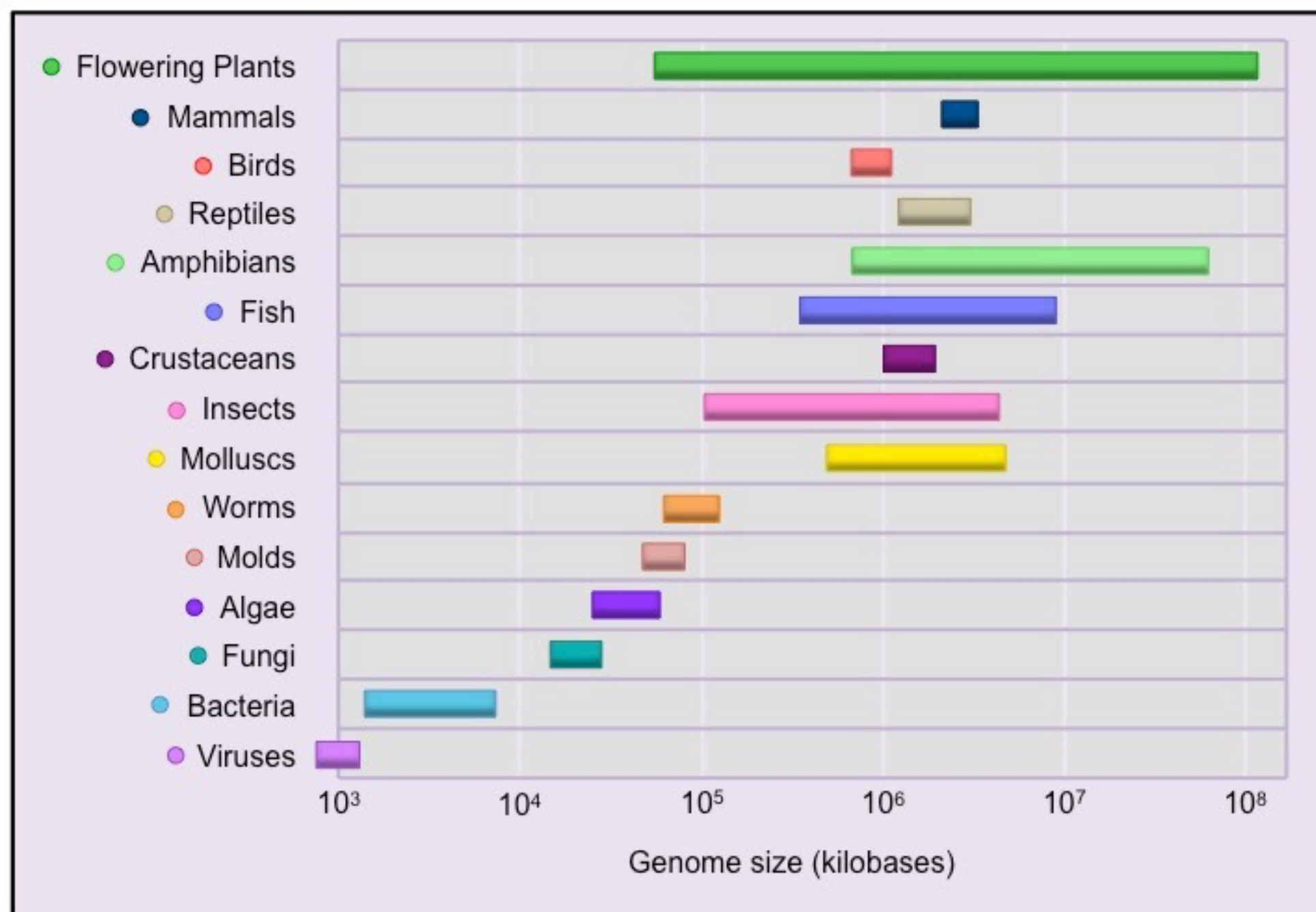


Genome

- **DNA:** string of complex molecules called nucleotides. It contains the genetic information and acts as a set of instructions for how to build and maintain you
- **Genome:** complete set of DNA
- **Gene:** DNA is organized into little chunks of information that each carry a specific set of instructions for how to make a certain aspect of you

Genome

- The complexity of an organism increases from the lower single-celled organisms to higher multicellular organisms
- Would an onion or human have a larger genome size?
 - C-Value Paradox: genome size fails to correlate well with apparent complexity
 - Onion: 16 billion bases, Human: 3.2 billion bases
- Size of the genome varies across different groups of organisms



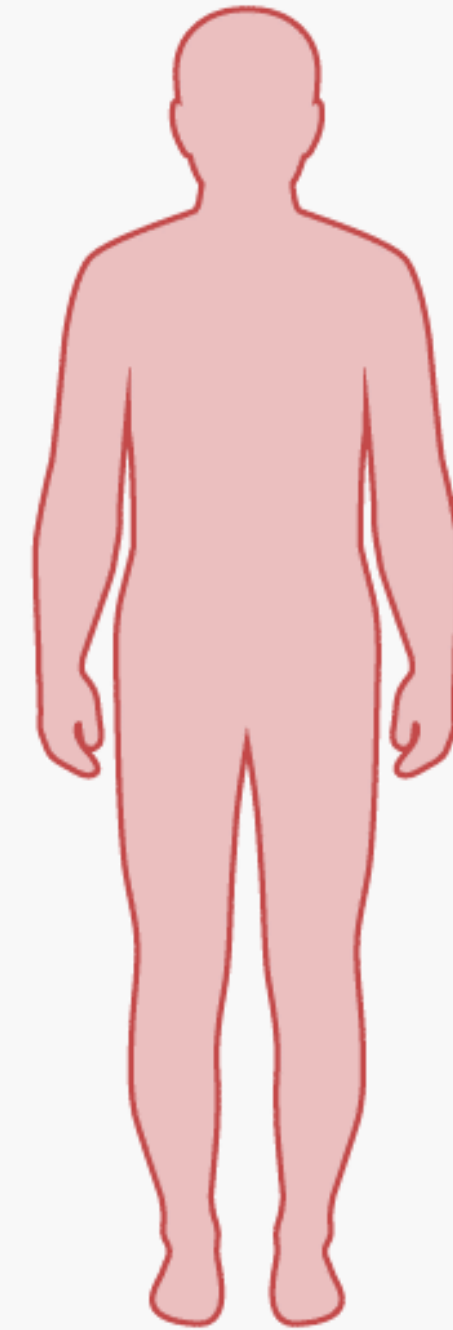
Fun Fact!



