



Current Topics in Health Sciences Librarianship

William H. Welch Medical Library — Johns Hopkins Medical Institutions

Information Management: The Key to the Human Genome Project

Robert J. Robbins

Fred Hutchinson Cancer Research Center

rrobbins@fhcrc.org

Abstract

Information Management: The Key to the Human Genome Project

The Human Genome Project (HGP), the first "big science" project in biology, now stands past the five-year mark in its 15-year plan to map and sequence the entire human genome. Often described as being ahead of schedule and under budget, the project's discoveries have already revolutionized many areas of biomedical research and promise to improve patient care. A critical part of the project is collecting, organizing, and making available for retrieval and analysis the massive amount of complex data that describe the 50,000-100,000 genes and three billion bases of sequence that make up the human genome. This session will first provide an overview of the basic biology behind the HGP and the techniques being used to accomplish its scientific goals. Then the information infrastructure of the project will be discussed, with emphasis on the worldwide network of databases in which data of many types are being stored. Finally, a larger information infrastructure for biology and the potential for electronic data publishing to become truly a new form of scientific communication will be discussed.

IT is transforming biology and the relentless effects of Moore's Law is transforming that transformation.

The Example of Genomics

Key Points

- Information Technology is a key enabling technology that allows the genome project to occur.
- Genetic information is passed from parent to child in a form that is truly, not metaphorically, digital.
- The immediate goal of the Human Genome Project (HGP) is to obtain a copy of that digital information for humans and for several selected model organisms. The ultimate result of the HGP will be an understanding of that digital information.
- Along the way, tremendous amounts of information must be collected, analyzed, stored, and managed.
- Electronic Data Publishing (EDP) is a new kind of scientific literature.
- Ensuring the continued utility of EDP will require the active participation of information management professionals.

Introduction

Effect of Information Technology

IT reduces the effects of:

- *distance*
- *time*
- *complexity*

All of these significantly affect biological research...

Effect of Information Technology

Effect of IT on tasks:

- *accomplishment*
- *coordination*
- *possibility*

This improves both efficiency and effectiveness, and even allows new strategies to be pursued.

IT-Biology Synergism

IT is Special

Information Technology:

- *affects both the performance and the management of tasks*
- *is incredibly plastic*
(programming and poetry are both exercises in pure thought)
- *improves exponentially*

Biology is Special

Life is Characterized by:

- *individuality*
- *historicity*
- *contingency*
- *high information content*

No law of large numbers...

IT - Biology Synergism

- *Physics needs calculus, the method for manipulating information about infinite numbers of vanishingly small, independent, equivalent things.*
- *Biology needs information technology, the method for manipulating information about large numbers of dependent, historically contingent, individual things.*

Biology Transformed by IT

Paradigm Shift in Biology

There are two kinds of scientists: those who read the literature and those who create the literature.

Walter Gilbert, 1977?

Paradigm Shift in Biology

[I]n the current paradigm, the attack on the problems of biology is viewed as being solely experimental. The 'correct' approach is to identify a gene by some direct experimental procedure determined by some property of its product or otherwise related to its phenotype to clone it, to sequence it, to make its product and to continue to work experimentally so as to seek an understanding of its function.

Walter Gilbert. 1991. Towards a paradigm shift in biology. *Nature*, 349:99.

Paradigm Shift in Biology

The new paradigm, now emerging, is that all the 'genes' will be known (in the sense of being resident in databases available electronically), and that the starting point of a biological investigation will be theoretical. An individual scientist will begin with a theoretical conjecture, only then turning to experiment to follow or test that hypothesis.

Walter Gilbert. 1991. Towards a paradigm shift in biology. *Nature*, 349:99.

Paradigm Shift in Biology

The next tenfold increase in the amount of information in the databases will divide the world into haves and have nots, unless each of us connects to that information and learns how to sift through it for the parts we need. This is not more difficult than knowing how to access the scientific literature as it is at present, for even that skill involves more than a traditional reading of the printed page, but today involves a search by computer.

Walter Gilbert. 1991. Towards a paradigm shift in biology. *Nature*, 349:99.

Paradigm Shift in Biology

We must hook our individual computers into the worldwide network that gives us access to daily changes in the database and also makes immediate our communications with each other. The programs that display and analyze the material for us must be improved and we must learn how to use them more effectively. Like the purchased kits, they will make our life easier, but also like the kits, we must understand enough of how they work to use them effectively.

Walter Gilbert. 1991. Towards a paradigm shift in biology. *Nature*, 349:99.

Paradigm Shift in Biology

To use this flood of knowledge, which will pour across the computer networks of the world, biologists not only must become computer literate, but also change their approach to the problem of understanding life.

Walter Gilbert. 1991. Towards a paradigm shift in biology. *Nature*, 349:99.

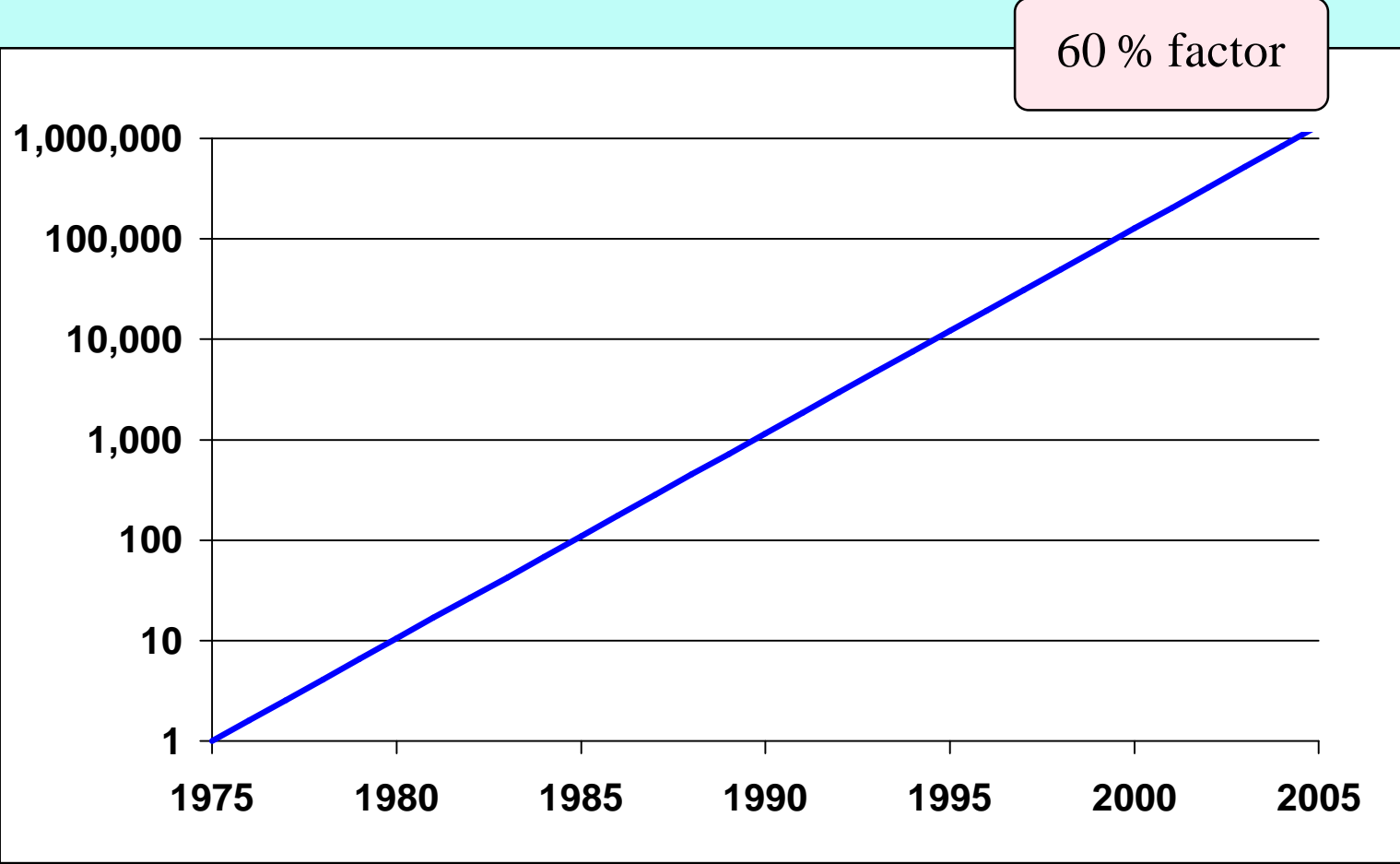
IT Transformed by Moore's Law

Moore's Law

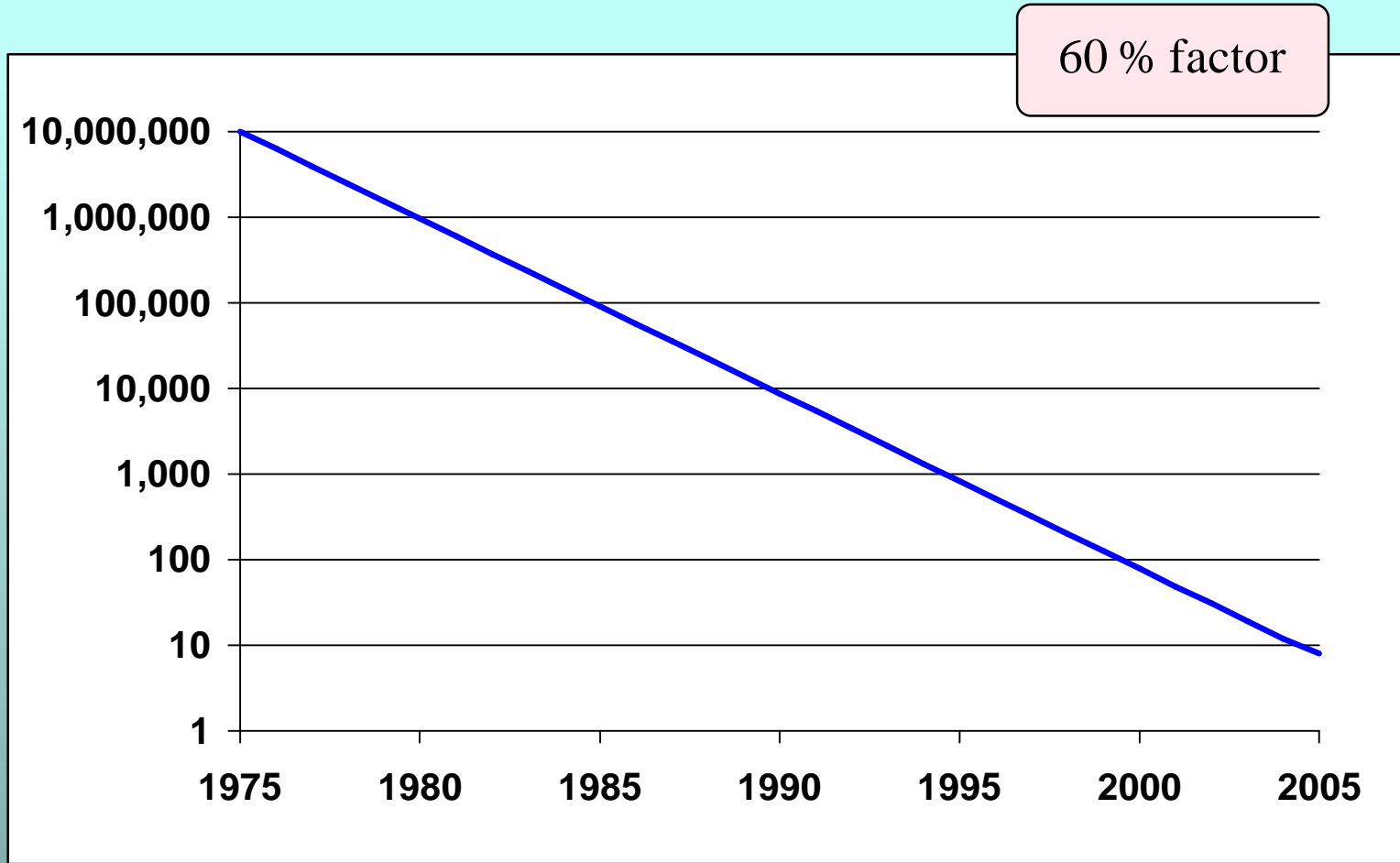
Every eighteen months, the number of transistors that can be placed on a chip doubles.

Gordon Moore, co-founder of Intel...

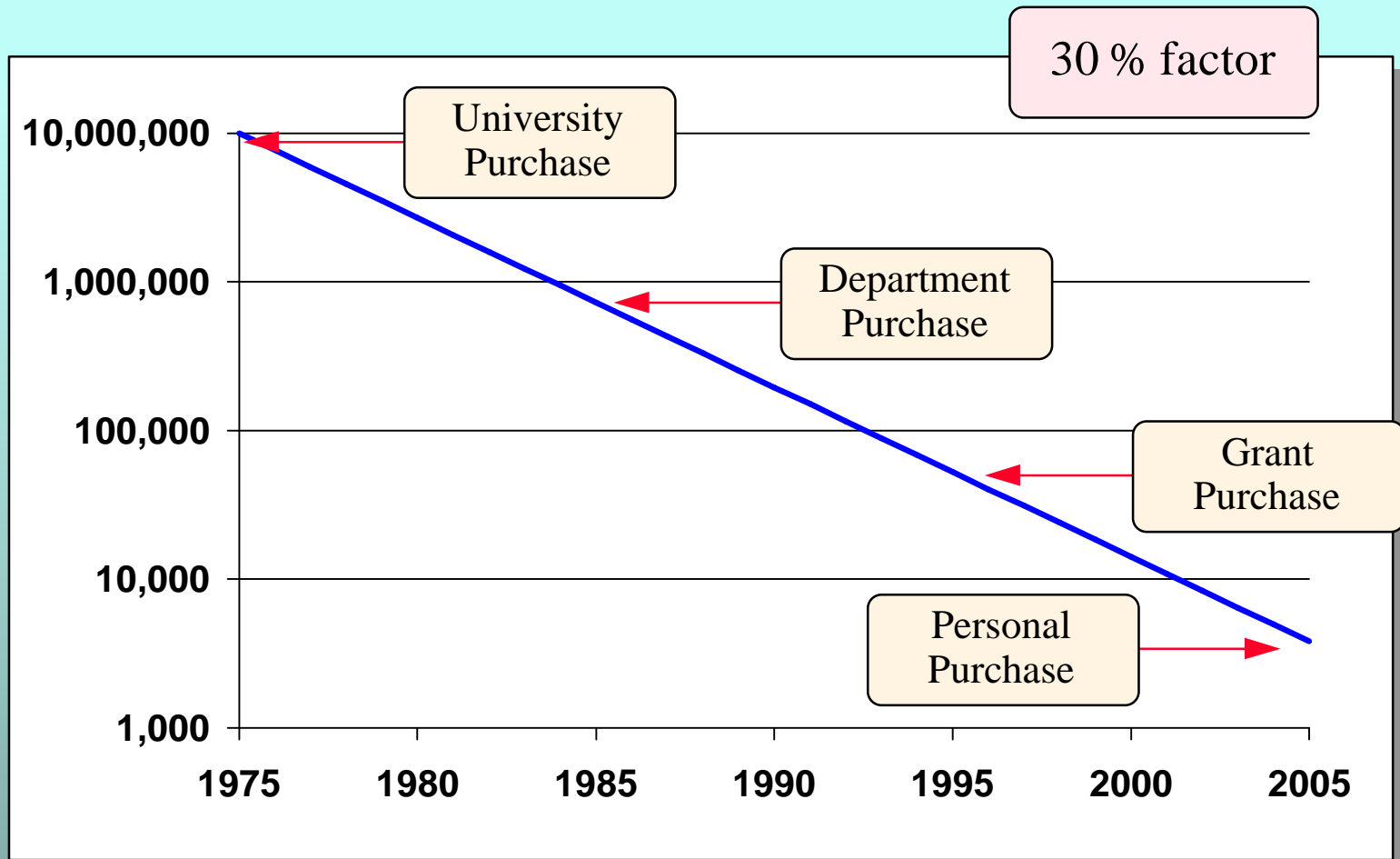
Performance (at constant cost)



Cost (at constant performance)



Cost (at constant performance)



Biology is Special

Biology is Special

For it is in relation to the statistical point of view that the structure of the vital parts of living organisms differs so entirely from that of any piece of matter that we physicists and chemists have ever handled in our laboratories or mentally at our writing desks.

Erwin Schroedinger. 1944. *What is Life*.

Genetics as Code

[The] chromosomes ... contain in some kind of code-script the entire pattern of the individual's future development and of its functioning in the mature state. ... [By] code-script we mean that the all-penetrating mind, once conceived by Laplace, to which every causal connection lay immediately open, could tell from their structure whether [an egg carrying them] would develop, under suitable conditions, into a black cock or into a speckled hen, into a fly or a maize plant, a rhodo-dendron, a beetle, a mouse, or a woman.

Erwin Schroedinger. 1944. *What is Life*.

Genomics As an Example

Infrastructure and the HGP

Progress towards all of the [Genome Project] goals will require the establishment of well-funded centralized facilities, including a stock center for the cloned DNA fragments generated in the mapping and sequencing effort and a data center for the computer-based collection and distribution of large amounts of DNA sequence information.

National Research Council. 1988. *Mapping and Sequencing the Human Genome*. Washington, DC: National Academy Press. p. 3

Databases and the Genome Project

[The] database developer should provide, in some real sense, an intellectual focus for the interpretation of genomic data.

NIH-DOE Ad Hoc Committee on Genome Databases

Goals of the Genome Project

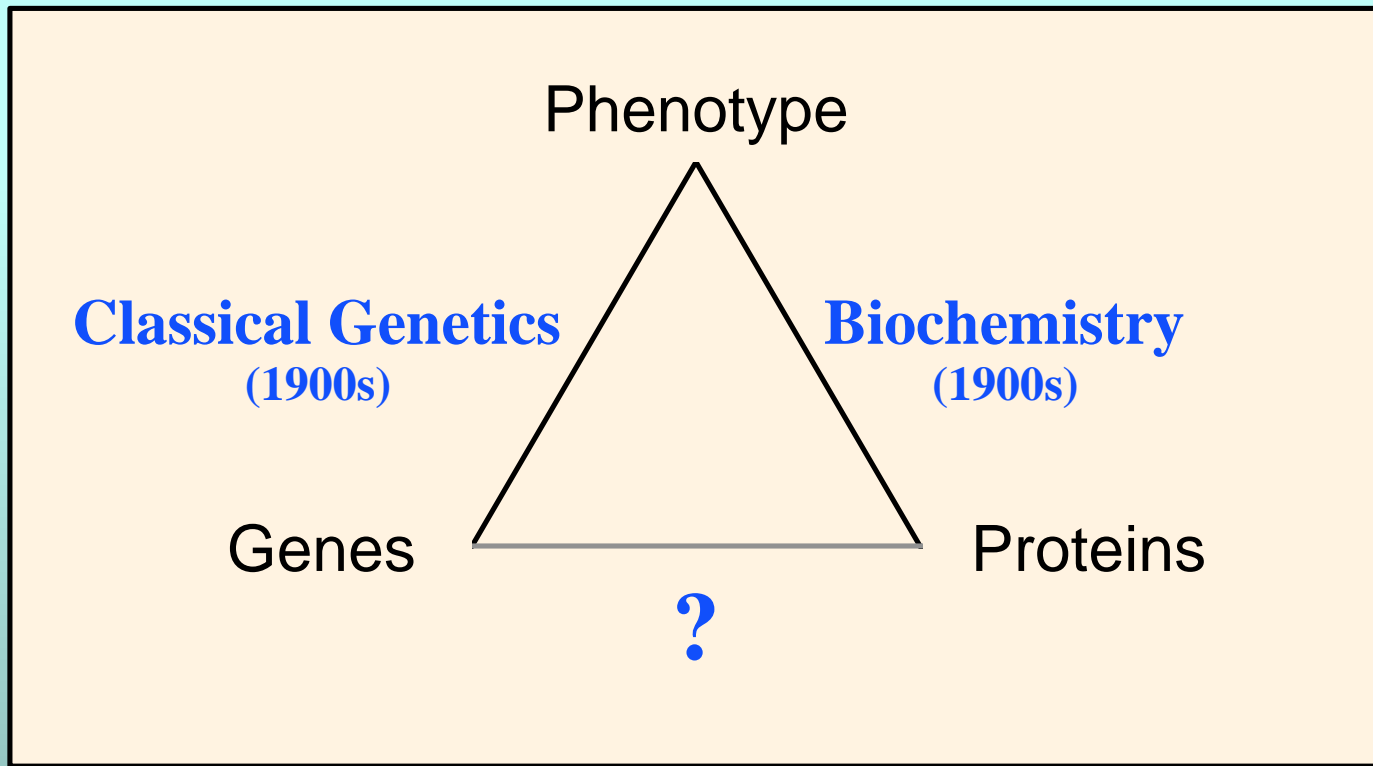
HGP - Overall Goals

- construction of a high-resolution genetic map of the human genome;
- production of a variety of physical maps of all human chromosomes and of the DNA of selected model organisms;
- determination of the complete sequence of human DNA and of the DNA of selected model organisms;
- development of capabilities for collecting, storing, distributing, and analyzing the data produced;
- creation of appropriate technologies necessary to achieve these objectives.

USDOE. 1990. *Understanding Our Genetic Inheritance. The U.S. Human Genome Project: The First Five Years.*

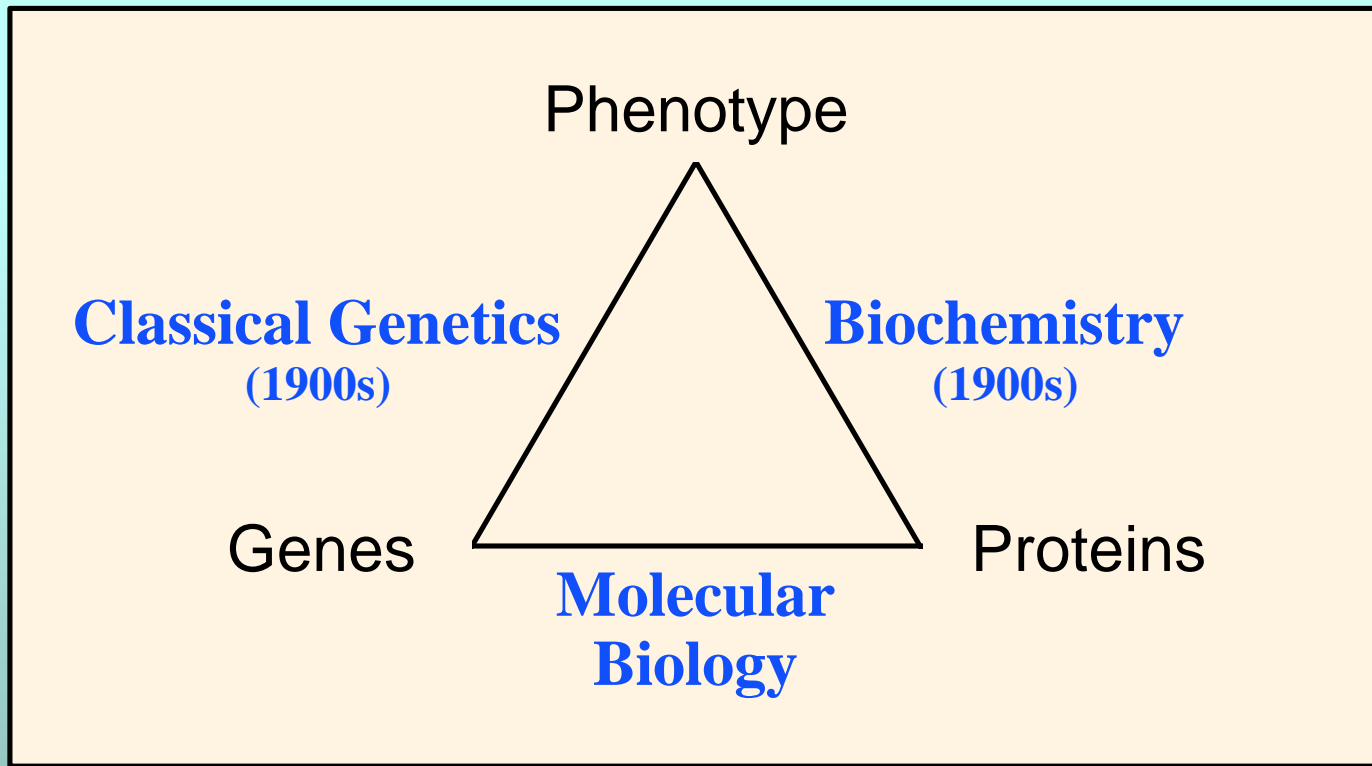
Biological Background

The Origins of Molecular Biology



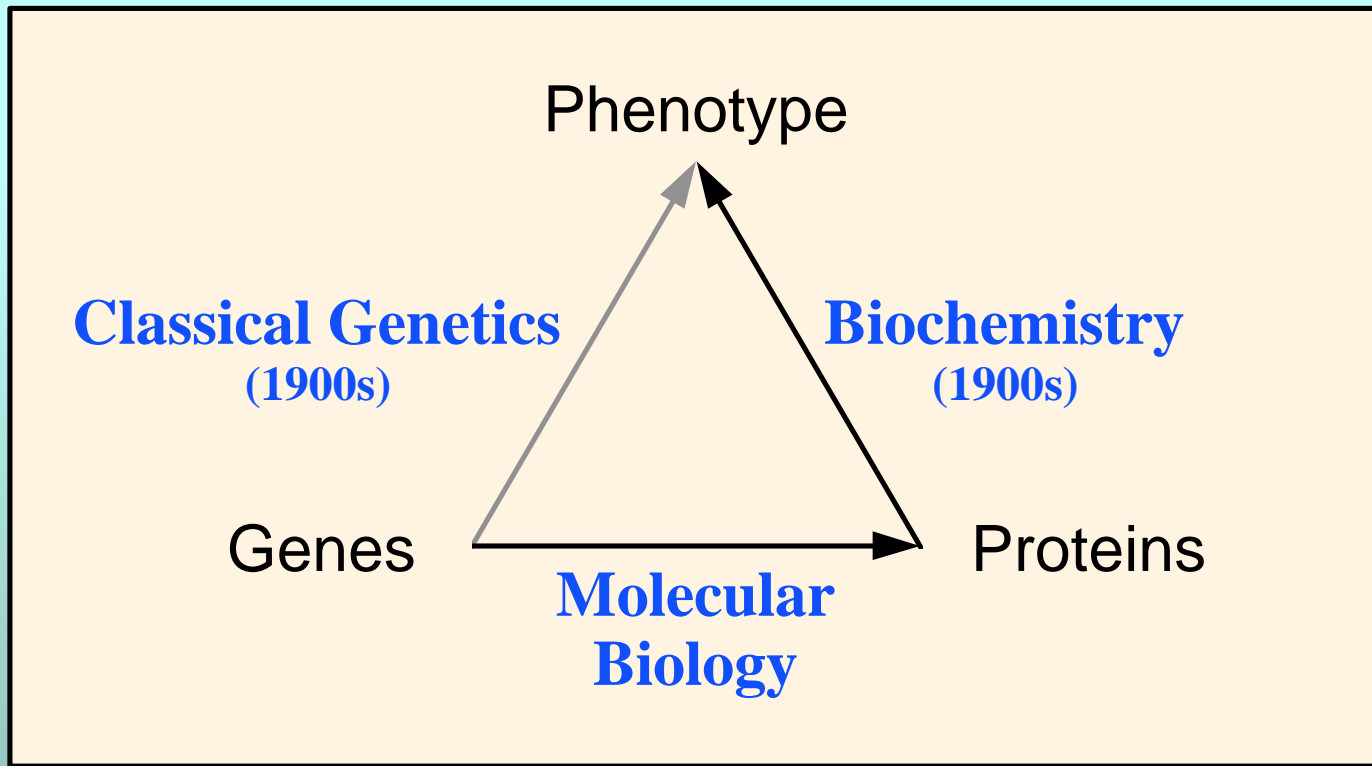
The *phenotype* of an organism denotes its external appearance (size, color, intelligence, etc.). *Classical genetics* showed that genes control the transmission of phenotype from one generation to the next. *Biochemistry* showed that within one generation, *proteins* had a determining effect on phenotype. For many years, however, the relationship between genes and proteins was a mystery.

