CS 364 COMPUTATIONAL BIOLOGY

Sara Mathieson

Haverford College

Admin

EVERYONE:

- *Sign in*
- *Pick up a handout*
- *Pick up a notecard*
- *Pick up a construction paper sheet*

Outline

- Introductions
- Computational biology overview
- Syllabus highlights
- First algorithms: string search

Introductions

Notecard and Name card

- Instructor: Sara Mathieson
- TA: Ella Manning

- Notecard:
  - *Preferred first name*
  - *Pronouns (optional)*
  - *One topic you're hoping we'll cover in CS364*

    (be ready to share with the class)
  - *Anything else that would be helpful for me to know*

- Name card ("tent")
  - *Preferred first name*
  - *Pronouns (optional)*
  - *(sharpies going around!)*

With a partner briefly discuss…

1. **How long is the human genome in base pairs?**

   (A) 3 thousand
   (B) 3 million
   (C) 3 billion

2. **If I compare two human genomes, approximately how often will there be a difference?**

   (A) Every 10 bases
   (B) Every 100 bases
   (C) Every 1000 bases

3. **How much does it cost to sequence a human genome?**

   (A) $100
   (B) $1,000
   (C) $10,000

4. **When did humans and chimp last share a common ancestor?**

   (A) 1 thousand years ago
   (B) 1 million years ago
   (C) 10 million years ago

5. **How long has life on earth been evolving?**

With a partner briefly discuss…

1. How long is the human genome in base pairs?

   (A) 3 thousand (B) 3 million [✓ (C) 3 billion]

2. If I compare two human genomes, approximately how often will there be a difference?

   (A) Every 10 bases (B) Every 100 bases (C) Every 1000 bases

3. How much does it cost to sequence a human genome?

   (A) $100 (B) $1,000 (C) $10,000

4. When did humans and chimp last share a common ancestor?

   (A) 1 thousand years ago (B) 1 million years ago (C) 10 million years ago

5. How long has life on earth been evolving?

With a partner briefly discuss…

1. How long is the human genome in base pairs?

   (A) 3 thousand (B) 3 million [✓ (C) 3 billion]

2. If I compare two human genomes, approximately how often will there be a difference?

   (A) Every 10 bases (B) Every 100 bases [✓ (C) Every 1000 bases]

3. How much does it cost to sequence a human genome?

   (A) $100 (B) $1,000 (C) $10,000

4. When did humans and chimp last share a common ancestor?

   (A) 1 thousand years ago (B) 1 million years ago (C) 10 million years ago

5. How long has life on earth been evolving?

With a partner briefly discuss…

1. How long is the human genome in base pairs?

   (A) 3 thousand (B) 3 million [✓ (C) 3 billion]

2. If I compare two human genomes, approximately how often will there be a difference?

   (A) Every 10 bases (B) Every 100 bases [✓ (C) Every 1000 bases]

3. How much does it cost to sequence a human genome?

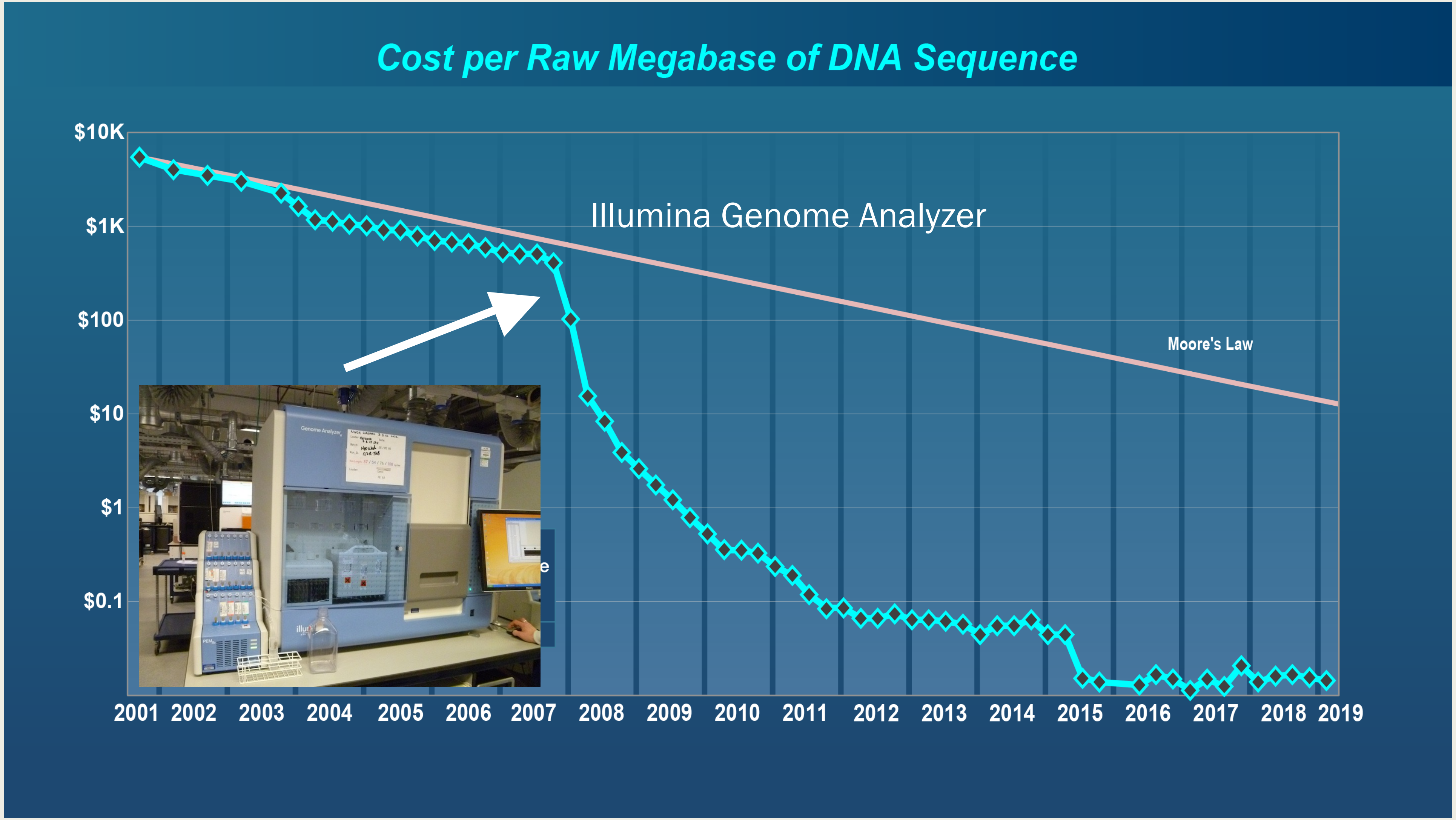
   [✓ (A) $100] ← important variants [✓ (B) $1,000] ← full sequence (C) $10,000

4. When did humans and chimp last share a common ancestor?

   (A) 1 thousand years ago (B) 1 million years ago (C) 10 million years ago

5. How long has life on earth been evolving?

Next Generation sequencing



With a partner briefly discuss…

1. How long is the human genome in base pairs?

   (A) 3 thousand   (B) 3 million   [✓ (C) 3 billion]

2. If I compare two human genomes, approximately how often will there be a difference?

   (A) Every 10 bases   (B) Every 100 bases   [✓ (C) Every 1000 bases]

3. How much does it cost to sequence a human genome?

   [✓ (A) $100] ← important variants   [✓ (B) $1,000] ← full sequence   (C) $10,000

4. When did humans and chimp last share a common ancestor?

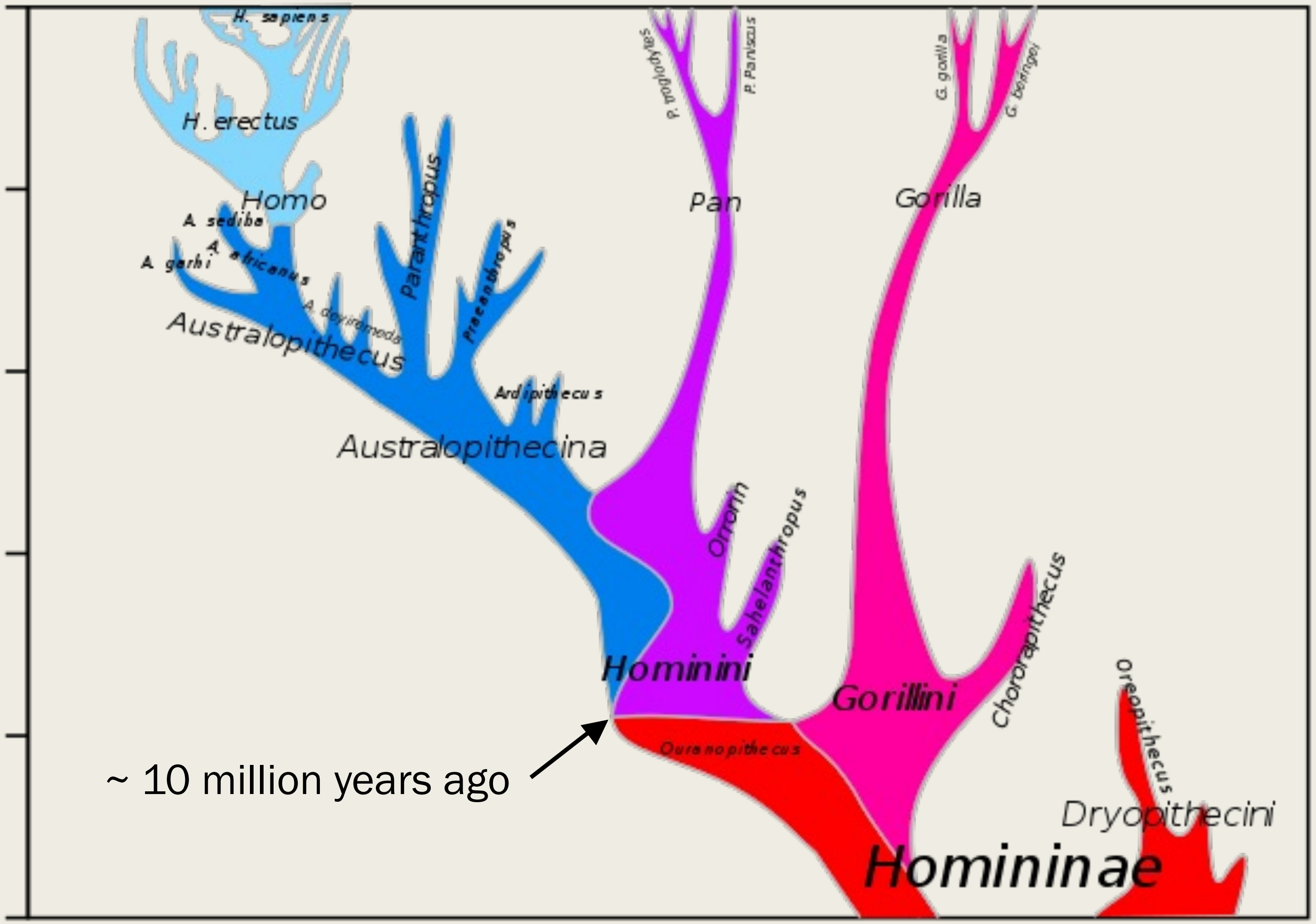
   (A) 1 thousand years ago   (B) 1 million years ago   [✓ (C) 10 million years ago]

