Leadership by Example:
Biomedical and Genome Informatics

http://www.esp.org/rjr/aaas2000.pdf

Robert J. Robbins
Fred Hutchinson Cancer Research Center
1100 Fairview Avenue North, DM-120
Seattle, Washington 98109

rrobbins@fhcrc.org
(206) 667 2920

Abstract

Over the next few years, the relentless exponential effect of Moore's Law will profoundly affect nearly all areas of science and technology. By 2005, analytical power previously available only at supercomputer centers will exist on every desktop and the volume of electronic data will be enormous. Even now, a standard Intel computer delivers more computational power than the first supercomputer and GenBank acquires more data every ten weeks than it did in its first ten years.

For the first time, bioinformatics (and logistics) have become a critical part of the ***practice*** of biology, not just a support service. Biology of the 21st Century will require an adequate information infrastructure. Those with access will participate in the transformation of science; those without may become irrelevant.

Topics

- Biotechnology and information technology (IT) will be the “magic” technologies of the 21st Century.

Topics

- Biotechnology and information technology (IT) will be the “magic” technologies of the 21st Century.
- Moore’s Law transforms IT (and everything else).

Topics

- Biotechnology and information technology (IT) will be the “magic” technologies of the 21st Century.
- Moore’s Law transforms IT (and everything else).
- IT has a special relationship with biology.

Topics

- Biotechnology and information technology (IT) will be the “magic” technologies of the 21st Century.
- Moore’s Law transforms IT (and everything else).
- IT has a special relationship with biology.
- 21st-Century biology will be based on bioinformatics.

Topics

- Biotechnology and information technology (IT) will be the “magic” technologies of the 21st Century.
- Moore’s Law transforms IT (and everything else).
- IT has a special relationship with biology.
- 21st-Century biology will be based on bioinformatics.
- Bioinformatics is becoming an independent discipline.

Introduction

Magical Technology

Magic

To a person from 1897, much current technology would seem like magic.

Magic

To a person from 1897, much current technology would seem like magic.

What technology of 2097 would seem magical to a person from 1997?

Magic

To a person from 1897, much current technology would seem like magic.

What technology of 2097 would seem magical to a person from 1997?

Candidate: Biotechnology so advanced that the distinction between living and non-living is blurred.

Magic

To a person from 1897, much current technology would seem like magic.

What technology of 2097 would seem magical to a person from 1997?

Candidate: Biotechnology so advanced that the distinction between living and non-living is blurred.

Information technology so advanced that access to information is immediate and universal.