The Origins of Molecular Biology



Then, it was found that genes contain digitally encoded instructions that direct the synthesis of proteins. The crucial insight of *molecular biology* is that hereditary information is passed from parent to progeny in a form that is truly, not just metaphorically, digital. Understanding how that digital code directs the processes of life is the goal of molecular biology.

The Origins of Molecular Biology

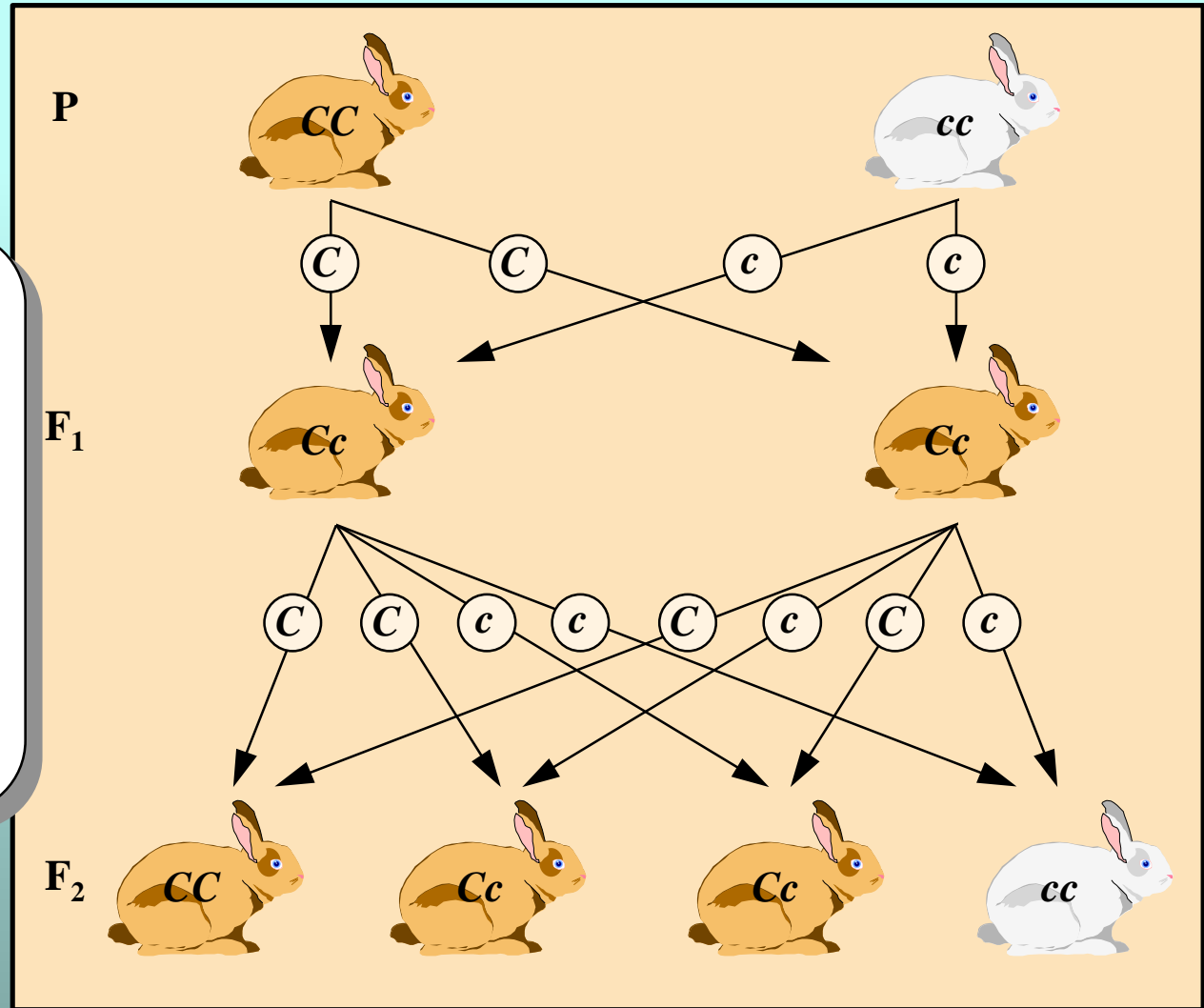


Modern molecular biology recognizes that genes control phenotypes indirectly, acting directly through control over the process of *DNA directed protein synthesis*.

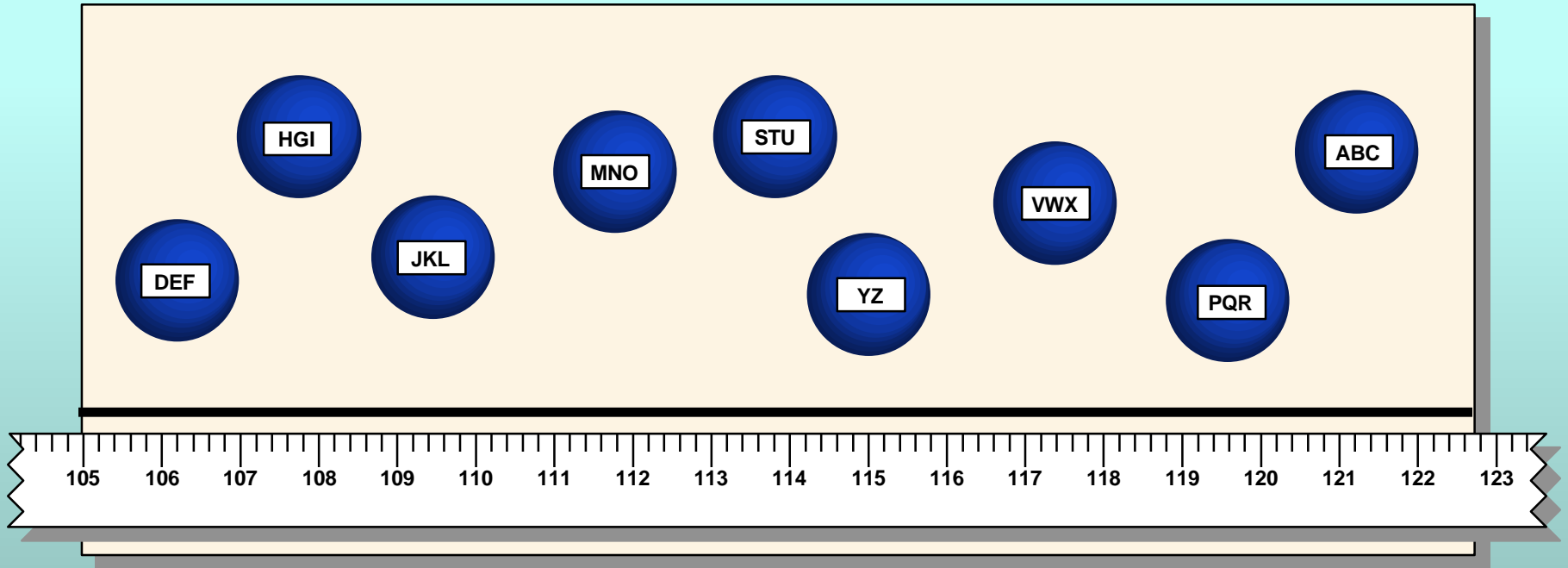
Classical Genetics

Mendel's Work

Regular numerical patterns of inheritance showed that the passage of traits from one generation to the next could be explained with the notion that hypothetical particles, or *genes*, were carried in pairs in adults, but transmitted singly to progeny.

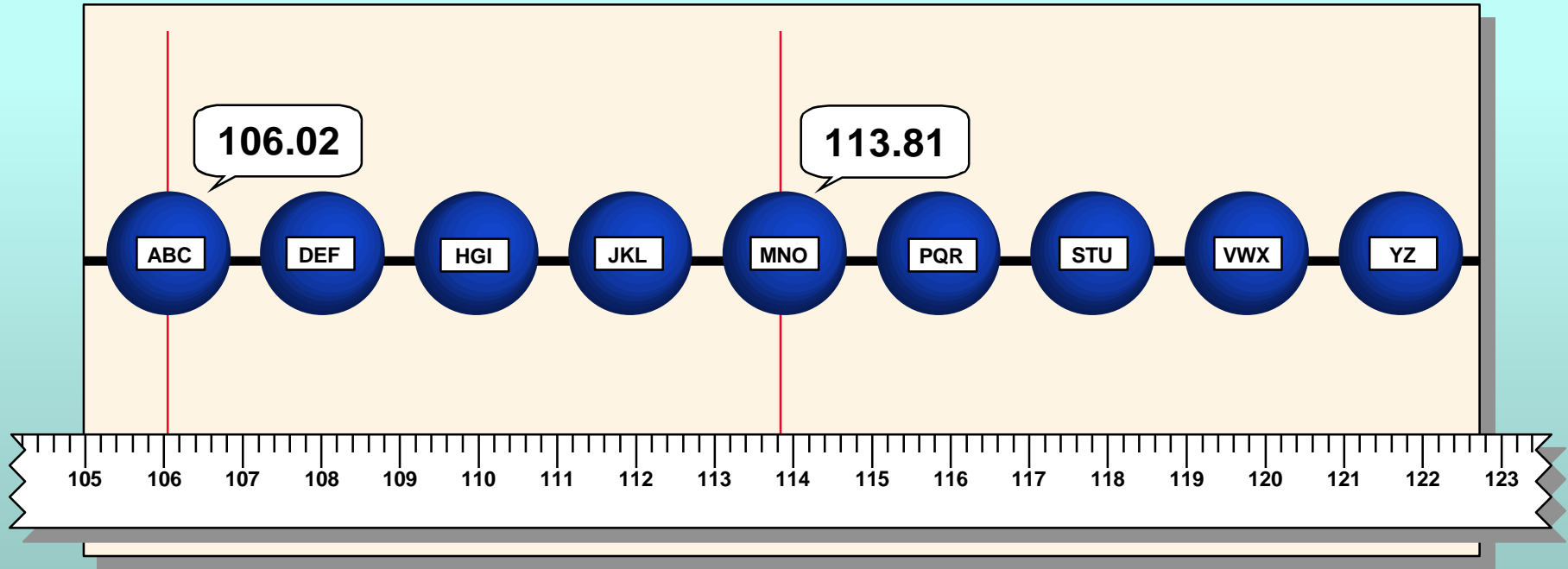


Classical Genetics



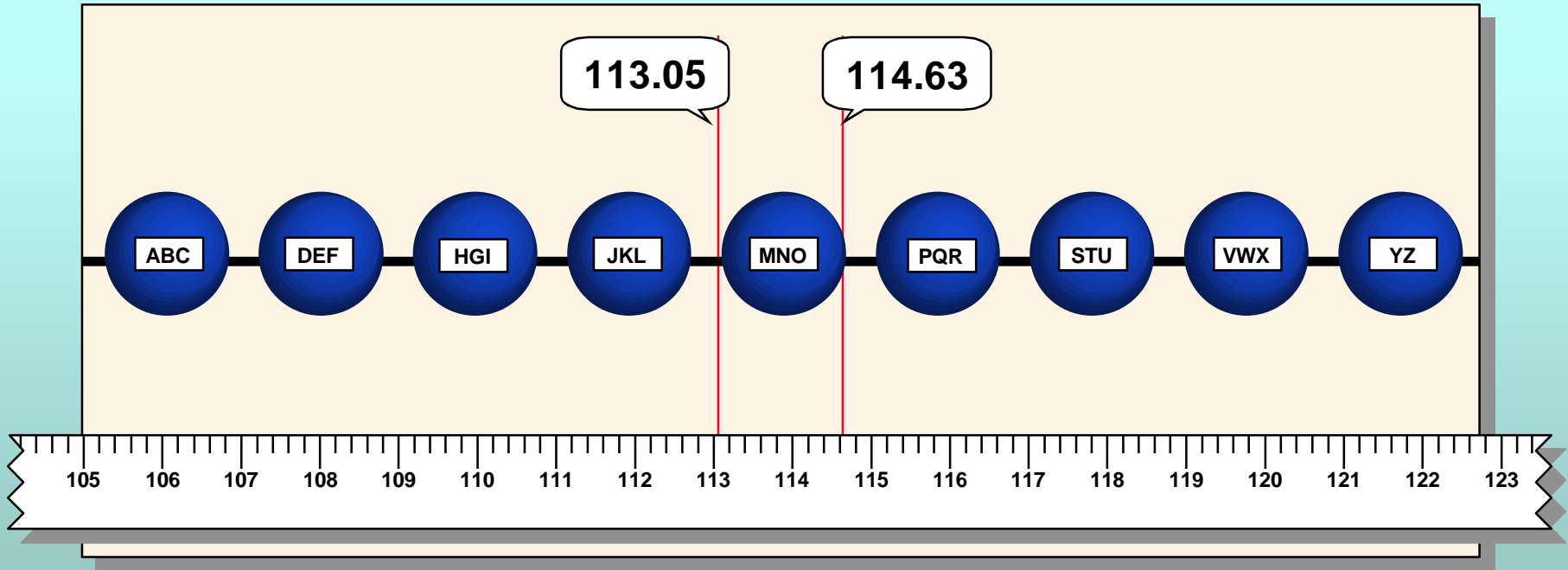
The beads can be conceptually separated from the string, which has “addresses” that are independent of the beads.

Classical Genetics



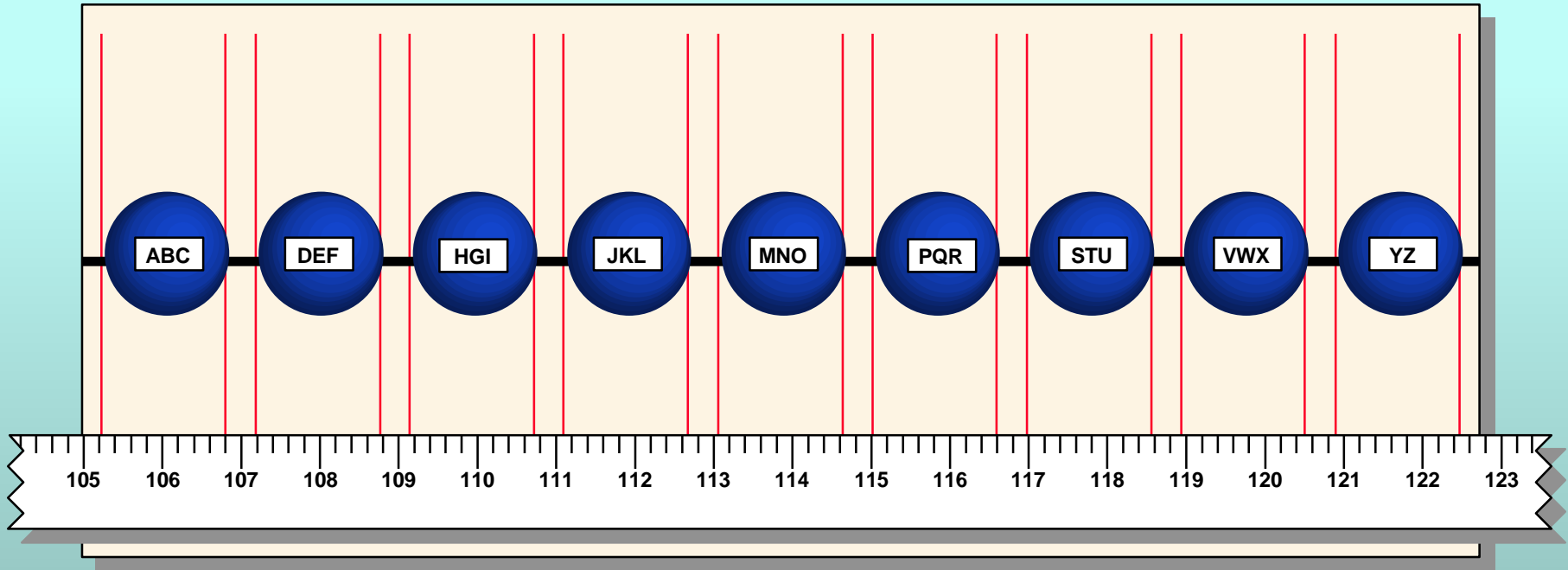
Mapping involves placing the beads in the correct order and assigning a correct address to each bead. The address assigned to a bead is its locus.

Classical Genetics



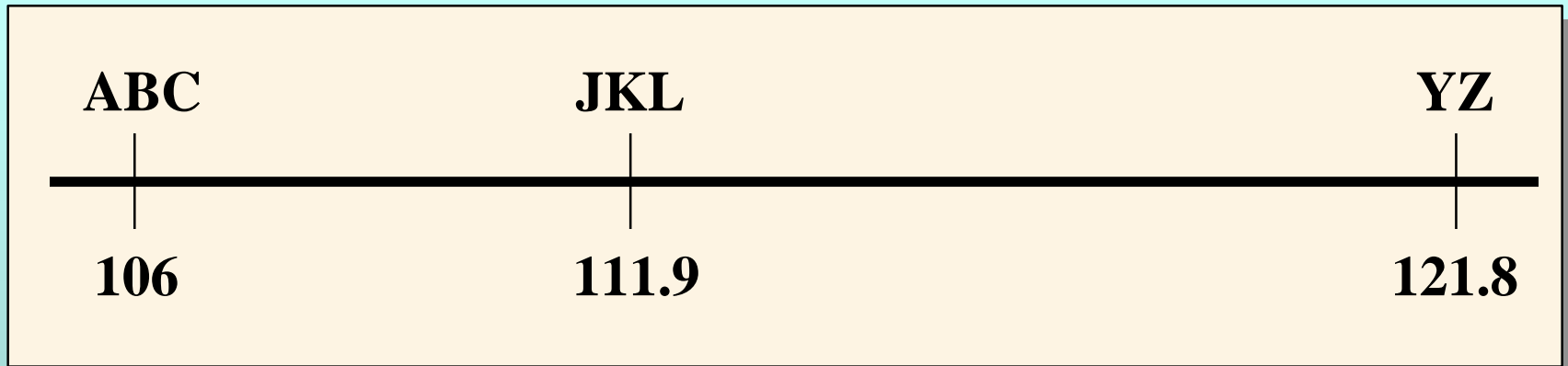
Recognizing that the beads have width, mapping could be extended to assigning a pair of numbers to each bead so that a locus is defined as a region, not a point.

Classical Genetics



In this model, genes are independent, mutually exclusive, non-overlapping entities, each with its own absolute address.

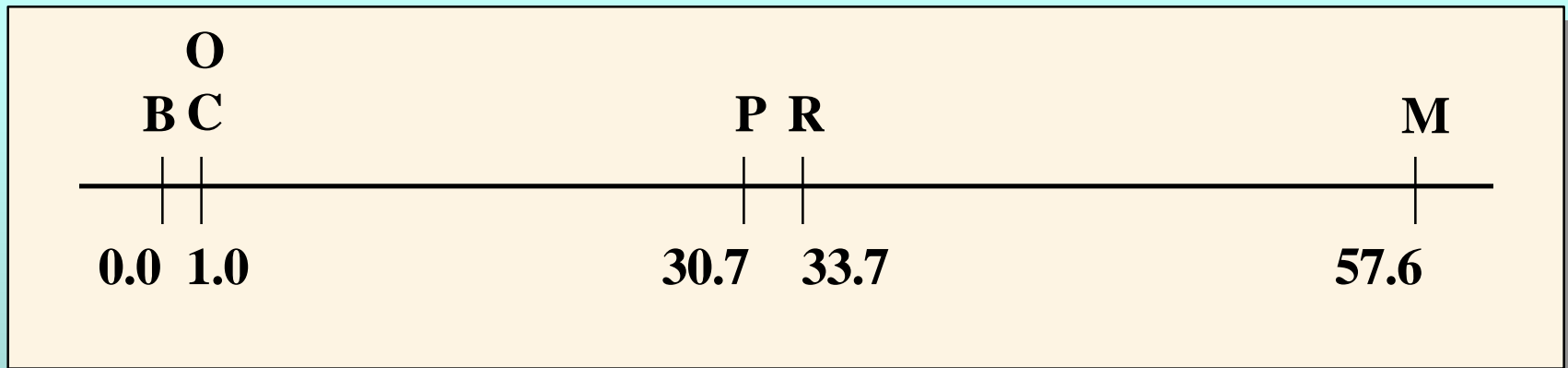
Classical Genetics



In principle, maps of a few genes might be represented by showing the gene names in order, with their relative positions indicated.

Classical Genetics

Drosophila melanogaster



B = yellow body

P = vermilion eye

M = miniature wing

C = white eye

R = rudimentary wing

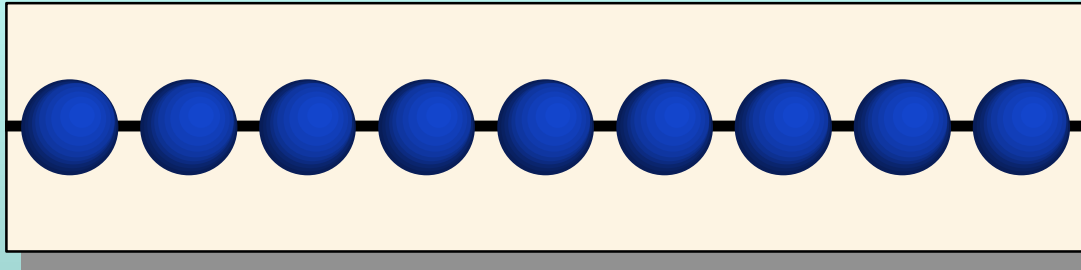
O = eosin eye

And, in fact, the first genetic map ever published was of just that type. Sturtevant, A.H., 1913, The linear arrangement of six sex-linked factors in *Drosophila* as shown by their mode of association, *Journal of Experimental Zoology* , 14:43-59.

Classical Genetics

The genes are arranged in a manner similar to beads strung on a loose string.

Sturtevant, A.H., and Beadle, G.W., 1939. *An Introduction to Genetics*. W. B. Saunders Company, Philadelphia, p. 94.



During the first half of this century, classical investigation of the gene established that theoretical objects called genes were the fundamental units of heredity. According to the classical model of the gene:

Genes behave in inheritance as independent particles.

Genes are carried in a linear arrangement in the chromosome, where they occupy stable positions.

Genes recombine as discrete units.

Genes can mutate to stable new forms.

Basically, genes seemed to be particulate objects, arranged on the chromosome like “beads on a string.”

The Classical Gene

Genes behave in inheritance as independent particles.

Genes are carried in a linear arrangement in the chromosome, where they occupy stable positions.

Genes recombine as discrete units.

Genes can mutate to stable new forms.

Biochemistry

Biochemistry

The aim of modern biology is to interpret the properties of the organism by the structure of its constituent molecules.

Jacob, F. 1973. *The Logic of Life*. New York: Pantheon Books.

Understanding the molecular basis of life had its beginnings with the advent of biochemistry. Early in the nineteenth century, it was discovered that preparations of fibrous material could be obtained from cell extracts of plants and animals. Mulder concluded in 1838 that this material was:

without doubt the most important of the known components of living matter, and it would appear that without life would not be possible. This substance has been named *protein*.

Later, many wondered whether chemical processes in living systems obeyed the same laws as did chemistry elsewhere. Complex carbon-based compounds were readily synthesized in cells, but seemed impossible to construct in the laboratory.

By the beginning of the twentieth century, chemists had been able to synthesize a few organic compounds, and, more importantly, to demonstrate that complex organic reactions could be accomplished in non-living cellular extracts. These reactions were found to be catalyzed by a class of proteins called *enzymes*.

Early biochemistry, then, was characterized by (1) efforts to understand the structure and chemistry of proteins themselves, and (2) efforts to discover, catalog, and understand enzymatically catalyzed biochemical reactions.

