Life @ Lehigh

Lehigh Life Days • April 12, 2006

Dan Lopresti

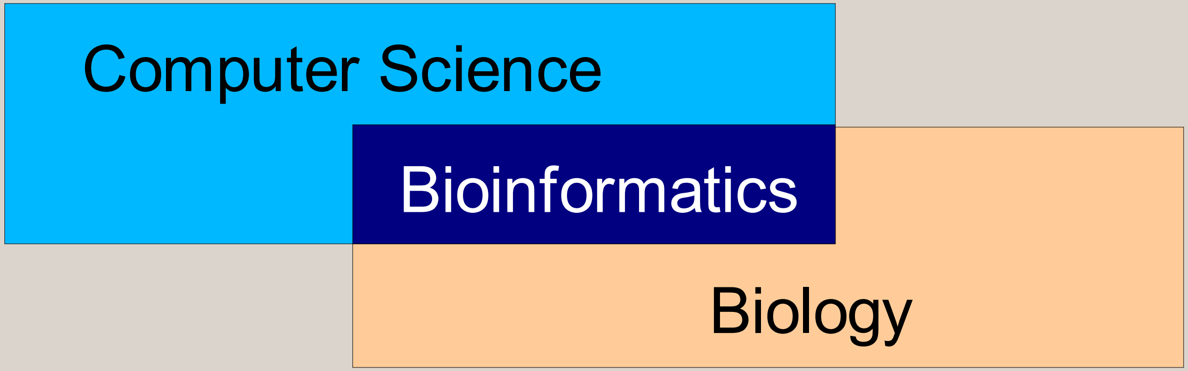
Associate Professor

Office PL 404B

dal9@lehigh.edu

Bioinformatics

What is bioinformatics? Application of techniques from computer science to problems from biology.

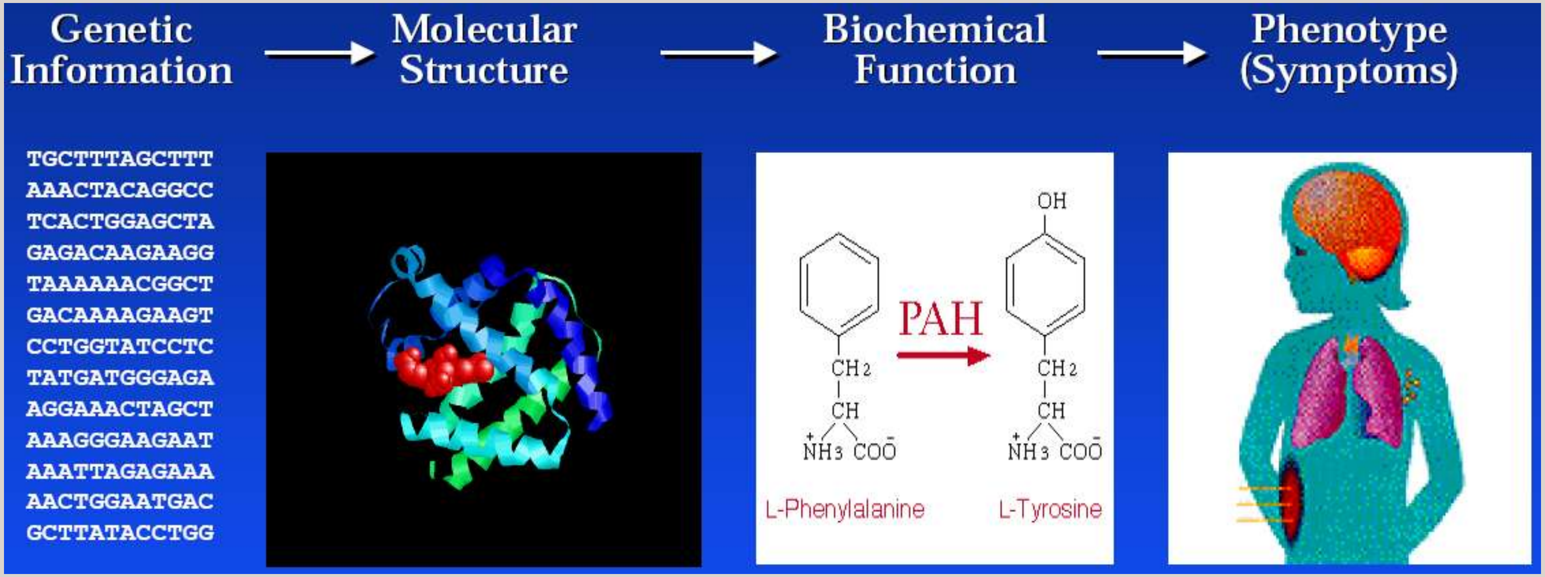


Why is it interesting?

- Important problems.
- Massive quantities of data.
- Desperate need for efficient solutions.
- Success is rewarded.

Motivation

"Biology easily has 500 years of exciting problems to work on."
Donald Knuth



By developing techniques for analyzing this data, we can attempt to understand genetic nature of diseases, evolution of life itself, etc.

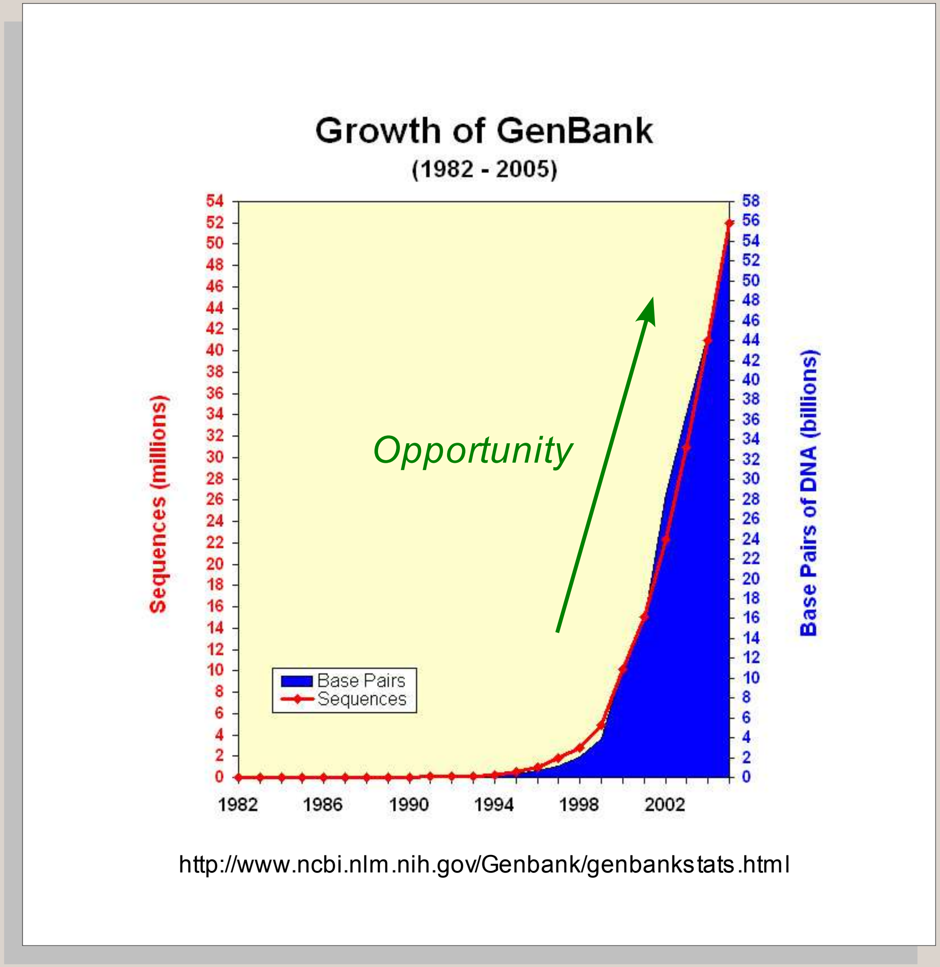
http://cmgm.stanford.edu/biochem218/

Opportunity

Genetic identity of most organisms is encoded in long molecules made up of four basic units, the nucleic acids:

(1) *Adenine*,
(2) *Cytosine*,
(3) *Guanine*,
(4) *Thymine*.

To first approximation, DNA is language over 4 character alphabet, {A, C, G, T}.