Moore's Law

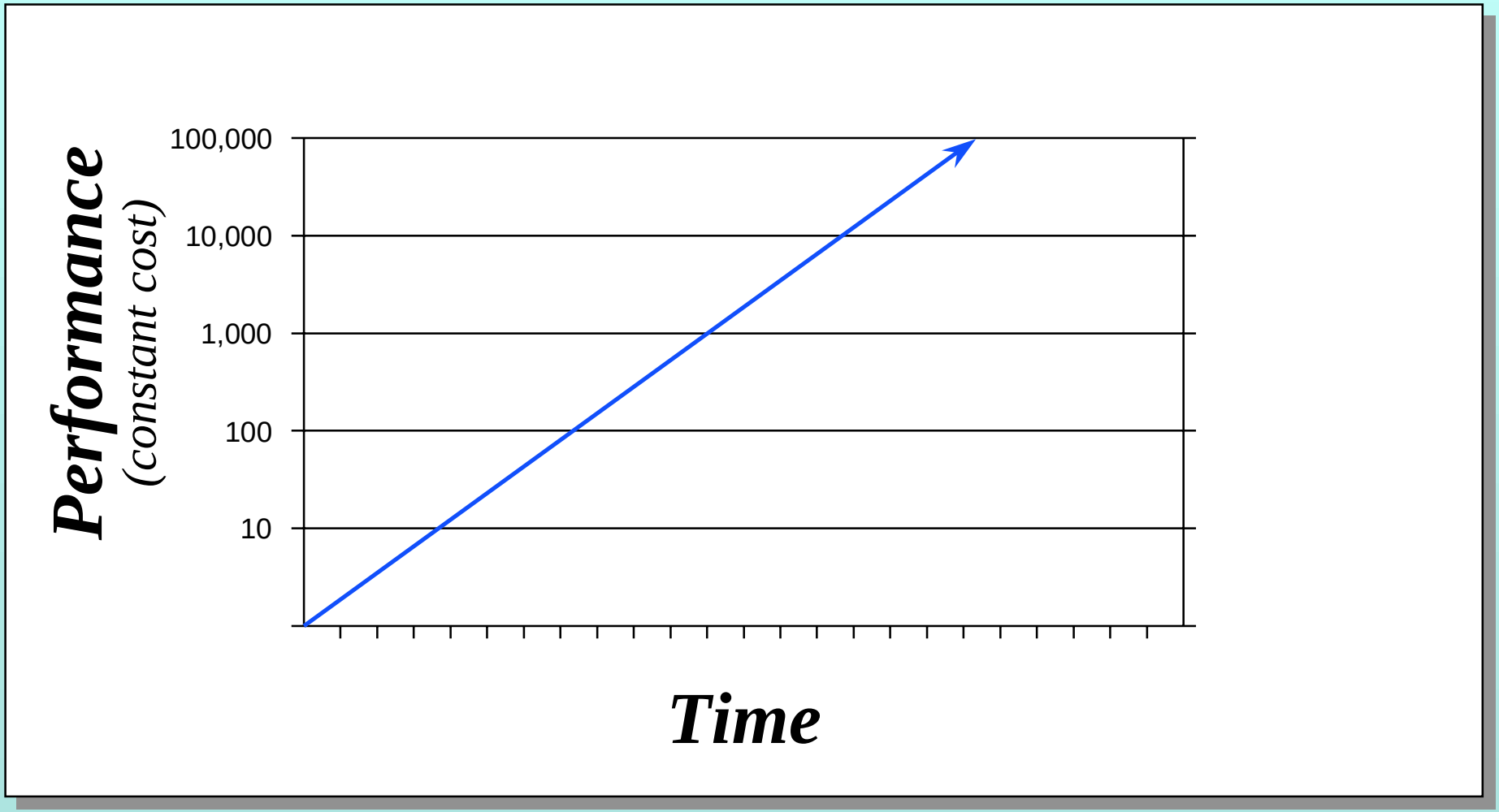
Transforms InfoTech (and everything else)

Moore's Law: *The Statement*