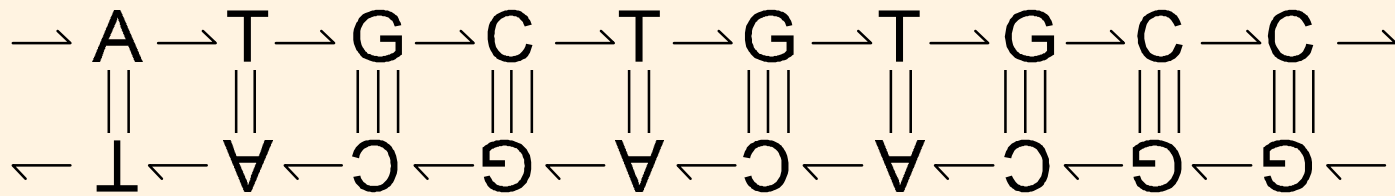
Molecular Biology

Origins of Molecular Biology

Key Discoveries:

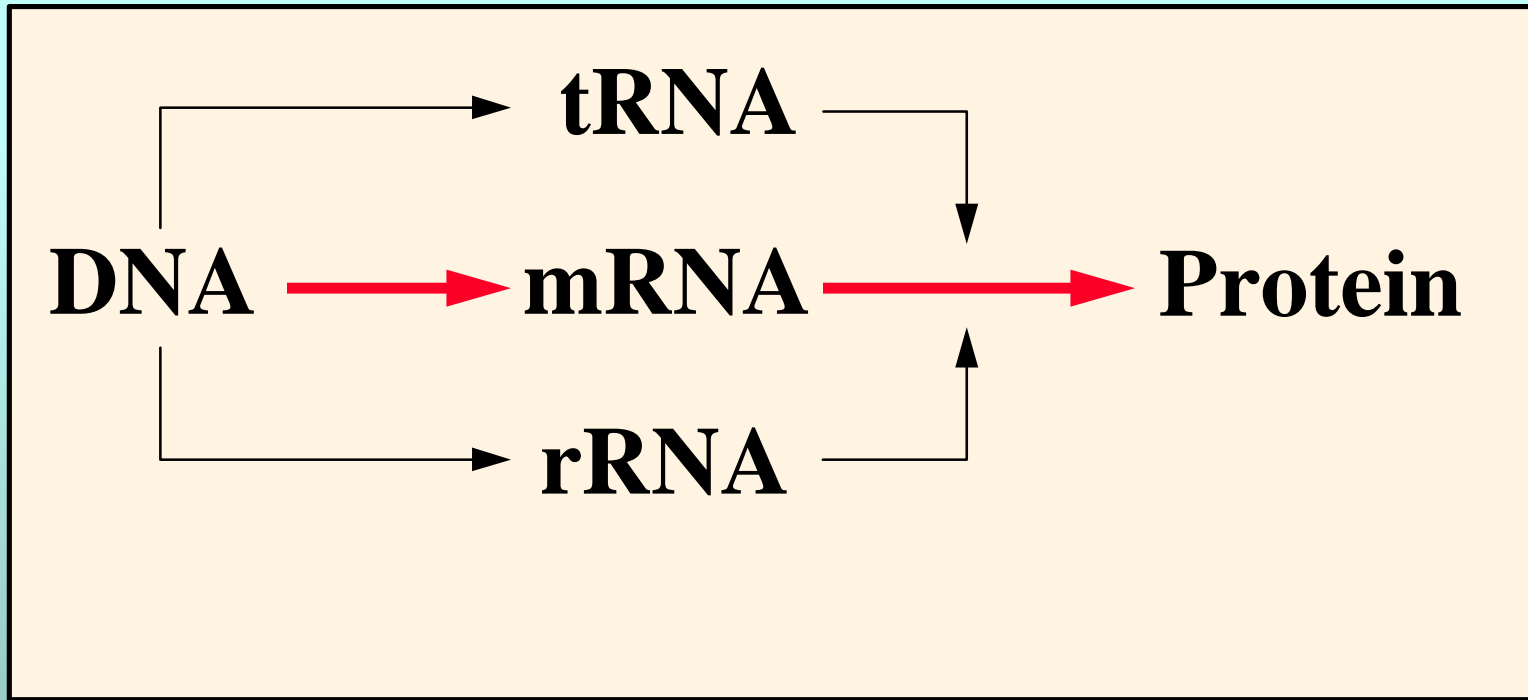
- 1928 Heritable changes can be transmitted from bacterium to bacterium through a chemical extract (the ***transforming factor***) taken from other bacteria.
- 1944 The transforming factor appears to be DNA.
- 1950 The tetranucleotide hypothesis of DNA structure is overthrown.
- 1953 The structure of DNA is established to be a double helix.

Molecular Biology



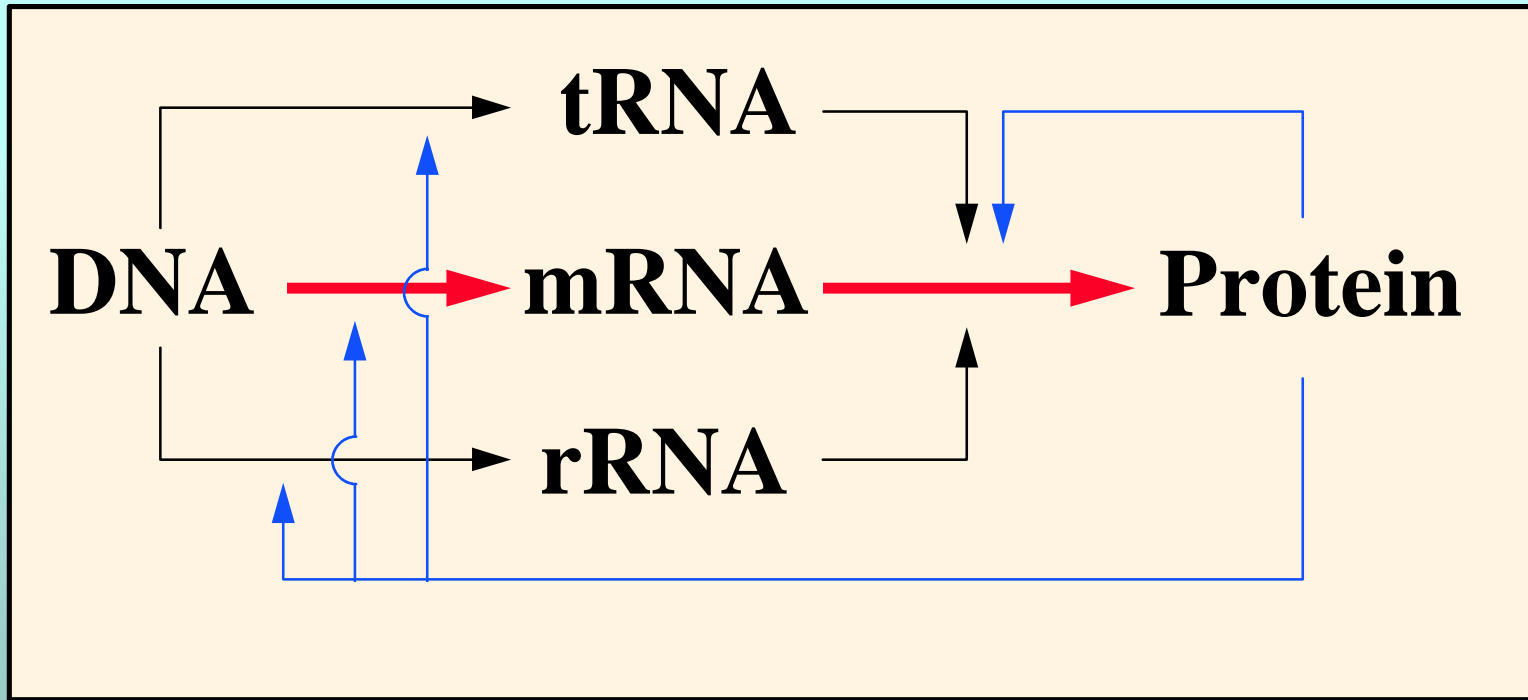
DNA is constructed as a double-stranded molecule, with absolutely no constraints upon the linear order of subcomponents along each strand, but with the pairing between strands totally constrained according to complementarity rules: A always pairs with T and C always pairs with G.

The Fundamental Dogma



DNA controls the synthesis of RNA which in turn directs the synthesis of protein.

The Fundamental Dogma



The whole system is recursive, in that certain proteins are required for the synthesis of RNAs, as well as for the synthesis of DNA itself.

mRNA to Amino Acid Dictionary

		U	C	A	G			
5'	U	phe	ser	tyr	cys	U	3'	
		phe	ser	tyr	cys			C
		leu	ser	STOP	STOP			A
		leu	ser	STOP	trp			G
	C	leu	pro	his	arg	U		
		leu	pro	his	arg			C
		leu	pro	gln	arg			A
		leu	pro	gln	arg			G
	A	ile	thr	asn	ser	U		
		ile	thr	asn	ser			C
		ile	thr	lys	arg			A
		met	thr	lys	arg			G
	G	val	ala	asp	gly	U		
		val	ala	asp	gly			C
		val	ala	glu	gly			A
		val	ala	glu	gly			G

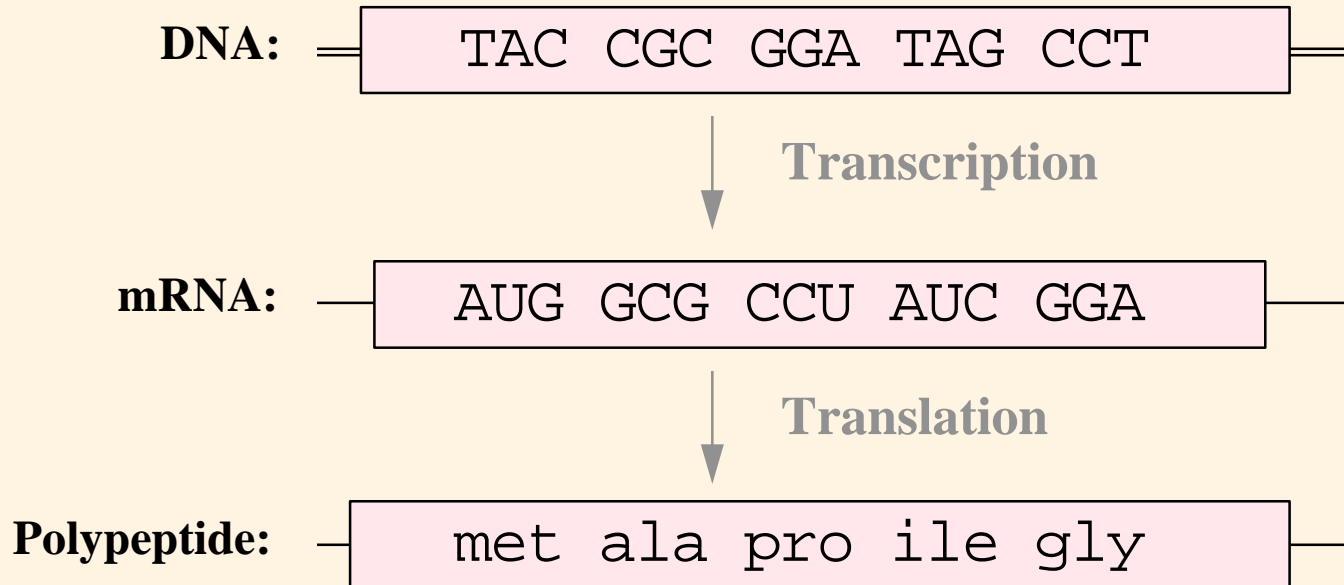
mRNA to Amino Acid Dictionary

This dictionary gives the sixty four different mRNA codons and the amino acids (or stop signals) for which they code. The 5' nucleotides are given along the left hand border, the middle nucleotides are given across the top, and the 3' nucleotides are given along the right hand border. The decoded meaning of a particular codon is given by the entry in the table.

For example, the meaning of the codon 5'AUG3' is determined as follows:

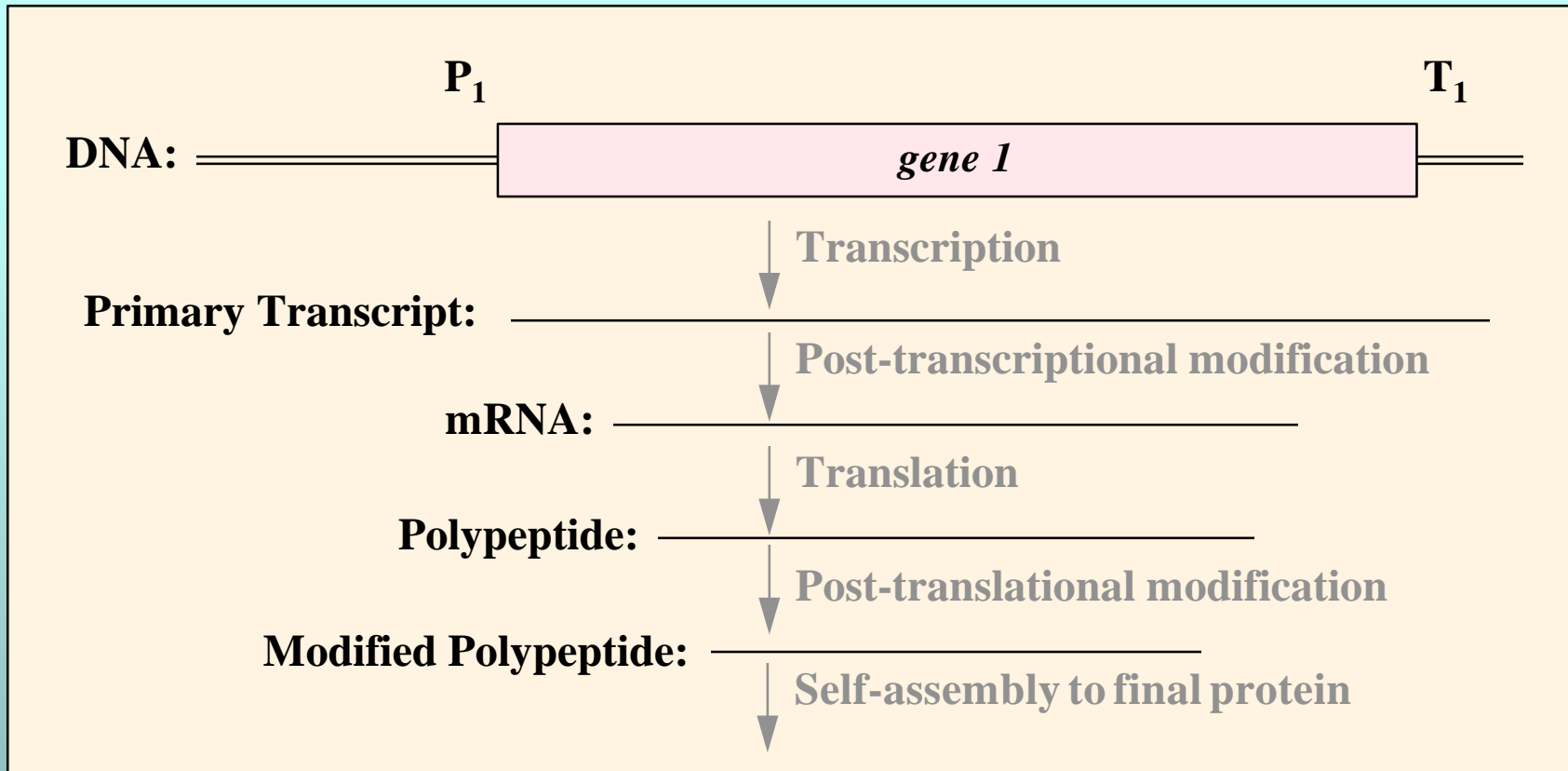
1. Examine the entries along the left hand side of the table to locate the horizontal block corresponding to the sixteen codons that have A in the 5' position.
2. Examine the entries along the top of the table to locate the vertical block corresponding to the sixteen codons that have U in the middle position.
3. Find the intersection of these two blocks. This intersection represents the four codons that have A in the 5' position and U in the middle position.
4. Examine the entries along the right hand side of the table to find the entry for the one codon that has A in the 5' position, U in the middle position, and G in the 3' position. The “met” indicates that the decoded meaning of the codon 5'AUG3' is methionine. That is, the codon 5'AUG3' codes for the amino acid methionine.

DNA Directed Protein Synthesis



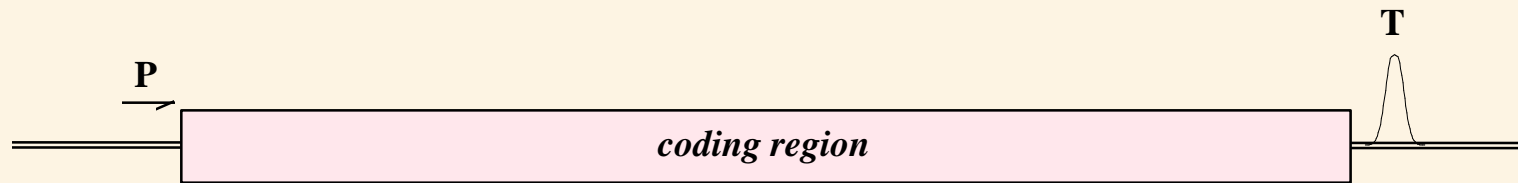
DNA directs protein synthesis through a multi-step process. First, DNA is copied to mRNA through the process of transcription. The rules governing transcription are the same as the rules governing the interstrand constraint in DNA. Then translation produces a polypeptide with an amino-acid sequence that is completely specified by the sequence of nucleotides in the RNA. A simple code, the same for all living things on this planet, governs the synthesis of protein from mRNA instructions.

DNA Directed Protein Synthesis



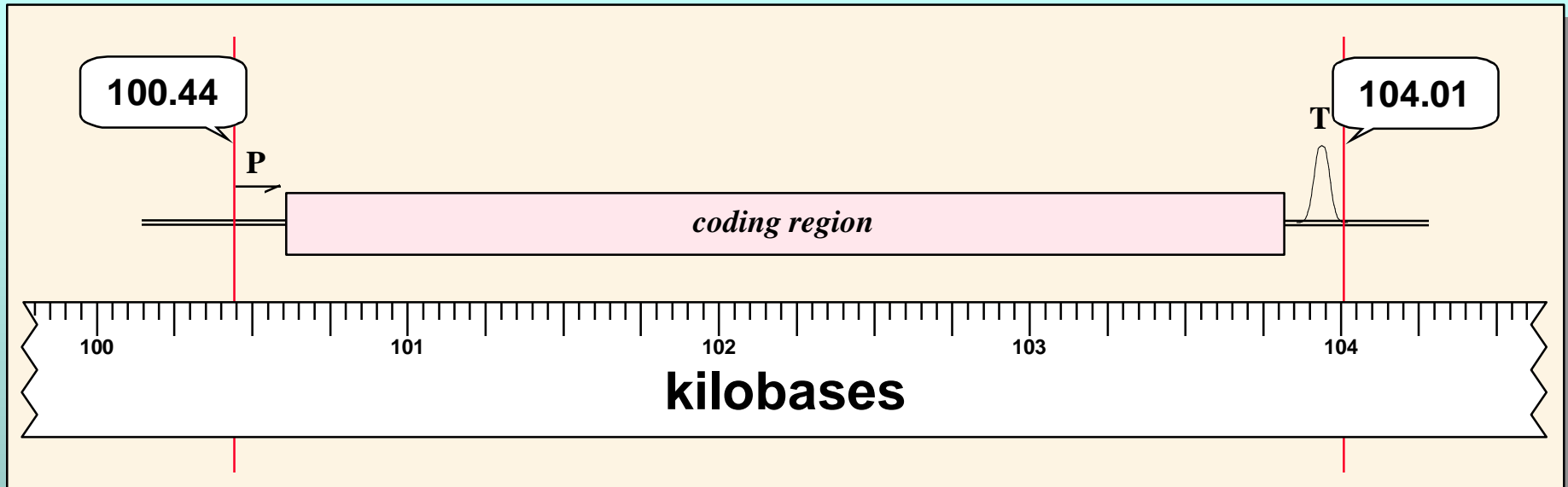
Some post-transcriptional processing of the immediate RNA transcript is necessary to produce a finished RNA, and post-translational processing of polypeptides can be needed to produce a final protein.

The (Simplistic) Molecular View of a Gene



A gene is a transcribed region of DNA, flanked by upstream start regulatory sequences and downstream stop regulatory sequences.

The (Simplistic) Molecular View of a Gene



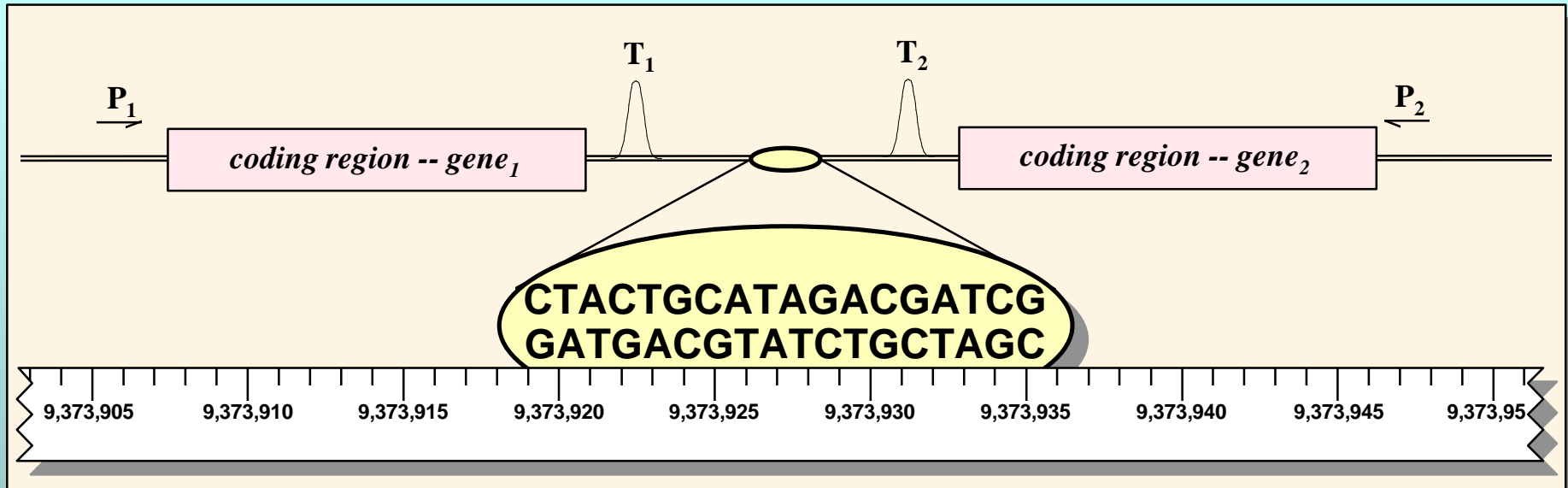
The location of a gene can be designated by specifying the base-pair location of its beginning and end.

The (Simplistic) Molecular View of a Gene



DNA may be transcribed in either direction. Therefore, fully specifying a gene's position requires noting its orientation as well as its start and stop positions.

The (Simplistic) Molecular View of a Gene



A naive view holds that a genome can be represented as a continuous linear string of nucleotides, with landmarks identified by the chromosome number followed by the offset number of the nucleotide at the beginning and end of the region of interest. This simplistic approach ignores the fact that human chromosomes may vary in length by tens of millions of nucleotides.

Restated Genome Project Goals

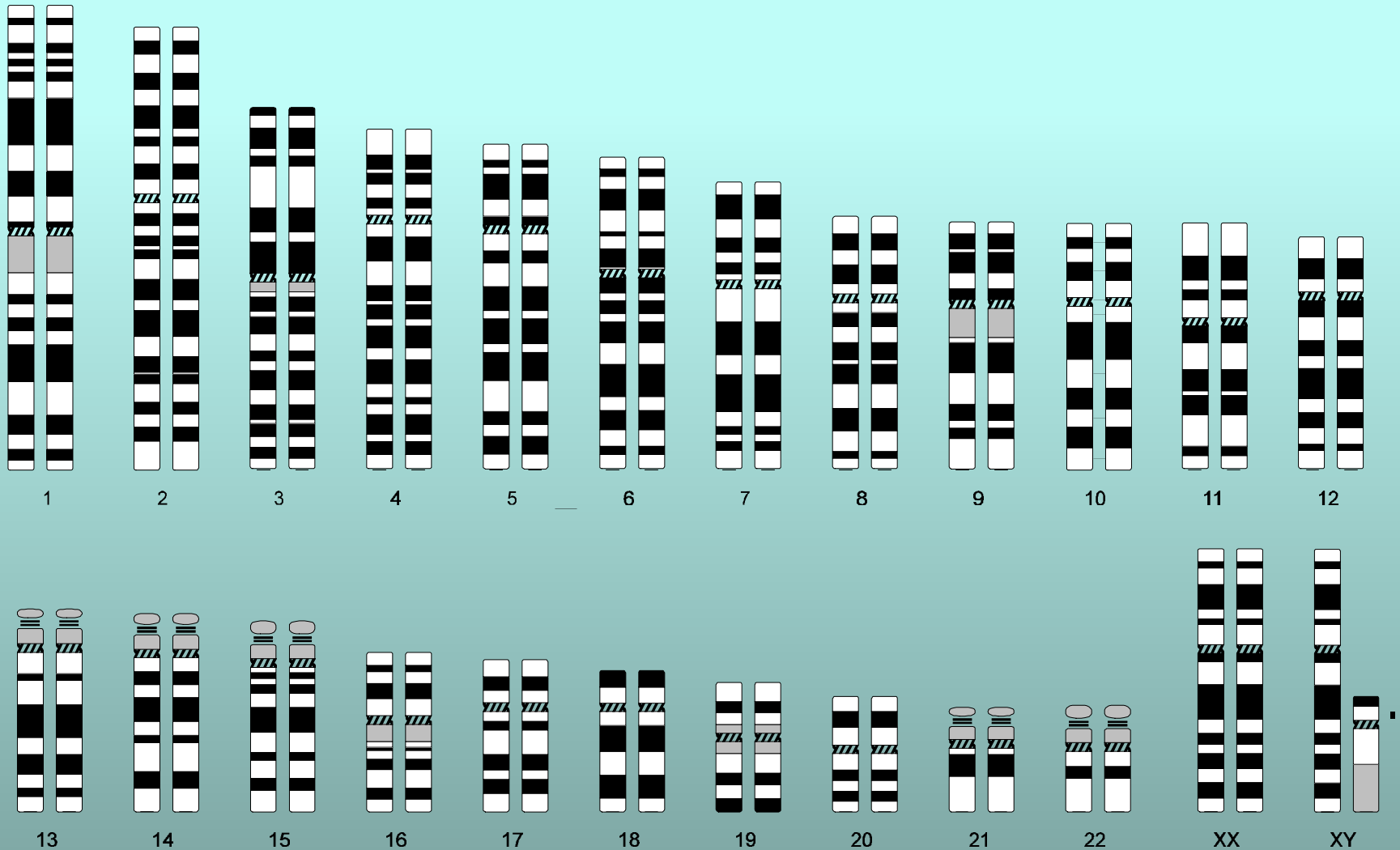
The Human Genome Project

The human genome is believed to consist of 50,000 to 100,000 genes encoded in 3.3 billion base pairs of DNA, which are packaged into 23 chromosomes.