The genetic similarity
between a human
and a mouse is:

85%

Source: National Human Genome Research Institute

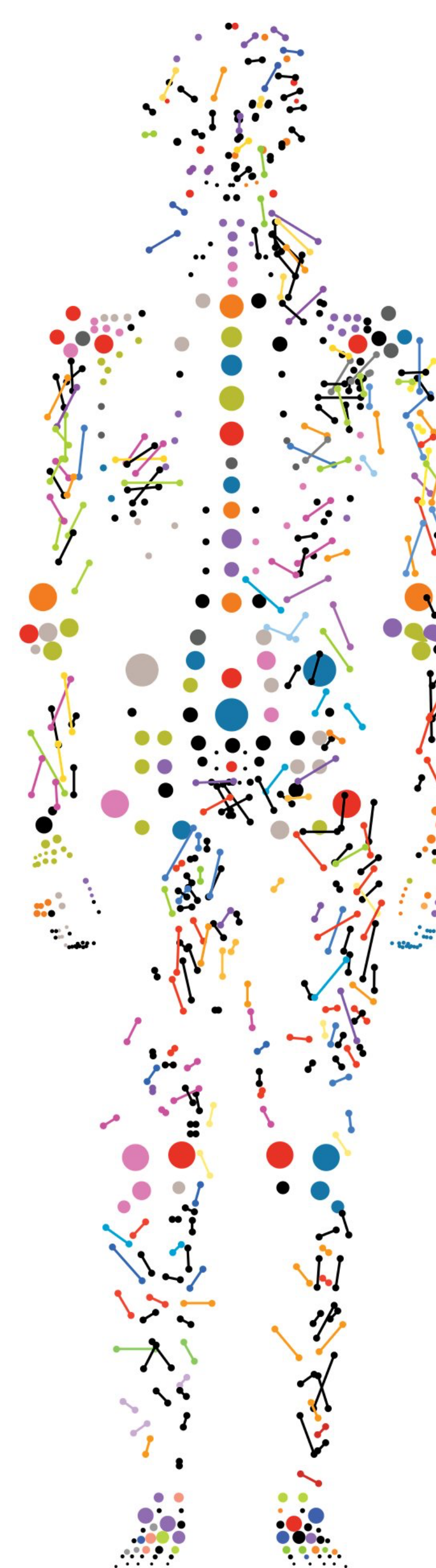


BUSINESS INSIDER



Human Genome Project

- An international scientific research project with the goal of determining the base pairs that make up human DNA
- Launched in October 1990 and completed in April 2003
- Tells us a lot about our genes and how they are organized!



Bioinformatics: Genomic Analysis

How does bioinformatics allow us to understand the similarity in genes?

Algorithms will scan past both ends of the matching sequence



Mouse

. . . A T G C G T A G C C A T A T C C G A A T C G A . . .

Similarities in sequences:
Analyze those genes and see how they translate into similar traits

Differences in sequences:
Analyze those genes and see how they translate into different traits



Human

. . . A T G C G T A G C C A T A T C C G A A C T T T . . .

Bioinformatics: Genomic Analysis



Cinderella

. . . A T G C G T A G C C A C A T C C G A A T C G A . . .

Is this **base difference C/T**
significant for disease?



Belle

. . . A T G C G T A G C C A T A T C C G A A T C G A . . .

Conduct a Study

Is this **base difference** significant for disease?



Group A: 100 Healthy Subjects

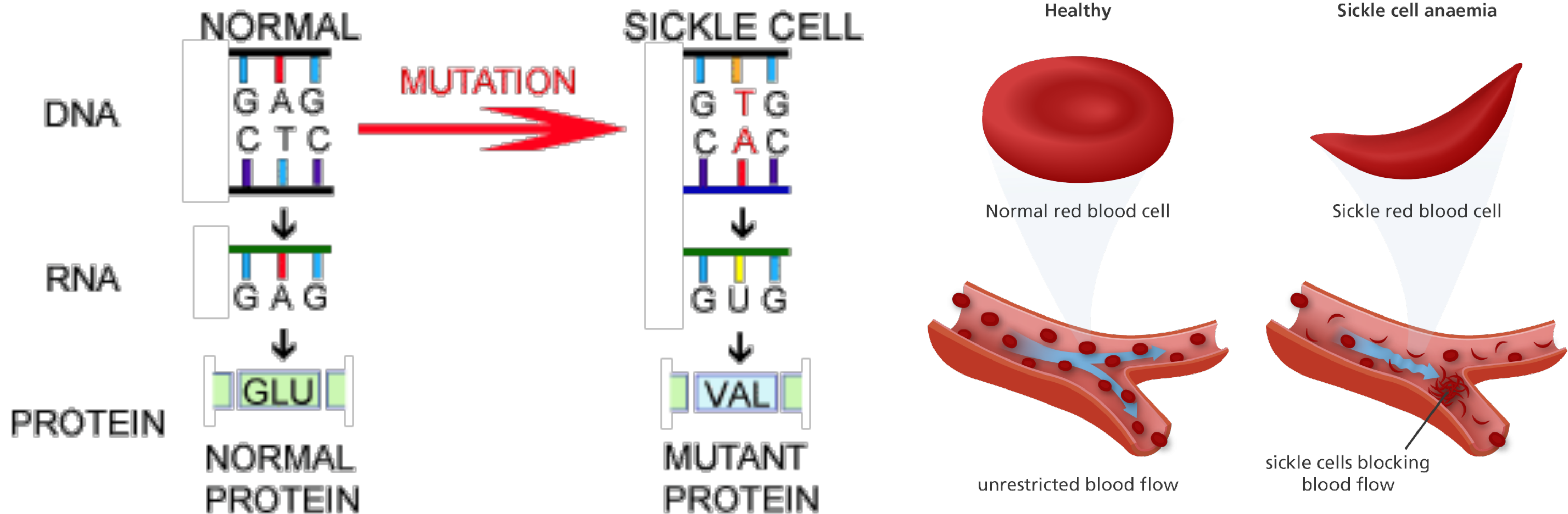


Group B: 100 Diabetic Subjects

Hypothetical Results: **4/100** of group A have a **T** and **98/100** of group B have **T**

Base Substitution Sickle Cell Disease

- Sickle cell disease is an inherited disease in which red blood cells contort into a sickle shape and die early, leaving a shortage of healthy red blood cells
- Discovered through genomic analysis, the genetic basis of sickle cell disease is an **A-to-T transversion** in the sixth codon of the HBB gene

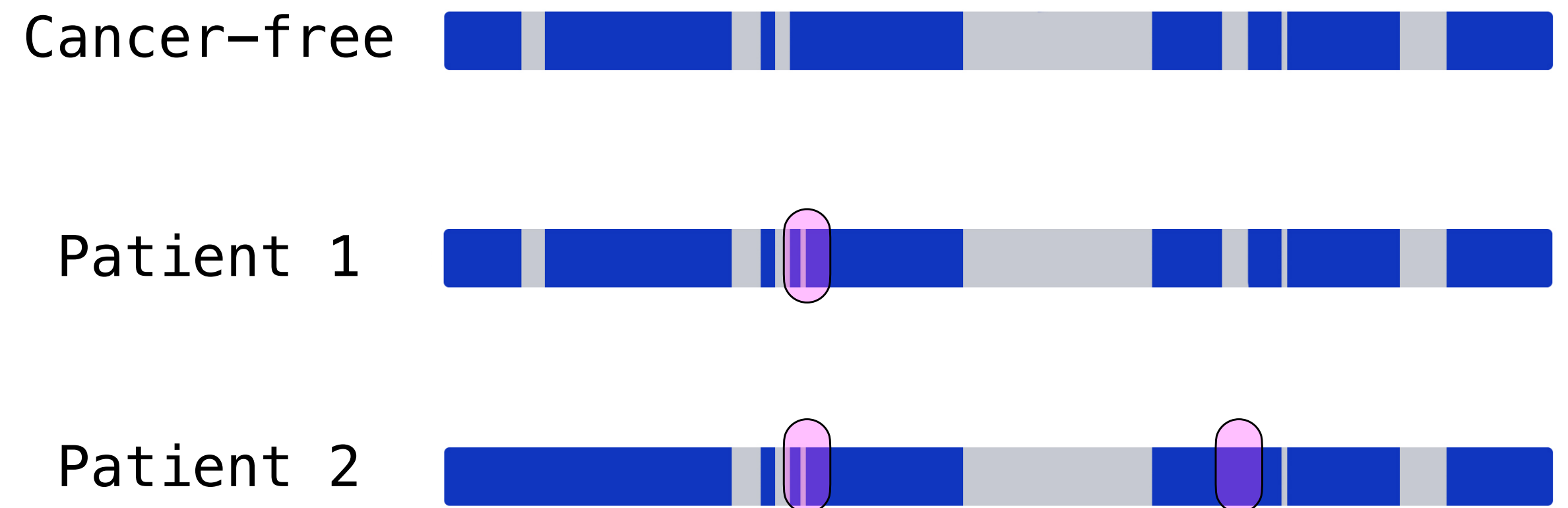


Applications in Neuroscience

- Stroke is a leading cause of death in the US and 87% of strokes are ischemic strokes which cause a lot of irreversible damage
- Ischemic stroke: blood supply to part of the brain is interrupted/reduced
- Arctic ground squirrels: their brain is incredibly resilient!
- **Provides us clues for stroke treatment**