5. How long has life on earth been evolving?

Human-chimp divergence

Human

Chimp



~ 10 million years ago

Image: modified from wikipedia

With a partner briefly discuss…

1. How long is the human genome in base pairs?

   (A) 3 thousand (B) 3 million [✓ (C) 3 billion]

2. If I compare two human genomes, approximately how often will there be a difference?

   (A) Every 10 bases (B) Every 100 bases [✓ (C) Every 1000 bases]

3. How much does it cost to sequence a human genome?

   [✓ (A) $100] ← important variants [✓ (B) $1,000] ← full sequence (C) $10,000

4. When did humans and chimp last share a common ancestor?

   (A) 1 thousand years ago (B) 1 million years ago [✓ (C) 10 million years ago]

5. How long has life on earth been evolving? ~ 4 billion years

Earliest life on earth

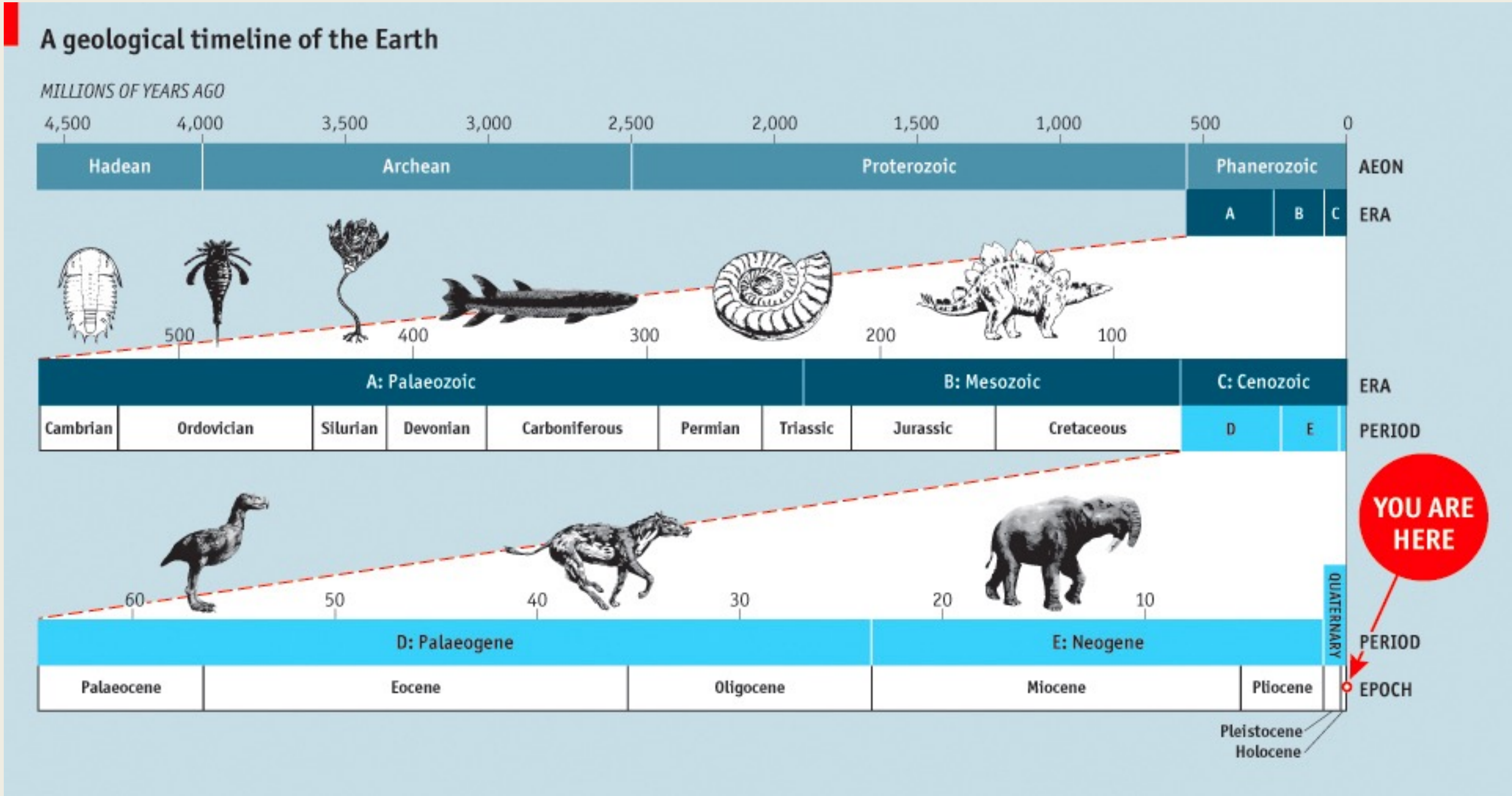


Image: *The Economist*

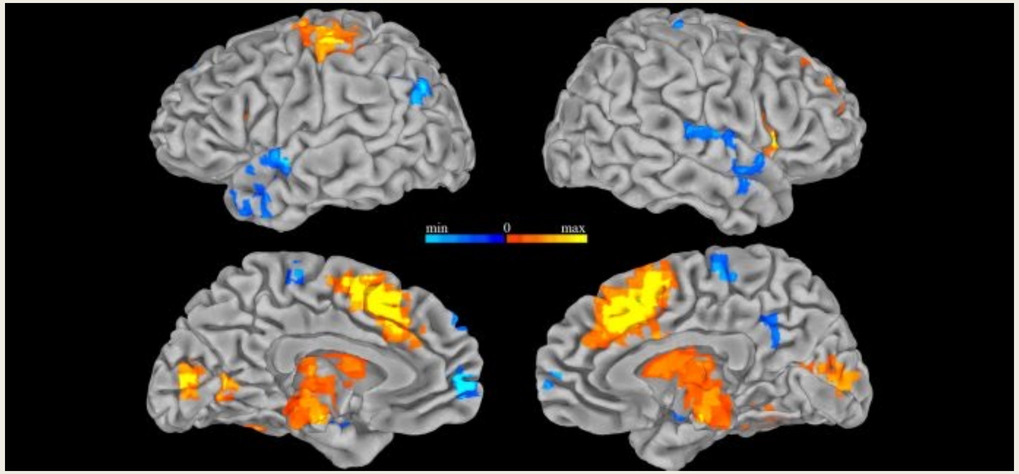
Computational Biology Overview

Why take Computational Biology?

- In the last 25 years, genome sequencing costs have plummeted and as a result, we have amazing “big data”
- We have data from tens of thousands of species and hundreds of thousands of individuals from our species
- We are now in a position to answer biological questions with this data, but algorithms for analyzing and learning from this data have not developed at the same pace
- CS364 is an opportunity to learn how biological data has driven algorithm development, and how existing algorithms have been repurposed for biology
- We will also discuss the future of computational biology, genomic privacy and ethics, and challenging problems that remain unsolved

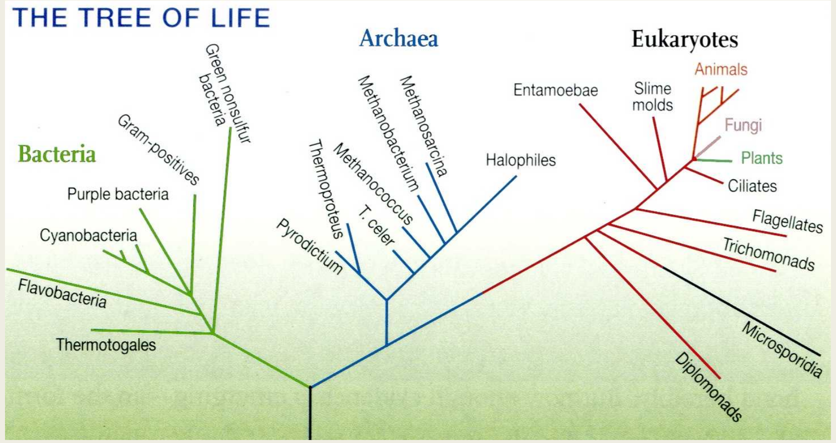
Computational Biology Problems

Discovering genetic variants that increases disease risk



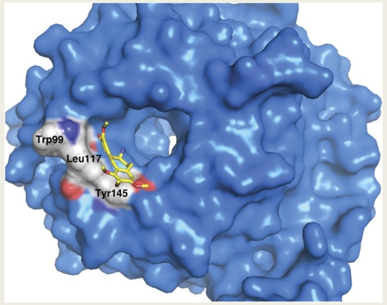
Finding out which parts of the brain are involved in playing and appreciating music