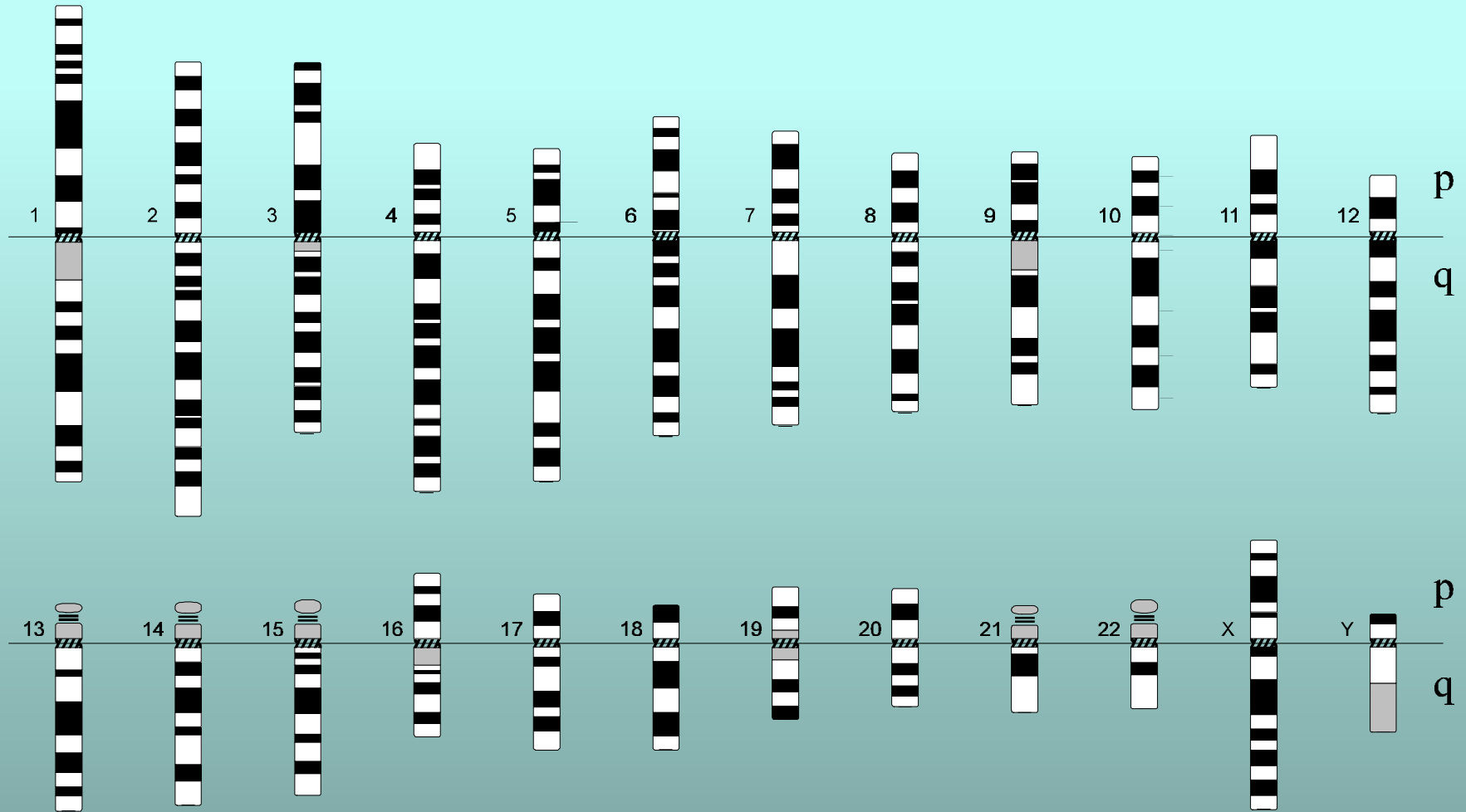
The goal of the Human Genome Project is learning the specific order of those 3.3 billion base pairs and of identifying and locating all of the genes encoded by that DNA.

Basic Genomics

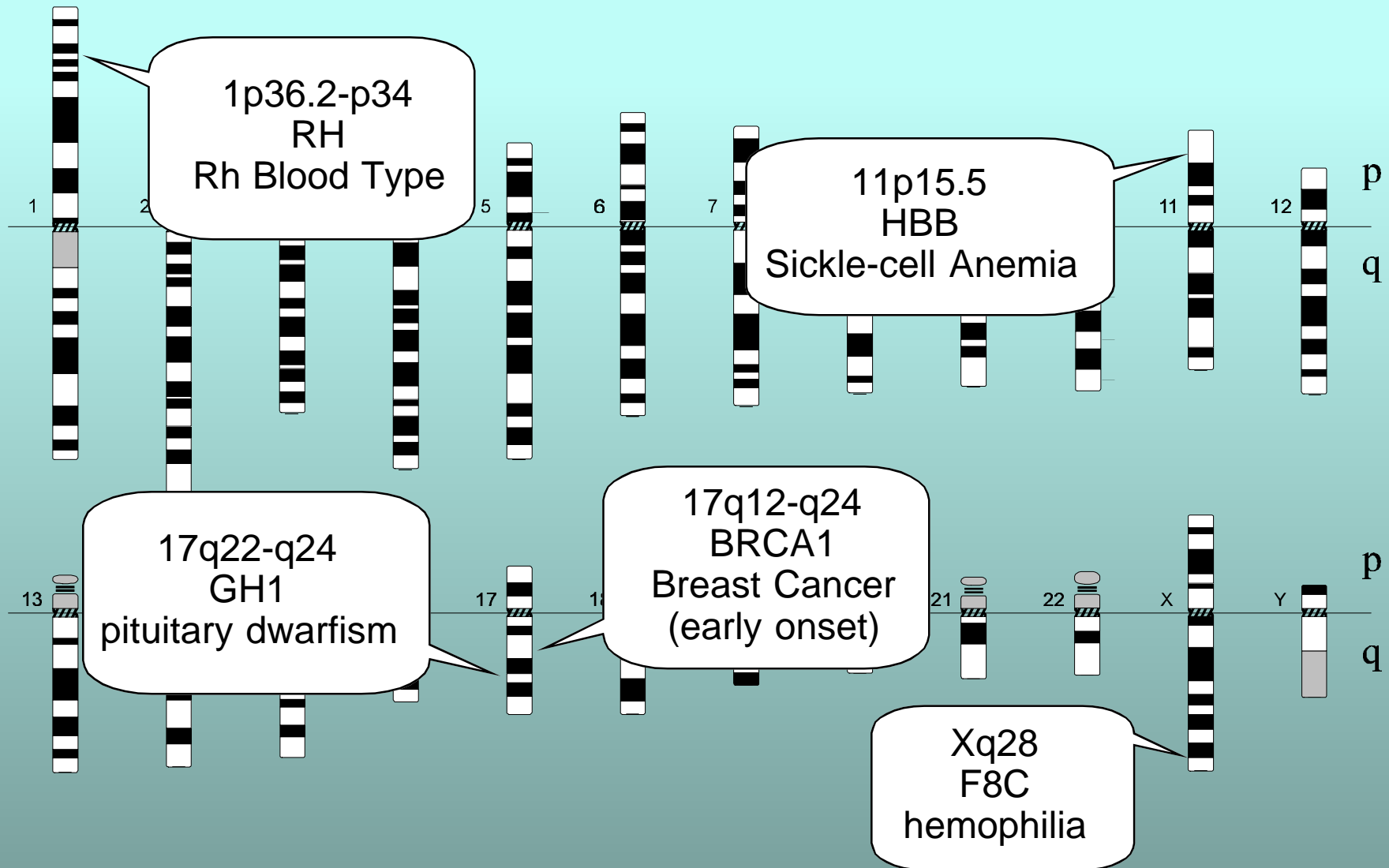
Human Chromosomes



Human Chromosomes



Human Chromosomes

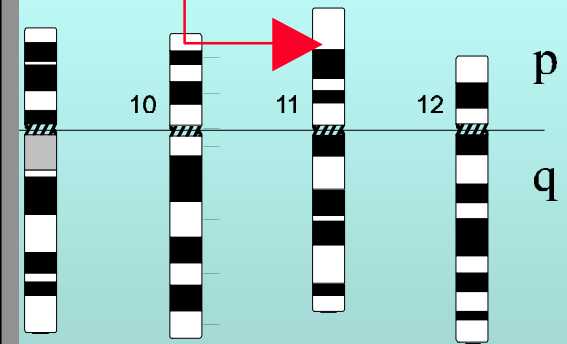


Beta Hemoglobin

```

1 ccctgtggag ccacacccta gggttggcca atctactccc aggagcaggg agggcaggag
61 ccagggctgg gcataaaaagt cagggcagag ccatctattg ctt  acatttg cttctgacac
121 aactgtgttc actagcaacc tcaaacagac accATGGTGC ACCTGACTCC TGAGGAGAAG
181 TCTGCCGTTA CTGCCCTGTG GGGCAAGGTG AACGTGGATG AAGTTGGTGG TGAGGCCCTG
241 GGCAGGttgg tatcaaggtt acaagacagg ttaaggaga ccaatagaaa ctgggcatgt
301 ggagacagag aagactcttg ggtttctgat aggcactgac tctctctgcc tattggteta
361 ttttcccacc cttagg CTGC TGGTGGTCTA CCCTGGACC CAGAGGTTCT TTGAGTCCTT
421 TGGGGATCTG TCCACTCCTG ATGCTGTTAT GGGCAACCCT AAGGTGAAGG CTCATGGCAA
481 GAAAGTGCTC GGTGCCTTTA GTGATGGCCT GGCTCACCTG GACAACCTCA AGGGCACCTT
541 TGCCCACTG AGTGAGCTGC ACTGTGACAA GCTGCACGTG GATCCTGAGA ACTTCAGGt
601 gagtctatgg gacccttgat gttttctttc cctttctttt ctatggttaa gttcatgtca
661 taggaagggg agaagtaaca gggtagactt tagaatggga aacagacgaa tgattgcatc
721 agtgtggaag tctcaggatc gttttagttt cttttatttg ctgttcataa caattgtttt
781 cttttgttta attcttgctt tctttttttt tctttctcgc aatttttact attatactta
841 atgccttaac attgtgtata acaaaaggaa atatctctga gatacattaa gtaacttaa
901 aaaaaacttt acacagtctg cctagtagat tactatttgg aatatatgtg tgcttatttg
961 catattcata atctccctac tttattttct tttattttta attgatacat aatcattata
1021 catatthtat ggttaaagtg taatgtttta atatgtgtac acatattgac caaatcaggg
1081 taatthtgca tttgtaattt taaaaaatgc tttcttcttt taatatactt ttttgthtat
1141 cttatthcta atactthccc taatctcttt ctttcagggc aataatgata caatgtatca
1201 tgcctctthg caccattcta aagaataaca gtgataaatt ctgggthtaag gcaatagcaa
1261 taththtgca tataaatatt tctgcatata aathgthact gatgthagag gththcatatt
1321 gthaatagca gthacaatcc agthaccatt ctgctthtat ththagthtg gthaataggt
1381 gthaththct gththcaagc ththgcttht ththgthcat gththcatct cththctthc
1441 ththcag CT CCTGGGCAAC GTGCTGGTCT GTGTGCTGGC CCATCACTTT GGCAAGAAT
1501 TCACCCACC AGTGCAGGCT GCCTATCAGA AAGTGGTGGC TGGTGTGGCT AATGCCCTG
1561 CCCACAAGTA TCACTAAgct cgththcttg ctgtccaatt tctathaaag gthctththg
1621 thctthagtc caactactaa actgggggat aththgaggg gcctthgagca ththgathth
1681 gctthataaa aacaththt ththctthg atgaththt thaththth thgaththt
1741 ththataaag gthaththg agthctgthc ththataaaca thaththth gthgctgth
1801 caaacctthg gthaththac ththctthaa actthctgaa agthgthgag gthgcaacca
1861 gthaththc attgthcaaca gctthctgth cththgctth ththctthct cagthaththg
1921 thctththgag gctthgathth gthgththaa gthththgth gthgththth acaththth
1981 thgthththg ththctthct aaththctth ththctthct thgththct thgthctthc
2041 thgctththg ct

```



If we could zoom in on the HBB gene on chromosome 11, we could see the DNA sequence for beta-hemoglobin.

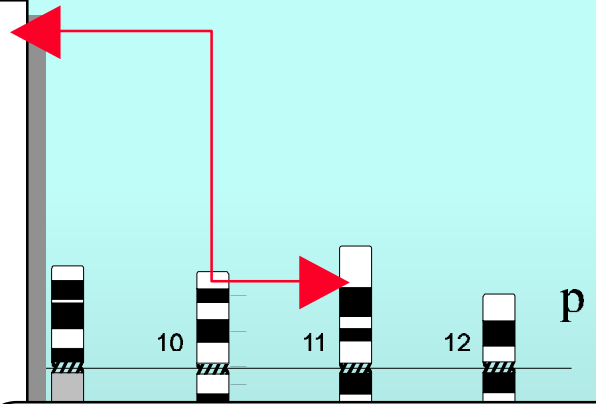


Beta Hemoglobin

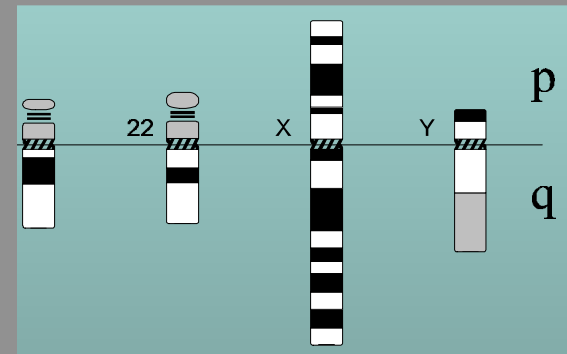
```

1 ccctgtggag ccacacccta gggttggcca atctactccc agggcaggag agggcaggag
61 ccagggctgg gcataaaagt cagggcagag ccatctattg cttacatttg cttctgacac
121 aactgtgttc actagcaacc tcaaacagac accATGGTGC ACCTGACTCC TGAGGAGAAG
181 TCTGCCGTTA CTGCCCTGTG GGGCAAGGTG AACGTGGATG AAGTTGGTGG TGAGGCCCTG
241 GGCAGGttgg tatcaaggtt acaagacagg ttaagagaga ccaatagaaa ctgggcatgt
301 ggagacagag aagactcttg ggtttctgat aggcactgac tctctctgcc tattggtcta
361 ttttcccacc cttagg CTGC TGGTGGTCTA CCCTTGGACC CAGAGGTTCT TTGAGTCCTT
421 TGGGGATCTG TCCACTCCTG ATGCTGTTAT GGGCAACCCT AAGGTGAAGG CTCATGGCAA
481 GAAAGTGCTC GGTGCCTTTA GTGATGGCCT GGCTCACCTG GACAACCTCA AGGGCACCTT
541 TGCCACACTG AGTGAGCTGC ACTGTGACAA GCTGCACGTG GATCCTGAGA ACTTCAGgt
601 gagtctatgg gacccttgat gttttctttc cctttctttt ctatggtaa gttcatgtca
661 taggaagggg agaagtaaca gggtagactt tagaatggga aacagacgaa tgattgcatc
721 agtgtggaag tctcaggatc gttttagttt cttttatttg ctgttcataa caattgtttt
781 cttttgttta attcttgctt tctttttttt tctttctcgc aatttttact attatactta
841 atgccttaac attgtgtata acaaaaggaa atatctctga gatacattaa gtaacttaa
901 aaaaaacttt acacagtctg cctagtagat tactatttgg aatatatgty tgcttatttg
961 catattcata atctccctac tttattttct tttattttta attgatacat aatcattata
1021 catatttatg ggtaaagtg taatgtttta atatgtgtac acatattgac caaatcaggg
1081 taattttgca tttgtaattt taaaaaatgc tttcttcttt taatatactt tttgtttat
1141 cttatttcta atactttccc taatctcttt ctttcagggc aataatgata caatgtatca
1201 tgcctctttg caccattcta aagaataaca gtgataaatt ctgggntaag gcaatagcaa
1261 tattttctgca tataaatatt tctgcatata aattgtaact gatgtaagag gtttcatatt
1321 gctaatagca gctacaatcc agctaccatt ctgcttttat tttatggttg ggataaggct
1381 ggattattct gagtccaagc taggcccttt tgctaatacat gttcatacct cttatcttcc
1441 tcccacag CT CCTGGGCAAC GTGCTGGTCT GTGTGCTGGC CCATCACTTT GCGAAAGAAT
1501 TCACCCCACC AGTGCAGGCT GCCTATCAGA AAGTGGTGGC TGGTGTGGCT AATGCCCTGG
1561 CCCACAAGTA TCACTAAGct cgctttcttg ctgtccaatt tctattaagag gttcctttgt
1621 tccctaagtc caactactaa actgggggat attatgaagg gccttgagca tctggattct
1681 gcctaataaa aacattttat tttcattga atgatgtatt taattattt ctgaatattt
1741 tactaaaaag ggaatgtggg aggtcagtgc attttaaaca taaagaatg atgagctggt
1801 caaaccttgg gaaaatacac tatatcttaa actccatgaa agaaggtgag gctgcaacca
1861 gctaatagcac attggcaaca gcccctgatg cctatgcctt attcatcctt cagaaaagga
1921 ttcttgtaga ggcttgattt gcagggttaa gttttgctat gctgtatttt acattactta
1981 ttgtttttagc tgcctcatg aatgtctttt cactacccat ttgcttatcc tgcactctc
2041 tcagccttga ct

```



The letters in red are the introns that are spliced together after initial transcription. The UPPER CASE RED letters are the actual coding region that specify the amino-acid sequence for beta-hemoglobin.

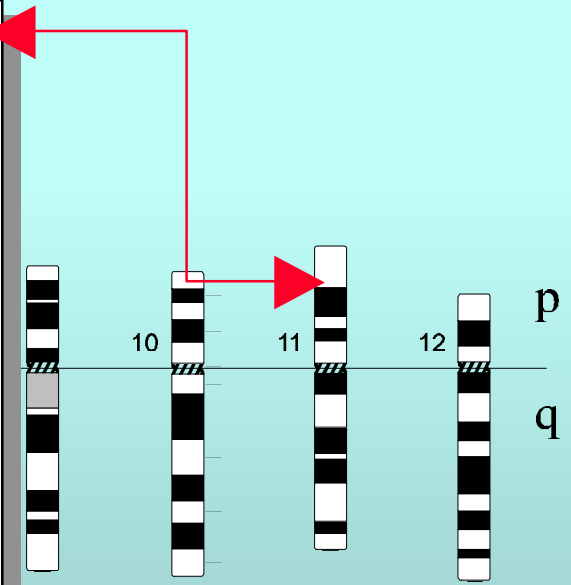


Beta Hemoglobin

```

1 ccctgtggag ccacacccta gggttggcca atctactccc aggagcaggg agggcaggag
61 ccagggctgg gcataaaagt cagggcagag ccatctattg cttacatttg cttctgacac
121 aactgtgttc actagcaacc tcaaacagac accATGGTGC ACCTGACTCC TGAGGAGAAG
181 TCTGCCGTTA CTGCCCTGTG GGGCAAGGTG AACGTGGATG AAGTTGGTGG TGAGGCCCTG
241 GGCAGGttgg tatcaaggtt acaagacagg ttaagagaga ccaatagaaa ctgggcatgt
301 ggagacagag aagactcttg ggtttctgat aggcactgac tctctctgcc tattggtcta
361 ttttcccacc cttagg CTGC TGGTGTCTA CCCTTGGACC CAGAGGTTCT TTGAGTCCTT
421 TGGGGATCTG TCCACTCCTG ATGCTGTTAT GGGCAACCCT AAGGTGAAGG CTCATGGCAA
481 GAAAGTGCTC GGTGCCTTTA GTGATGGCCT GGCTCACCTG GACAACCTCA AGGGCACCTT
541 TGCCACACTG AGTGAGCTGC ACTGTGACAA GCTGCACGTG GATCCTGAGA ACTTCAGgt
601 gagtctatgg gacccttgat gttttctttc cctttctttt ctatggtaa gttcatgtca
661 taggaagggg agaagtaaca gggtagactt tagaatggga aacagacgaa tgattgcatc
721 agtgtggaag tctcaggatc gttttagttt cttttatttg ctgttcataa caattgtttt
781 cttttgttta attcttgctt tctttttttt tctttctcgc aatttttact attatactta
841 atgccttaac attgtgtata acaaaaggaa atatctctga gatacattaa gtaacttaa
901 aaaaaacttt acacagtctg cctagtagat tactatttgg aatatatgtg tgcttatttg
961 catattcata atctccctac tttattttct tttattttta attgatacat aatcattata
1021 catatttatg ggtaaagtg taatgtttta atatgtgtac acatattgac caaatcaggg
1081 taattttgca tttgtaattt taaaaaatgc tttcttcttt taatatactt ttttgtttat
1141 cttatttcta atactttccc taatctcttt ctttcagggc aataatgata caatgtatca
1201 tgcctctttg caccattcta aagaataaca gtgataatth ctgggttaag gcaatagcaa
1261 tattttctgca tataaatatt tctgcatata aattgtaact gatgtaagag gtttcatatt
1321 gctaatagca gctacaatcc agctaccatt ctgcttttat tttatggttg ggataagget
1381 ggattattct gagtccaagc taggcccttt tgctaatacat gttcatacct cttatcttcc
1441 tcccacag CT CCTGGGCAAC GTGCTGGTCT GTGTGCTGGC CCATCACTTT GCGAAAGAAT
1501 TCACCCACAC AGTGCAGGCT GCCTATCAGA AAGTGGTGGC TGGTGTGGCT AATGCCCTGG
1561 CCCACAAGTA TCACTAagct cgctttcttg ctgtccaatt tctattaag gttcctttgt
1621 tccctaagtc caactactaa actgggggat attatgaagg gccttgagca tctggattct
1681 gcctaataaa aacattttat tttcttga atgatgtatt taattattt ctgaatattt
1741 tctaataaag ggaatgtggg aggtcagtg acattaaaaca taaagaatg atgagctggt
1801 caaaccttgg gaaaatacac tatatcttaa actccatgaa aqaaggtgag qctqcaacca
1861 gctaatagca attggcaaca gcccctg
1921 ttcttgtaga ggcttgattt gcaggtt
1981 ttgttttagc tgcctcatg aatgtct
2041 tcagccttga ct

```



The coding region is excerpted from the transcript and is shown below.



```

ATG GTG CAC CTG ACT CCT GAG GAG AAG TCT GCC GTT ACT GCC CTG TGG GGC AAG GTG
AAC GTG GAT GAA GTT GGT GGT GAG GCC CTG GGC AGG CTG CTG GTG GTC TAC CCT TGG
ACC CAG AGG TTC TTT GAG TCC TTT GGG GAT CTG TCC ACT CCT GAT GCT GTT ATG GGC
AAC CCT AAG GTG AAG GCT CAT GGC AAG AAA GTG CTC GGT GCC TTT AGT GAT GGC CTG
GCT CAC CTG GAC AAC CTC AAG GGC ACC TTT GCC ACA CTG AGT GAG CTG CAC TGT GAC
AAG CTG CAC GTG GAT CCT GAG AAC TTC AGG CTC CTG GGC AAC GTG CTG GTC TGT GTG
CTG GCC CAT CAC TTT GGC AAA GAA TTC ACC CCA CCA GTG CAG GCT GCC TAT CAG AAA
GTG GTG GCT GGT GTG GCT AAT GCC CTG GCC CAC AAG TAT CAC TAA

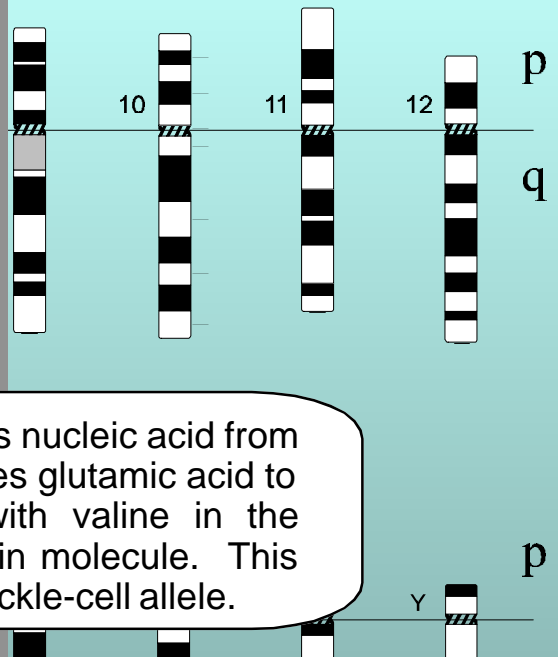
```

Beta Hemoglobin

1 ccctgtggag ccacacccta gggttggcca atctactccc aggagcaggg agggcaggag
 61 ccagggctgg gcataaaagt cagggcagag ccatctattg **cttacatttg ctctcgacac**
 121 **aactgtgttc actagcaacc tcaaacagac accATGGTGC ACCTGACTCC TGAGGAGAAG**
 181 **TCTGCCGTTA CTGCCCTGTG GGGCAAGGTG AACGTGGATG AAGTTGGTGC TGAGGCCCTG**
 241 **GGCAGG**ttgg tatcaaggtt acaagagag ttttagggga ggaatagaa ggggcatgt
 301 ggagacagag aagactcttg **CTCCTT**
 361 ttttcccacc cttagg **CTGCGCAA**
 421 **TGGGGATCTG TCCACTCC**
 481 **GAAAGTGCTC GGTGCCTT**
 541 **TGCCACACTG AGTGAGCTT**
 601 gagtctatgg gacccttgat
 661 taggaagggg agaagtaaca
 721 agtgtggaag tctcaggatc
 781 cttttgttta attcttgctt tctttttttt tcttttttgc aatctttact accatactta
 841 atgccttaac atttgttata acaaaaggaa atatctctga gatacattaa gtaacttaa
 901 aaaaaacttt acacagctc cctagtagat tactatttgg aatatatgtg tgcttatttg
 961 catattcata atctccctac tttattttct tttattttta attgatacat aatcattata
 1021 catatttatg ggtaaagtg taatgtttta atatgtgtac acatattgac caaatcaggg
 1081 taattttgca tttgtaattt taaaaaatgc tttcttcttt taatatactt ttttgtttt
 1141 cttatttcta atactttccc taactctctt ctttcagggc aataatgata caatgtatca
 1201 tgcctctttg caccattcta aagaataaca gtgataatct ctgggttaag gcaatagcaa
 1261 tatttctgca tataaatatt tctgcatata aattgtaact gatgtaagag gtttctatatt
 1321 gctaatagca gctacaatcc agctaccatt ctgcttttat tttatggttg
 1381 ggattattct gagtccaagc taggcccttt tgctaatacat gttcataacc
 1441 tcccacag **CT CCTGGGCAAC GTGCTGGTCT GTGTGCTGGC CCATCA**
 1501 **TCACCCCACC AGTGCAGGCT GCCTATCAGA AAGTGGTGGC TGGTGT**
 1561 **CCCACAAGTA TCACTAAGct cgctttcttg ctgtccaatt tctatt**
 1621 **tccctaagta caactactaa actgggggat attatgaagg gcctt**
 1681 **gcctaataaa aacatttat tttcattga** atgatgtatt taattat
 1741 tactaaaaag ggaatgtggg aggtcagtgc attttaaaca taagaagtg aag
 1801 caaaccttgg gaaaatacac tatatcttaa actccatgaa aqaaggtgag actgtaacca
 1861 gctaatagca attggcaaca gccctg
 1921 ttcttgtaga ggcttgattt gcaggtt
 1981 ttgtttttagc tgcctcatg aatgtct
 2041 tcagccttga ct

Changing just one nucleotide out of 3,000,000,000 is enough to produce a lethal gene, just as one incorrect bit can crash an operating system.