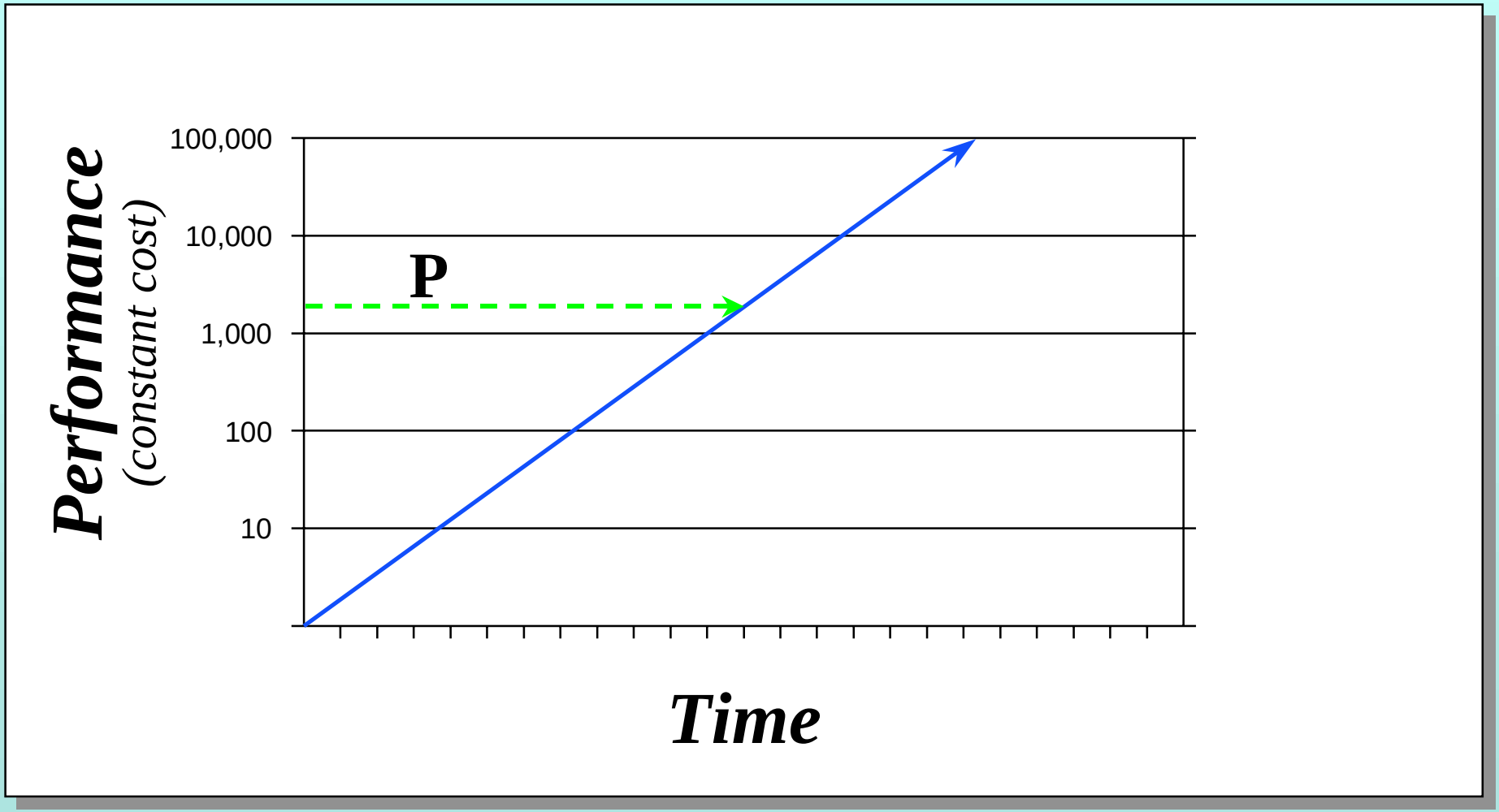
Every eighteen months, the number of transistors that can be placed on a chip doubles.