Understanding how climate change will affect forests



Explaining the origin of life

Images: NIH; PLOS Blogs; Science; NJSAS



Biological processes

Experimental Data

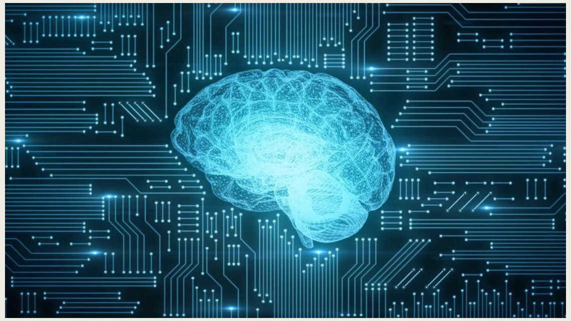
Algorithms

Sequence genomes

Base calling
Error correction
Genome assembly
Read mapping

$$

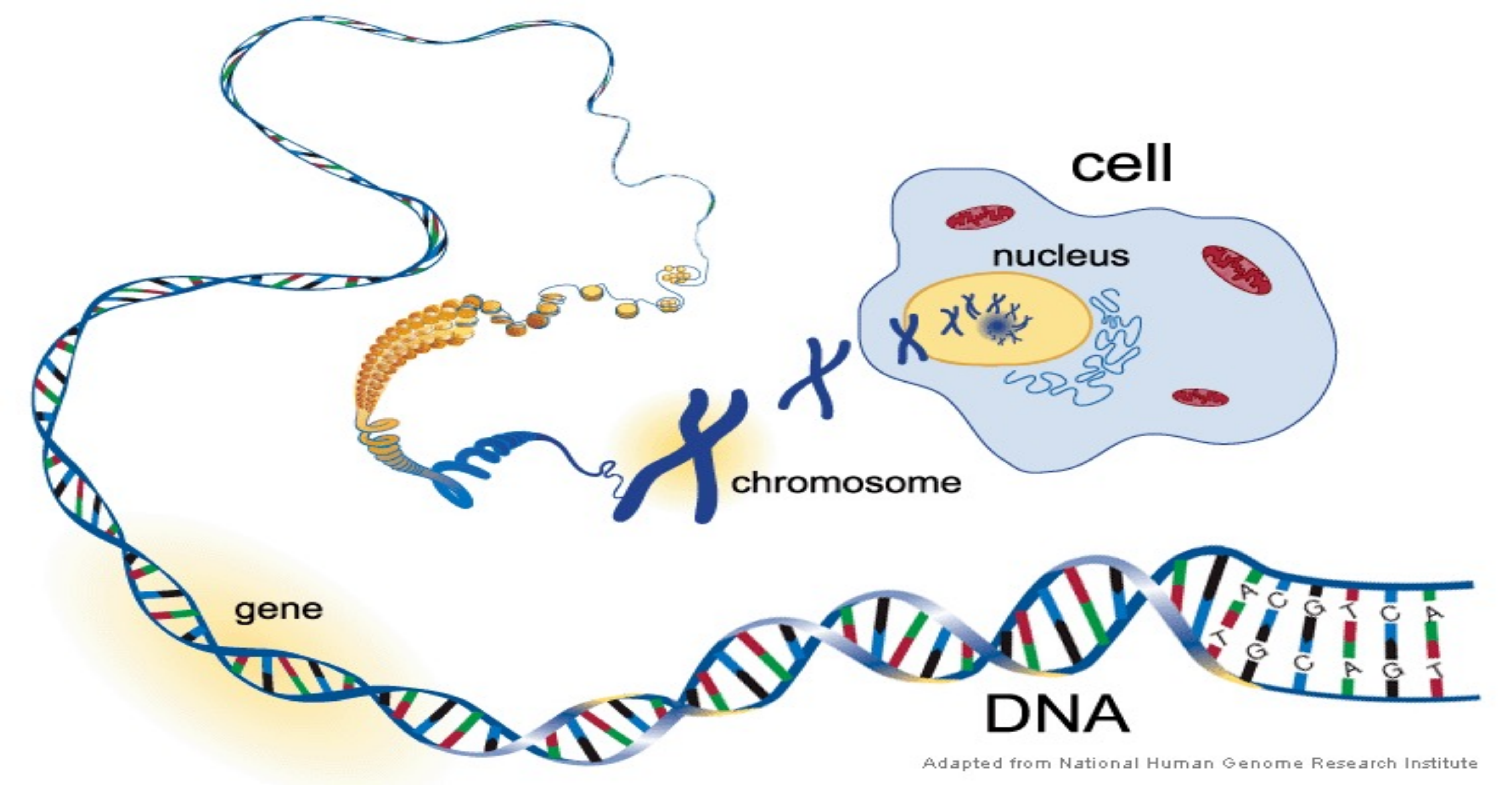
Computational Thinking



How to solve them?

Images: BBC, Illumina

What data are we sequencing?



Adapted from National Human Genome Research Institute

Obtaining the data

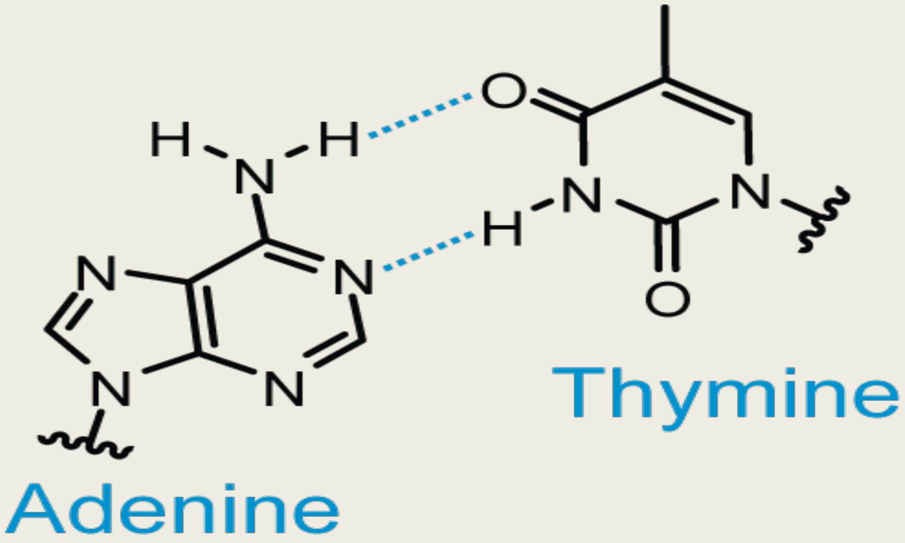
Illumina

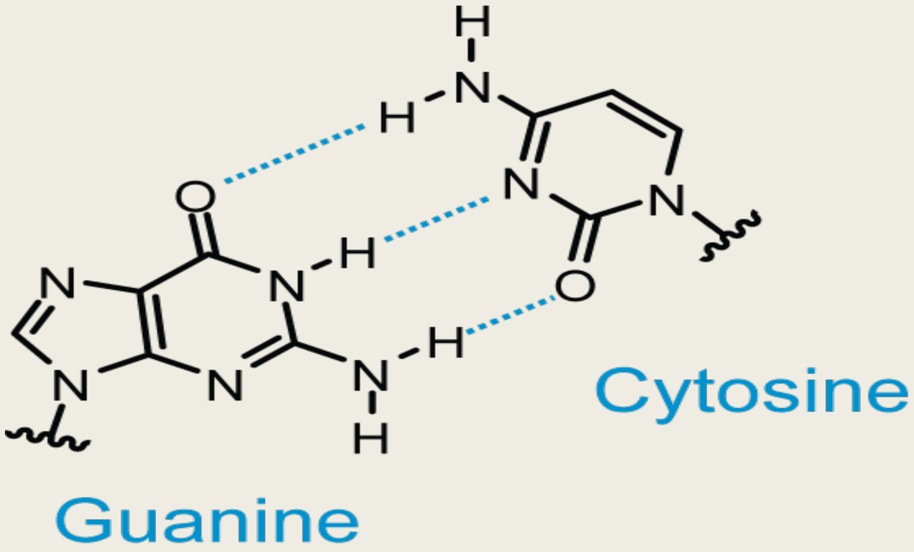
```
AGCCCTAATCTAACCCTGGCCAACCTGTCTCTCAACTTACCCTCCATTACCCTGCCTCCACTCGTTACCCTGTCCCATTCAAC
CATACCACTCCGAACCACCATCCATCCCTCTACTTACTACCACTCACCCACCGTTACCCTCCAATTACCCATATCCAACCCACT
GCCACTTACCCTACCATTACCCTACCATCCACCATGACCTACTCACCATACTGTTCTTCTACCCACCATATTGAAACGCTAACA
AATGATCGTAAATAACACACACGTGCTTACCCTACCACTTTATACCACCACCACATGCCATACTCACCCTCACTTGTATACTGA
TTTTACGTACGCACACGGATGCTACAGTATATACCATCTCAAACTTACCCTACTCTCAGATTCCACTTCACTCCATGGCCCATC
TCTCACTGAATCAGTACCAAATGCACTCACATCATTATGCACGGCACTTGCCTCAGCGGTCTATACCCTGTGCCATTTACCCA
TAACGCCCATCATTATCCACATTTTGATATCTATATCTCATTCGGCGGTCCCAAATATTGTATAACTGCCCTTAATACATACGTT
ATACCACTTTTGCACCATATACTTACCACTCCATTTATATACACTTATGTCAATATTACAGAAAAATCCCCACAAAAATCACCTA
AACATAAAAATATTCTACTTTTCAACAATAATACATAAACATATTGGCTTGTGGTAGCAACACTATCATGGTATCACTAACGTAA
AAGTTCCTCAATATTGCAATTTGCTTGAACGGATGCTATTTCAGAATATTTCGTACTTACACAGGCCATACATTAGAATAATAT
GTCACATCACTGTCGTAACACTCTTTATTCACCGAGCAATAATACGGTAGTGGCTCAAACTCATGCGGGTGCTATGATACAAT
TATATCTTATTTCCATTCCCATATGCTAACCGCAATATCCTAAAAGCATAACTGATGCATCTTTAATCTTGTATGTGACACTACT
CATACGAAGGGACTATATCTAGTCAAGACGATACTGTGATAGGTACGTTATTTAATAGGATCTATAACGAAATGTCAAATAATT
TTACGGTAATATAACTTATCAGCGGCGTATACTAAAACGGACGTTACGATATTGTCTCACTTCATCTTACCACCCTCTATCTTAT
TGCTGATAGAACACT GATATTTT
ACGTGTCAAAAAATC ATCTTGTT
CTTAGAAGTGACGCA TTATTTAA
GGACAAAGGTTGCGAAGCCGCACATTTCCAATTTCATTGTTGTTTATTGGACATACACTGTTAGCTTTATTACCGTCCACGTT
TTTTCTACAATAGTGTAGAAGTTTCTTTCTTATGTTCATCGTATTCATAAAATGCTTCACGAACACCGTCATTGATCAAATAGG
TCTATAATATTAATATACATTTATATAATCTACGGTATTTATATCATCAAAAAAAAGTAGTTTTTTATTTTATTTTGTTCGTTAAT
TTTCAATTTCTATGGAAACCCGTTCGTAAAATTGGCGTTTGTCTCTAGTTTGCGATAGTGTAGATACCGTCCTTGGATAGAGC
ACTGGAGATGGCTGGCTTTAATCTGCTGGAGTACCATGGAACACCGGTGATCATTCTGGTCACTTGGTCTGGAGCAATACCG
GTCAACATGGTGGTGAAGTCACCGTAGTTGAAAACGGCTTCAGCAACTTCGACTGGGTAGGTTTCAGTTGGGTGGGCGGCT
TGGAACATGTAGTATTGGGCTAAGTGAGCTCTGATATCAGAGACGTAGACACCCAATTCCACCAAGTTGACTCTTTCGTCAG
ATTGAGCTAGAGTGGTGGTTGCAGAAGCAGTAGCAGCGATGGCAGCGACACCAGCGGCGATTGAAGTTAATTTGACCATTG
TATTTGTTTTGTTTGTTAGTGCTGATATAAGCTTAACAGGAAAGGAAAGAATAAAGACATATTCTCAAAGGCATATAGTTGAAG
CAGCTCTATTTATACCCATTCCCTCATGGGTTGTTGCTATTTAAACGATCGCTGACTGGCACCAGTTCCTCATCAAATATTCTC
TATATCTCATCTTTCACACAATCTCATTATCTCTATGGAGATGCTCTTGTTTCTGAACGAATCATAAATCTTTCATAGGTTTCGT
ATGTGGAGTACTGTTTTATGGCGCTTATGTGTATTCGTATGCGCAGAATGTGGGAATGCCAATTATAGGGGTGCCGAGGTGC
CTTATAAAACCCTTTTCTGTGCCTGTGACATTTCCTTTTTCGGTCAAAAAGAATATCCGAATTTTAGATTTGGACCCTCGTACA
GAAGCTTATTGTCTAAGCCTGAATTCAGTCTGCTTTAAACGGCTTCCGCGGAGGAAATATTTCCATCTCTTGAATTCGTACAA
CATTAAACGTGTGTTGGGAGTCGTATACTGTTAGGGTCTGTAAACTTGTGAACTCTCGGCAAATGCCTTGGTGCAATTACGT
AATTTAGCCGCTGAGAAGCGGATGGTAATGAGACAAGTTGATATCAAACAGATACATATTTAAAAGAGGGTACCGCTAATTT
AGCAGGGCAGTATTATTGTAGTTTGATATGTACGGCTAACTGAACCTAAGTAGGGATATGAGAGTAAGAACGTTCGGCTACTC
TTCTTTCTAAGTGGGATTTTTCTTAATCCTTGGATTCTTAAAAGGTTATTAAAGTTCCGCACAAAGAACGCTTGGAAATCGCA
```