Genomic Analysis for Cancer Treatment and Diagnosis

- Clinician's can order genome sequencing of their patients
- The patient's cancer cells are compared with the normal genome and genome of many other patients with cancer
 - Pinpoint mutations that are allowing the cancer cells to grow uncontrollably
 - Choose the best treatment



Sequence Alignment

- **Sequence alignment** is a way of arranging the DNA sequences to identify regions of similarity that may be a consequence of functional, **structural**, or **evolutionary** relationships between the sequences
- Aligned sequences are typically represented as rows within a **matrix**
- Two alignment types are used: **global** and **local**

Insulin Gene Sequence Database

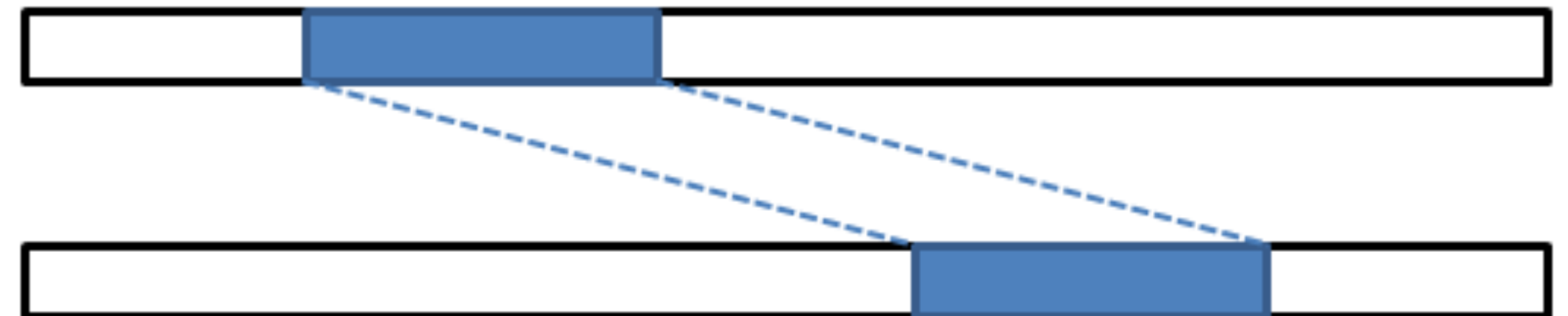
```
Mouse ) ----MRIMAVITQERKIAKWKIEEVKELEOKLREYHTIIIANIEGFPADKLHDIRKKMRGM-AEIKVTKNTLFGIAAKNAG----LDVS
Rat   ) ----MKRLALALKQRKVASWKLEEVKELTELKNSNTILIGNLEGFPADKLHEIRKKLRGK-ATIKVTKNTLFKIAAKNAG----IDIE
      ) MSVVS LVGQMYKREKPIPEWKTMLRELEELFSKHRVVLFAADLTGTPTFVVQRVRKKLWKK-YPMMAVAKKRIILRAMKAAGLE---LDDN
      ) -MMLAIGKRRYVRTROYPARKVKIVSEATELLQKYPYVFLFDLHGLSSRILHEYRYRLRRY-GVIKI IKPTLFKIAFTKVYGG---IPAE
      ) -----MAEERHHTEHIPQWKKDEIENIKELIQSHKVFGMVGIEGILATKMOKIRRDLDKDV-AVLKVSNTLTERALNQLG-----ETIP
      ) -----MAEERHHTEHIPQWKKDEIENIKELIQSHKVFGMVRIEGILATKIQKIRRDLDKDV-AVLKVSNTLTERALNQLG-----ESIP
      ) -----MAAVRGS---PPEYKVRAVEEIKRMIS SKPVVAIVSFRNVPAGOMOKIRREFRGK-AEIKVVKNTLLE RALDALG-----GDYL
      ) MAVKAKGQPPSGYE PKVAEWKRREVKELKELMDEYENVGLVDLEGIPAPQLQEIRAKLRERDTIIRMSRNTLMR IALEEKLDER--PELE
      ) -----MAHVAEWKKKEVQELHDLIKGYEVVGIANLADIPARQLQKMRQTLRDS-ALIRMSKKTLLISLALAEKAGREL--ENV
      ) -----MITAESEHKIAPWKIEEVNKLKELLKNGQIVALVDMMEVPAVQLQEIRDKIR-GTMTLKMSRNTLIERAIKEVAEETGNPEFA
      ) -----MIDAKSEHKIAPWKIEEVNALKELLKSNVIALIDMMEVPAVQLQEIRDKIR-DQMTLKMSRNTLIKRAVEEVAEETGNPEFA
      ) -----METKVKAHVAPWKIEEVKTLKGLIKSKPVVAIVDMMDVPAPQLQEIRDKIR-DKVKLRMSRNTLIIRALKEAAEELNNPKLA
      ) -----MAHVAEWKKKEVEELANLIKSYVIALVDVSSMPAYPLSQMRRLLIRENGGLLRVSRNTLIELAIKKAQELGKPELE
```

Global & Local Alignment

- The global approach compares one whole sequence with other entire sequences
- The output of a global alignment is a one-to-comparison of two sequences
 - Used when comparing two genes of similar function
- The local method uses a subset of a sequence and attempts to align it to subset of other sequences
- Local regions are aligned with the **highest level of similarity**
- Looking for conserved patterns in DNA



Global Alignment



Local Alignment

BLAST: Basic Local Alignment Search Tool

- Identifies similarities between sequences by comparing it with database of sequences