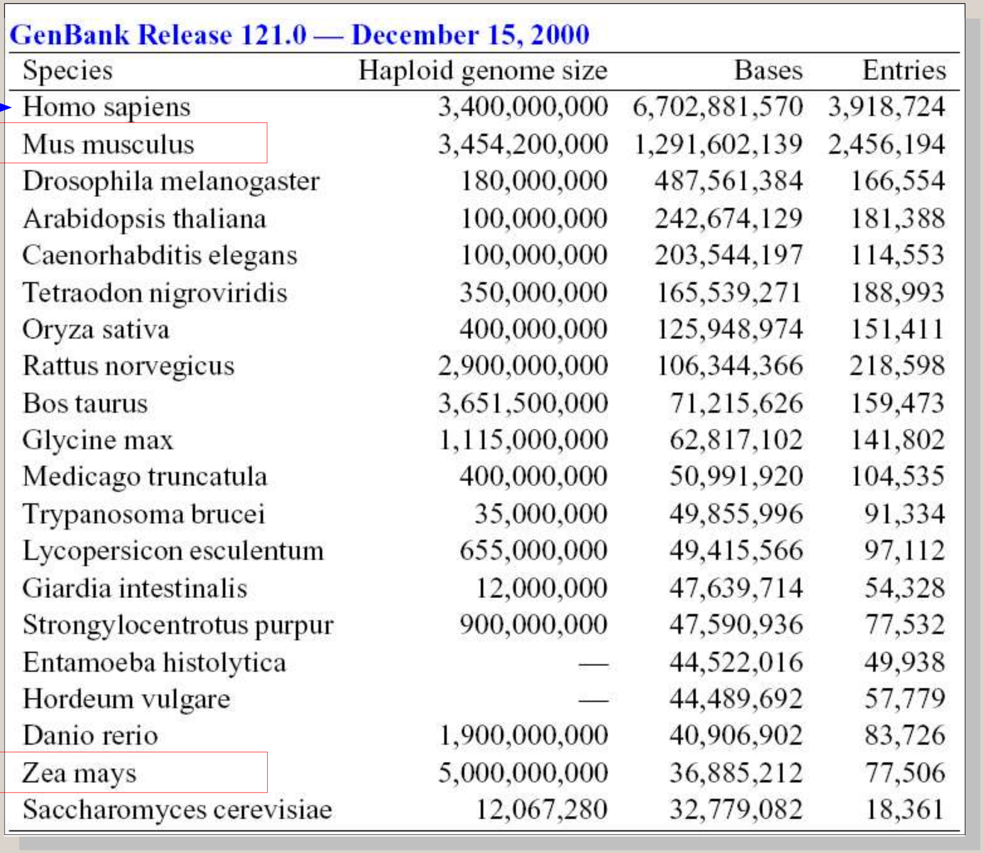


http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html

Genomes

Complete set of chromosomes that determines an organism is known as its *genome*.

Us



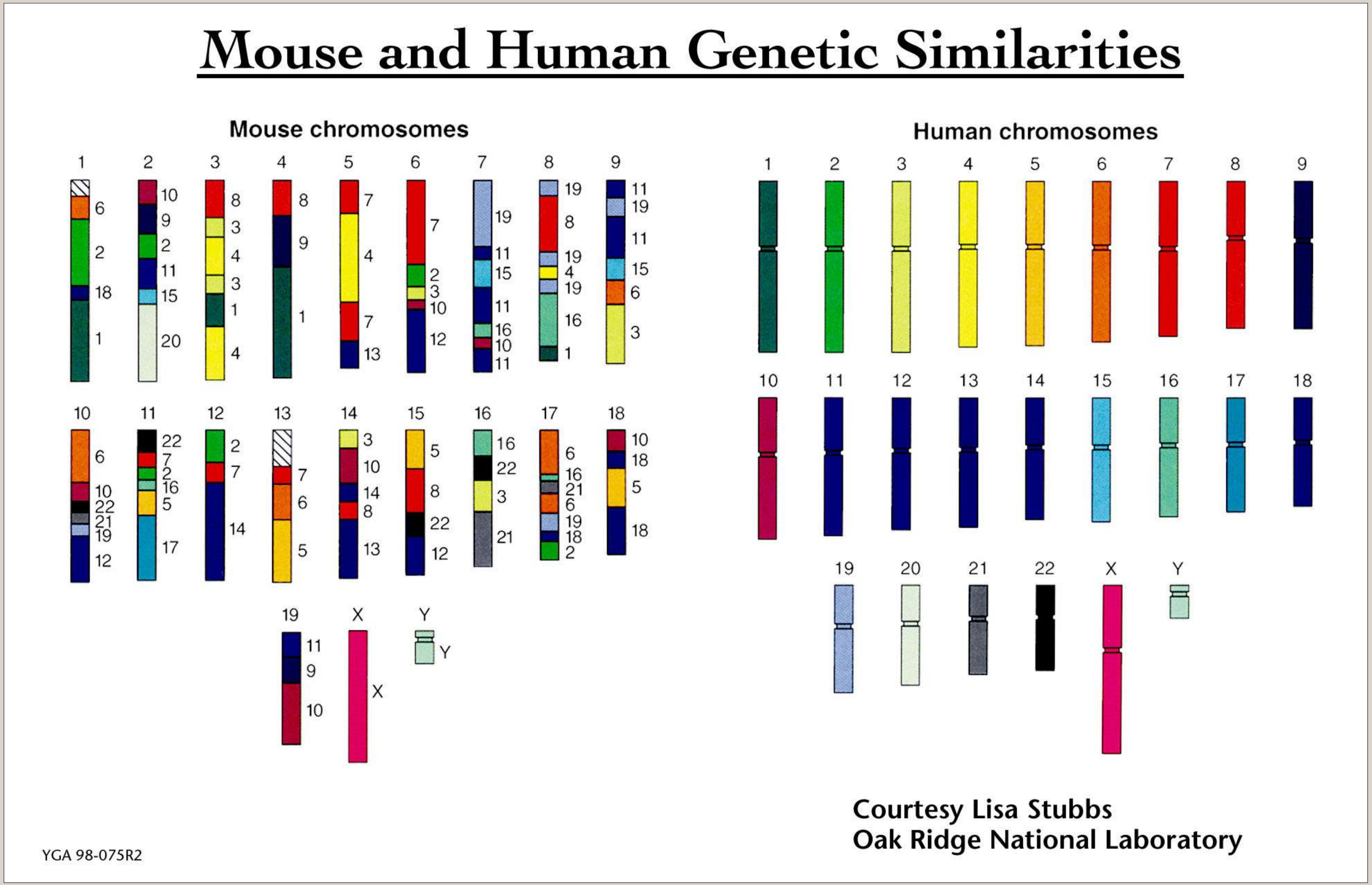
GenBank Release 121.0 — December 15, 2000

Species	Haploid genome size	Bases	Entries
Homo sapiens	3,400,000,000	6,702,881,570	3,918,724
Mus musculus	3,454,200,000	1,291,602,139	2,456,194
Drosophila melanogaster	180,000,000	487,561,384	166,554
Arabidopsis thaliana	100,000,000	242,674,129	181,388
Caenorhabditis elegans	100,000,000	203,544,197	114,553
Tetraodon nigroviridis	350,000,000	165,539,271	188,993
Oryza sativa	400,000,000	125,948,974	151,411
Rattus norvegicus	2,900,000,000	106,344,366	218,598
Bos taurus	3,651,500,000	71,215,626	159,473
Glycine max	1,115,000,000	62,817,102	141,802
Medicago truncatula	400,000,000	50,991,920	104,535
Trypanosoma brucei	35,000,000	49,855,996	91,334
Lycopersicon esculentum	655,000,000	49,415,566	97,112
Giardia intestinalis	12,000,000	47,639,714	54,328
Strongylocentrotus purpur	900,000,000	47,590,936	77,532
Entamoeba histolytica	—	44,522,016	49,938
Hordeum vulgare	—	44,489,692	57,779
Danio rerio	1,900,000,000	40,906,902	83,726
Zea mays	5,000,000,000	36,885,212	77,506
Saccharomyces cerevisiae	12,067,280	32,779,082	18,361

Mus musculus

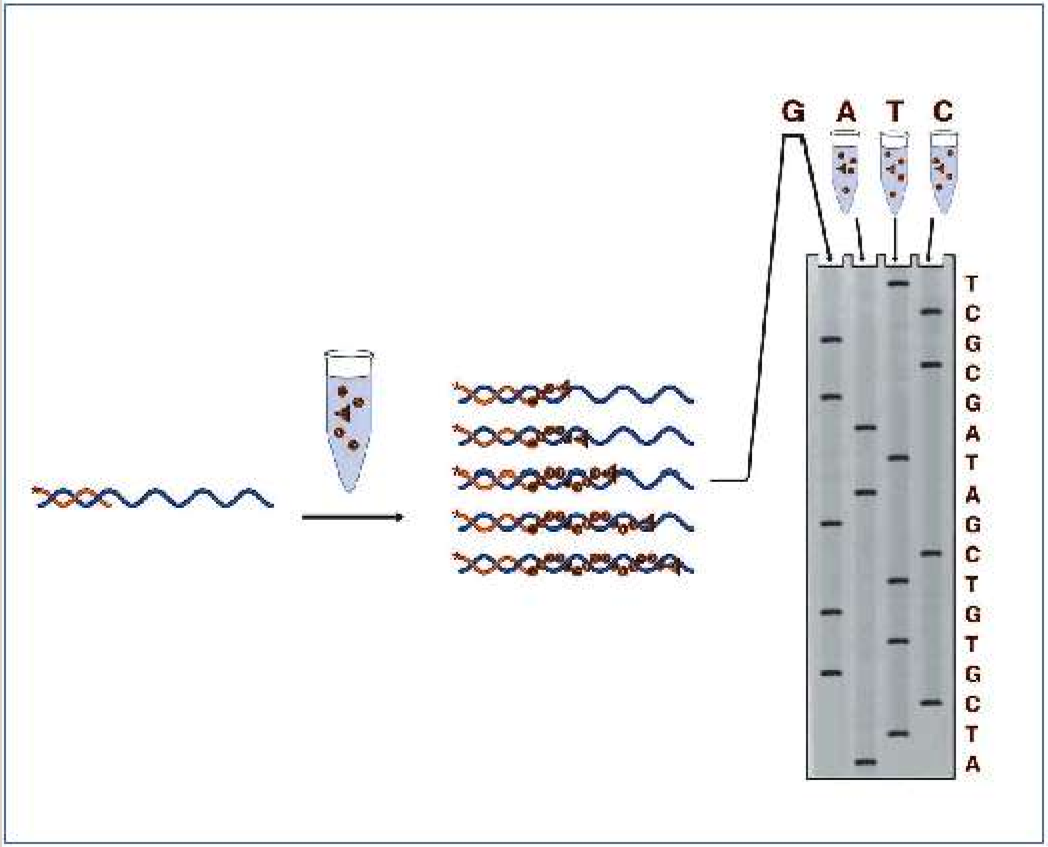
http://www.cbs.dtu.dk/databases/DOGS/
http://www.nsrl.ttu.edu/tmot1/mus_musc.htm
http://www.oardc.ohio-state.edu/seedid/single.asp?strID=324