We obtain a “string” of *bases* (A,C,G,T)

Base-pairing

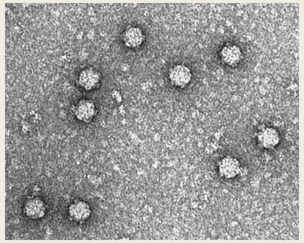
- “A” with “T”
- “G” with “C”
- Humans: 3 billion base pairs (bp)





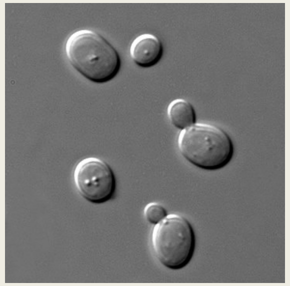
Images: MSOE Center for BioMolecular Modeling

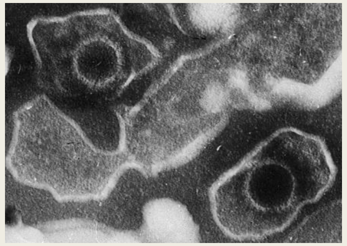
Early Genome Sequences



ΦX174 1977 (5 kb)

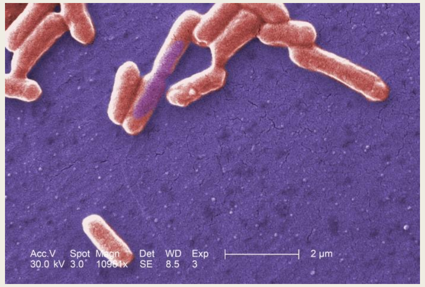
S. *cerevisiae* 1996 (12 Mb)

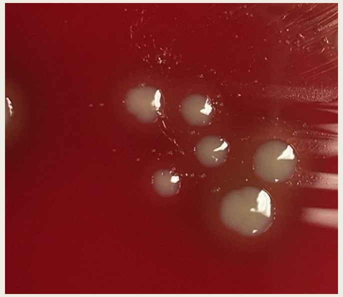




Epstein-Barr virus 1984 (170kb)

E. coli 1997 (4.6 Mb)



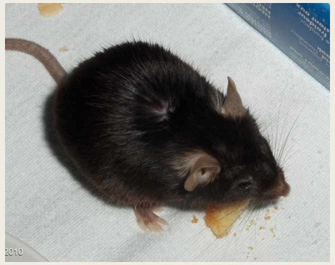


H. influenzae 1995 (1.8Mb)

Early Eukaryotic Sequences

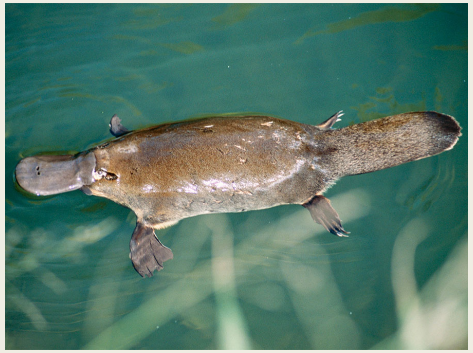
A. thaliana 2001 (135 Mb)

Brown Rat 2004
(2.8 Gb)



Mouse 2002
(2.6 Gb)

Platypus 2007
(2.3 Gb)



Chimpanzee 2005
(3 Gb)

“Clint”

And many other species…

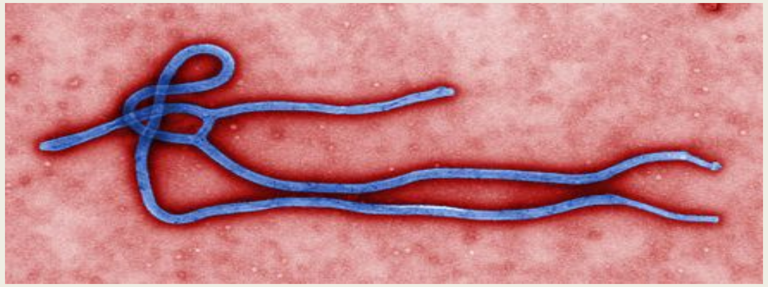
Chimp

Melitaea cinxia

Buffalo

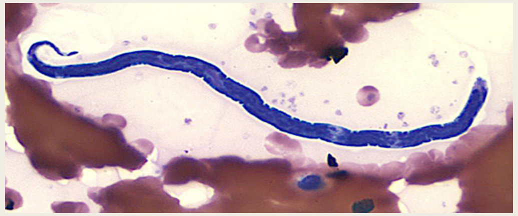
Images: wikipedia

Chinese liver worm



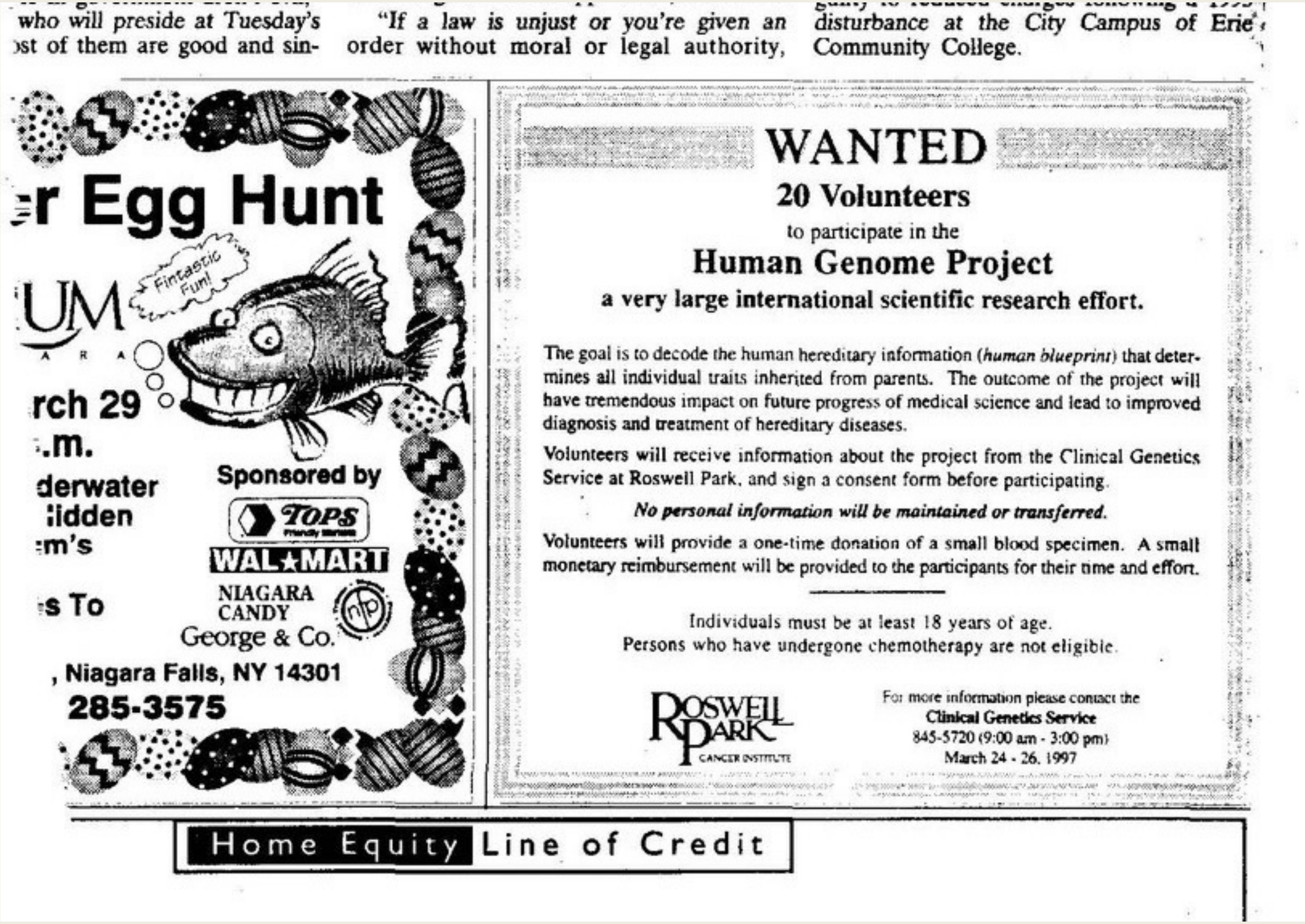
Ebola

Maize



Loa loa (eye worm)

Human genome project



who will preside at Tuesday's
ost of them are good and sin-

"If a law is unjust or you're given an order without moral or legal authority,

disturbance at the City Campus of Erie Community College.

r Egg Hunt

Fintastic Fun!

UM
A R A

rch 29
.m.

derwater
idden
m's

s To

Sponsored by
TOPS
Friendly Markets
WAL★MART
NIAGARA
CANDY
George & Co.