Insulin Gene Sequence Database

```

Mouse -----MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKOMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
Rat -----MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKOMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
-----MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKOMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
-----MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKOMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
-----MPREDRATWKSNYFMKIIQLLDDYPKCFVVGADNVGSKOMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
-----MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKOMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--SALE
-----MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKOMQTIIRLSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
-----MPREDRATWKSNYFLKIIQLLNDYPKCFIVGADNVGSKOMQTIIRLSLRGK-AIVLMGKNTMMRKAIRGHLENN--PALE
-----MVRENKAAWKAQYFIKVVLELDFPKCFIVGADNVGSKOMQNIIRTSLRGL-AVVLMGKNTMMRKAIRGHLENN--PQLE
-----MSGAG-SKRKKLFIEKATKLFTTYDKMIVAEADNVGSQLOKIRKSIRGI-GAVLMGKNTMIRKVIRDLADSK--PELD
-----MSGAG-SKRKNVIEKATKLFTTYDKMIVAEADNVGSQLOKIRKSIRGI-GAVLMGKNTMIRKVIRDLADSK--PELD
-----MAKLSKQKKQMYIEKLSLIQQYSKILIVHVDNVGSNOMASVRKSLRGK-ATIILMGKNTRIRRTALKKNLQAV--PQIE
----MIGLAVTTTKKIAKWKVDEVAELTEKLRKTHKTIIANIEGFPADKLHEIRKKLRGK-ADIKVTKNNLNFNIALKNAG----YDTK
---MRIMAVITQERKIAKWKIEEVKELEQKLREYHTIIIANIEGFPADKLHDIRKKMRGM-AEIKVTKNTLFGIAAKNAG----LDVS
---MKRLALALKQRKVASWVKELELTIKNSNTILIGNLEGFPADKLHEIRKKLRGK-ATIKVTKNTLFGIAAKNAG----IDIE
MSVVSIVGQMYKREKPIPEWKTLMLELELFSKHRVVLADLTGTPTFVVQVRVKKLWKK-YPMVAKKRIILRAMKAAGLE---LDDN
-MMLAIGKRRYVRTROYAPARKVKIVSEATELLQKYPYVFLFDLHGLSSRILHEYYRRLRY-GVIKIIKPTLFGIAAKNAG---IPAE
-----MAERHTEHIPQWKKDEIENIKELIQSHKVFVGMVRIEGILATKIQKIRRDLDV-AVLKVSNTLTERALNQLG----ETIP
-----MAERHTEHIPQWKKDEIENIKELIQSHKVFVGMVRIEGILATKIQKIRRDLDV-AVLKVSNTLTERALNQLG----ESIP
-----MAAVRGS--PPEYKVRAVEEIKRMISKPVVAIVSFRNVPAGOMQKIRREFRGK-AEIKVVKNTLLERALDALG----GDYL
MAVKAKGQPPSGYEPKVAEWKRREVKELKELMDEYENVGLVDLEGIPAPQLOEIRAKLRERTIIRMSRNTLMRIALEEKLDER--PELE
-----MAHVAEWKKKEVQELHDLIKGYEVVGIANLADIPAROLOKMRQTLRDS-ALIRMSKKTLLISLAEKAGREL--ENVD
-----MITAESEHKIAPWKIEEVNKLKELLKNGQIVALVDMMEVPAQLOEIRDKIR-GTMTLKMSRNTLIERAIKEVAEETGNPEFA
-----MIDAKSEHKIAPWKIEEVNALKELLKSANVIALIDMMEVPAVQLOEIRDKIR-DQMTLKMSRNTLIKRAVEEVAEETGNPEFA
-----METKVKAHVAPWKIEEVKTLKGLIKSKPVVAIVDMMDVPAPQLOEIRDKIR-DKVKLRMSRNTLIIRALKEAAEELNPNKLA
-----MAHVAEWKKKEVEELANLIKSYPVIALVDVSSMPAYPLSQMRRLLIRENGGLLRVSNTLIE LAIKKAAQELGKPELE
-----MAHVAEWKKKEVEELAKLIKSYPVIALVDVSSMPAYPLSQMRRLLIRENGGLLRVSNTLIE LAIKKAAQELGKPELE
-----MAHVAEWKKKEVEELANLIKSYPVVALVDVSSMPAYPLSQMRRLLIRENGLLRVSNTLIE LAIKKVAQELGKPELE
-----MAHVAEWKKKEVEELANIKSYPVIALVDVAGVPAYPLSKMRDKLR-GKALLRVSNTLIE LAIKRAAQELGQPELE
----MSAESERKTETIPEWKQEEVDAIVEMIESYESVGVVNIAGIPSRQLODMRRDLHGT-AELRVSNTLTERALDDVD----DGLE
----MSESEVRQTEVIPQWKREEVDELVDFIESYESVGVVGVAGIPSRQLOSMRRE LHGS-AAVRMSRNTLVNRRALDEVN----DGFE
----MSAEEQRTTEEVPEWKRQEVAVELVDLLETYDSVGVVNVGTGIPSKOLODMRRGLHGQ-AALRMSRNTLLVRALEEAG----DGLD
-----MKEVSQKKKELVNEITQRIKASRSVAIVDTAGIRTRQIQDIRGKNRGK-INLKVIKKTLLFKALENLGD---EKLS
-----MRKINPKKKEIVSELAODITKSKAVAIVDIKGVTRROMODIRAKNRDK-VKIKVVKKTLLFKALDSIND---EKLT
-----MTEPAQWKIDFVKNLENEINSRKVAAIVSIKGLRNNFQKIRNSIRDK-ARIKVSRRARLLRLAIENTGK---NNIV
1.....10.....20.....30.....40.....50.....60.....70.....80.....90