Comparative Genomics



http://www.ornl.gov/sci/techresources/Human_Genome/graphics/slides/ttmousehuman.shtml

Reading DNA



This is known as *Sanger sequencing*.

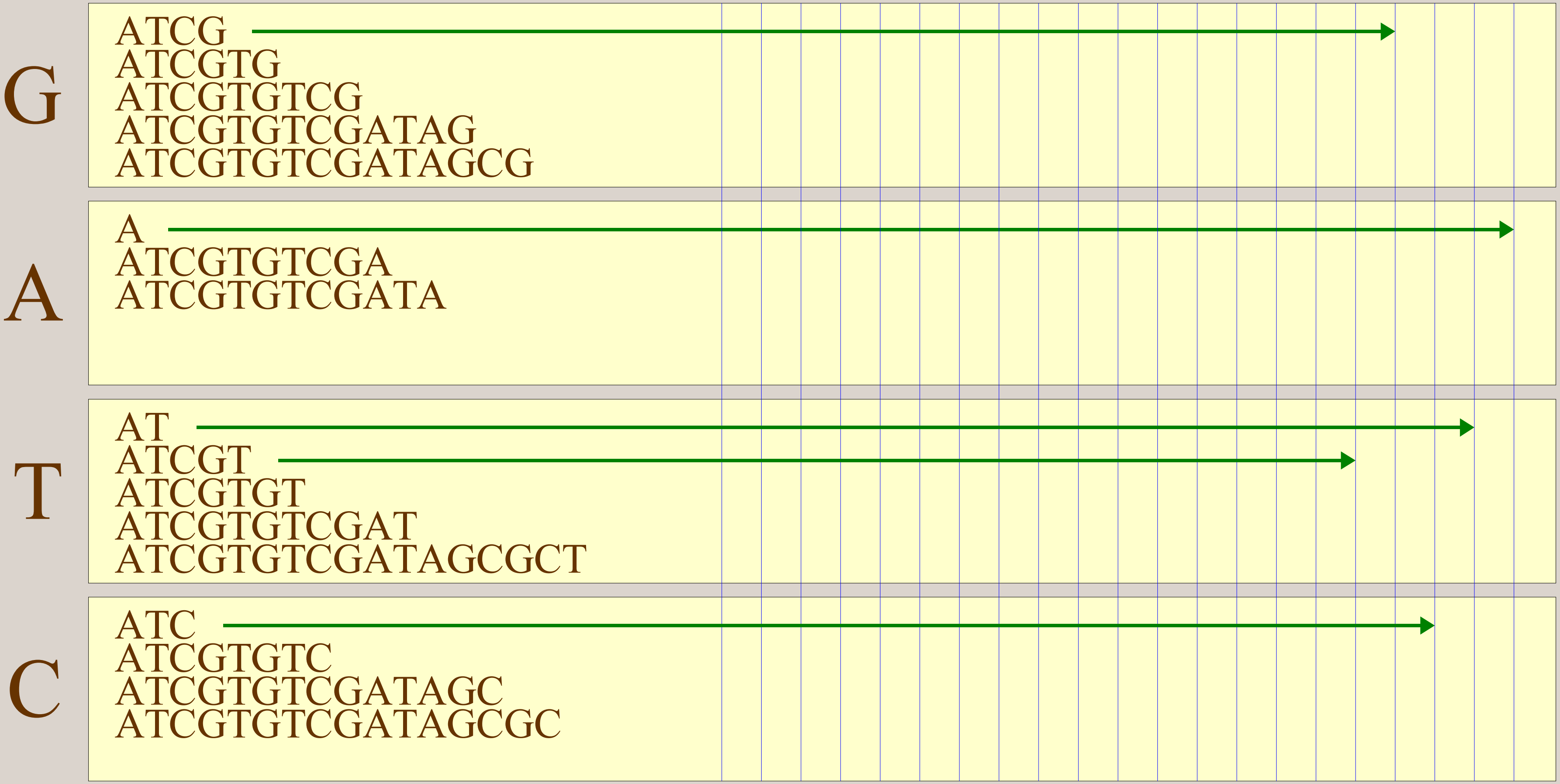
http://www.apelex.fr/anglais/applications/sommaire2/sanger.htm
http://www.iupui.edu/~wellsctr/MMIA/htm/animations.htm

Gel electrophoresis is process of separating a mixture of molecules in a gel media by application of an electric field. In general, DNA molecules with similar lengths will migrate same distance.

First cut DNA at each base: A, C, G, T. Then run gel and read off sequence: ATCGTG ...

Reading DNA

Original sequence: ATCGTGTCGATAGCGCT



G

ATCG
ATCGTG
ATCGTGTCG
ATCGTGTCGATAG
ATCGTGTCGATAGCG

A

A
ATCGTGTCGA
ATCGTGTCGATA

T

AT
ATCGT
ATCGTGT
ATCGTGTCGAT
ATCGTGTCGATAGCGCT

C

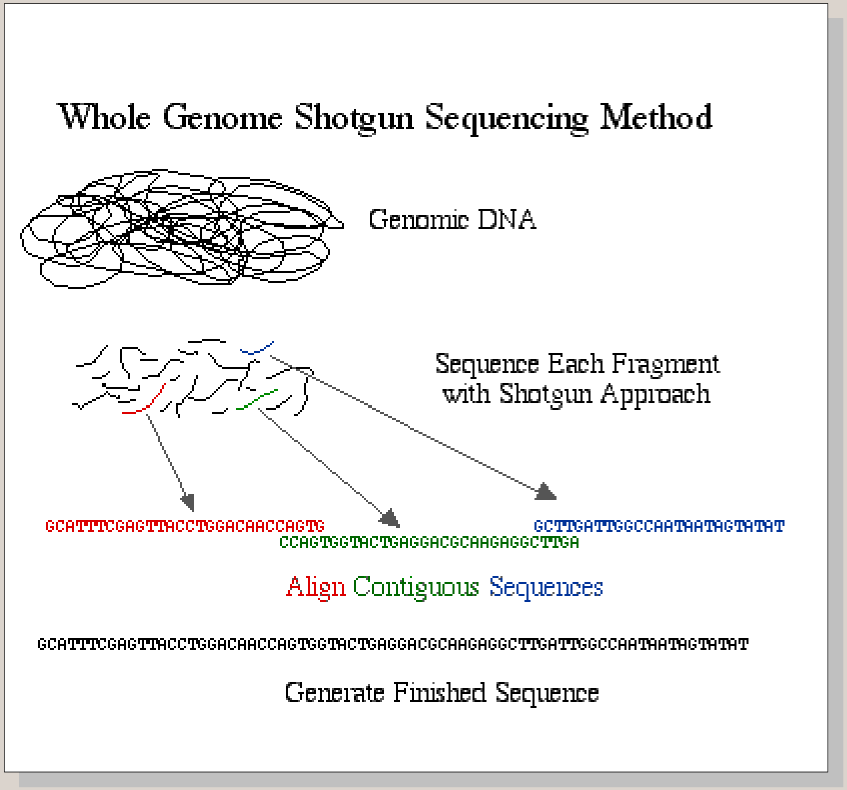
ATC
ATCGTGTC
ATCGTGTCGATAGC
ATCGTGTCGATAGCGC

Sequencing a Genome

Unfortunately, current sequencing technologies can only read 700 nucleotides at a time.