, Niagara Falls, NY 14301
285-3575

WANTED

20 Volunteers

to participate in the

Human Genome Project

a very large international scientific research effort.

The goal is to decode the human hereditary information (*human blueprint*) that determines all individual traits inherited from parents. The outcome of the project will have tremendous impact on future progress of medical science and lead to improved diagnosis and treatment of hereditary diseases.

Volunteers will receive information about the project from the Clinical Genetics Service at Roswell Park, and sign a consent form before participating.

No personal information will be maintained or transferred.

Volunteers will provide a one-time donation of a small blood specimen. A small monetary reimbursement will be provided to the participants for their time and effort.

Individuals must be at least 18 years of age.
Persons who have undergone chemotherapy are not eligible.

ROSWELL PARK
CANCER INSTITUTE

For more information please contact the
Clinical Genetics Service
845-5720 (9:00 am - 3:00 pm)
March 24 - 26, 1997

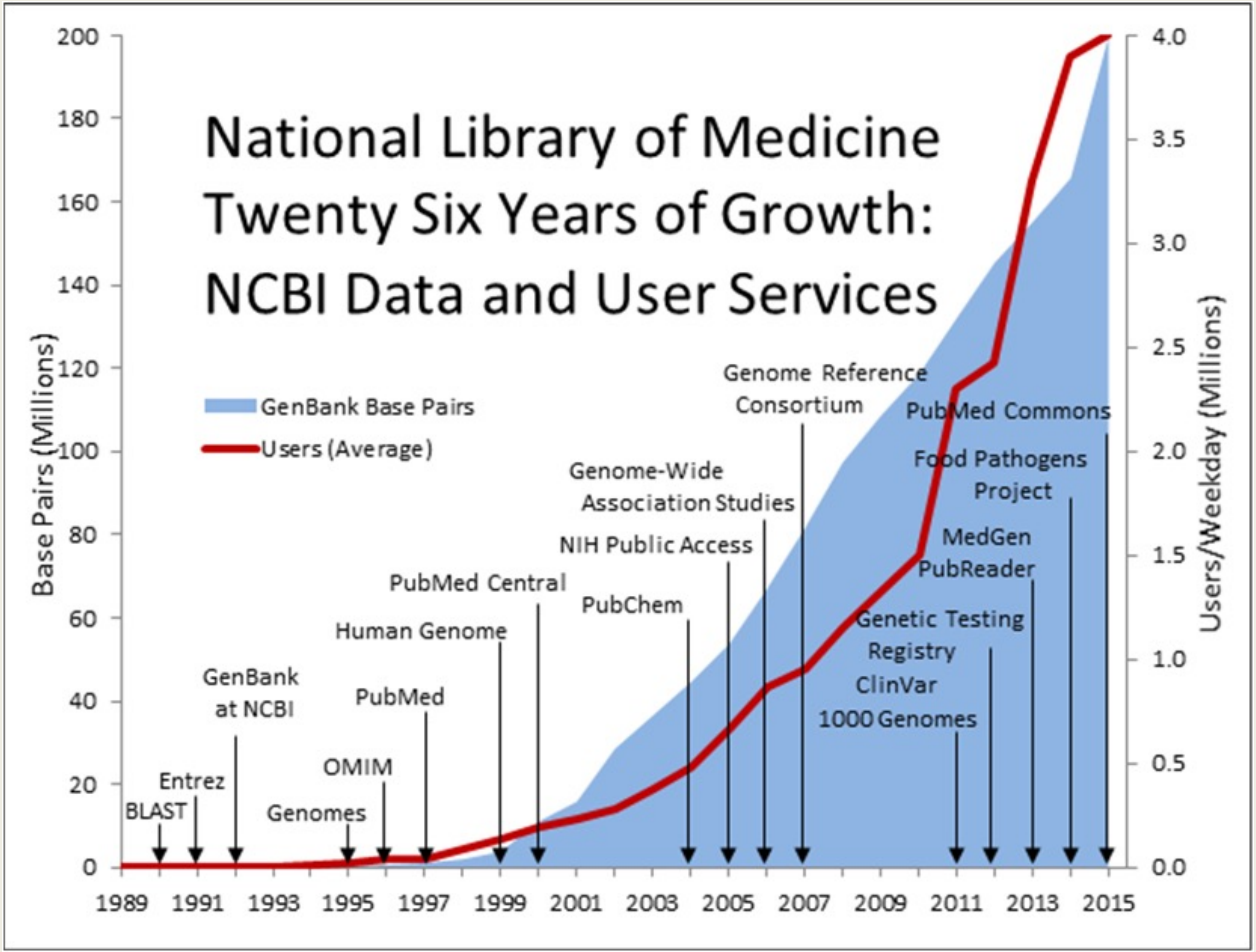
Home Equity Line of Credit

Buffalo News, 3/23/1997

Human genome project

nature
the human genome
Reprinted from Nature, Vol. 409, No. 6822, pp. 860-921, 15 February 2001
© Nature Publishing Group, 2001
16 February 2001
Science
Vol. 291 No. 5507
Pages 1145–1434 $9
THE HUMAN GENOME
American Association for the Advancement of Science

Draft sequence, 2001. Cost: 3 Billion dollars (around $1 per base…)



National Library of Medicine Twenty Six Years of Growth: NCBI Data and User Services

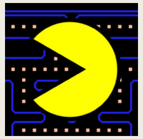
First goal of this course:

Algorithms to analyze biological data

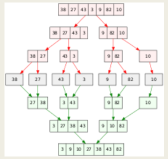
* How can we assemble genomes from short reads?

* How can we infer the relationships between species based on genomes?

Types of algorithms



Greedy – at each stage of the algorithm, improve the solution as much as we can



Divide-and-Conquer – break the problem up into smaller parts them combine the solutions

Brute-force – write out all the possible solutions and pick the best one

Randomized – include some randomness in the solution

Recursive – solve a series of smaller problems first

Biology is computation

DNA is a means of **storing and transmitting information**

Turning that information into biological objects (e.g. you and me) is a **computational process**

So much of biology, much of what is going on inside your body, is **computation**

Second goal of this course:

Using biology as inspiration

* Can we make a computer think like a brain?

* Can we learn how to design algorithms from evolution?

Syllabus (highlights)

Course Goals

- Handle “real-world” data sets (large and noisy)
- Connect core bioinformatics algorithms to CS
- Understand, implement, and apply core bioinformatics algorithms
- Learn to model uncertainty using probability
- Communicate ideas effectively
- Understand the scientific method (asking a biological question, forming a hypothesis, designing a computational experiment, implementing and applying algorithms, iterating the process, drawing conclusions and communicating the results)
- Develop an appreciation for questions that require interdisciplinary skills to answer

Topics (tentative)

- String search
- Read mapping and Burrows-Wheeler
- Genome assembly
- Sequence alignment (dynamic programming)
- Phylogenetic tree algorithms (clustering)
- Ancestral reconstruction
- Population genetics and sequence diversity
- Hidden Markov models (HMMs)
- Deep learning in biology
- Cancer genomics
- RNA folding and non-linear structures
- Genomic privacy and ethics

Prerequisites

- No biology prerequisite
- CS260 Foundations of Data Science
- (helpful but not required) Linear Algebra

Course Components

- Labs (8 total): 35%
- Midterms (2 in-class): 40% (20% each)
- Final project: 15% (includes an oral presentation and “lab notebook”)
- Participation: 10%

My expectations

- Come to class (Tu/Th) and lab (Th), and actively participate during both
- Complete the weekly reading *before* lab
- Come to office hours (tentatively Mondays 4-5pm)
- Post questions on Piazza

Syllabus highlights

(Note: you are responsible for reading the entire syllabus on the course webpage)

- Notes and slides will be posted *after* class on the course webpage
- Durbin *et al* book is required, but could be shared or borrowed from the library
- Class and lab attendance will be taken every day and absences will quickly affect your participation grade
- You will get <u>**2 late days**</u> during the semester (counts for both partners if pair lab)
- Extensions beyond these two days must be arranged with your class dean
- Email: allow 24 hours for a response
- Piazza: should be used for all content/logistics questions

First algorithms: string search

String Terminology

Symbols/characters: Fundamental units
e.g., {A,C,T,G} for DNA

Alphabets: Finite set of symbols
e.g., {A,C,T,G}, {A-Z, 0-9, “,”, “.”, “ “, “-”}

Words/strings: sequence of symbols from an alphabet set:
e.g., AAC, “This is a string”

k-tuple, k-mer: An ordered set of k symbols

Substring:
A contiguous subset of symbols

```
S   = “ACGTACGTA”
S′  = “ACGT”
S′′ =      “GTACG”
```

Prefixes and suffixes note: in class we'll start from 1 but in Python start from 0 (sorry!)

S = ACGACGCGAC

Prefixes, S[1..i]	Suffixes, S[i..|S|]
A:1	**ACGACGCGAC:1**
AC:2	**CGACGCGAC:2**
ACG:3	**GACGCGAC:3**
ACGA:4	**ACGCGAC:4**
ACGAC:5	**CGCGAC:5**
ACGACG:6	**GCGAC:6**
ACGACGC:7	**CGAC:7**
ACGACGCG:8	**GAC:8**
ACGACGCGA:9	**AC:9**
ACGACGCGAC:10	**C:10**

First problem: Exact search

Find the string “whale”

I stuffed a shirt or two into my old carpet-bag, tucked it under my arm, and started for Cape Horn and the Pacific. Quitting the good city of old Manhatto, I duly arrived in New Bedford. It was a Saturday night in December. Much was I disappointed upon learning that the little packet for Nantucket had already sailed, and that no way of reaching that place would offer, till the following Monday. As most young candidates for the pains and penalties of whaling stop at this same New Bedford, thence to embark on their voyage, it may as well be related that I, for one, had no idea of so doing. For my mind was made up to sail in no other than a Nantucket craft, because there was a fine, boisterous something about everything connected with that famous old island, which amazingly pleased me. Besides though New Bedford has of late been gradually monopolising the business of whaling, and though in this matter poor old Nantucket is now much behind her, yet Nantucket was her great original—the Tyre of this Carthage;—the place where the first dead American whale was stranded. Where else but from Nantucket did those aboriginal whalemen, the Red-Men, first sally out in canoes to give chase to the Leviathan? And where but from Nantucket, too, did that first adventurous little sloop put forth, partly laden with imported cobblestones—so goes the story—to throw at the whales, in order to discover when they were nigh enough to risk a harpoon from the bowsprit?