Errors in the genetic code lead to errors in protein synthesis, with potentially devastating effects. Here, the single change is illustrated that produces the gene for sickle-cell anemia.



A change in this nucleic acid from an A to T causes glutamic acid to be replaced with valine in the beta-hemoglobin molecule. This produces the sickle-cell allele.

ATG	GTG	CAC	CTG	ACT	CCT	GAG	GAG	AAG	TCT	GCC	GTT	ACT	GCC	CTG	TGG	GGC	AAG	GTG
AAC	GTG	GAT	GAA	GTT	GGT	GGT	GAG	GCC	CTG	GGC	AGG	CTG	CTG	GTG	GTC	TAC	CCT	TGG
ACC	CAG	AGG	TTC	TTT	GAG	TCC	TTT	GGG	GAT	CTG	TCC	ACT	CCT	GAT	GCT	GTT	ATG	GGC
AAC	CCT	AAG	GTG	AAG	GCT	CAT	GGC	AAG	AAA	GTG	CTC	GGT	GCC	TTT	AGT	GAT	GGC	CTG
GCT	CAC	CTG	GAC	AAC	CTC	AAG	GGC	ACC	TTT	GCC	ACA	CTG	AGT	GAG	CTG	CAC	TGT	GAC
AAG	CTG	CAC	GTG	GAT	CCT	GAG	AAC	TTC	AGG	CTC	CTG	GGC	AAC	GTG	CTG	GTC	TGT	GTG
CTG	GCC	CAT	CAC	TTT	GGC	AAA	GAA	TTC	ACC	CCA	CCA	GTG	CAG	GCT	GCC	TAT	CAG	AAA
GTG	GTG	GCT	GGT	GTG	GCT	AAT	GCC	CTG	GCC	CAC	AAG	TAT	CAC	TAA				

Genome Databases

Data Management Requirements

- Reagent data
- Genetic-map data
- Sequence data
- Structural data
- Comparative data
- Functional data
- Other data...

Human Genome

O M I M -- Beta Hemoglobin

Title
*141900 HEMOGLOBIN--BETA LOCUS
(HBB) SICKLE CELL ANEMIA,
D; BETA-THALASSEMIA,
D; HEINZ BODY ANEMIAS,
OBIN TYPE, ...]

alpha and beta loci determine the
of the 2 types of polypeptide
adult hemoglobin, Hb A. By
graphy using heavy-labeled
oin-specific messenger RNA,
l. (1972) found labeling of a
ome 2 and a group B chromo-
ey concluded, incorrectly as it
it, that the beta-gamma-delta
roup was on a group B
ome since the zone of labeling
er on that chromosome than on
ome 2 (which by this

P I R -- Beta Hemoglobin

DEFINITION
Hemoglobin beta chain
chimpanzee, pygmy
nzee, and gorilla

RY [SUM]
Protein
ular-weight 15867
n 146
sum 1242

CE
P E E K S A V T A L W G
D E V G G E A L G R L L
W T Q R F F E S F G D L
A V M G N P K V K A H G
G A F S D G L A H L D N
F A T L S E L H C D K L

GenBank -- Beta Hemoglobin

DEFINITION [DEF]
[HUMHBB] Human beta
globin region
[C]
HUMHBB
NO. [ACC]
93 J00094 J00096
59 J00160 J00161

[KEY]
e element; HPFH;
ve sequence; RNA
III; allelic
ternate cap site;

tctccctctcacta
aggagtgggtgctc
gtttgatataaaaa
tgggaggatccctt

G D B -- Beta Hemoglobin

** Locus Detail View **

Symbol: HBB
Name: hemoglobin, beta
MIM Num: 141900
Location: 11p15.5
Created: 01 Jan 86 00:00

** Polymorphism Table **

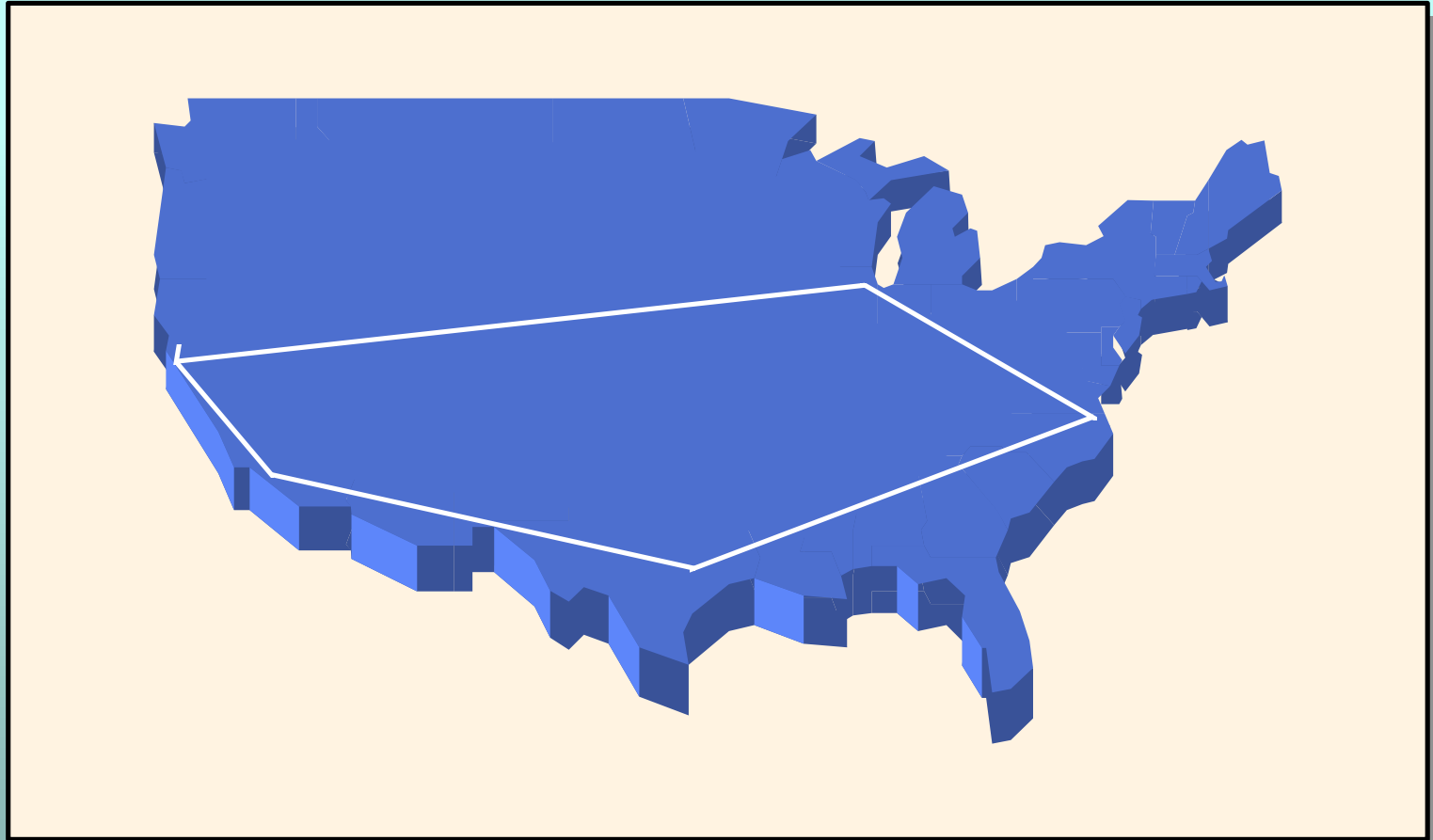
Probe	Enzyme
beta-globin cDNA	RsaI
beta-globin cDNA,JW10+	AvaI
Pstbeta,JW102,BD23,pB+	BamHI
pRK29,Unknown	HindII
beta-IVS2 probe	HphI
IVS-2 normal	HphI
Unknown	AvrII

**Knowledge Management
through
Electronic Data Publishing**

Data Management Challenges

- Size
- Complexity
- Audacity

Data Management Challenge: Size



CCCTGTGGAGCCACACCCTAGGGTTGGCCAATCTACTCCCAGGAGCAGGGAGGGCAGGAGC

Data Management Challenge: Complexity

Consider the DNA sequence of a human genome as equivalent to 3.3 gigabytes of files on the mass-storage device of some computer system of unknown design. Obtaining the sequence is equivalent to obtaining an image of the contents of that mass-storage device. Understanding the sequence is equivalent to reverse engineering that unknown computer system (both the hardware and the 3.3 gigabytes of software) all the way back to a full set of design and maintenance specifications.

Data Management Challenge: Audacity

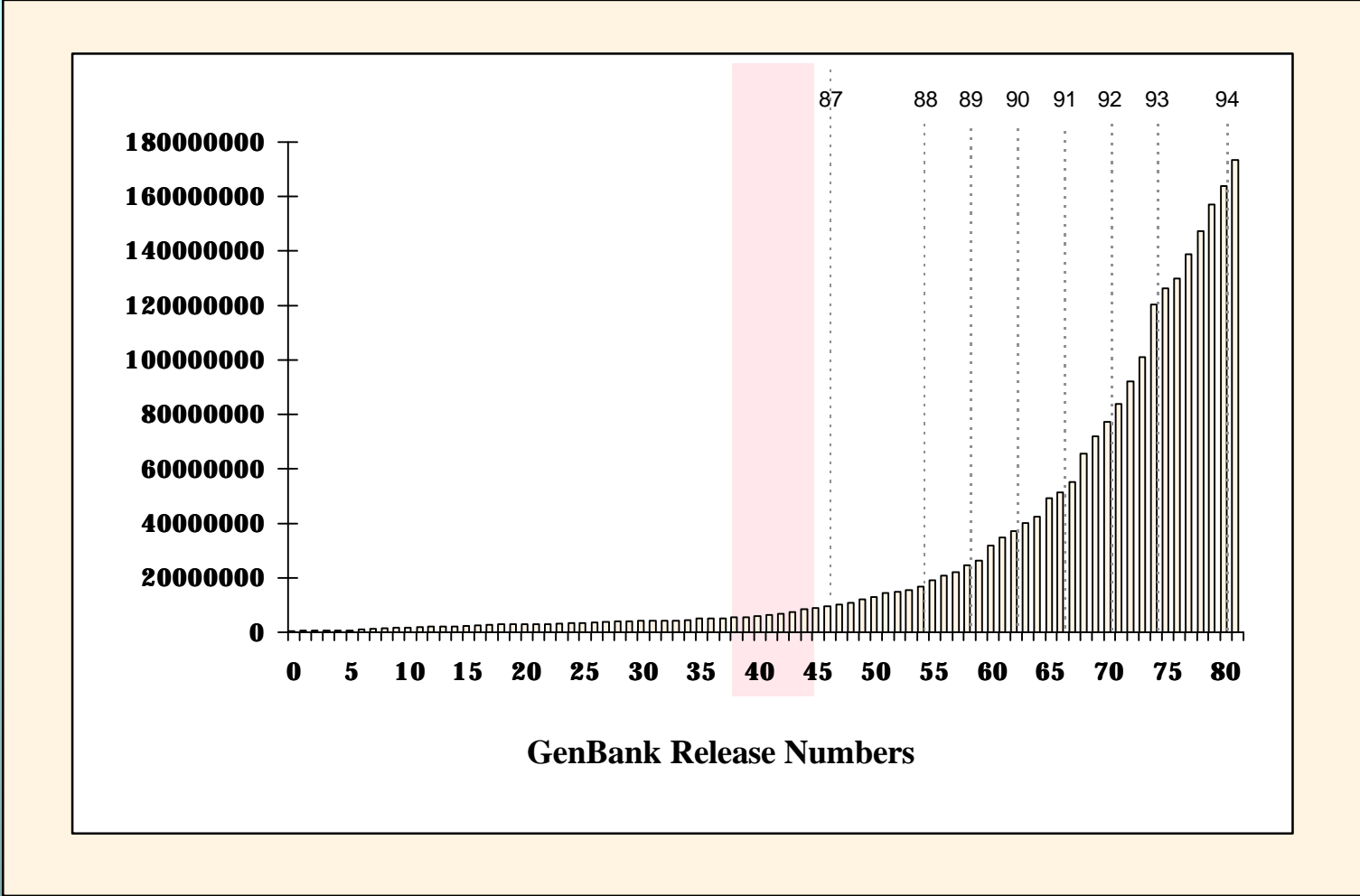
When the Human Genome Project is finished, many of the innovative laboratory methods involved in its successful conclusion will begin to fade from memory. What will remain, as the project's enduring contribution, is a vast amount of computerized knowledge. Seen in this light, the Human Genome Project is nothing but the effort to create the most important database ever attempted -- the database containing instructions for creating life.

Technical Challenges

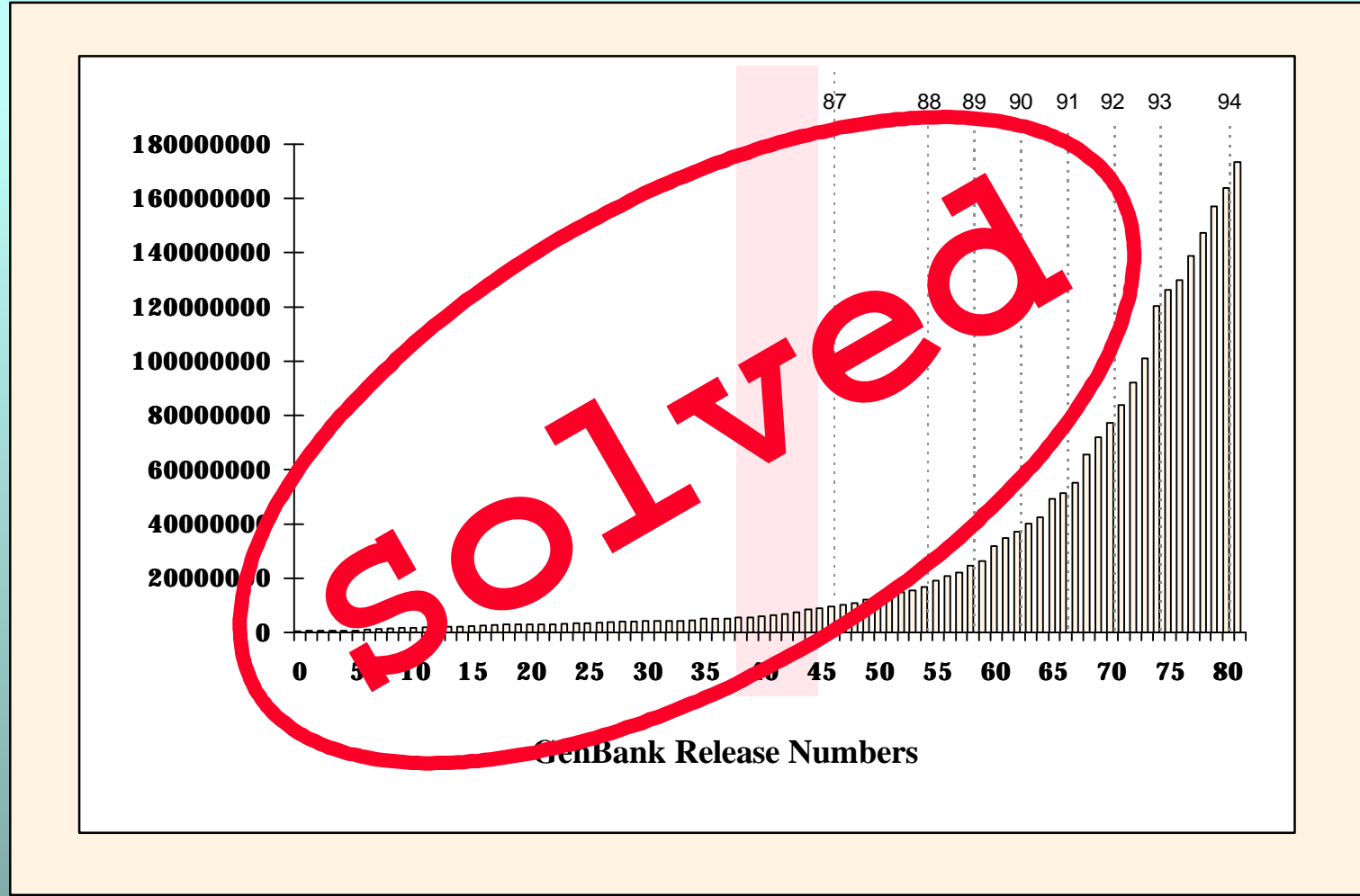
**1980s: Data Acquisition
Data Access**

1990s: Data Integration

Data Acquisition Crisis

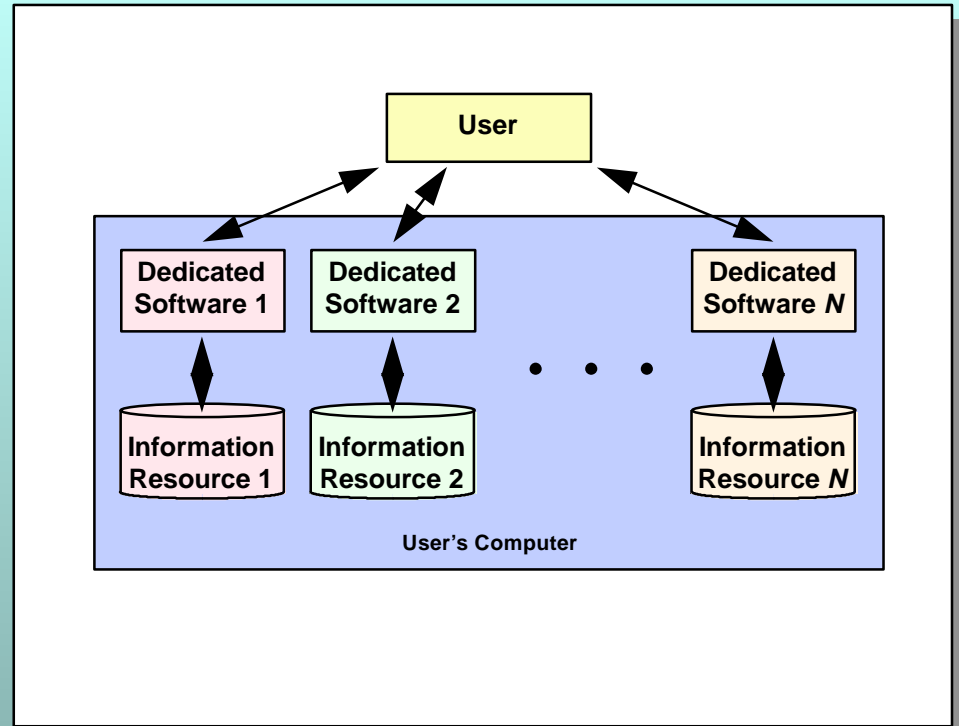


Data Acquisition Crisis



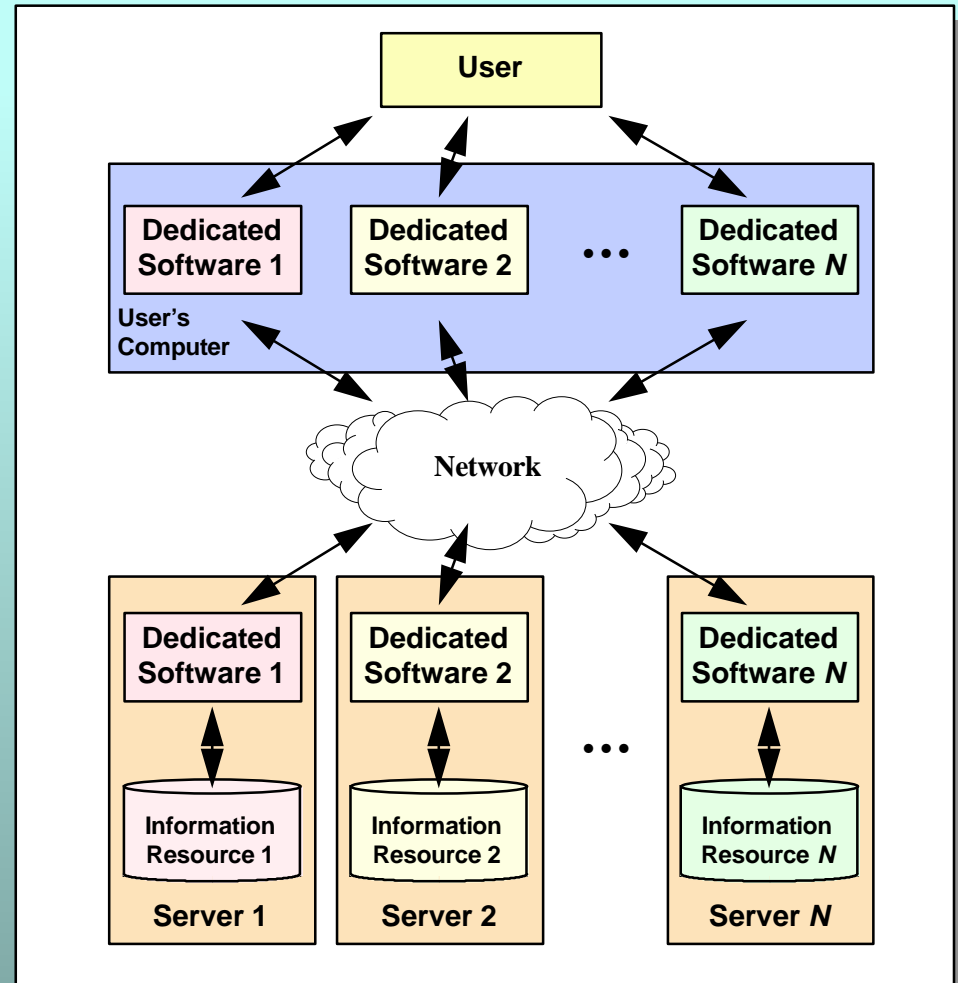
Data Access Crisis: Local Systems

In the early days of bioinformatics, computerized information systems occurred only as stand-alone that had to be completely installed locally, including both programs and data.



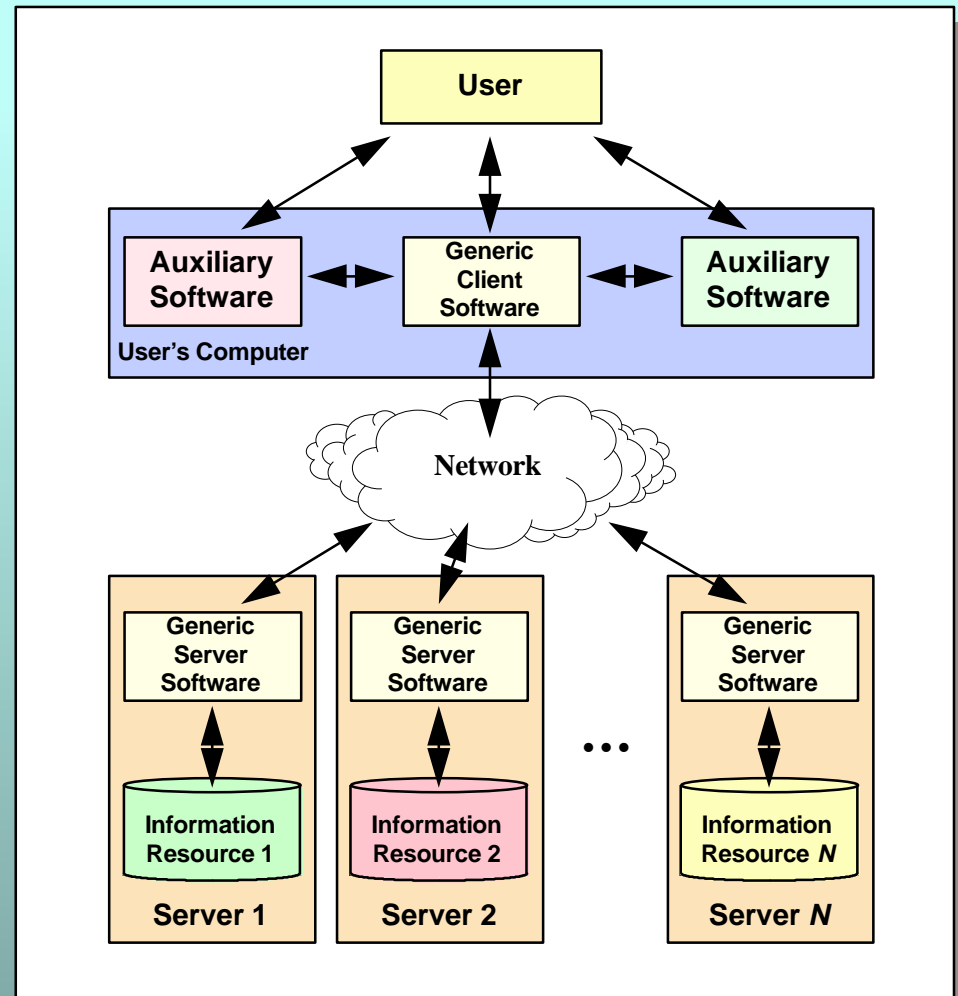
Data Access Crisis: Client Server Systems

The next step was the development of client-server systems, that made the data available remotely. However, custom client software was required for each such resource.