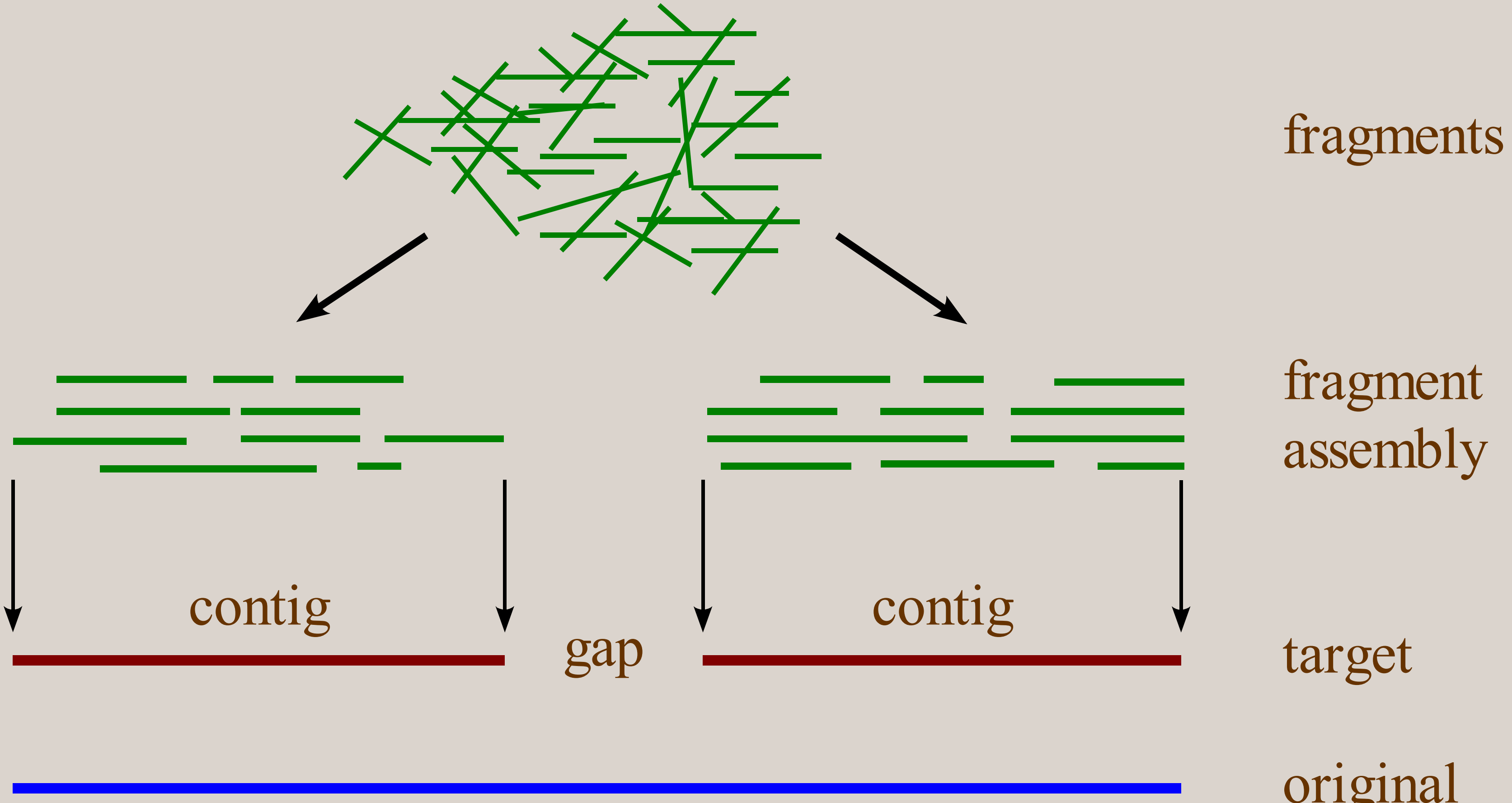
For genomes, we use *shotgun sequencing*, which breaks a chromosome into overlapping short sequences which must then be reassembled.

It's kind of like putting together a jigsaw puzzle with millions of pieces (a lot of which are “blue sky”).


Whole Genome Shotgun Sequencing Method
Genomic DNA
Sequence Each Fragment with Shotgun Approach
GCATTTCGAGTTACCTGGACAACCAGTG
CCAGTGGTACTGAGGACGCAAGAGGCTTGA
GCTTGATTGGCCAATAATAGTATAT
Align Contiguous Sequences
GCATTTCGAGTTACCTGGACAACCAGTGGTACTGAGGACGCAAGAGGCTTGATTGGCCAATAATAGTATAT
Generate Finished Sequence

http://occawlonline.pearsoned.com/bookbind/pubbooks/bc_mcampbell_genomics_1/medialib/method/shotgun.html

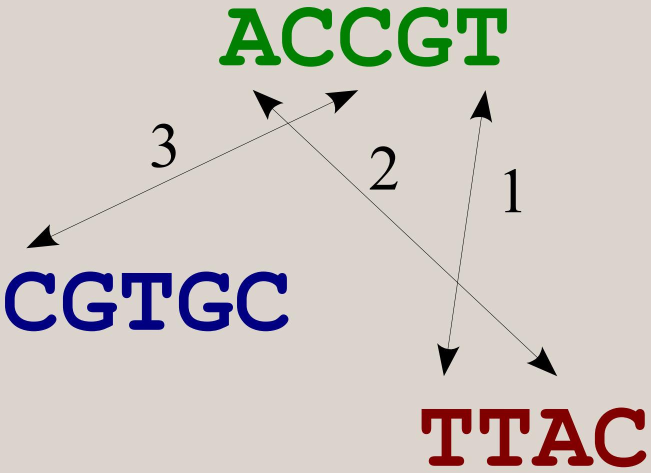
Sequence Assembly



Sequence Assembly

A simple model of DNA assembly is the *Shortest Supersequence Problem*: given a set of sequences, find the shortest sequence *S* such that each of original sequences appears as subsequence of *S*.

Look for overlap between *prefix* of one sequence and *suffix* of another:



```
--ACCGT--
----CGTGC
TTAC-----
---------
TTACCGTGC
```

Sequence Assembly

Sketch of algorithm (procedure for assembling fragments):

- Create an *overlap graph* in which every node represents a fragment and edges indicate overlap.
- Determine which overlaps will be used in the final assembly: find an optimal collection of paths in overlap graph.

W = AGTATTGGCAATC

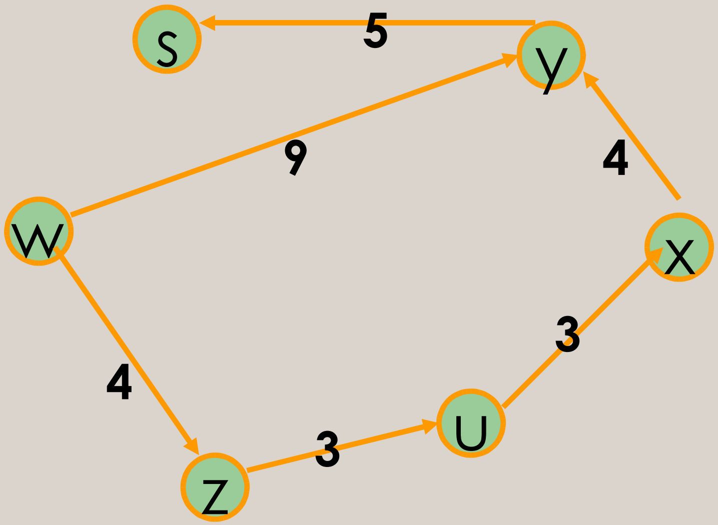
Z = AATCGATG

U = ATGCAAACCT

X = CCTTTTGG

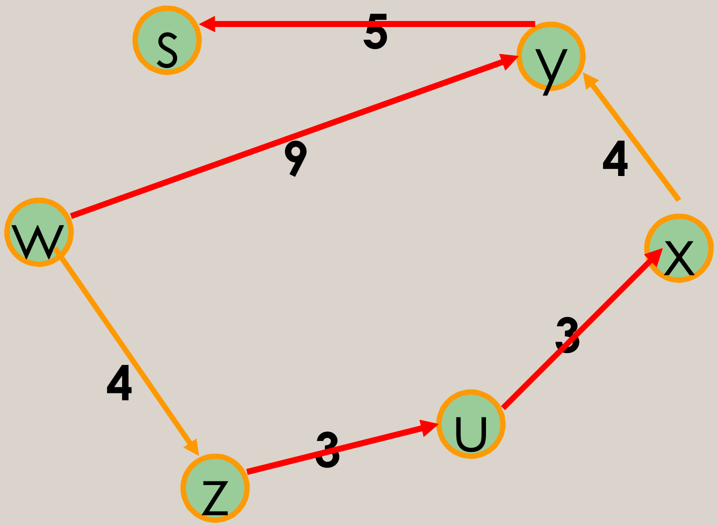
Y = TTGGCAATCA

S = AATCAGG



Sequence Assembly

- Look for paths of maximum weight: use *greedy* algorithm to select edge with highest weight at every step.
- Selected edge must connect nodes with in- and out-degrees <= 1.
- May end up with set of paths: each corresponds to a contig.