First problem: Exact search

Find the string “whale”

I stuffed a shirt or two into my old carpet-bag, tucked it under my arm, and started for Cape Horn and the Pacific. Quitting the good city of old Manhatto, I duly arrived in New Bedford. It was a Saturday night in December. Much was I disappointed upon learning that the little packet for Nantucket had already sailed, and that no way of reaching that place would offer, till the following Monday. As most young candidates for the pains and penalties of whaling stop at this same New Bedford, thence to embark on their voyage, it may as well be related that I, for one, had no idea of so doing. For my mind was made up to sail in no other than a Nantucket craft, because there was a fine, boisterous something about everything connected with that famous old island, which amazingly pleased me. Besides though New Bedford has of late been gradually monopolising the business of whaling, and though in this matter poor old Nantucket is now much behind her, yet Nantucket was her great original—the Tyre of this Carthage;—the place where the first dead American **whale** was stranded. Where else but from Nantucket did those aboriginal **whale**men, the Red-Men, first sally out in canoes to give chase to the Leviathan? And where but from Nantucket, too, did that first adventurous little sloop put forth, partly laden with imported cobblestones—so goes the story—to throw at the **whale**s, in order to discover when they were nigh enough to risk a harpoon from the bowsprit?

Second problem: Inexact search

Find a string that looks like “whale”

I stuffed a shirt or two into my old carpet-bag, tucked it under my arm, and started for Cape Horn and the Pacific. Quitting the good city of old Manhatto, I duly arrived in New Bedford. It was a Saturday night in December. Much was I disappointed upon learning that the little packet for Nantucket had already sailed, and that no way of reaching that place would offer, till the following Monday. As most young candidates for the pains and penalties of **whaling** stop at this same New Bedford, thence to embark on their voyage, it may as well be related that I, for one, had no idea of so doing. For my mind was made up to sail in no other than a Nantucket craft, because there was a fine, boisterous something about everything connected with that famous old island, which amazingly pleased me. Besides though New Bedford has of late been gradually monopolising the business of **whaling**, and though in this matter poor old Nantucket is now much behind her, yet Nantucket was her great original—the Tyre of this Carthage;—the place where the first dead American **whale** was stranded. Where else but from Nantucket did those aboriginal **whale**men, the Red-Men, first sally out in canoes to give chase to the Leviathan? And where but from Nantucket, too, did that first adventurous little sloop put forth, partly laden with imported cobblestones—so goes the story—to throw at the **whale**s, in order to discover when they were nigh enough to risk a harpoon from the bowsprit?

Third problem: Context search

Find a string that means “whale”

I stuffed a shirt or two into my old carpet-bag, tucked it under my arm, and started for Cape Horn and the Pacific. Quitting the good city of old Manhatto, I duly arrived in New Bedford. It was a Saturday night in December. Much was I disappointed upon learning that the little packet for Nantucket had already sailed, and that no way of reaching that place would offer, till the following Monday. As most young candidates for the pains and penalties of **whaling** stop at this same New Bedford, thence to embark on their voyage, it may as well be related that I, for one, had no idea of so doing. For my mind was made up to sail in no other than a Nantucket craft, because there was a fine, boisterous something about everything connected with that famous old island, which amazingly pleased me. Besides though New Bedford has of late been gradually monopolising the business of **whaling**, and though in this matter poor old Nantucket is now much behind her, yet Nantucket was her great original—the Tyre of this Carthage;—the place where the first dead American **whale** was stranded. Where else but from Nantucket did those aboriginal **whale**men, the Red-Men, first sally out in canoes to give chase to the **Leviathan**? And where but from Nantucket, too, did that first adventurous little sloop put forth, partly laden with imported cobblestones—so goes the story—to throw at the **whale**s, in order to discover when they were nigh enough to risk a harpoon from the bowsprit?

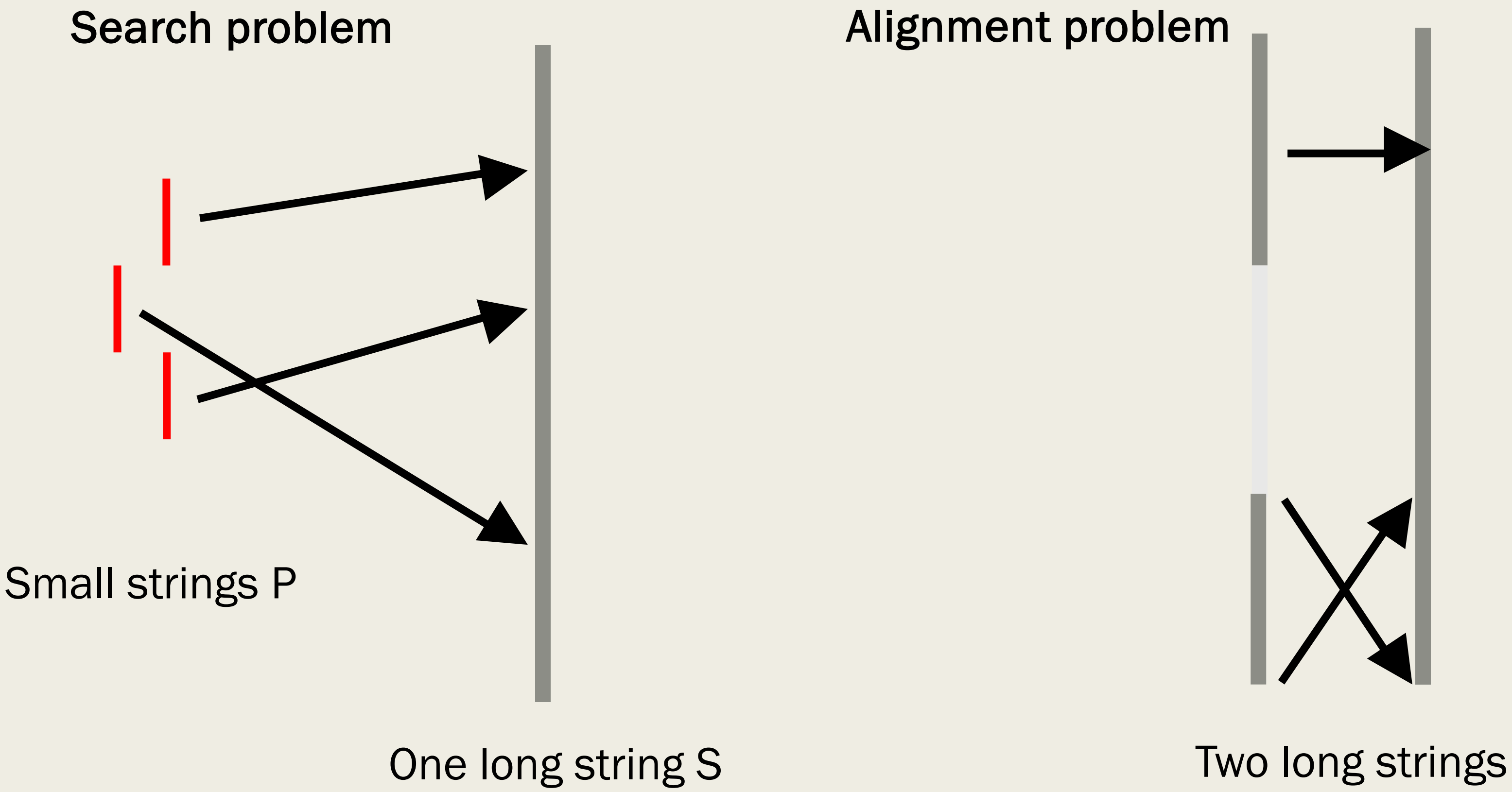
Types of Search Problems

	Literature	Genomics
I (exact)	Find the string “whale”	Find the pattern “GATTACA” in the human genome
II (inexact)	Find a string that looks like “whale” – e.g. “hale”, “whole”, “while”, “elahw”;	Find out where in the human genome this sequencing read came from
III (context)	Find a word that means “whale”	Find a sequence that stops transcription Find a sequence that codes for a protein
IV (meaning)	Find a sentence that expresses the desire to go to sea.	Find sequences that control olfaction Find parts of the genome make you more likely to suffer from diabetes

We will focus on searches in DNA sequences, but most of this applies to any sequence

Note: Generally we are trying to find a pattern (P) that is much shorter than the sequence (S) that we are finding it in.

The problem where we are trying to match up two strings of similar lengths is called alignment (to be covered later).



Types of search problems

1. Search a document once. e.g. Ctrl-F on a website to find a specific phrase

   This week: Boyer-Moore algorithm – preprocess search pattern

2. Perform many searches – e.g. mapping next generation sequence reads

   Next week: suffix trees and hash tables – preprocess search string

Basic search problem

Given some alphabet set (e.g., A, C, G, T), string S from the alphabet, find the first occurrence of a string P (Pattern) in S.

P = GAC

S = AACGACTACGGGACTAACGATCAGATC

Naïve Algorithm

```
S = AACGACTACGGGACTAACGATCAGATC
    GAC
     GAC
      GAC
       GAC
```

Move right 1 position and assess match

Is this a good algorithm?

Correct?

Fast?

Flexible?

Small?

How long does the naive search algorithm take?

How long does this algorithm take?

- We want to know how long this algorithm will take
- We can't say how long in minutes, because that depends on all sorts of things
- More interested in questions like:

  "How much slower will it be if we double the length of the pattern"
  "Can we run it on the human genome in less than a year"

```
S  =  AACGACTACGGGACTAACGATCAGATC
      GAC
       GAC
        GAC
         GAC
          .........
```

Move right 1 position and assess match

How long does this algorithm take?

Suppose P is *M* symbols long and S is *N* symbols long

P = GAC *M=3*

S = AAC**GAC**TACGG**GAC**TAACGATCAGATC *N=27*

Comparing two symbols to see if they are equal is one operation

How many operations do we need?

```
S = AACGACTACGGGACTAACGATCAGATC
    GAC
     GAC
      GAC
       GAC
        .........
```

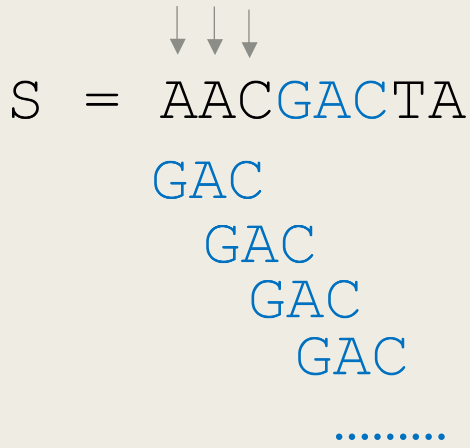
Move right 1 position and assess match

How long does this algorithm take?

Suppose P is *M* symbols long and S is *N* symbols long

Comparing two symbols to see if they are equal is one operation
How many operations do we need?