```

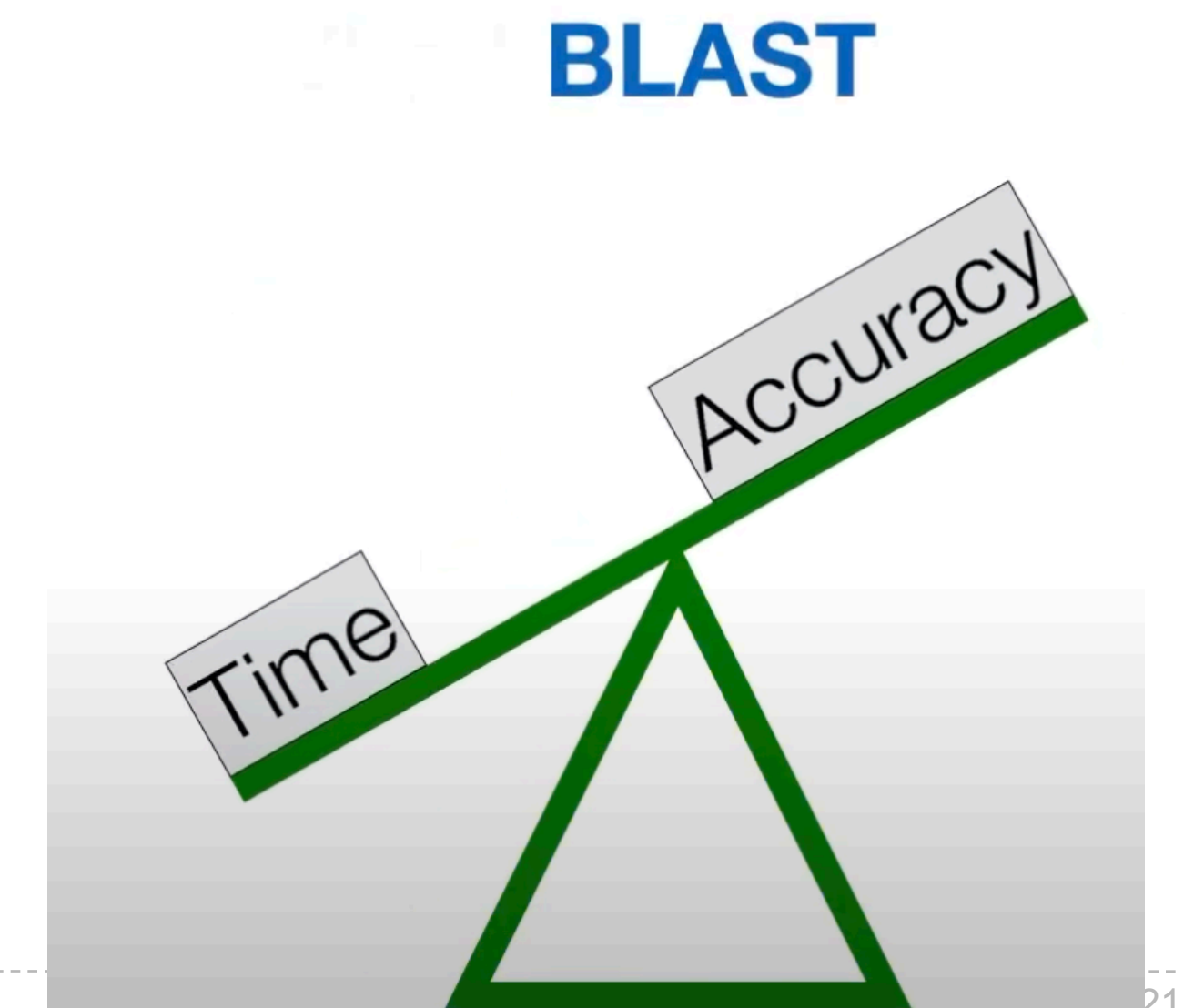
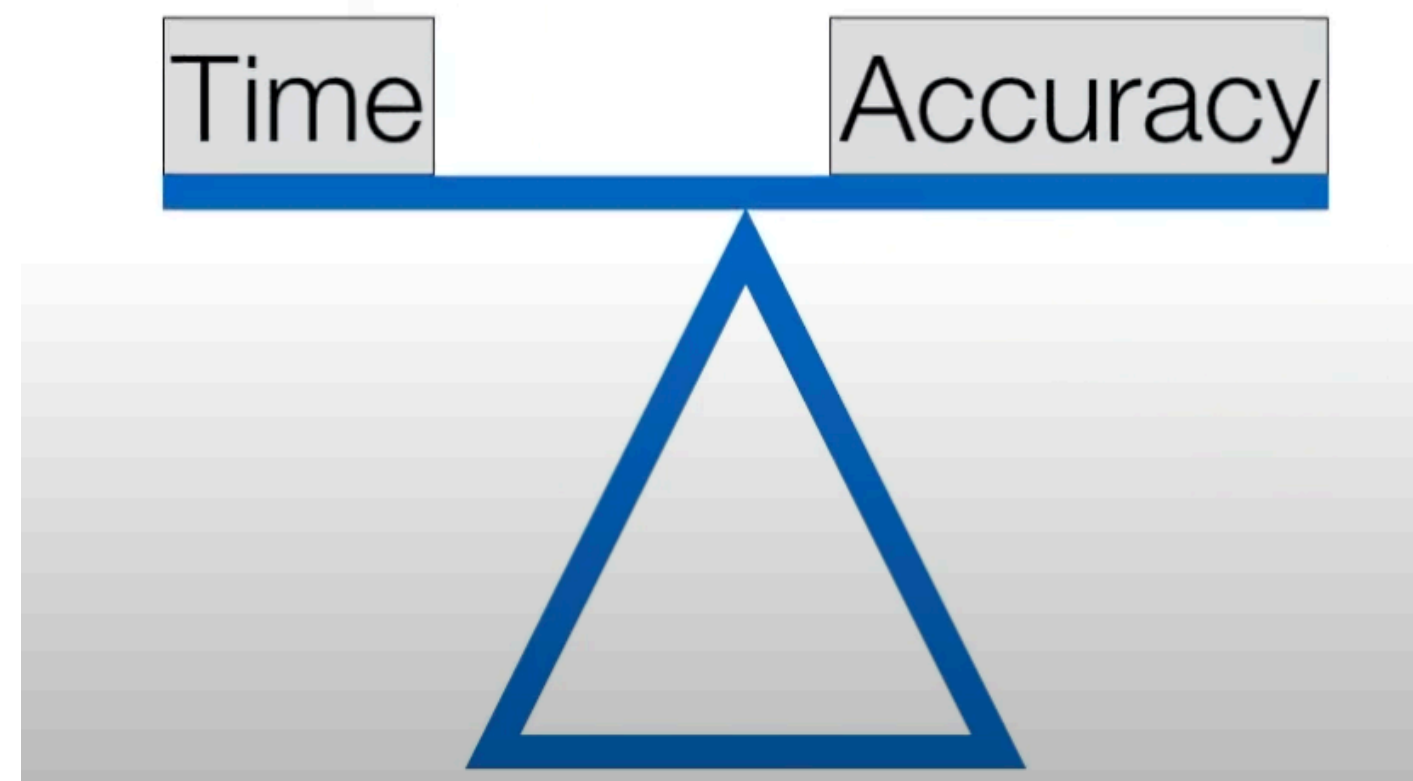
Query Sequence
(Human Insulin Gene)