W→Y→S

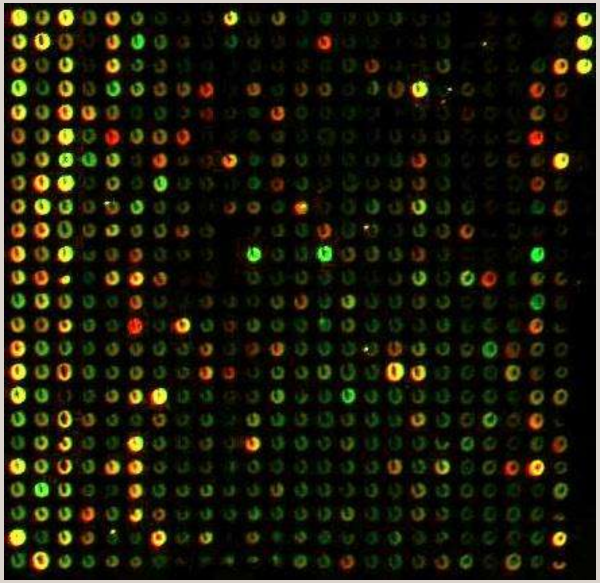
```
AGTATTGGCAATC
    TTGGCAATCA
         AATCAGG
AGTATTGGCAATCAGG
```

Z→U→X

```
AATCGATG
     ATGCAAACCT
            CCTTTTGG
AATCGATGCAAACCTTTTGG
```

DNA Microarrays

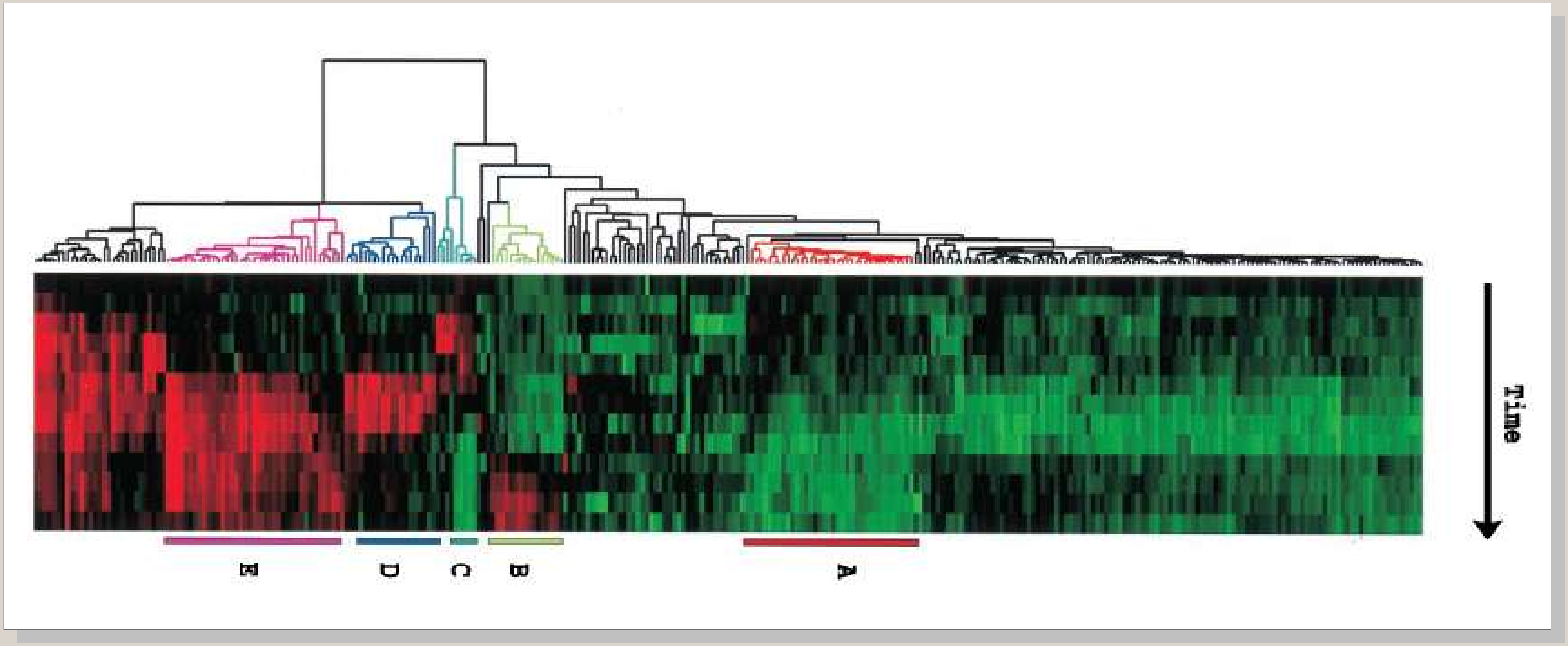
- Allows simultaneous measurement of the level of transcription for every gene in a genome (gene expression).
- Differential expression, changes over time.
- Single microarray can test ~10k genes.
- Data obtained faster than can be processed.
- Want to find genes that behave similarly.
- A pattern discovery problem.



green = repressed

red = induced

Visualizing Microarray Data



From "Cluster analysis and display of genome-wide expression patterns" by Eisen, Spellman, Brown, and Botstein, Proc. Natl. Acad. Sci. USA, Vol. 95, pp. 14863–14868, December 1998

Clustering Microarray Data

K-means clustering is one way to organize this data:

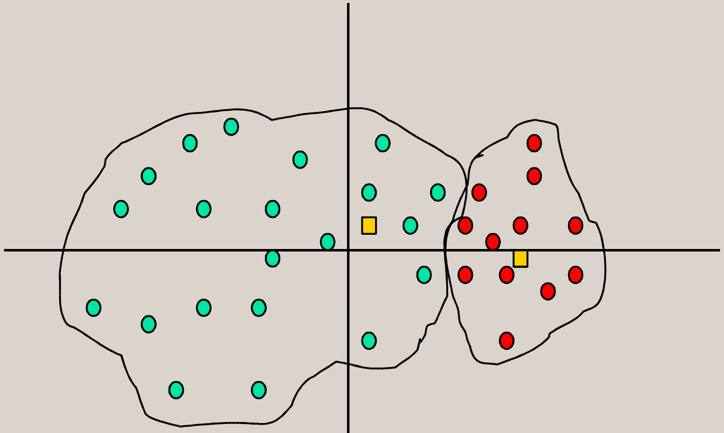
- Given set of n data points and an integer k.
- We want to find set of k center points that minimizes mean-squared distance from each data point to its nearest cluster center.

Sketch of algorithm:

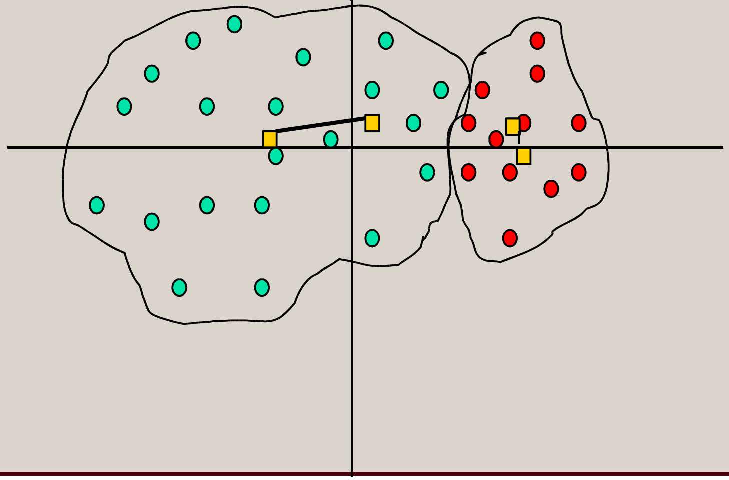
- Choose k initial center points randomly and cluster data.
- Calculate new centers for each cluster using points in cluster.
- Re-cluster all data using new center points.
- Repeat second two steps until no data points are moved from one cluster to another or some other convergence criterion is met.

Clustering Microarray Data

- Pick $k = 2$ centers at random.
- Cluster data around these center points.



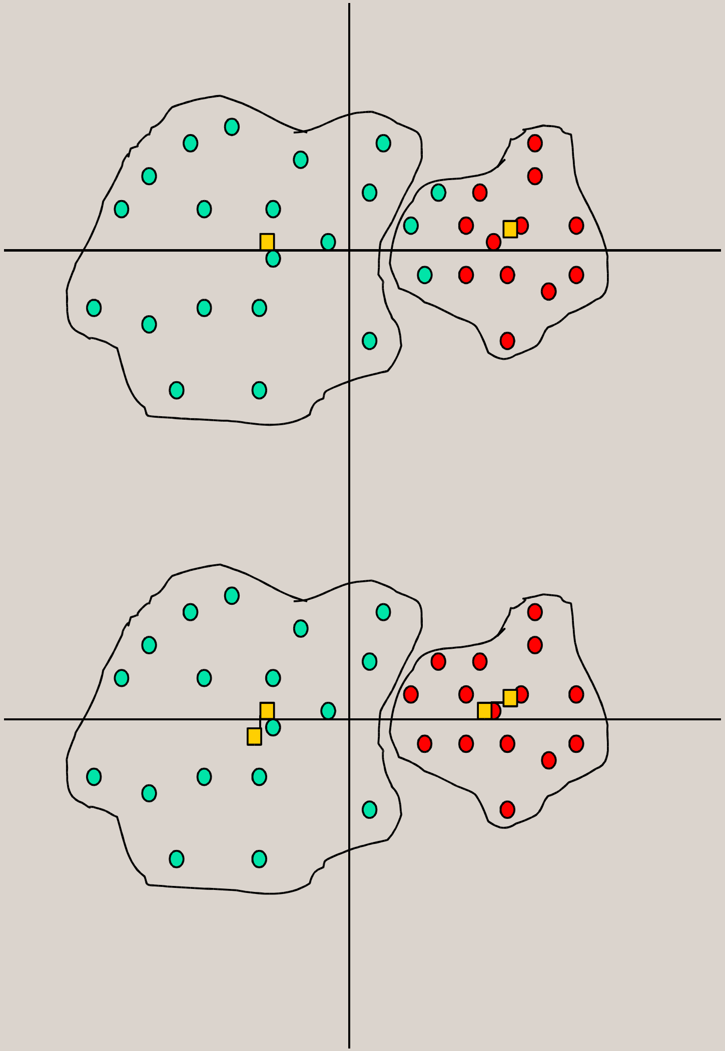
- Re-calculate centers based on current clusters.