M comparisons at each position

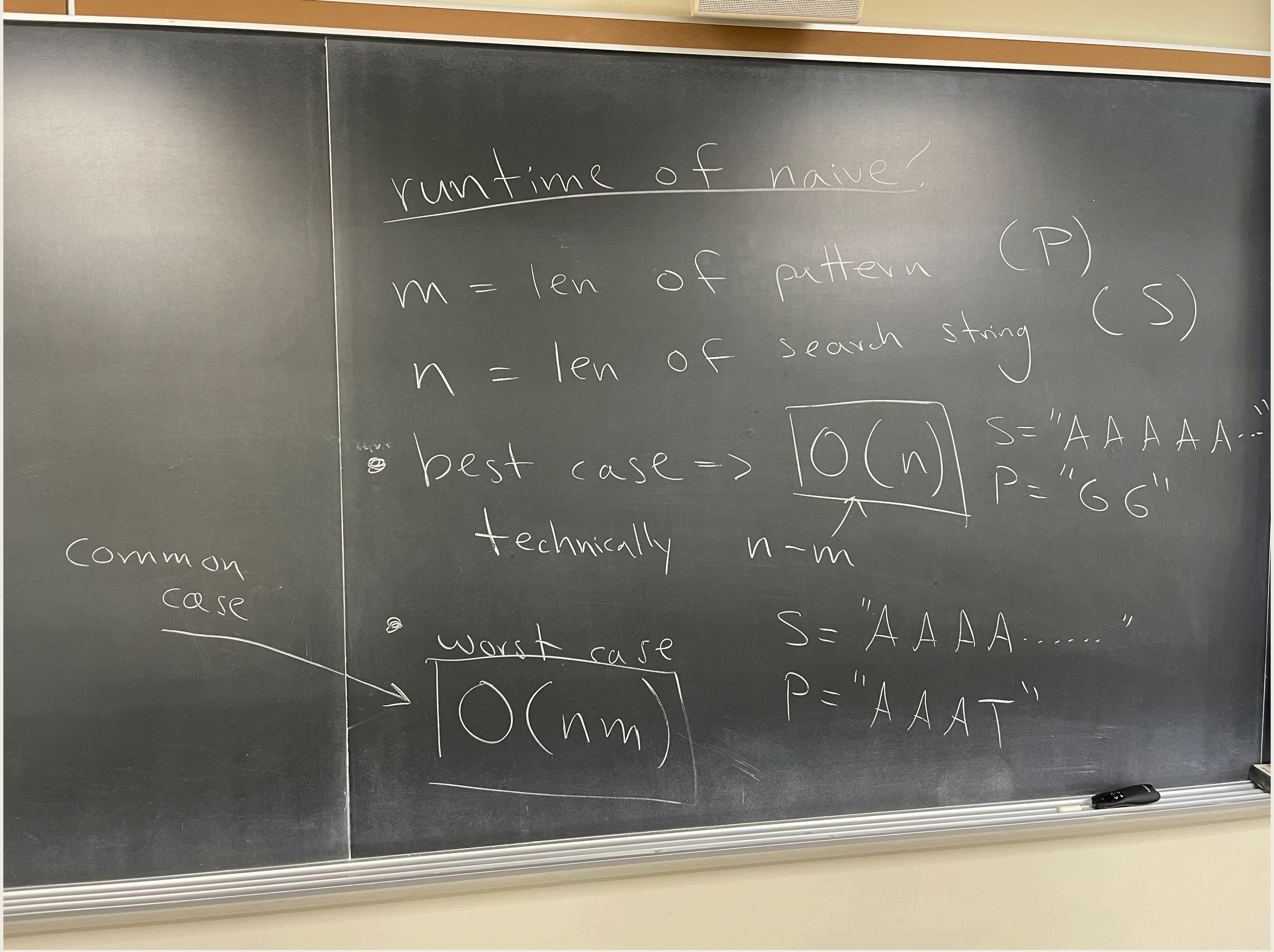


```
S = AACGACTACGGGACTAACGATCAGATC
    GAC
      GAC
        GAC
          GAC
            .........
```

N-M different positions

We need *M(N-M)* operations,
which is
about *NM* if *N<M*

We call this O(MN)



runtime of naive?

m = len of pattern (P)

n = len of search string (S)

- best case => $\boxed{O(n)}$ S="AAAAA..." P="GG"

  technically n-m

Common case

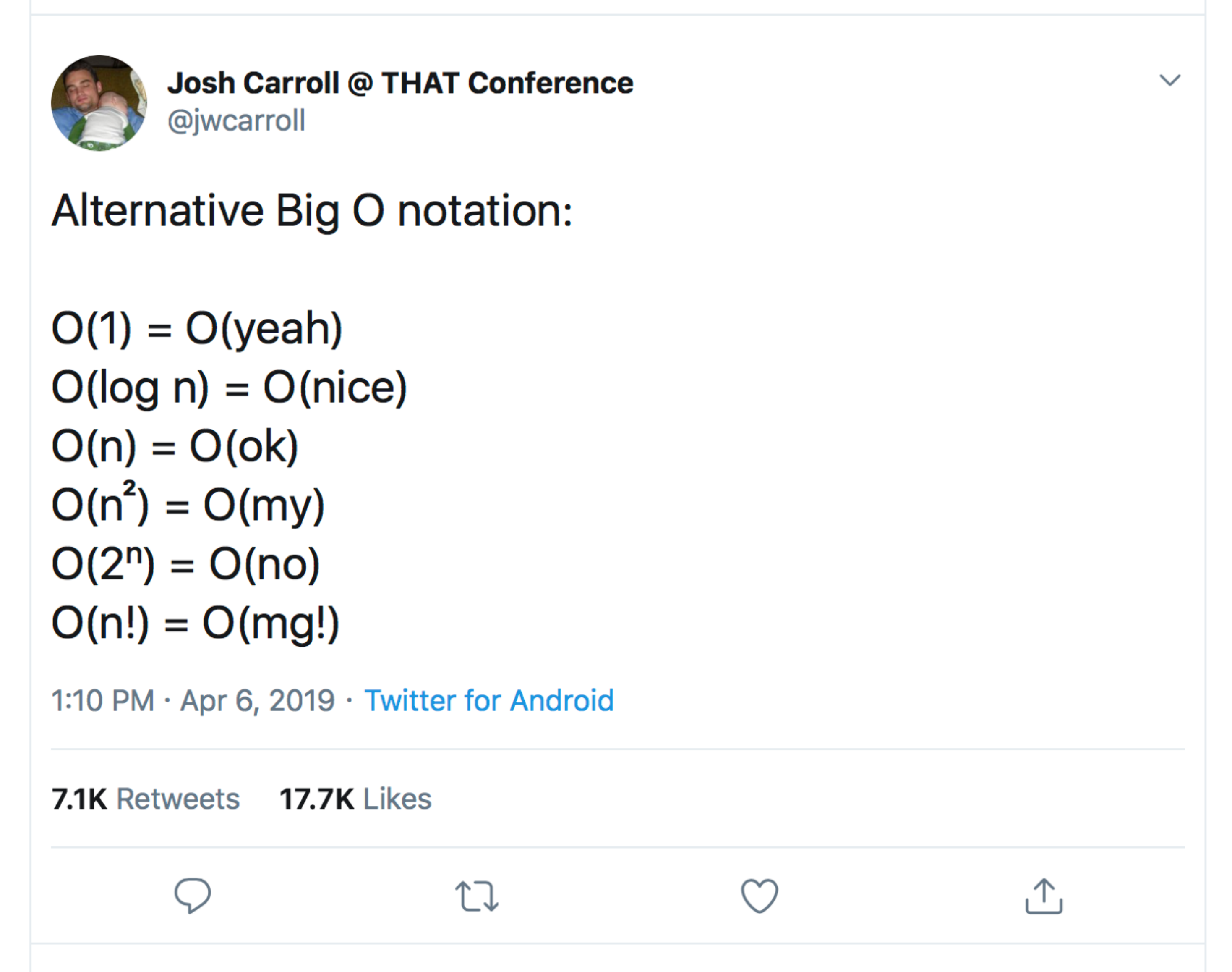
- worst case $\boxed{O(nm)}$ S="AAAA......" P="AAAT"

Big-O notation in practice

Suppose my algorithm works in 1 second on the human mitochondrial genome (16,000 bases). How long will it take to run on the autosomal genome (187 times larger)?

OK	O(1):	1 second
	O(log n):	5 seconds
	O(n):	3 minutes
If you have to	O(n^2):	10 hours
😱	O(1.1^n):	2 years
	O(1.2^n)	20 million years
	O(2^n):	40 orders of magnitude longer than the age of the universe
	O(n!):	Nope!

Big-O notation in practice



Josh Carroll @ THAT Conference
@jwcarroll

Alternative Big O notation:

$O(1)$ = O(yeah)
$O(\log n)$ = O(nice)
$O(n)$ = O(ok)
$O(n^2)$ = O(my)
$O(2^n)$ = O(no)
$O(n!)$ = O(mg!)

1:10 PM · Apr 6, 2019 · Twitter for Android

7.1K Retweets **17.7K** Likes

Note 1: time vs space complexity

Time complexity: How long will it take to run

Space complexity: How much storage do I need?

In practice, space complexity tends to be lower, but can be more of a problem. Quadratic (n^2) memory is a worse problem than quadratic time. More later

Often there is a tradeoff between time, space, speed and sometimes accuracy.

Note 2: average vs worst case

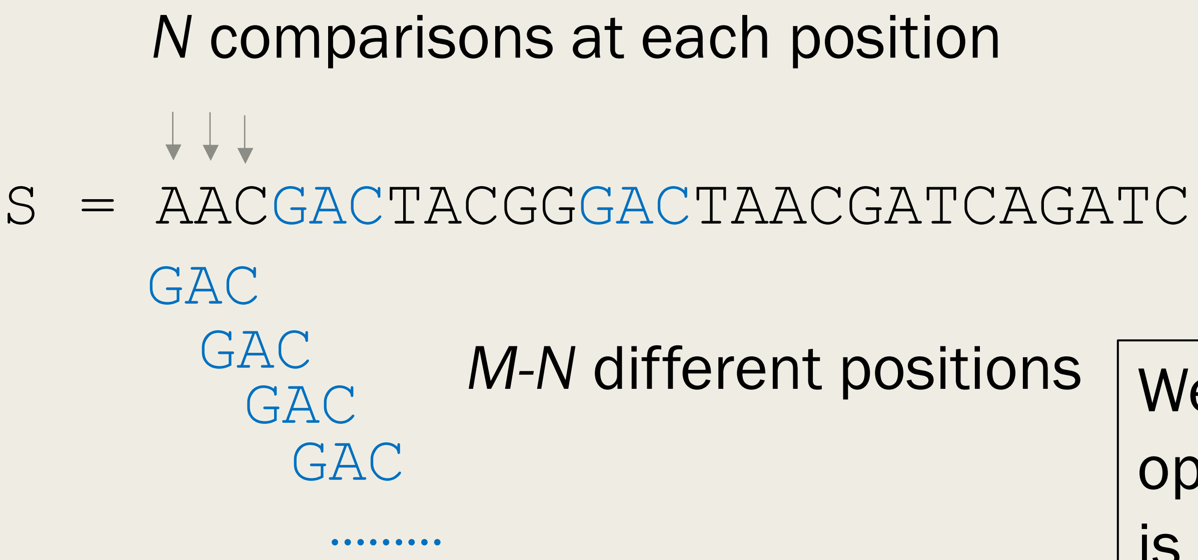
Average case: How fast is the algorithm “on average”

Worst case: How slow could it be if we were unlucky?