Gordon Moore, co-founder of Intel...

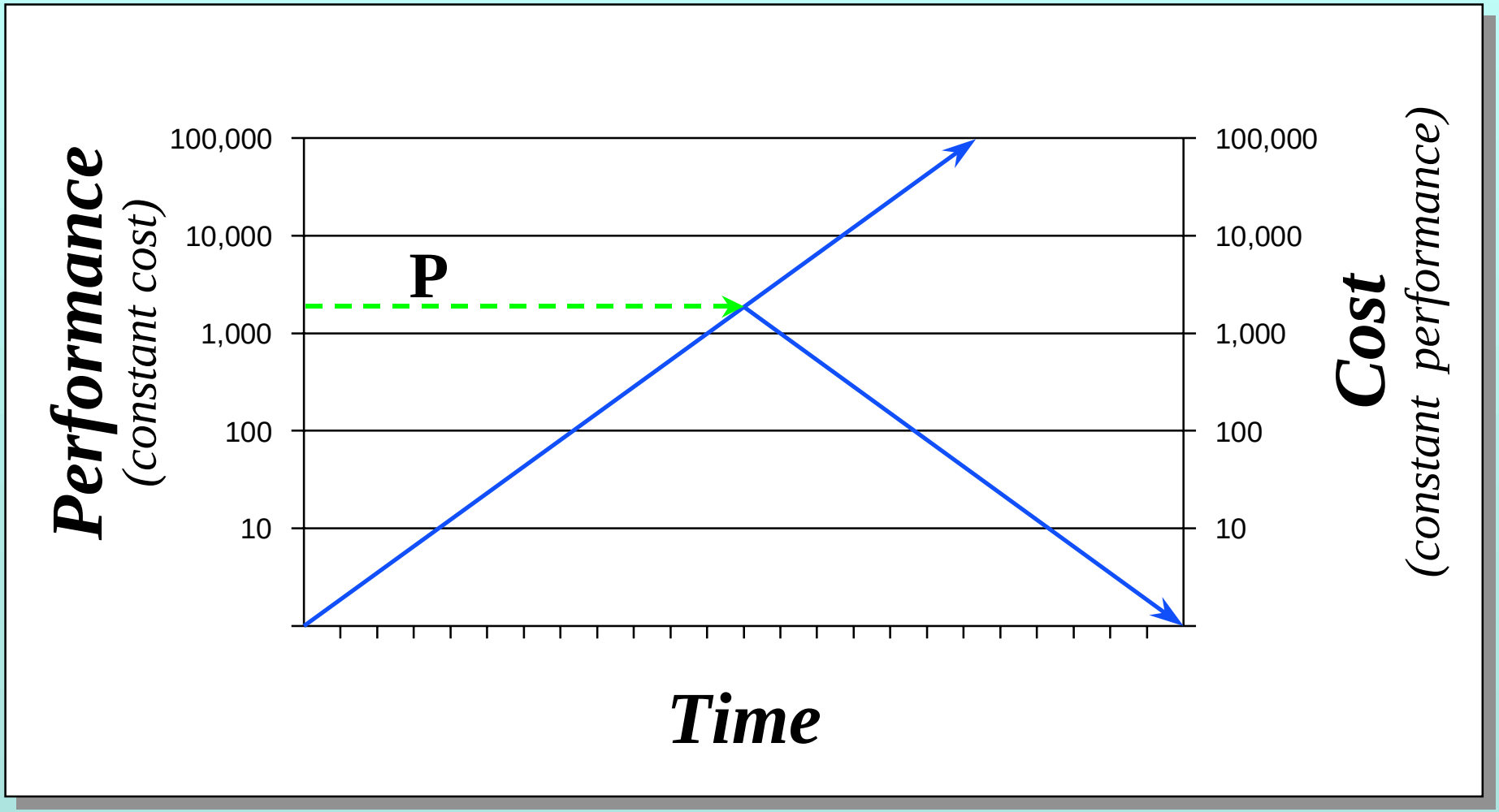
Moore's Law: *The Effect*



Moore's Law: *The Effect*



Moore's Law: *The Effect*