From "Data Analysis Tools for DNA Microarrays" by Sorin Draghici.

Clustering Microarray Data

- Re-cluster data around new center points.

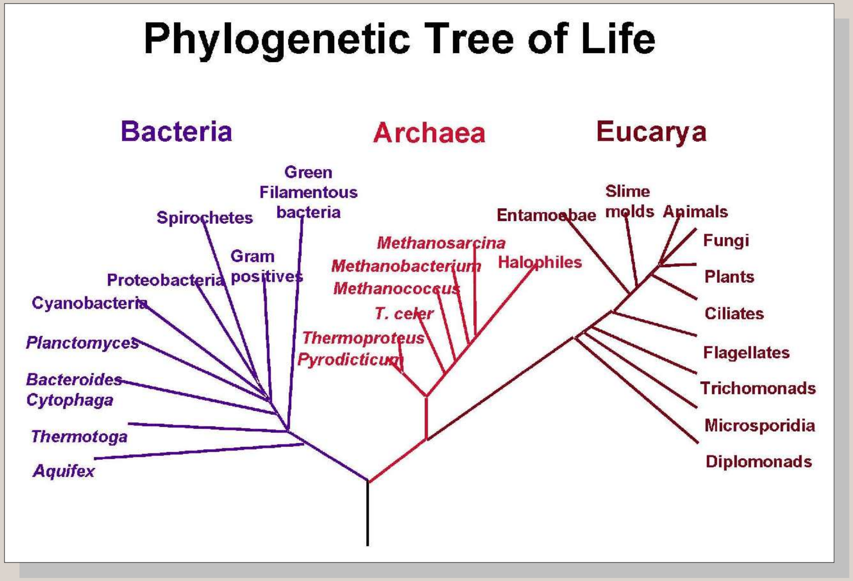


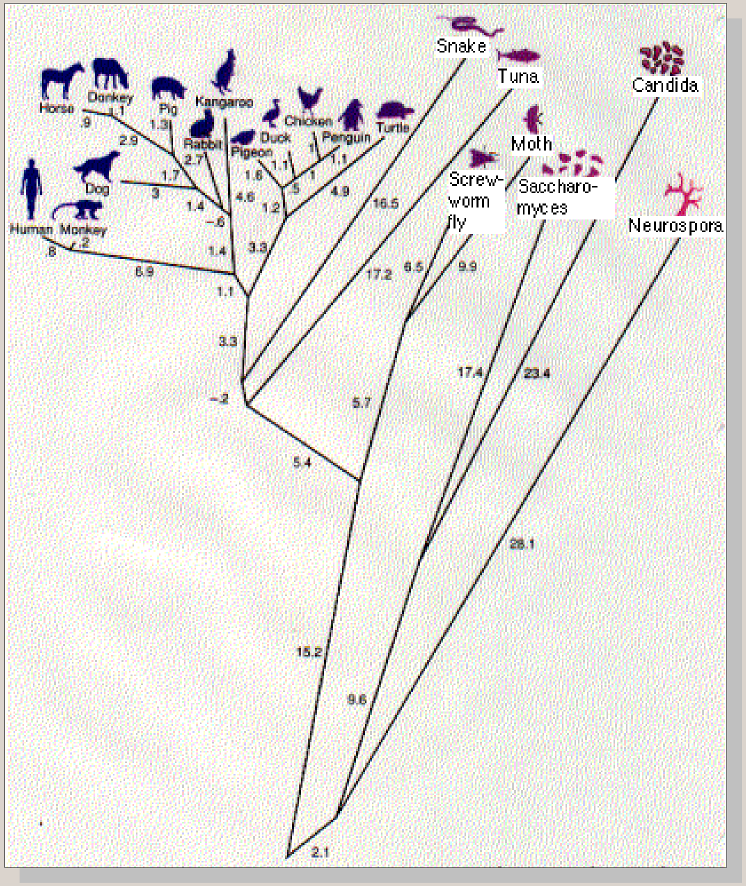
- Repeat last two steps until no more data points are moved into a different cluster.

From "Data Analysis Tools for DNA Microarrays" by Sorin Draghici.

Building the “Tree of Life”

Scientists build phylogenetic trees in an attempt to understand evolutionary relationships.





Note: these trees are “best guesses” and certainly contain some errors!

http://en.wikipedia.org/wiki/Phylogenetic_tree
http://users.rcn.com/jkimball.ma.ultranet/BiologyPages/T/Taxonomy.html

Why Study Bioinformatics?

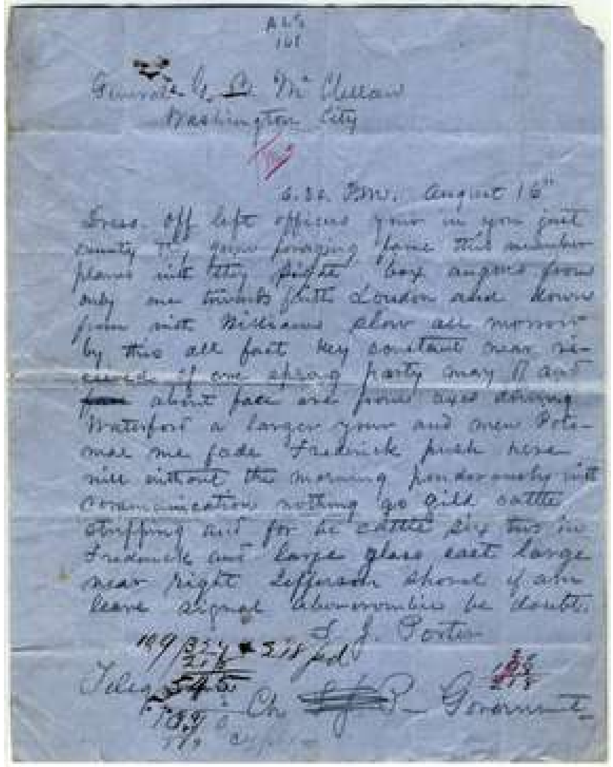
- Still many urgent open problems ⇒ lots of opportunities to make fundamental contributions (and become rich and famous).
- Stretch your creativity and problem-solving skills to the limit.
- Join a cross-disciplinary team – work with interesting people.
- Participate in unlocking the mysteries of life itself.
- Make the world a better place.

My Advice on Colleges

- Gather all the data. Then trust your instincts.
- Whatever your decision, college will probably be the best years of your life – take full advantage of it.
- Beyond the obvious criteria, look for opportunities to become engaged in cutting-edge research, whatever your major.
- Seek out projects that will attract attention and have an impact.

Breaking a Civil War Secret Code

The Civil War letter ...



ALS
161

General G. B. McClellan
Washington City

6.32 PM. August 16th

Truss off left officers your in you just county the grow foraging fame the mischiev- plains with tiny fifth bay anyone from only one twenty fifth [illegible] [illegible] knows from with Williams plow all memory by this all fact key constant near [illegible] ~~[illegible]~~ if one [illegible] party may [illegible] ~~[illegible]~~ about face and [illegible] eyes [illegible] Waterford a larger your and new Poto- mac me fade Frederick push here nil without the morning ponderously with communication nothing go gill battle stripping out for the cattle [illegible] two in Frederick and large glass east large near right Jefferson shone if [illegible] leave signal abominable be doubt.

F. J. Porter

http://digital.lib.lehigh.edu/remain

... encrypted – not yet broken.

The players ...