MTEPAQWKIDFVKNLENEINSRKVAAIVSIKGI

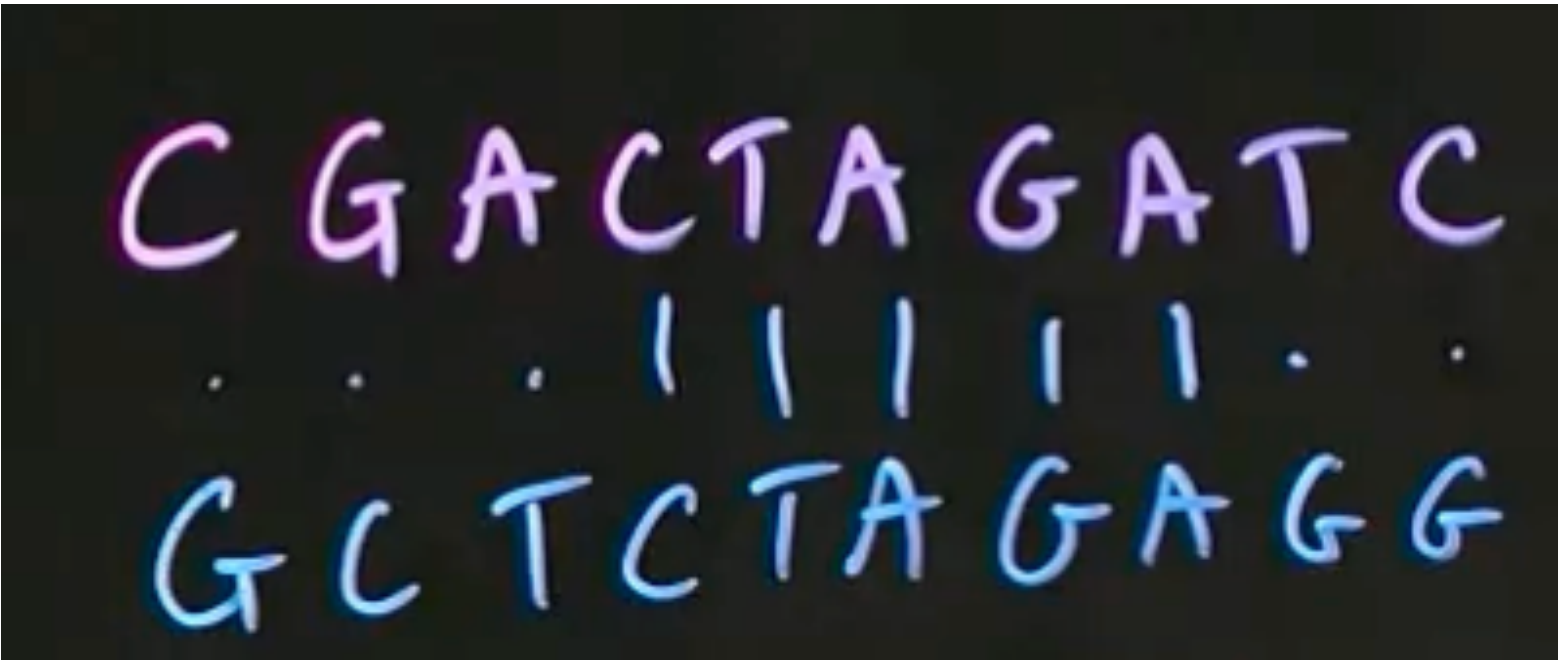
Break

BLAST Algorithm

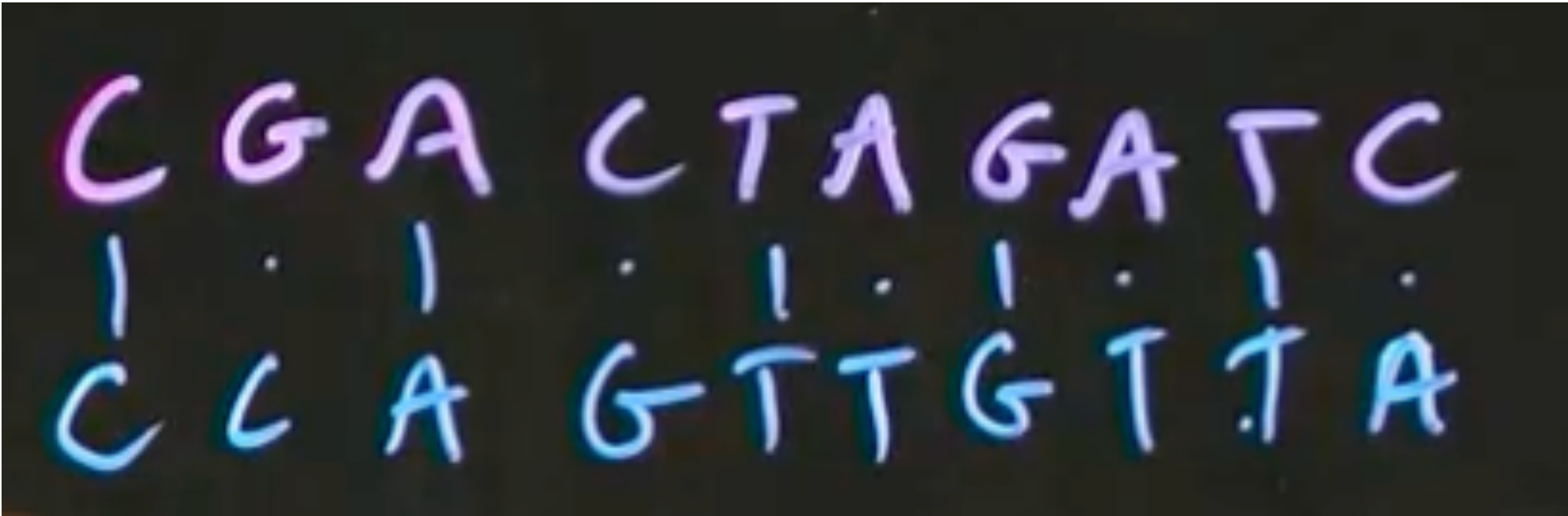
- BLAST uses a **seed and extend** algorithm
 - Scales with your query sequence and the size of the search database
 - <https://www.youtube.com/watch?v=jzSIC2UzxZ4>
- It is **heuristic**, based on trial and error and the process of elimination **NOT** precise mathematical formulations
 - BLAST does not look for exact matches because that would be computationally expensive
- Ctrl-F as a Tool for Scanning – is “BLAST” in “Composing Programs”
- BLAST – is something ~60–80% similar to “BLAST” in “Composing Programs”



Glance of the BLAST Algorithm



Query Sequence



Target Sequence in the Database

Query Sequence

GACAGC

Database Sequence

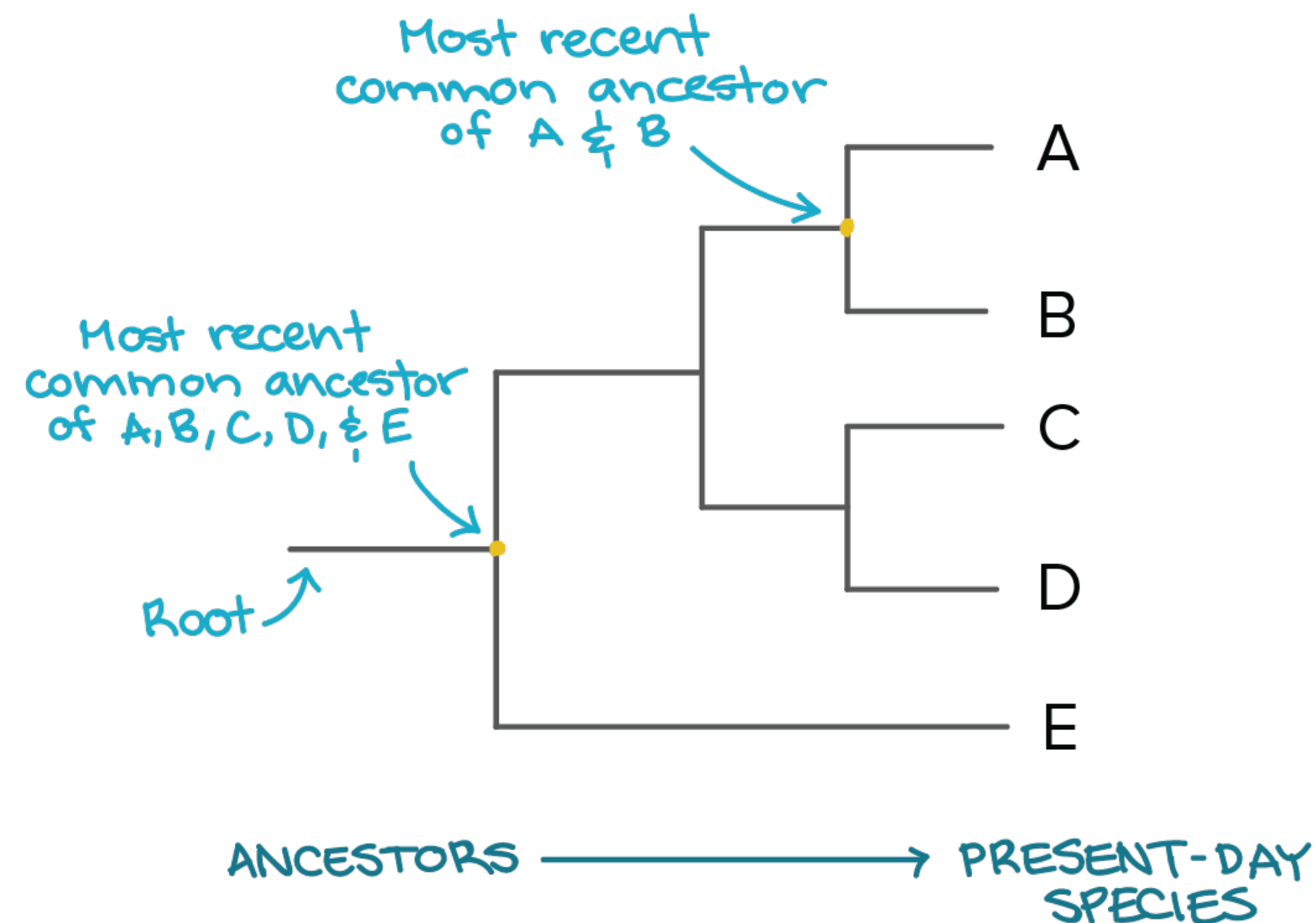
ACGGATTCCATAT

Scoring Scheme	
Match	1
Mismatch	-1
Gap Insertion	-1

		A	C	G	G	A	T	T	C	C	A	T	A	T
	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	0	1	1	0	0	0	0	0	0	0	0	0
A	0	1	0	0	0	2	1	0	0	0	1	0	1	0
C	0	0	2	1	0	1	1	0	1	1	0	0	0	0
A	0	1	1	1	0	1	0	0	0	0	2	1	1	0
G	0	0	0	2	2	1	0	0	0	0	1	1	0	0
C	0	0	1	1	1	1	0	0	1	1	0	0	0	0