Best case: How fast could it be if we were lucky?

Next: how can we improve the simple search algorithm?

Remember the simple string search algorithm. Can we make it faster?



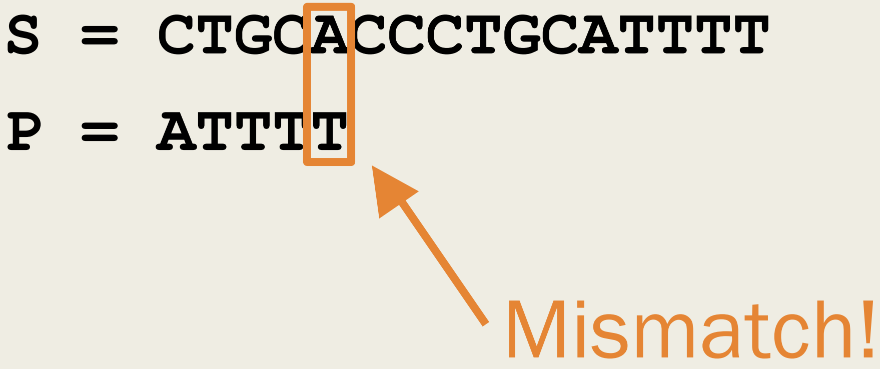
N comparisons at each position

```
S = AACGACTACGGGACTAACGATCAGATC
    GAC
      GAC
        GAC
          GAC
            .........
```

M-N different positions

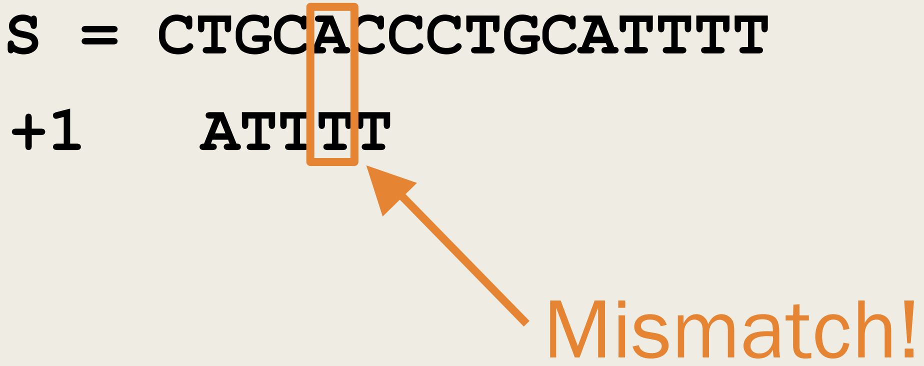
We need *N(M-N)* operations, which is about *NM* if *N<M*

Let’s think about a bad example



```
S = CTGCACCCTGCATTTTT
P = ATTTT
```

Let's think about a bad example



```
S = CTGCACCCTGCATTTT
+1   ATTTT
```

Mismatch!

Let’s think about a bad example

```
S = CTGCACCCTGCATTTT
+1    ATTTT
```

Mismatch!

Back to the first mismatch

```
S = CTGCACCCTGCATTTT
P = ATTTT
```

Look at the mismatch. The search string has an A, so there's no point checking any of the positions in the pattern that aren't also A

Back to the first mismatch

```
S = CTGCACCCTGCATTTT
+4      ATTTT
```

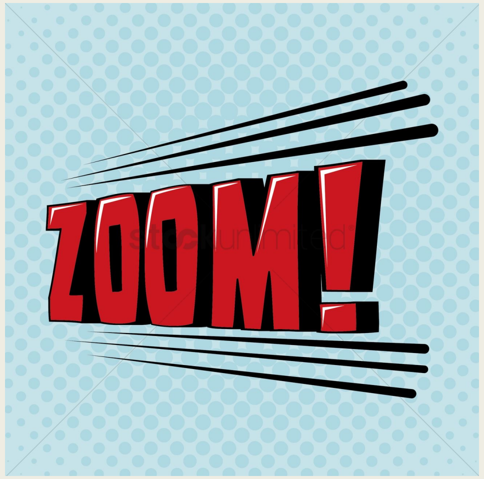
Look at the mismatch. The search string has an A, so there’s no point checking any of the positions in the pattern that aren’t also A

Idea: Scan right-to-left. If there is a mismatch, skip to the right-most matching character that is left of the current mismatch position

Bad character rule

```
   S= CTGCACCTCATTTT
1)    ATTTT
2) +1  ATTTT
3) +1   ATTTT
4) +1    ATTTT
5) +1     ATTTT
6) +1      ATTTT
7) +1       ATTTT
8) +1        ATTTT
9) +1         ATTTT
10)+1          ATTTT
```

```
   S= CTGCACCTCATTTT
1)    ATTTT
2) +4     ATTTT
3) +5          ATTTT
```



Idea: Scan right-to-left. If there is a mismatch, skip to the right-most matching character that is left of the current mismatch position

How much do we skip?

Key: can compute in advance. e.g P = “ATTTT”,

Which character in the pattern is mismatching?

	A	T	T	T	T
What is the mismatching Character in the search string?					

How do we compute this table?

Pre-compute: tradeoff space (memory) and time

How much do we skip?

Key: can compute in advance. e.g P = “ATTTT”,

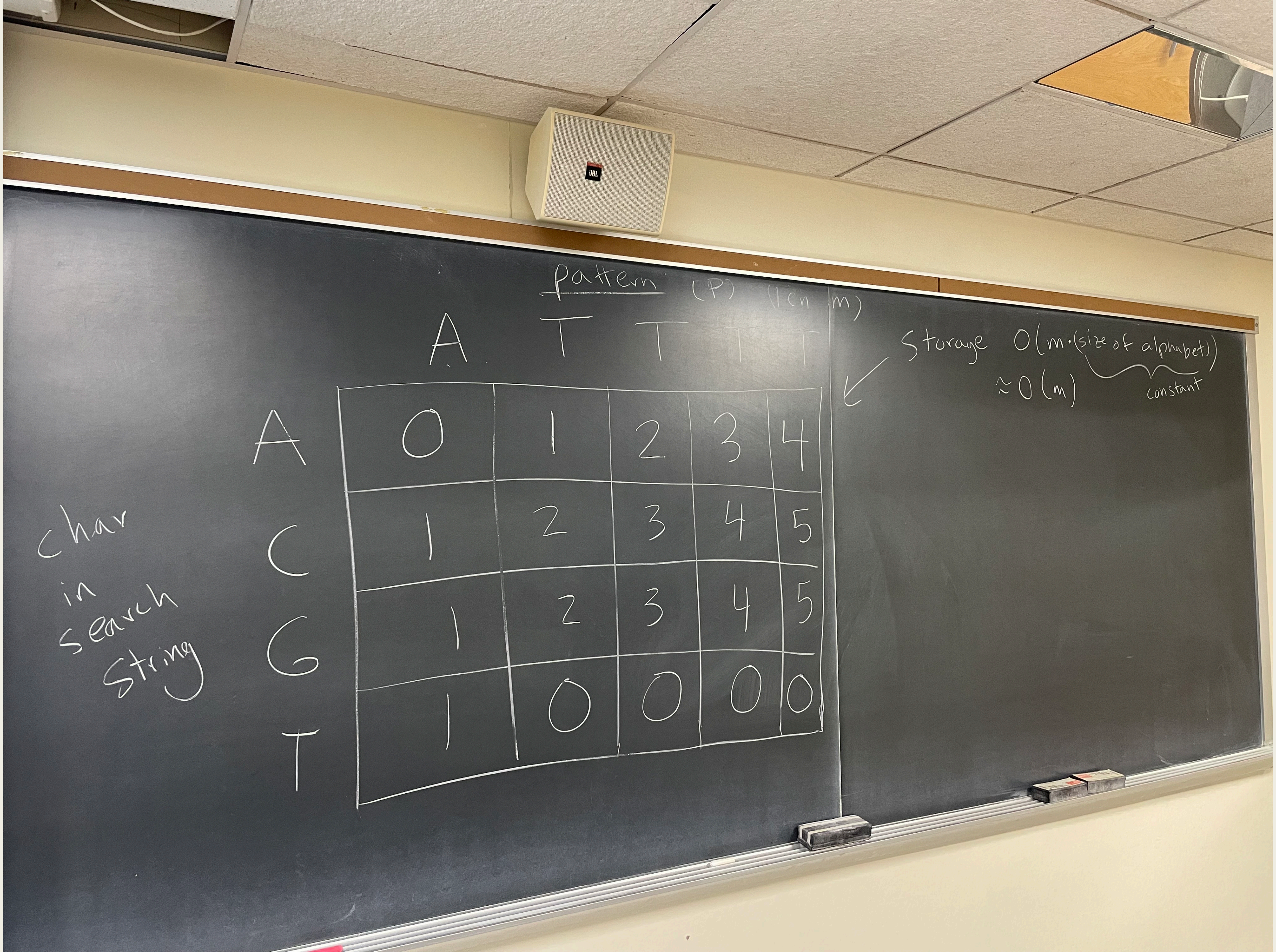
Which character in the pattern is mismatching?

What is the mismatching Character in the search string?

	A	T	T	T	T
	1	2	3	4	5
A	0	1	2	3	4
C	1	2	3	4	5
G	1	2	3	4	5
T	1	0	0	0	0

How do we compute this table?

Pre-compute: tradeoff space (memory) and time



pattern (P) (len m)

char in search string	A	T	T	T	T
A	0	1	2	3	4
C	1	2	3	4	5
G	1	2	3	4	5
T	1	0	0	0	0

Storage $O(m \cdot \underbrace{(\text{size of alphabet})}_{\text{constant}})$

$\approx O(m)$

Bad character rule worksheet