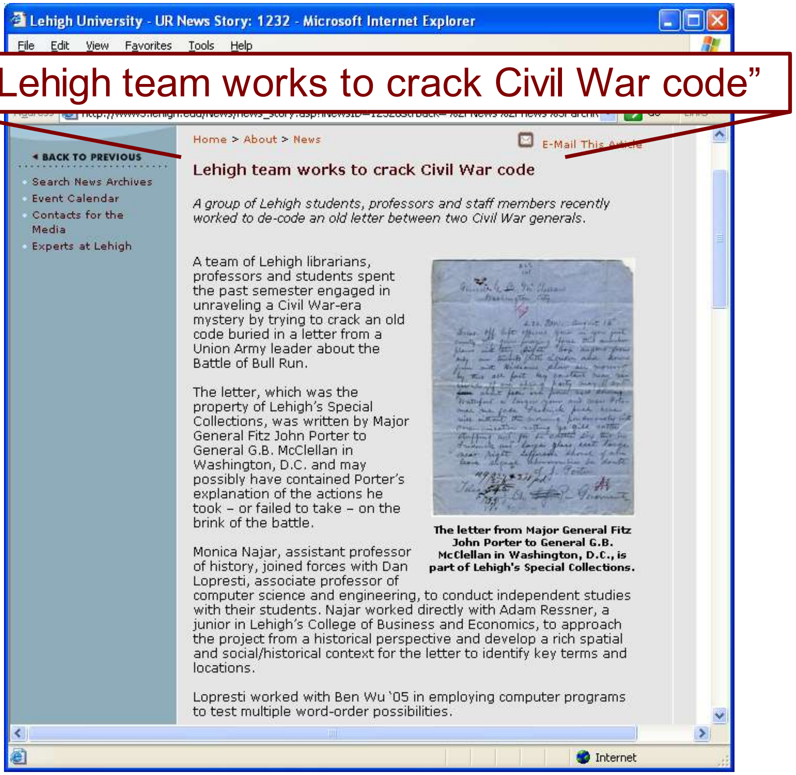
Major General
Fitz-John Porter
author of letter, blamed for Union loss at Second Bull Run, court-martialed in 1863

Major General
George McClellan
recipient of letter

Anson Stager
inventor of cypher system and later an early leader in U.S. telecommunications

Breaking a Civil War Secret Code

The news story ...



"Lehigh team works to crack Civil War code"

Lehigh team works to crack Civil War code

A group of Lehigh students, professors and staff members recently worked to de-code an old letter between two Civil War generals.

A team of Lehigh librarians, professors and students spent the past semester engaged in unraveling a Civil War-era mystery by trying to crack an old code buried in a letter from a Union Army leader about the Battle of Bull Run.

The letter, which was the property of Lehigh's Special Collections, was written by Major General Fitz John Porter to General G.B. McClellan in Washington, D.C. and may possibly have contained Porter's explanation of the actions he took – or failed to take – on the brink of the battle.

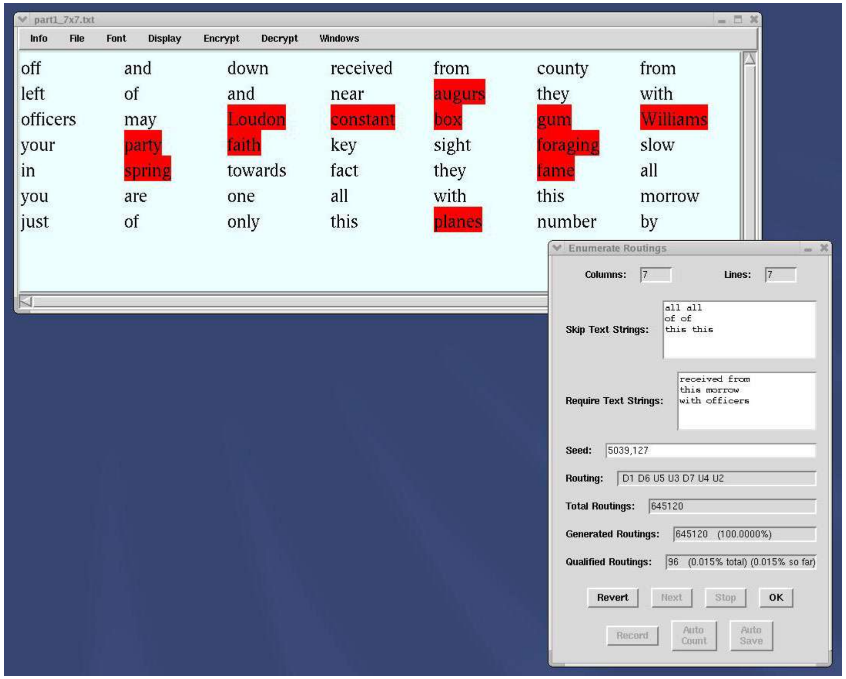
The letter from Major General Fitz John Porter to General G.B. McClellan in Washington, D.C., is part of Lehigh's Special Collections.

Monica Najar, assistant professor of history, joined forces with Dan Lopresti, associate professor of computer science and engineering, to conduct independent studies with their students. Najar worked directly with Adam Ressner, a junior in Lehigh's College of Business and Economics, to approach the project from a historical perspective and develop a rich spatial and social/historical context for the letter to identify key terms and locations.

Lopresti worked with Ben Wu '05 in employing computer programs to test multiple word-order possibilities.

... finale yet to be written.

The current software tool ...



off	and	down	received	from	county	from
left	of	and	near	augurs	they	with
officers	may	Loudon	constant	box	gum	Williams
your	party	faith	key	sight	foraging	slow
in	spring	towards	fact	they	fame	all
you	are	one	all	with	this	morrow
just	of	only	this	planes	number	by

Columns: 7 Lines: 7

Skip Text Strings:
```
all all
of of
this this
```

Require Text Strings:
```
received from
this morrow
with officers
```

Seed: 5039,127

Routing: D1 D6 U5 U3 D7 U4 U2

Total Routings: 645120

Generated Routings: 645120 (100.0000%)

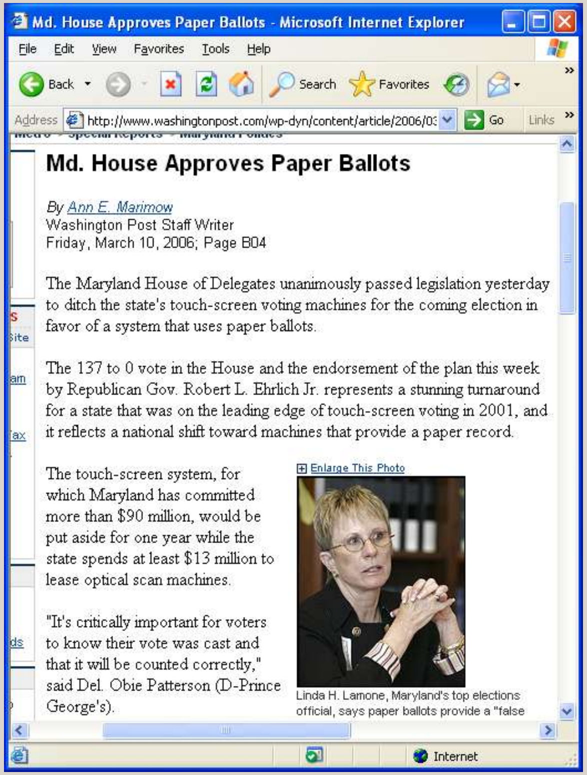
Qualified Routings: 96 (0.015% total) (0.015% so far)

... promising, but needs work.

Hint: I'm looking for help ...

Better Electronic Voting

E-voting has generated some controversy recently …


Md. House Approves Paper Ballots - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://www.washingtonpost.com/wp-dyn/content/article/2006/03 Go Links
Internet

Md. House Approves Paper Ballots

By Ann E. Marimow
Washington Post Staff Writer
Friday, March 10, 2006; Page B04

The Maryland House of Delegates unanimously passed legislation yesterday to ditch the state's touch-screen voting machines for the coming election in favor of a system that uses paper ballots.

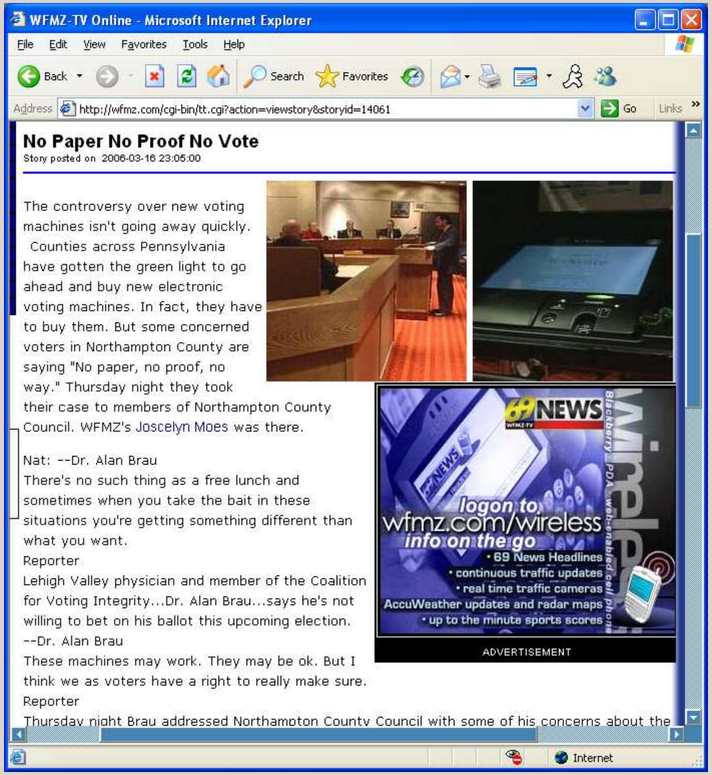
The 137 to 0 vote in the House and the endorsement of the plan this week by Republican Gov. Robert L. Ehrlich Jr. represents a stunning turnaround for a state that was on the leading edge of touch-screen voting in 2001, and it reflects a national shift toward machines that provide a paper record.