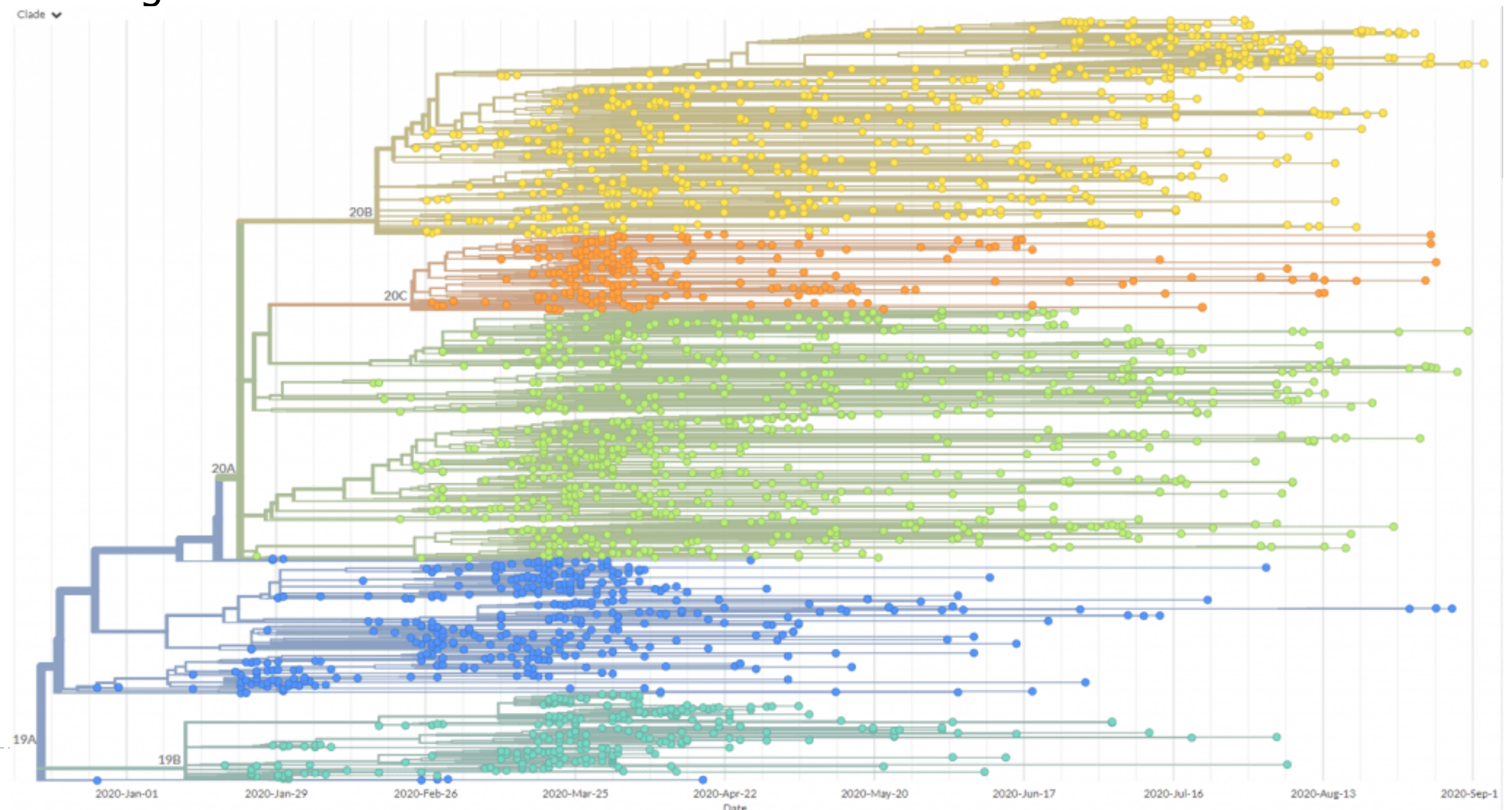
Understanding Evolution: Phylogeny

- How do we track the evolution of a virus? COVID-19 variants, for instance??
- Virus have a VERY HIGH rate of mutation
 - RNA viruses have high mutation rates—up to a **million times higher** than their hosts
- Through genomic analysis of virus samples, we can understand how the sequence of it changes over time
 - Phylogenetic trees allow us to visualize evolution