Moore's Law: *The Effect*

Three Phases of Novel IT Applications

- It's Impossible

Moore's Law: *The Effect*

Three Phases of Novel IT Applications

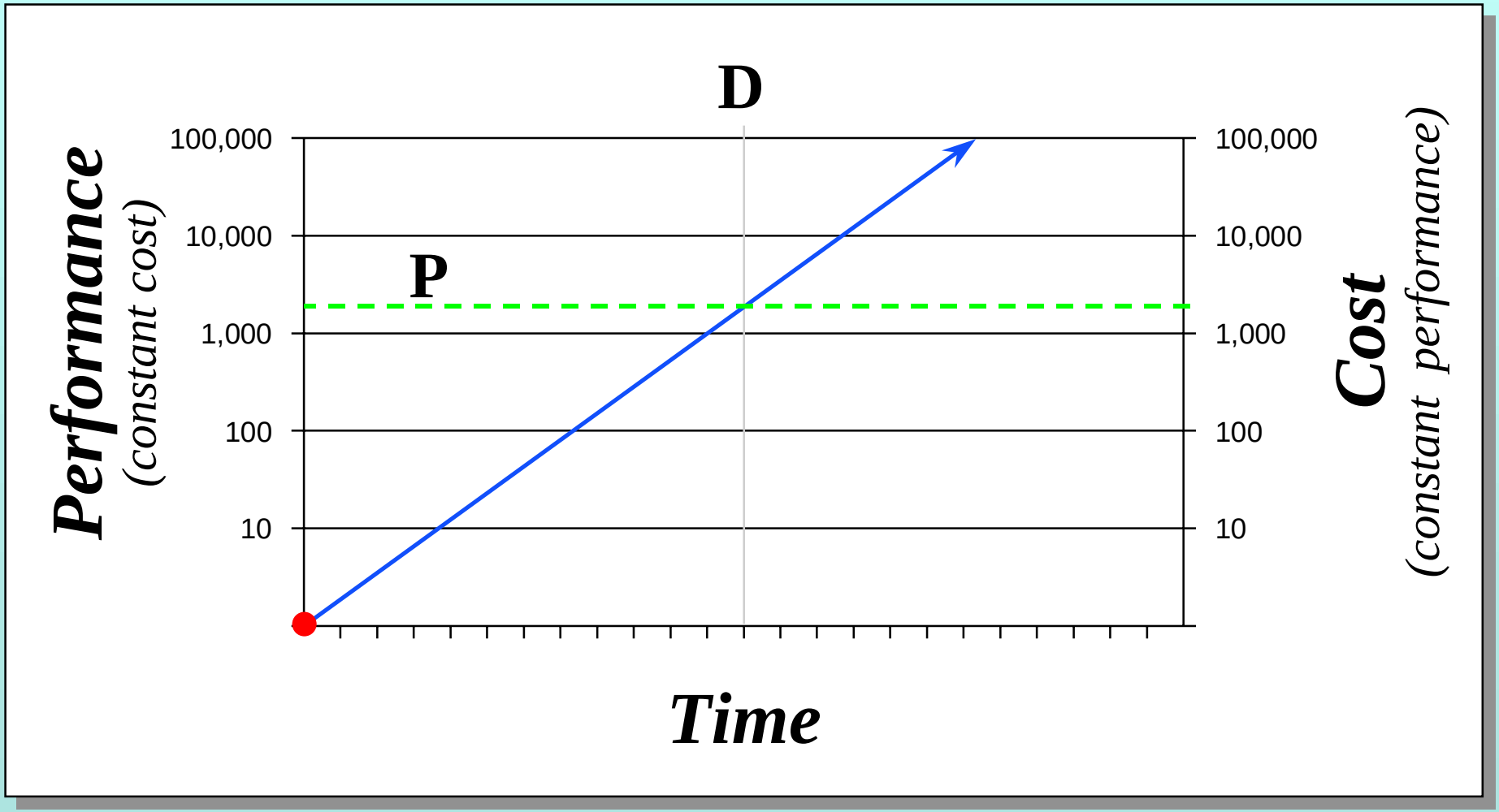
- It's Impossible
- It's Impractical

Moore's Law: *The Effect*

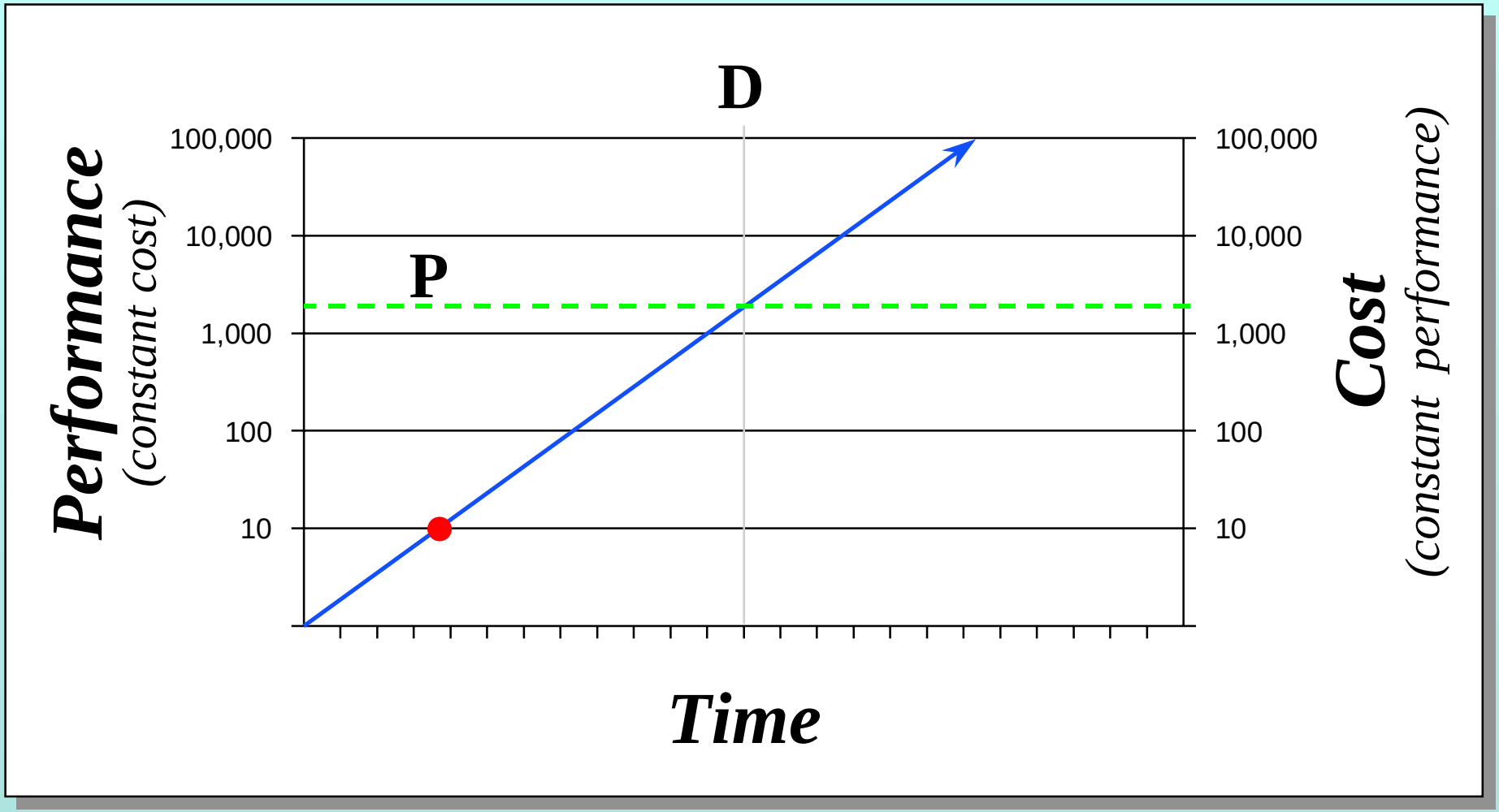