Data Access Crisis: Generic Client Server

The latest advance has been the development of generic client-server systems, so that the same client software can interact with many different servers. Once the generic client is installed, the user has access to any client that follows the generic protocols. *At this point, all the user needs is the name of the resource to be used.*



Data Integration Crisis

An embarrassment to the Human Genome Project is our inability to answer simple questions such as:

How many genes on the long arm of chromosome 21 have been sequenced?

Report of the Invitational DOE Workshop on Genome Informatics, 26-27 April 1993, Baltimore, Maryland

Data Integration Crisis

Adequate connections among data objects in different databases do not exist.

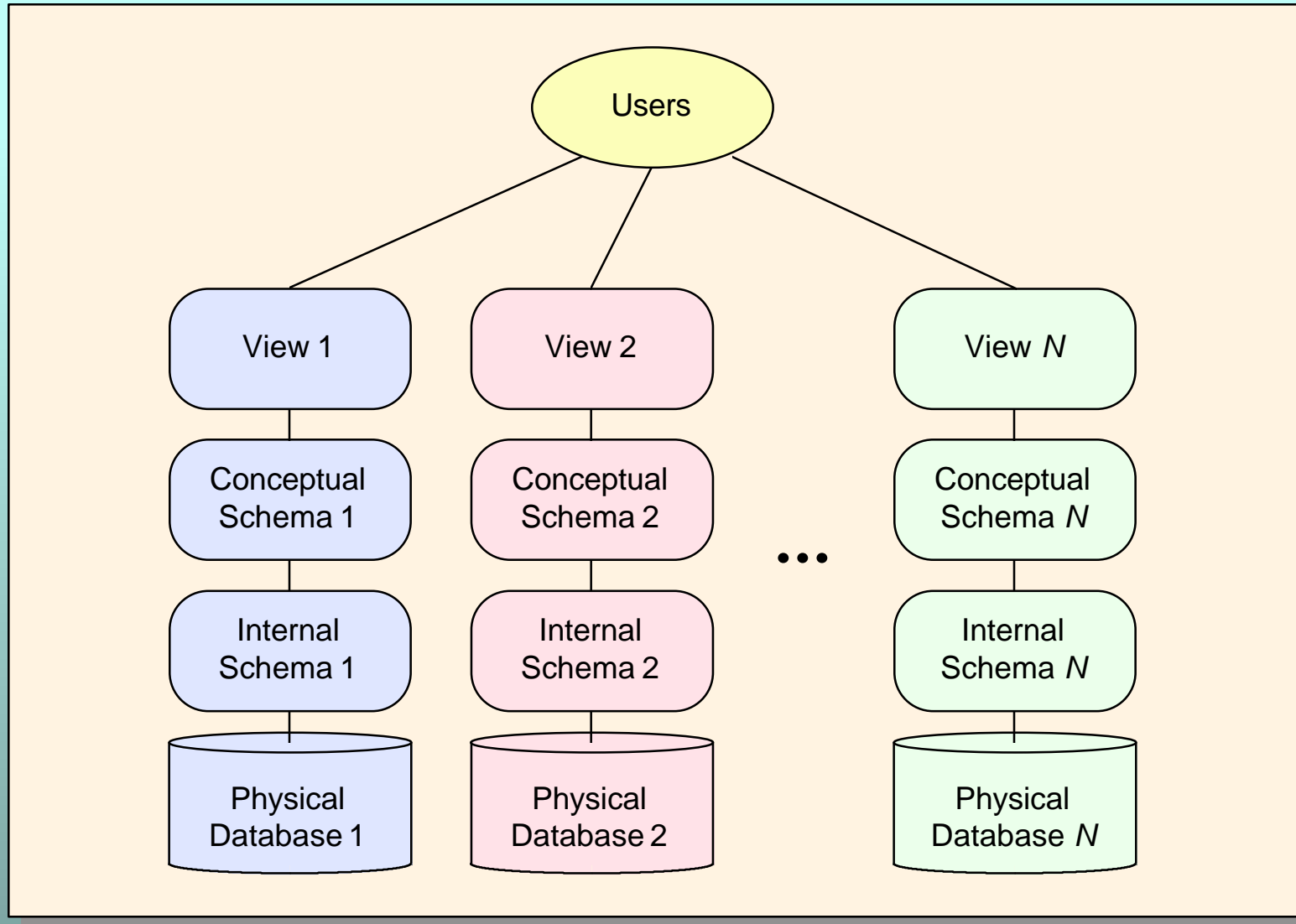
Without adequate connectivity, much of the value of the data will be lost.

Data Integration Goals

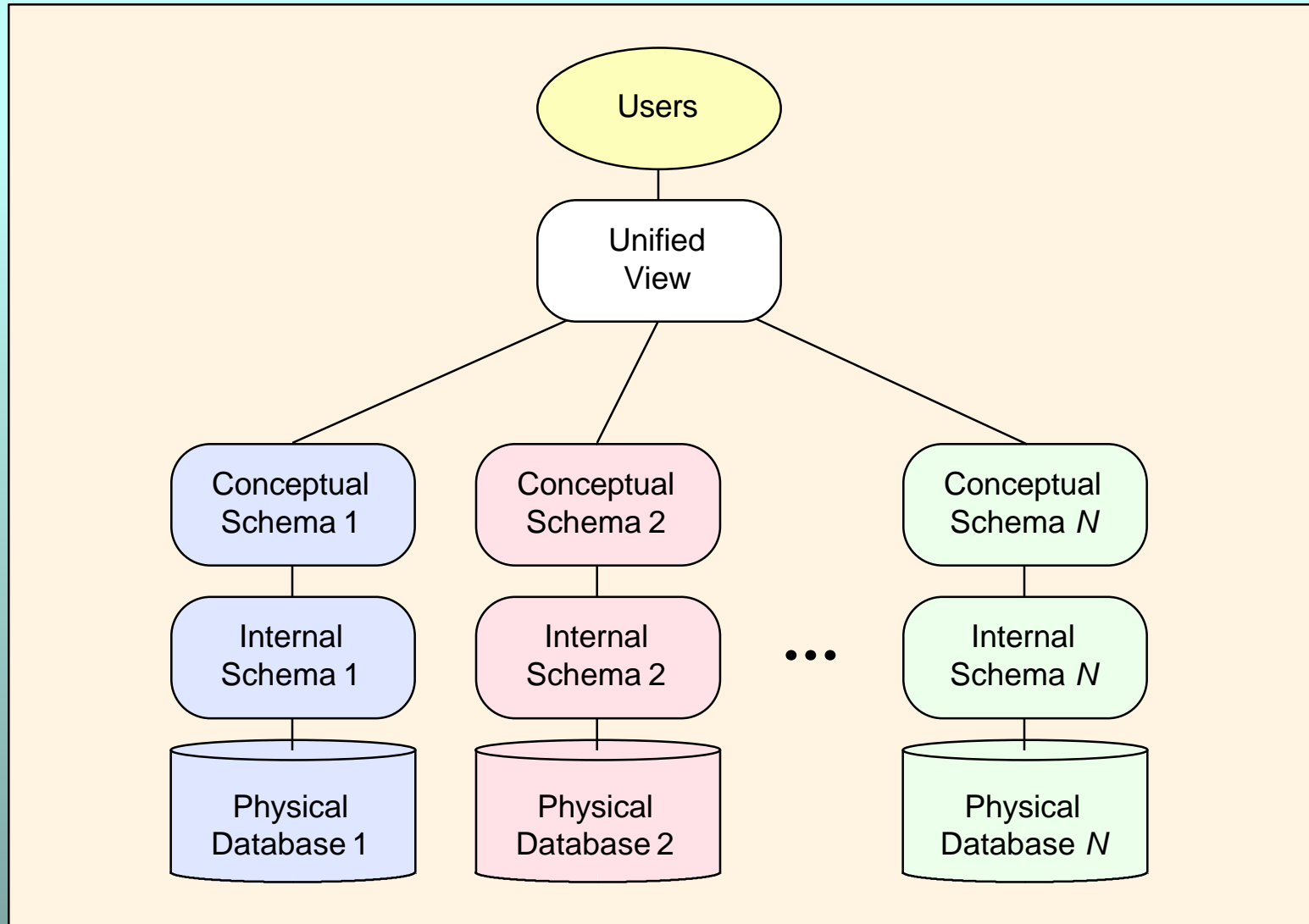
Achieve conceptual integration of genome data.

Provide technical integration of both data and analytical resources to facilitate conceptual integration.

Current Situation



Desired Situation



Data Integration Impediments

Technical: Integrating distributed, heterogeneous databases is not easy.

Sociological: Local incentives encourage competition, not cooperation.

Conceptual: Semantic mismatches exist among databases.

The Vision

We must begin to think of the computational infrastructure of genome research as a federated information infrastructure of interlocking pieces.

Report of the Invitational DOE Workshop on Genome Informatics, 26-27 April 1993, Baltimore, Maryland


An Ambitious Goal

Adding a new database to the federation should be no more difficult than adding another computer to the Internet.

Report of the Invitational DOE Workshop on Genome Informatics, 26-27 April 1993, Baltimore, Maryland

Federated Information Infrastructure

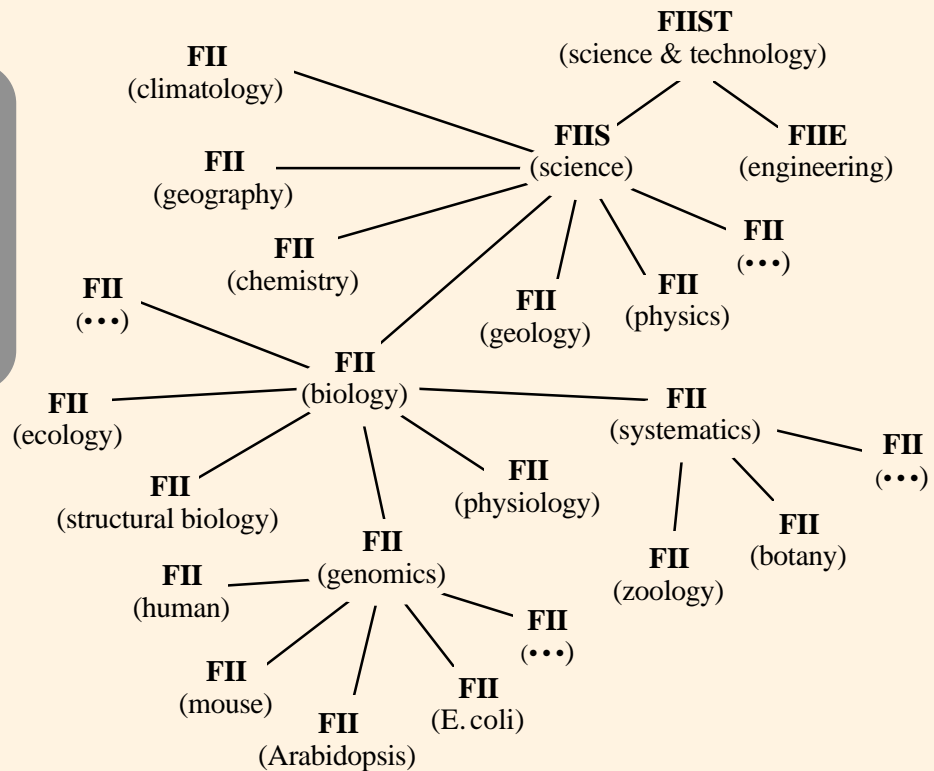
National Information Infrastructure

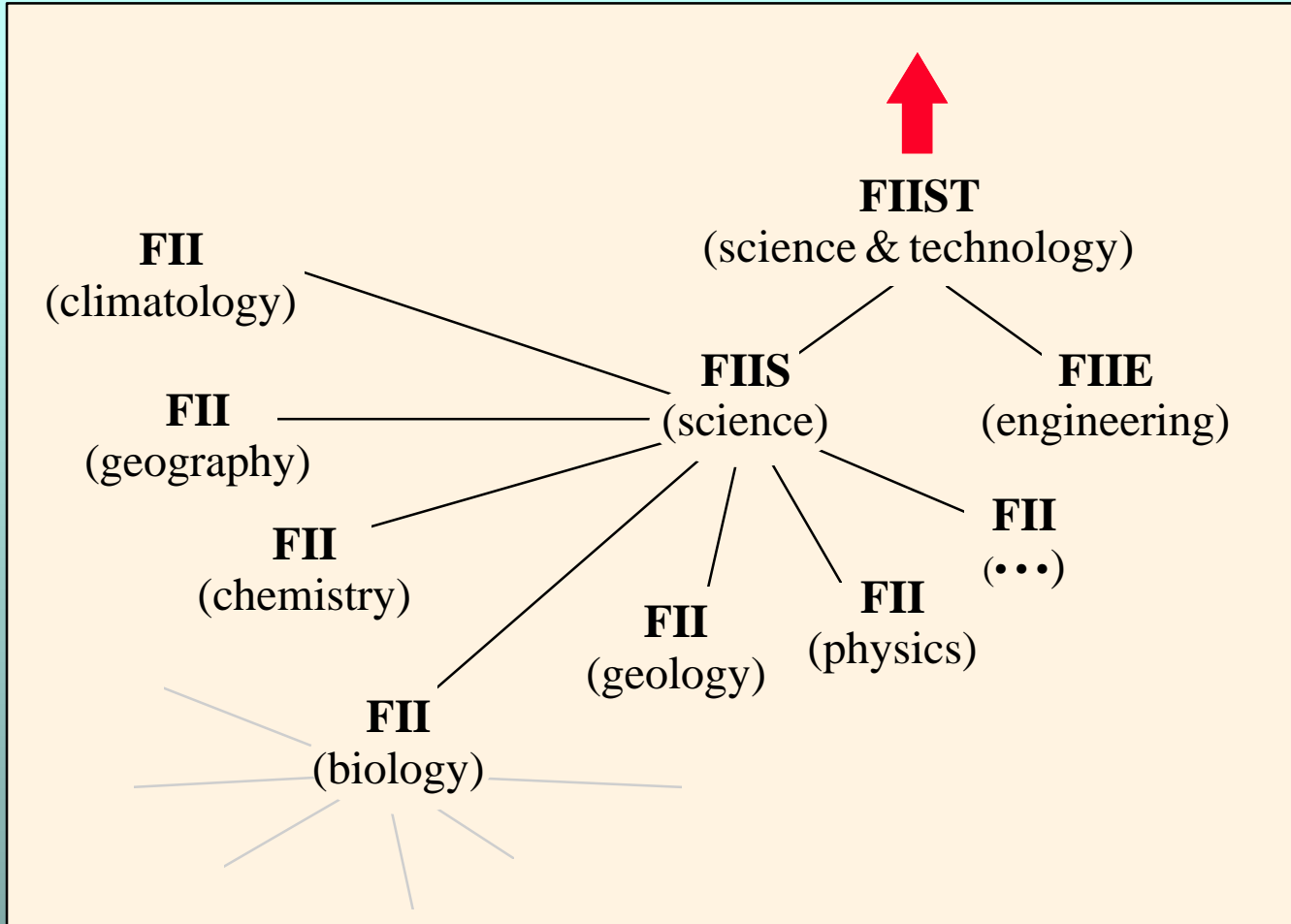
		commercial uses		non-commercial uses		
		ETC	other	Edu	Lib	Res
analog						
digital						

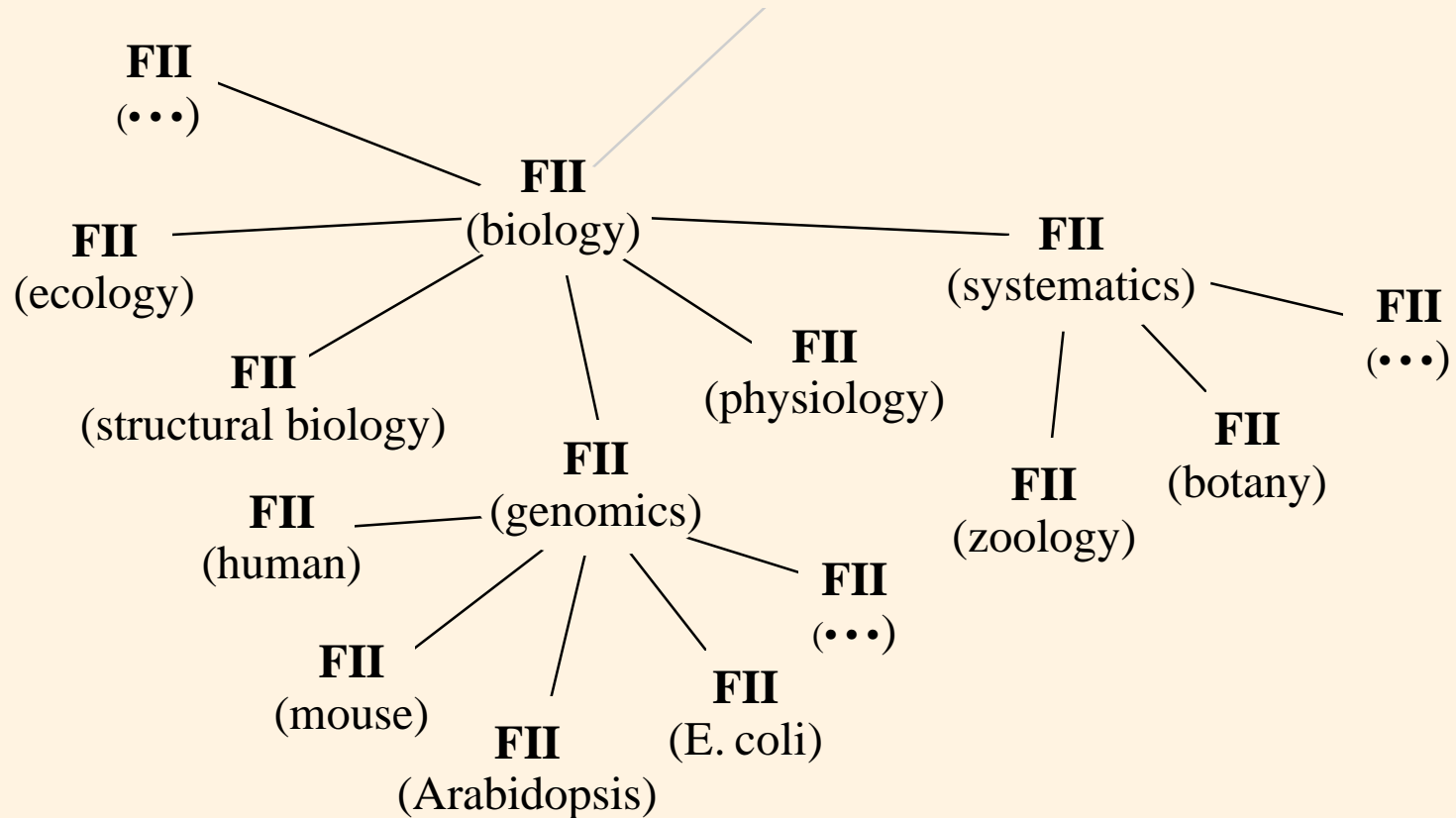
FIIST & NII

The research component of the NII contains a Federated Information Infrastructure for Science and Technology..

	commercial uses		non-commercial uses		
	ETC	other	Edu	Lib	Res
analog					
digital					❖

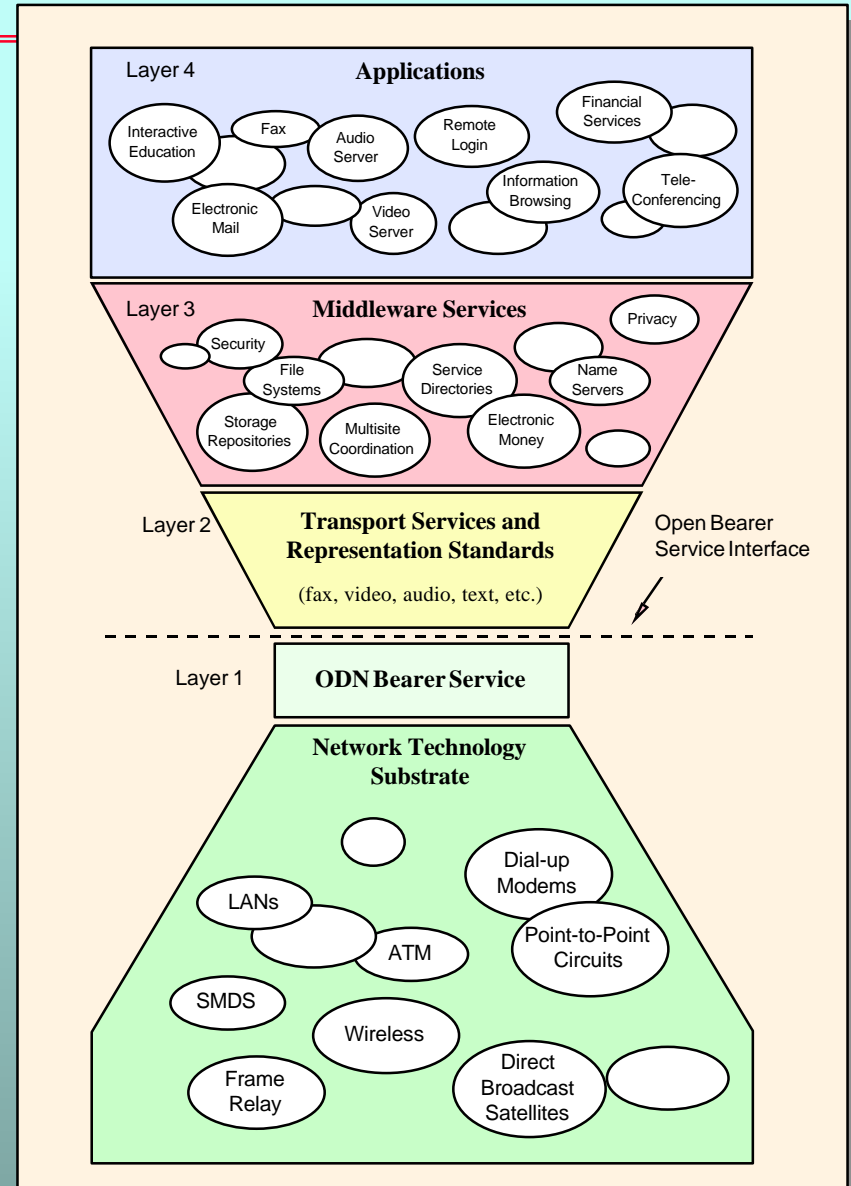






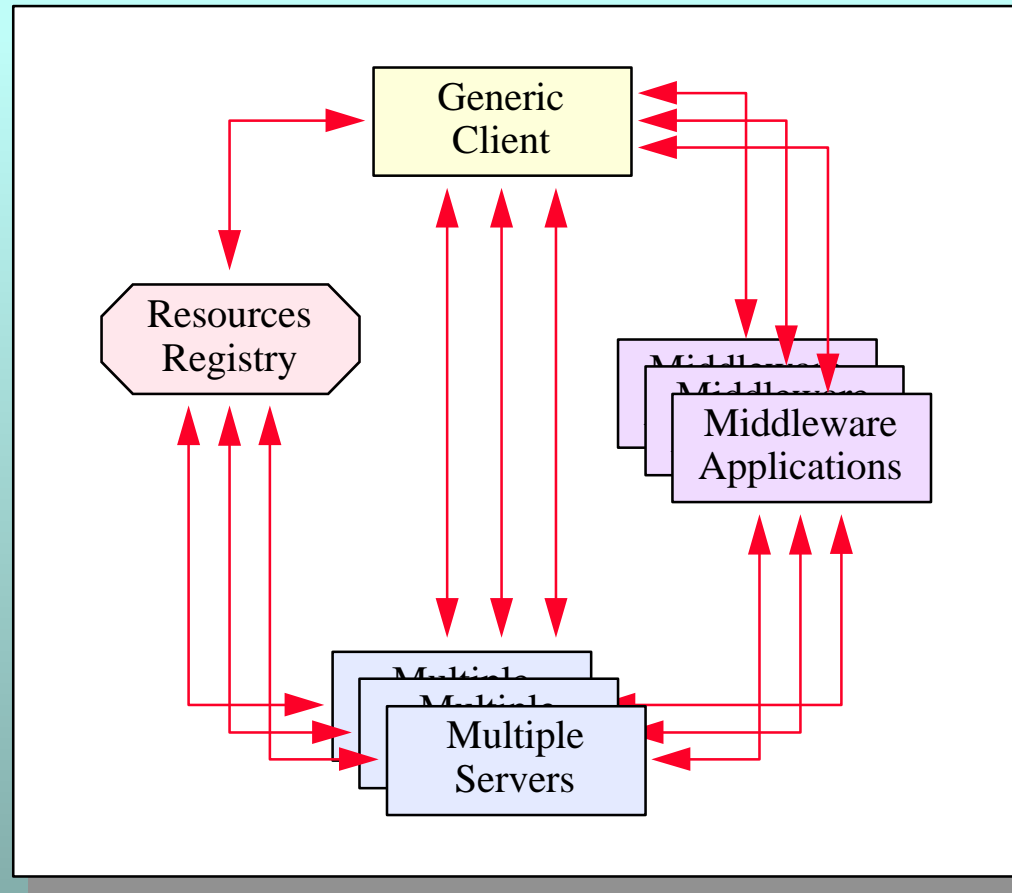
ODN Model

A recent NRC report, *Realizing the Information Future*, laid out a vision of an Open Data Network model, in which any information appliance could be operated over generic networking protocols...