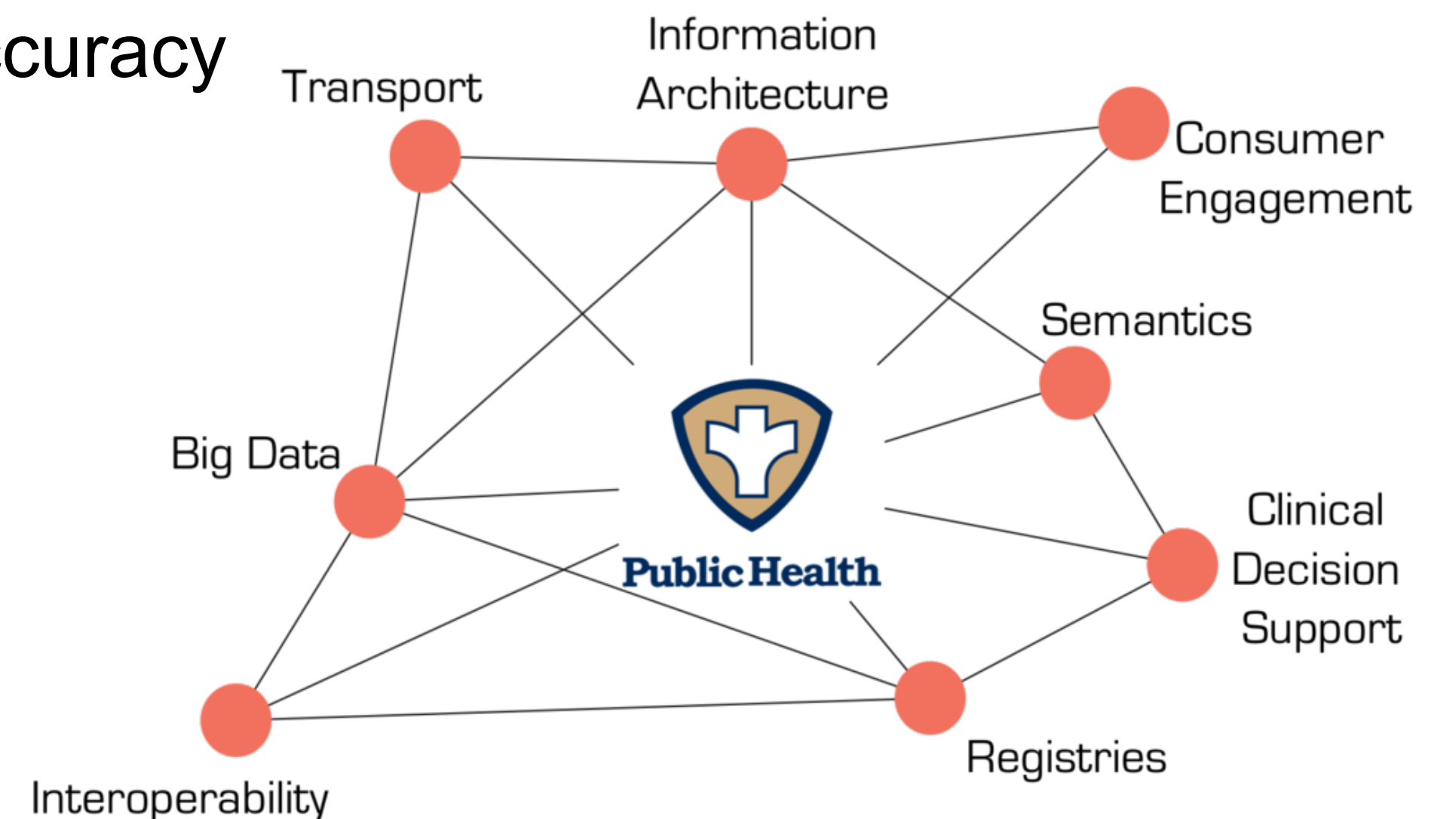
Phylogenetic Trees

- A branching diagram or tree showing the evolutionary relationship among various biological species
- Similarities and differences are based upon physical and genetic characteristics
- Two species are more related if they have a more recent common ancestor
- The root is the initial Wuhan SARS-CoV-2 genome



Public Health Informatics

- Capturing, managing and analyzing information to improve population-level health outcomes
- Transmit data to public health officials so they can better monitor and prevent disease
- Providers are already using AI algorithms to gain “unprecedented insights into diagnostics, care processes, treatment variability and patient outcomes”
 - 1 in 18 patients getting the wrong diagnosis in the ER department
 - According to the Society for the Improvement of Diagnosis in Medicine (SIDM) between 40,000 and 80,000 individuals die each year due to misdiagnoses
 - “Differential Diagnosis Tool” that had up to 96% diagnostic accuracy



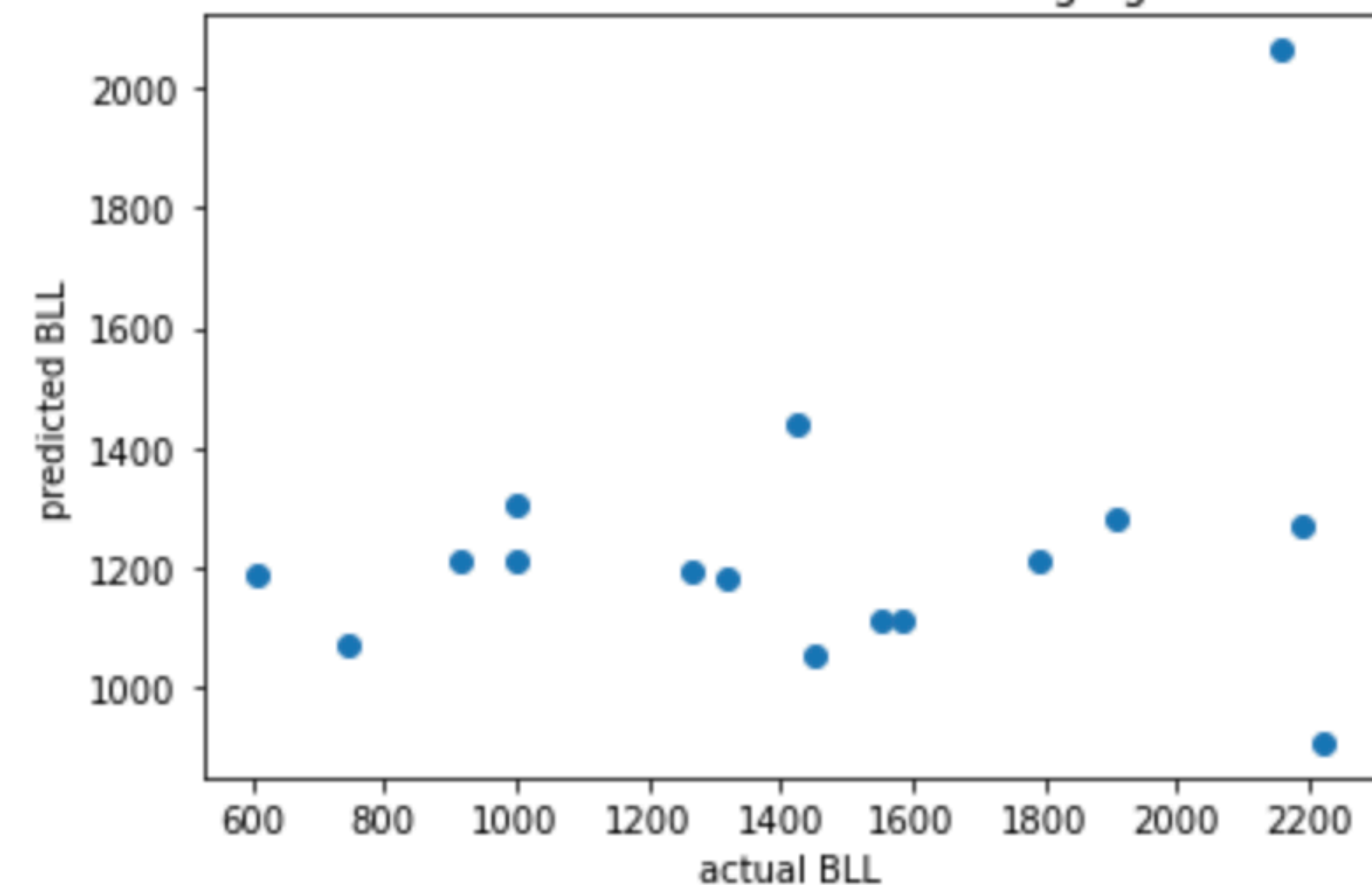
Lead Poisoning Research Project

- There is huge effort for prevention!
- Publicly available data on blood lead levels (BLL) from the Childhood Lead Poisoning Prevention Program (CLPPP)
- Why do some areas have more cases of lead poisoning than others?
 - Geographic, demographic, and socioeconomic factors!
 - For instance, I hypothesized there is a positive correlation between the number of severe cases (BLL > 4.5 μ g/dL) and house age due to likely use of lead paints

Modeling & Testing Hypothesis

- Geographic, demographic, and socioeconomic factors of a zip code can serve as reasonable features for a multiple regression model to predict number of cases in the future

BLL Prediction Based on Housing Ages



ZIP Code	Postal District Name	Number of BLLs > 4.5...	% of BLLs > 4.5 (0-6)	Total number of BLLs ...
95821	Sacramento	118	13.00%	908
95608	Carmichael	56	9.24%	606
94538	Fremont	39	4.76%	819
94087	Sunnyvale	22	4.53%	486
95051	Santa Clara	30	4.26%	705
94109	San Francisco	12	3.82%	314
94536	Fremont	29	3.61%	804
95670	Rancho Cordova	20	3.53%	566
90037	Los Angeles	47	3.24%	1450

Conclusion

- Bioinformatics is a fast-growing area with lots of exciting opportunities!
- **BIO ENG 145** Introduction to Machine Learning for Computational Biology
 - Using machine learning methods for genome-scale experimental data
- **BIO ENG 134** Genetic Design Automation
 - Use of software (lots of OOP) to design and manage genetics experiments
- **BIO ENG C131** Introduction to Computational Molecular and Cell Biology
 - Bioinformatics and Computational biology, with an emphasis on alignment, phylogeny, and ontologies
- Data Science Discovery Program for exposure working on these projects