Three Phases of Novel IT Applications

- It's Impossible
- It's Impractical
- It's Overdue

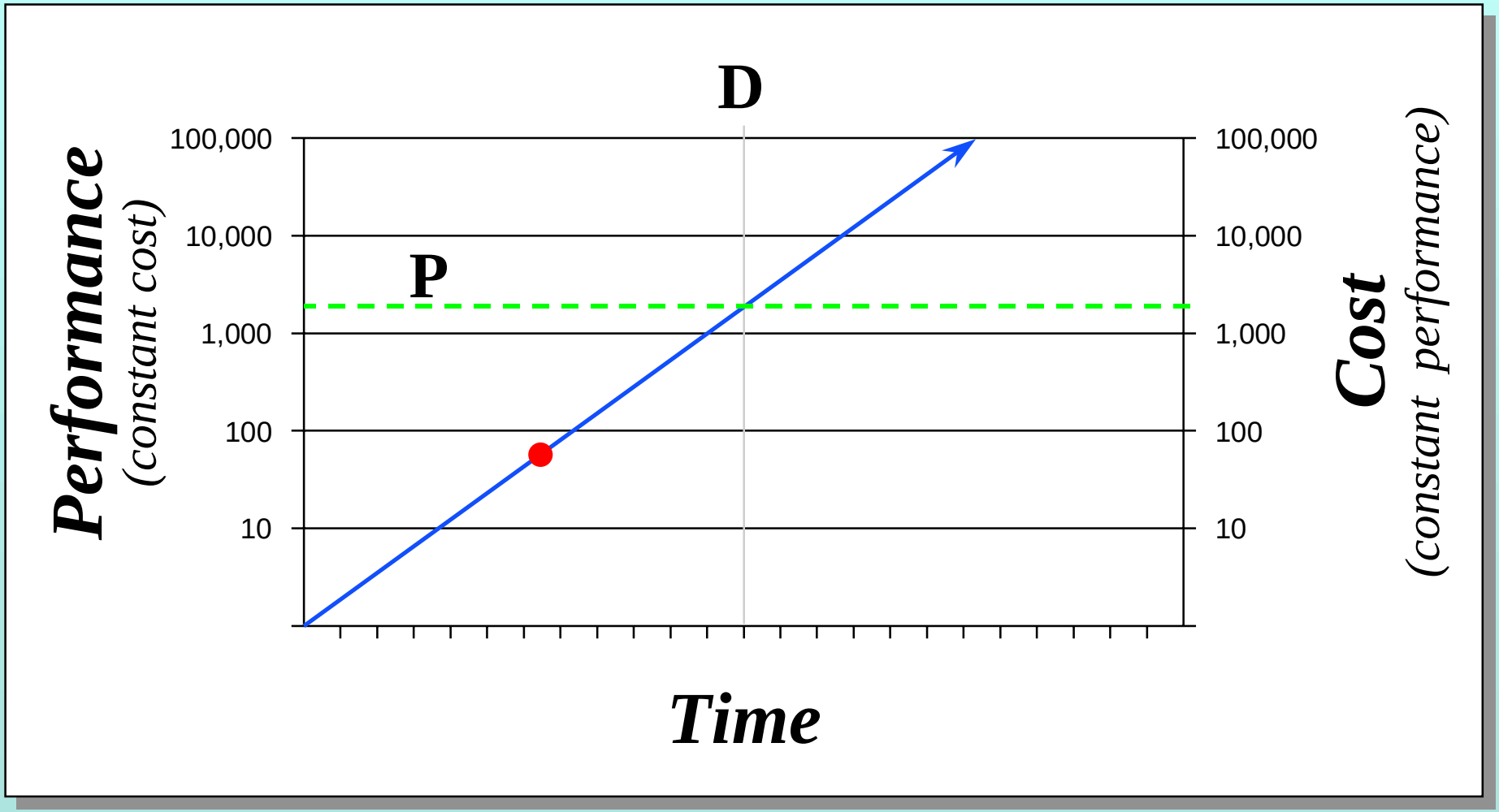
Moore's Law: *The Effect*