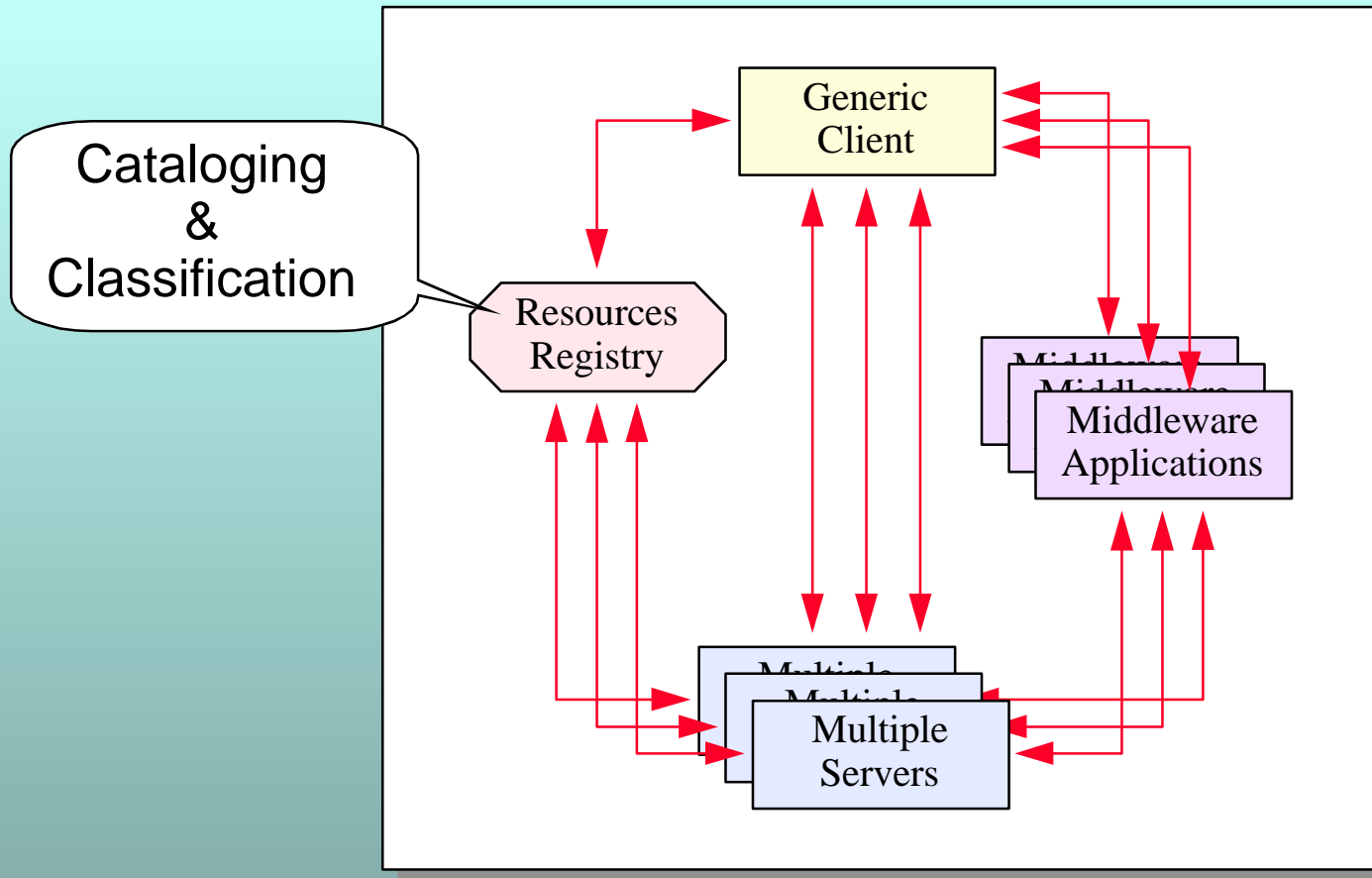


FOSM Reference Architecture

FOSM Reference Architecture

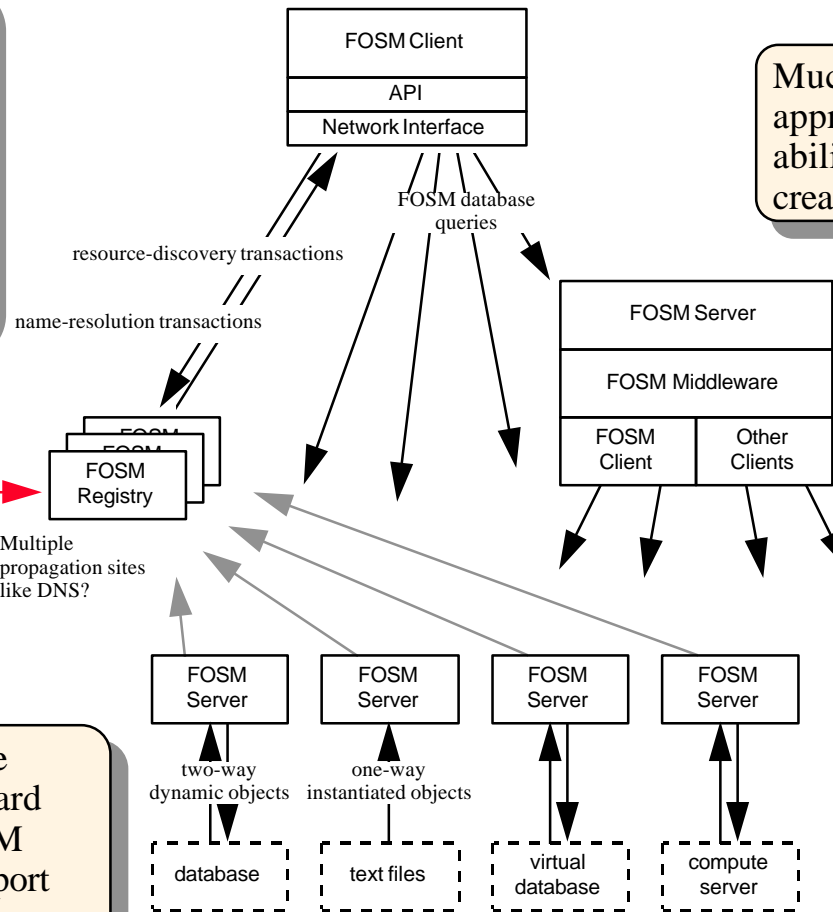


FOSM Reference Architecture



FOSM Reference Architecture

Registry of FOSM servers, FOSM objects (& versions & prunings), FOSM links, FOSM subfederations, FOSM editorial records, FOSM methods, FOSM names, FOSM cataloguing, etc.

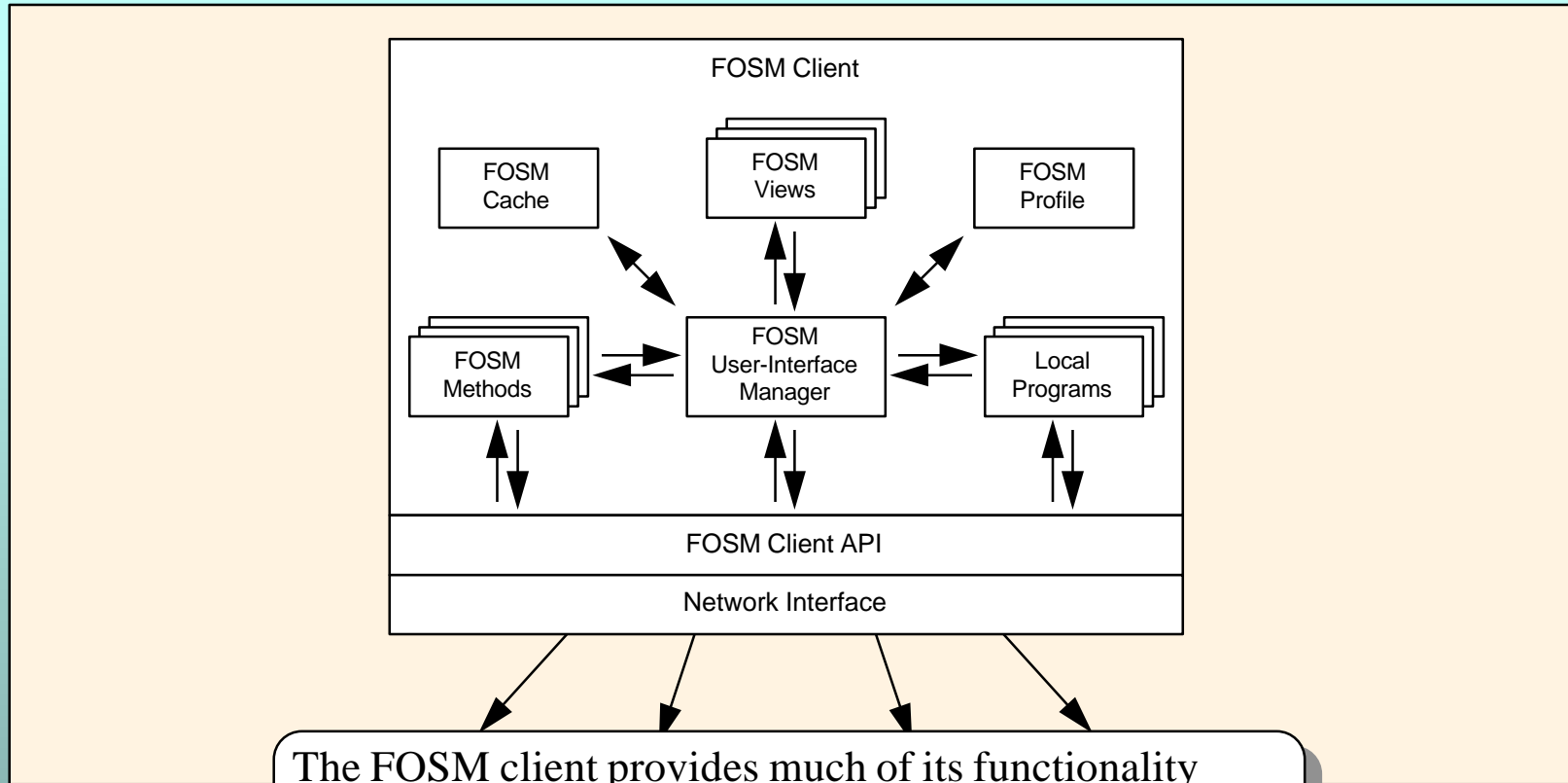


Much of the power of the FOSM approach will come from the ability of third-party developers to create interesting middleware.

Since FOSM servers should be able to provide different standard versions of their objects, FOSM naming conventions must support versioning.

Note multiple different kinds of resources behind each server. Some are actual databases, others modified text resources, and still others computer servers.

FOSM Client Architecture

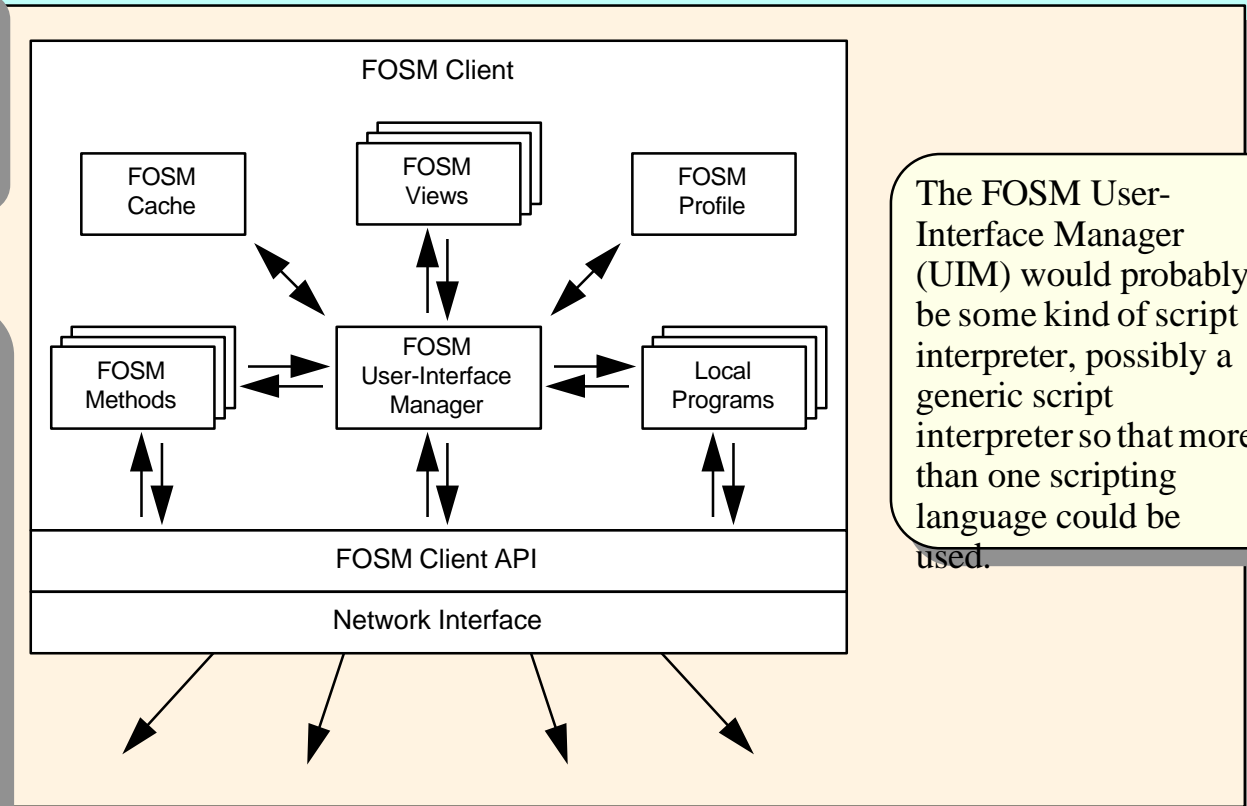


The FOSM client provides much of its functionality through its component-based design. All aspects of the FOSM system are intended to facilitate the value-adding activities of third-party developers.

FOSM Client Architecture

FOSM views will allow users to create local views on FOSM objects or to build virtual FOSM objects.

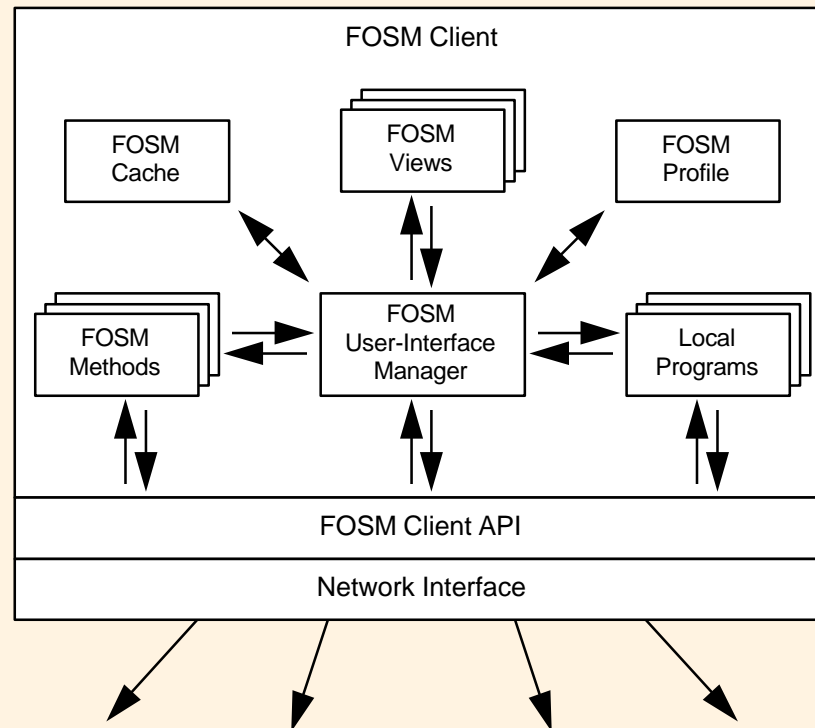
FOSM methods are local, hardware-specific software packages that are invoked to “view” objects obtained from FOSM servers. For example, one of the standard local methods would display and operate HTML documents; another would build, display, and operate query interfaces for FOSM objects.



The FOSM User-Interface Manager (UIM) would probably be some kind of script interpreter, possibly a generic script interpreter so that more than one scripting language could be used.

FOSM Client Architecture

To build a FOSM interface, the client must first query a server to obtain necessary type and format information. This, and other FOSM metadata, should be storable in a local cache. The size of the cache should be user-settable. Normally, the cache would be first-in, first-out, but the user should be able to specify certain cached elements that are never to be flushed.



A FOSM profile system will allow users to customize the behavior both of the local client and of remote servers without requiring servers to maintain registries of users and preferences.

The FOSM API should allow easy development of local programs that can interact directly with the client API, without requiring assistance from the user-interface manager. This would facilitate the development of third-party bulk-data-transaction modules for special markets: DNA sequences, finance, etc.

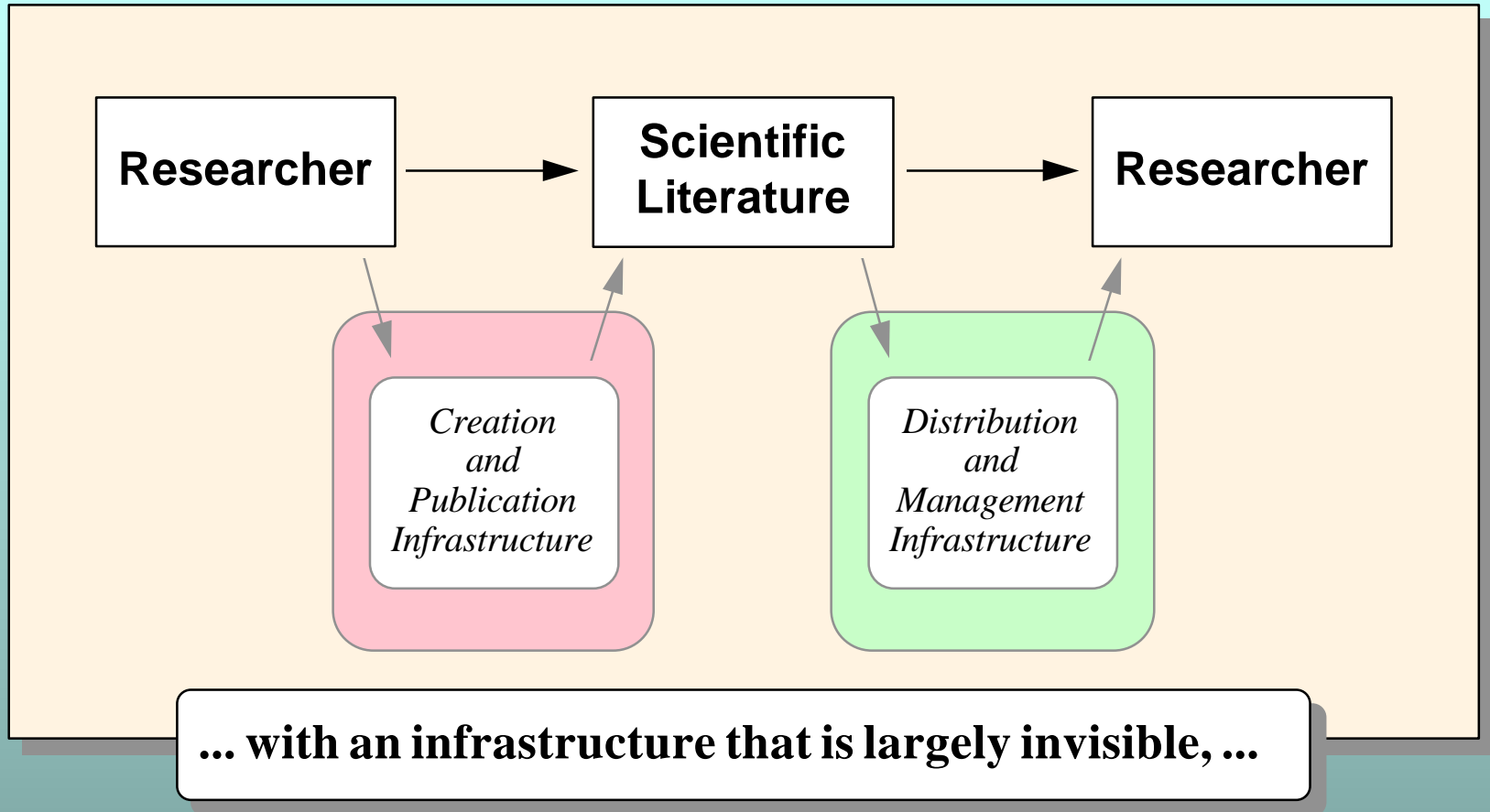
Electronic Data Publishing

Traditional Publishing...

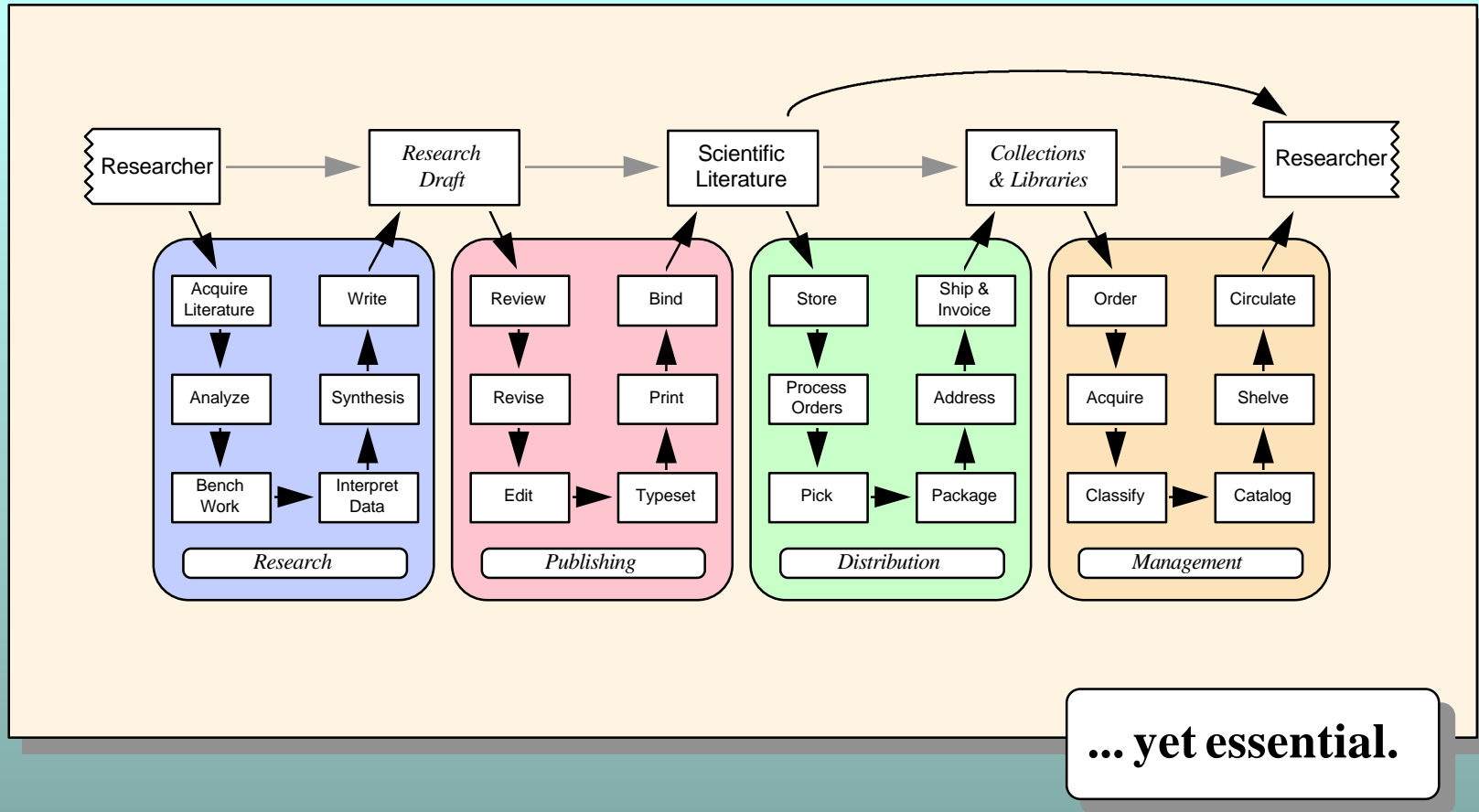


Print publication seems straightforward, ...

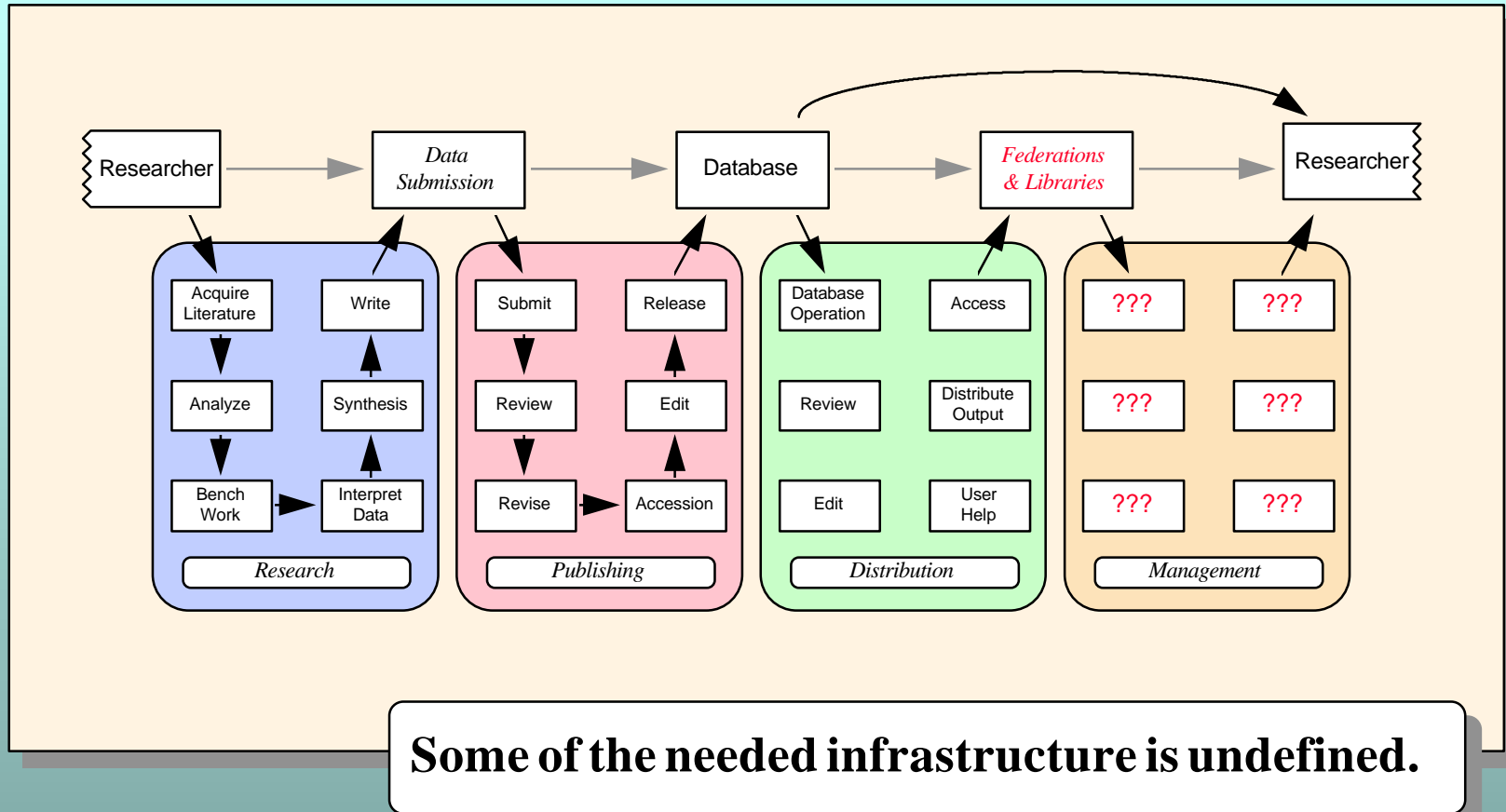
Traditional Publishing



Traditional Publishing



Electronic Publishing

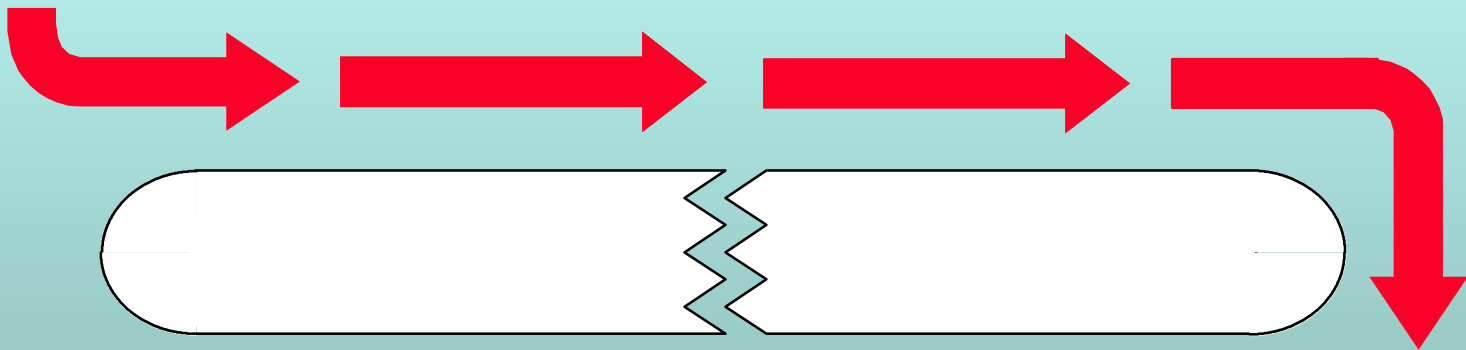


New Discipline of Informatics

What is Informatics?