The touch-screen system, for which Maryland has committed more than $90 million, would be put aside for one year while the state spends at least $13 million to lease optical scan machines.

"It's critically important for voters to know their vote was cast and that it will be counted correctly," said Del. Obie Patterson (D-Prince George's).

Enlarge This Photo

Linda H. Lamone, Maryland's top elections official, says paper ballots provide a "false


WFMZ-TV Online - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://wfmz.com/cgi-bin/tt.cgi?action=viewstory&storyid=14061 Go Links
Internet

No Paper No Proof No Vote

Story posted on 2006-03-16 23:05:00

The controversy over new voting machines isn't going away quickly. Counties across Pennsylvania have gotten the green light to go ahead and buy new electronic voting machines. In fact, they have to buy them. But some concerned voters in Northampton County are saying "No paper, no proof, no way." Thursday night they took their case to members of Northampton County Council. WFMZ's Joscelyn Moes was there.

Nat: --Dr. Alan Brau
There's no such thing as a free lunch and sometimes when you take the bait in these situations you're getting something different than what you want.
Reporter
Lehigh Valley physician and member of the Coalition for Voting Integrity...Dr. Alan Brau...says he's not willing to bet on his ballot this upcoming election.
--Dr. Alan Brau
These machines may work. They may be ok. But I think we as voters have a right to really make sure.
Reporter
Thursday night Brau addressed Northampton County Council with some of his concerns about the

Maryland votes "yes" for paper trail … while Pennsylvania votes "no."

Better Electronic Voting

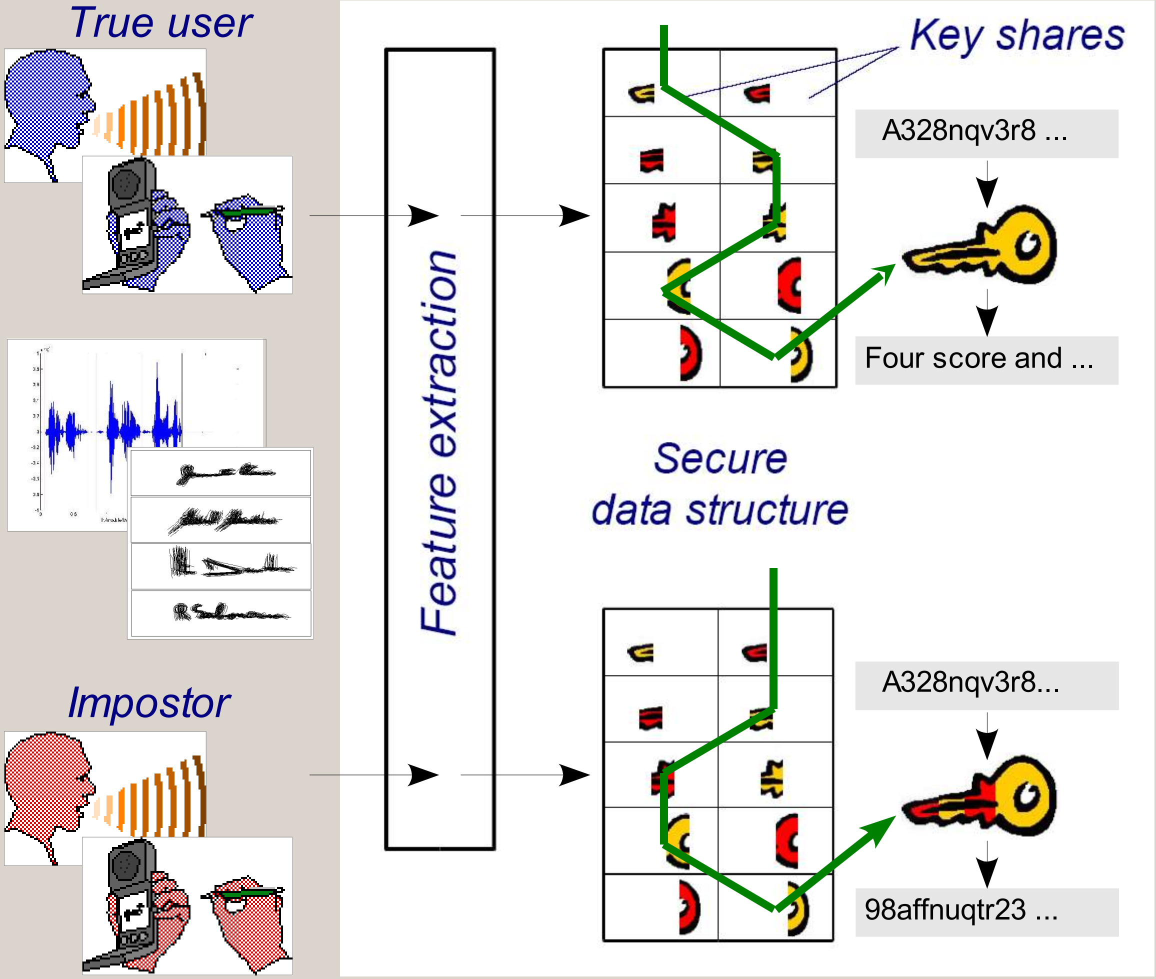
E-voting: what's the right answer?

- Take a critical look at all aspects of the problem.
- Examine both security and usability issues.
- Build a prototype of an e-voting system that includes a reliable Voter Verified Paper Audit Trail (VVPAT).
- Some critics claim it can't be done: we disagree.
- Yet another undergraduate project.
- Of fundamental importance because our demoncracy depends on fair and transparent elections.

Diebold e-voting system

Evaluating Biometric Security

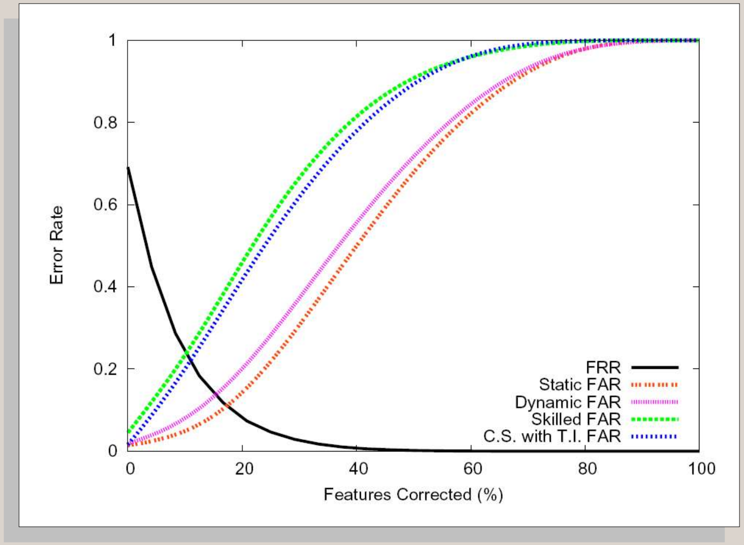
- Cryptographic key broken into shares and mixed with random data.
- Features extracted from user's speech or handwriting.
- Only input from true user will select correct shares to yield proper key.

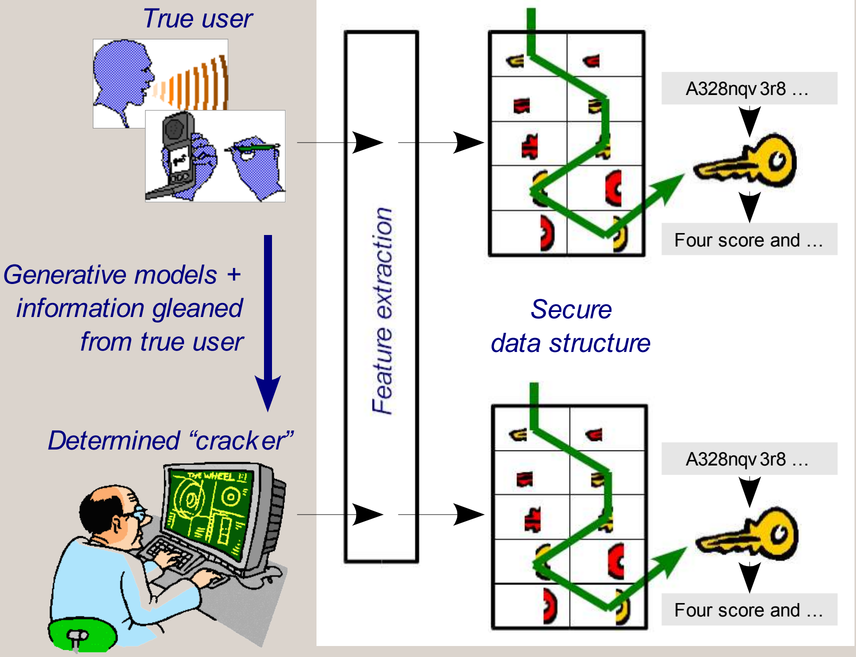


Evaluating Biometric Security

Biometrics may be vulnerable to attacks using generative models.

- Some current systems at risk.
- Results for handwriting show machine can equal performance of skilled human forger:





Use our experience to improve biometrics, increase security.

Last but not least …

Why attend Lehigh?

Because this is a great place to be a student!

Thank you!