Computer
Science
Research

----- Informatics -----



Biological
Application
Programs

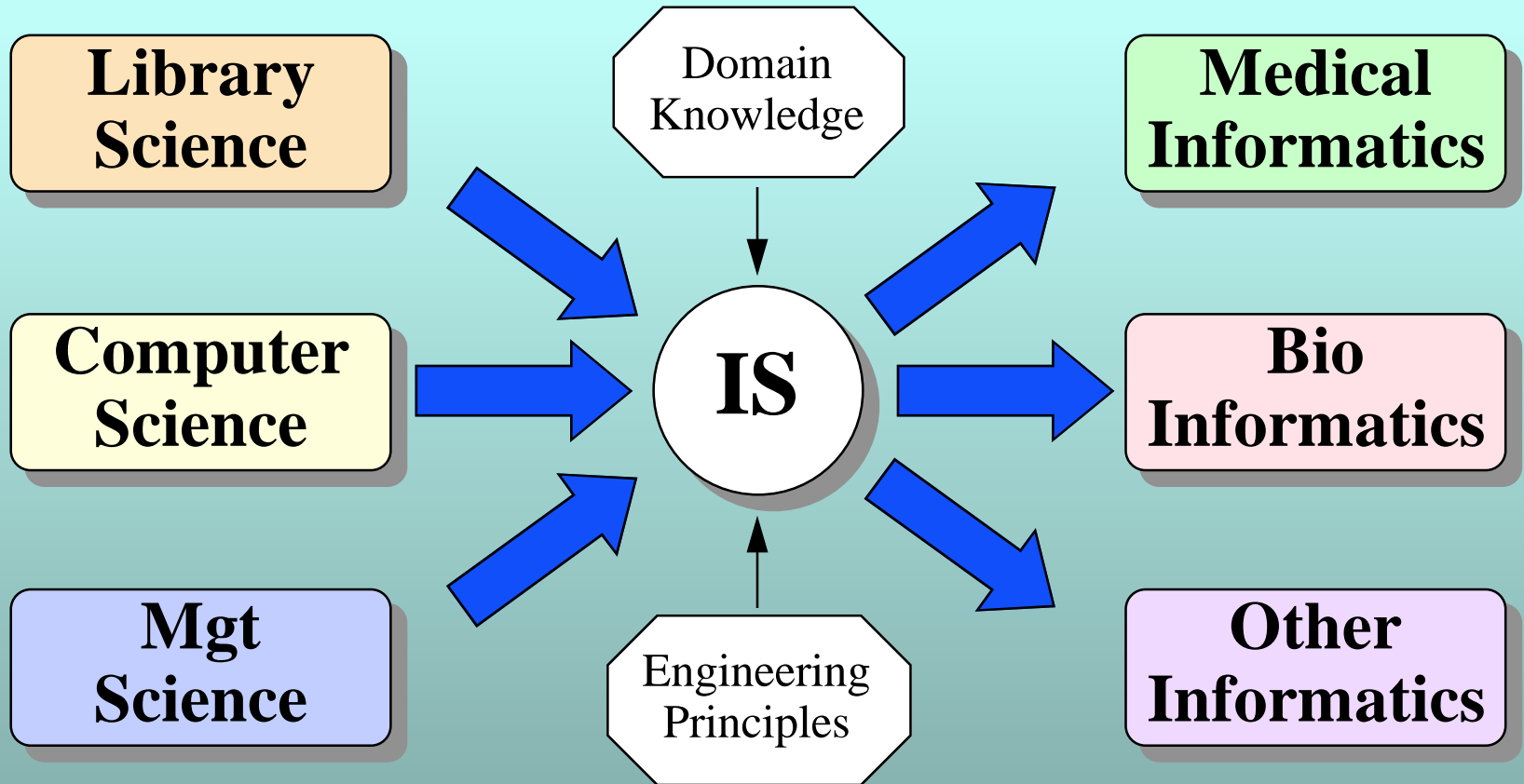
What is Informatics?

Informatics combines expertise from:

- *domain science (e.g., biology)*
- *computer science*
- *library science*
- *management science*

All tempered with an engineering mindset...

What is Informatics?



Engineering Mindset

Engineering is often defined as the use of scientific knowledge and principles for practical purposes. While the original usage restricted the word to the building of roads, bridges, and objects of military use, today's usage is more general and includes chemical, electronic, and even mathematical engineering.

Parnas, David Lorge. 1990. *Computer*, 23(1):17-22.

... or even information engineering.

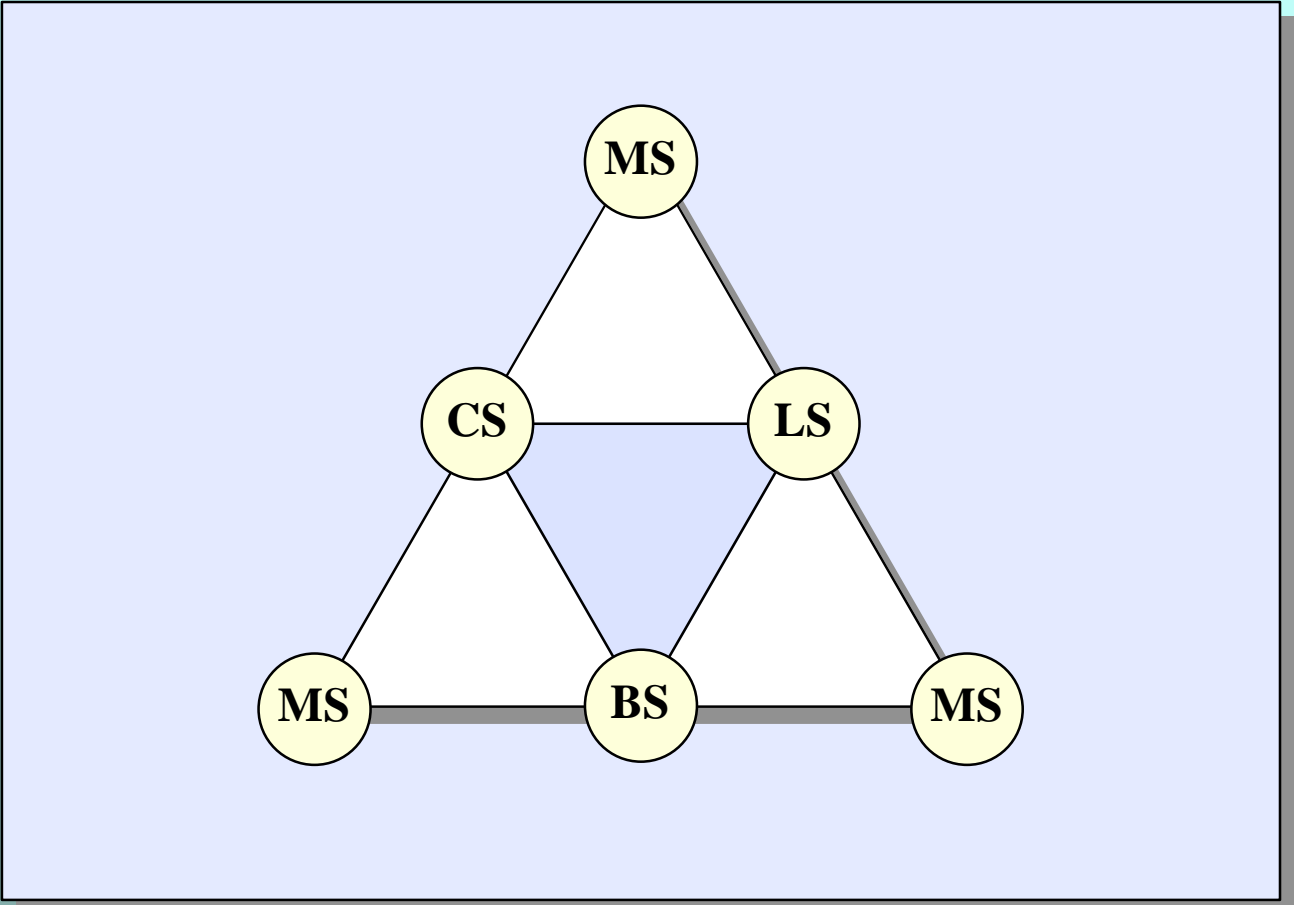
Engineering Mindset

Engineering education ... stresses finding good, as contrasted with workable, designs. Where a scientist may be happy with a device that validates his theory, an engineer is taught to make sure that the device is efficient, reliable, safe, easy to use, and robust.

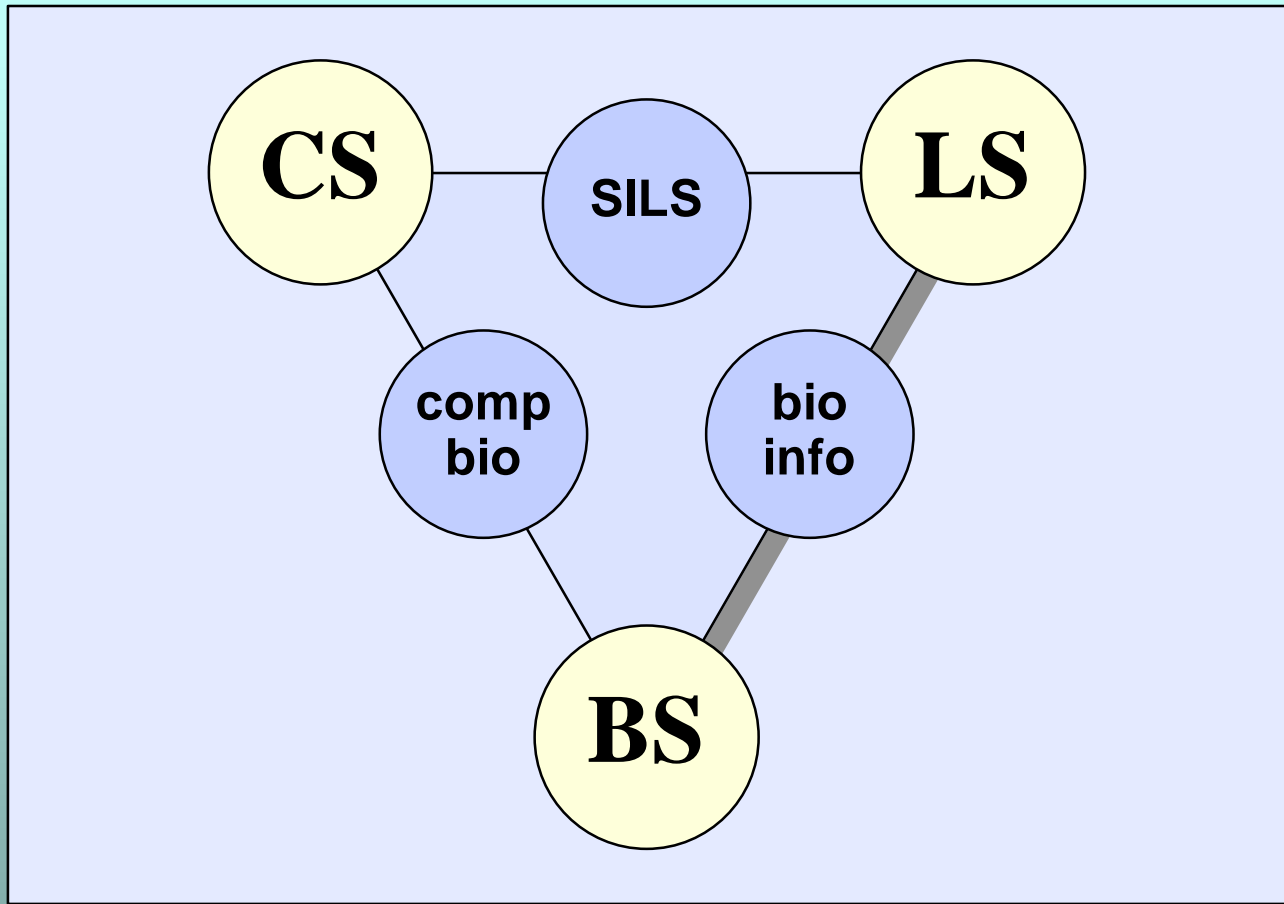
Parnas, David Lorge. 1990. *Computer*, 23(1):17-22.

The assembly of working, robust systems, on time and on budget, is the key requirement for a federated information infrastructure for biology.

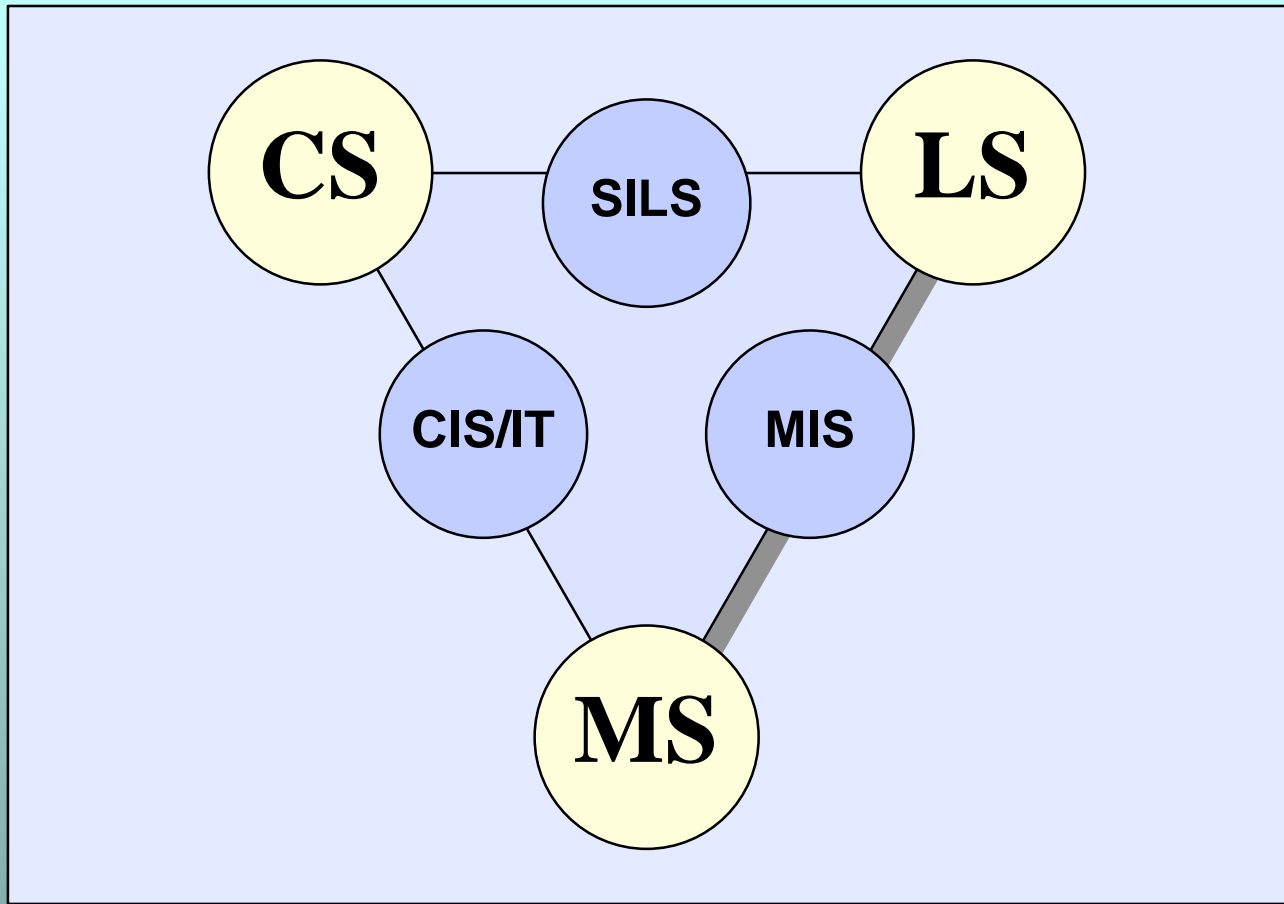
Informatics Triangle



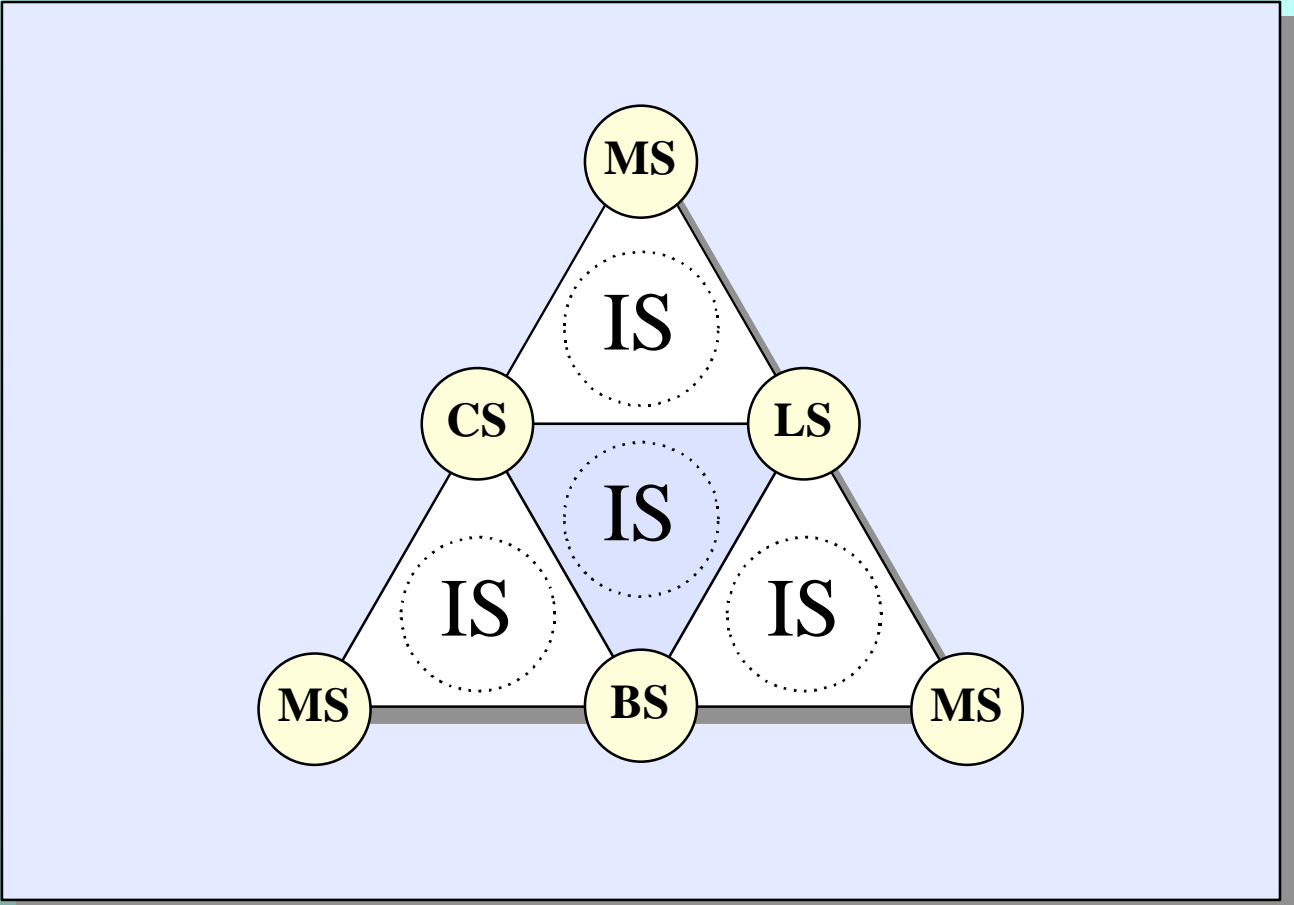
Informatics Triangle



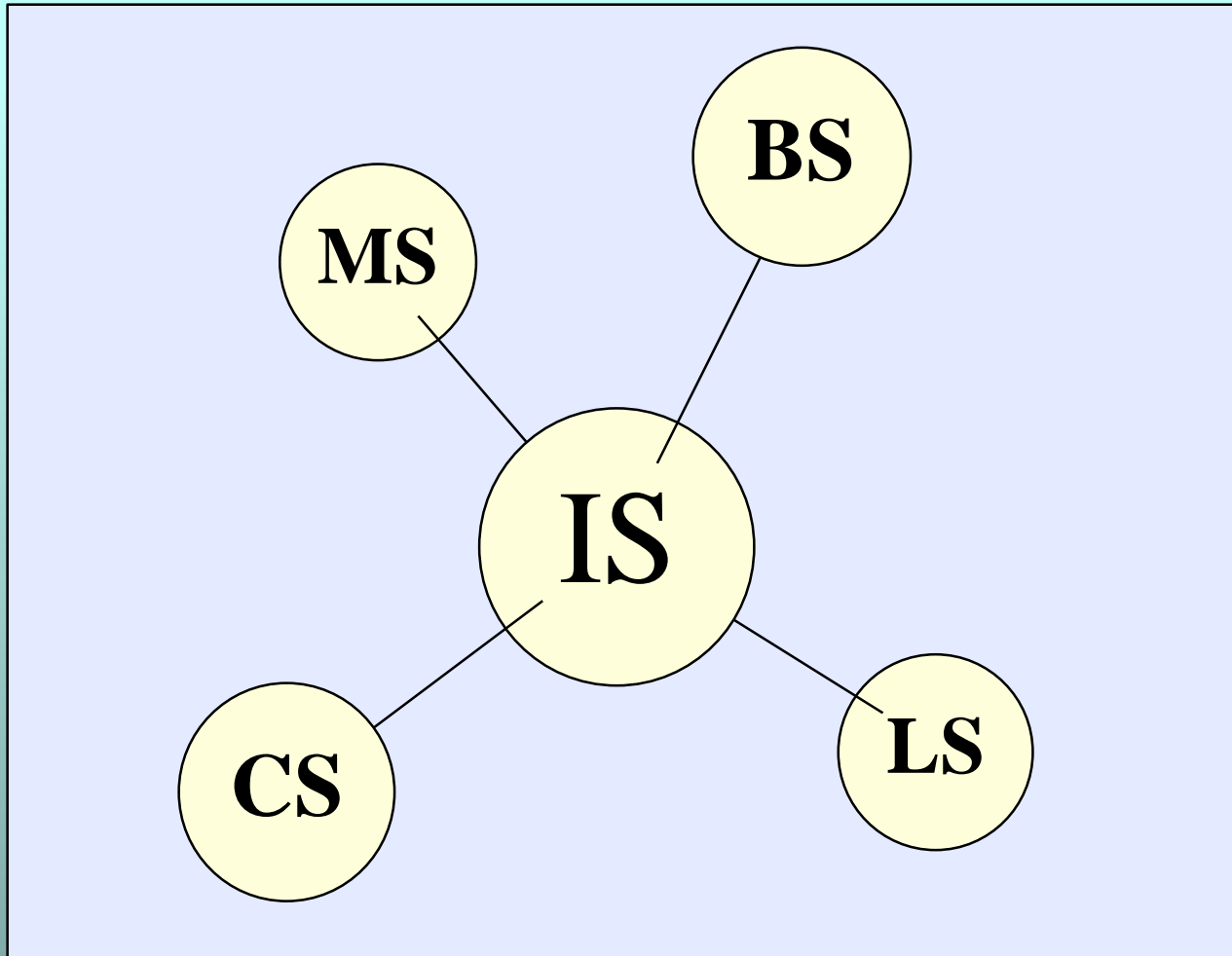
Informatics Triangle



Informatics Triangle



What is Informatics?



Computers as Scientific Instruments

Computers are not just tools for cataloging existing knowledge. They are instruments that change the way we can see the biological world. Computers allow us to see genomes, just as radio telescopes let us see quasars and microscopes let us see cells.

Slides available:

<http://www.gdb.org/rjr/cthsl.ppt>

<http://www.gdb.org/rjr/cthsl.pdf>