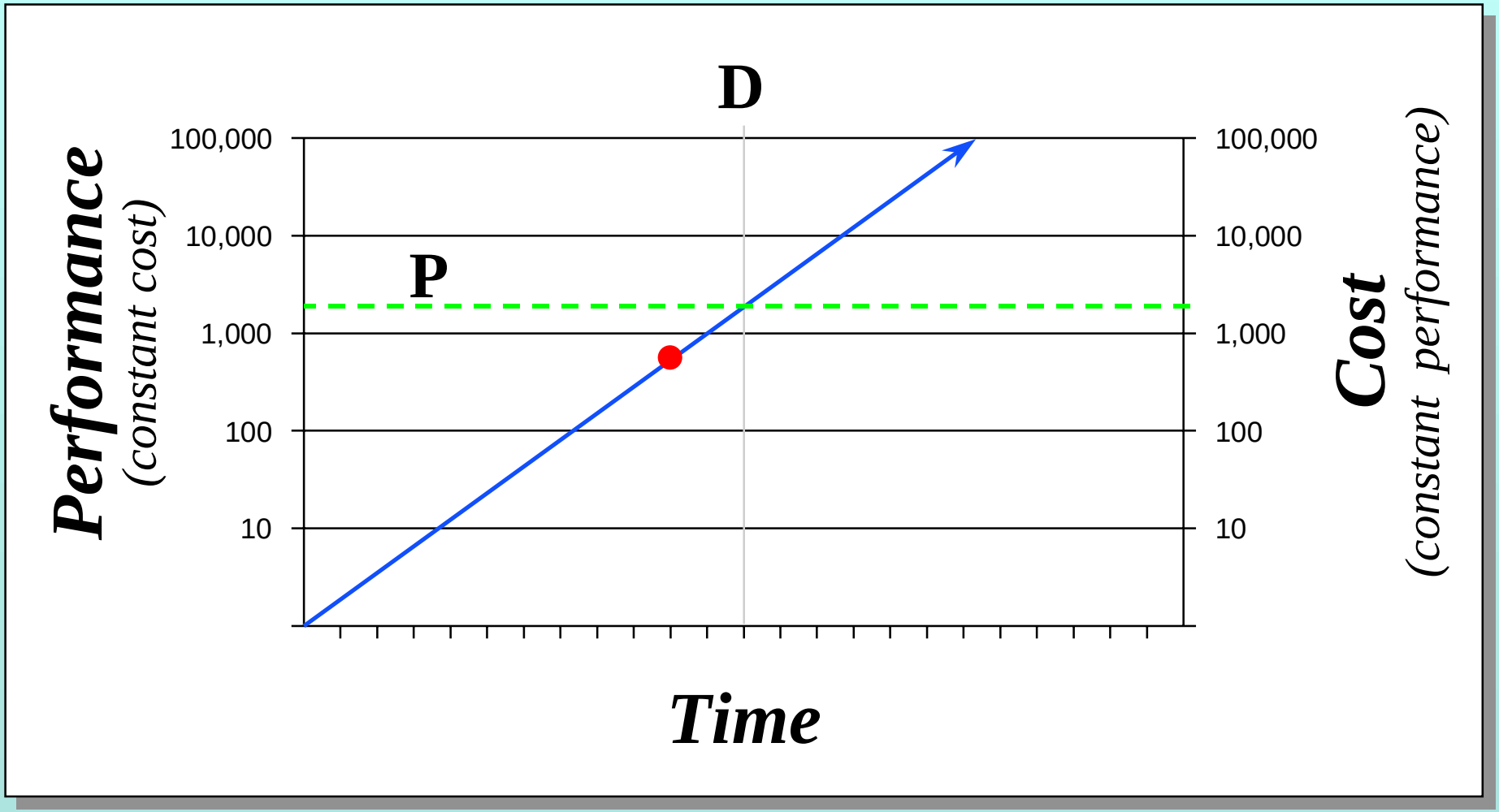
Moore's Law: *The Effect*



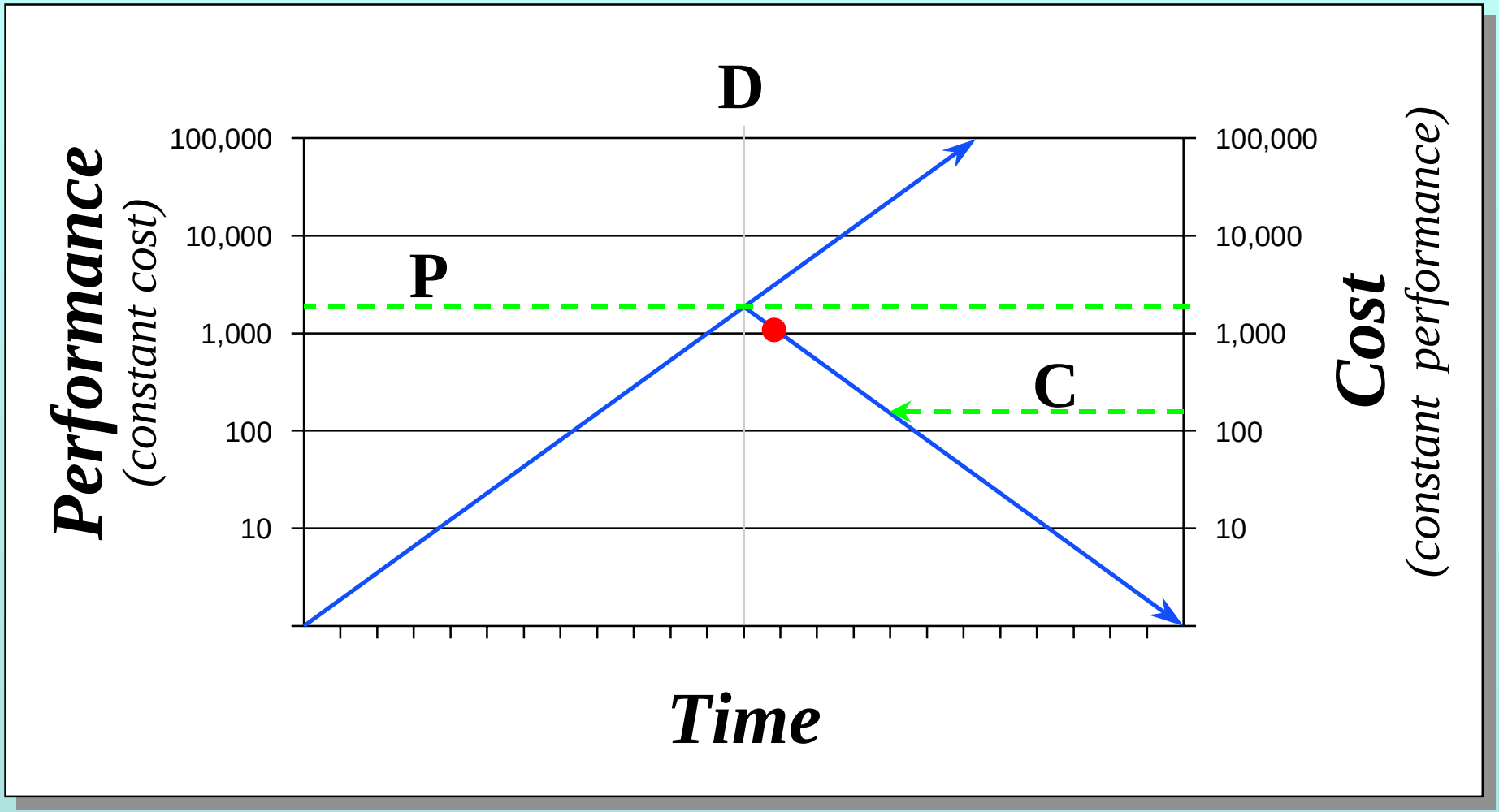
Moore's Law: *The Effect*



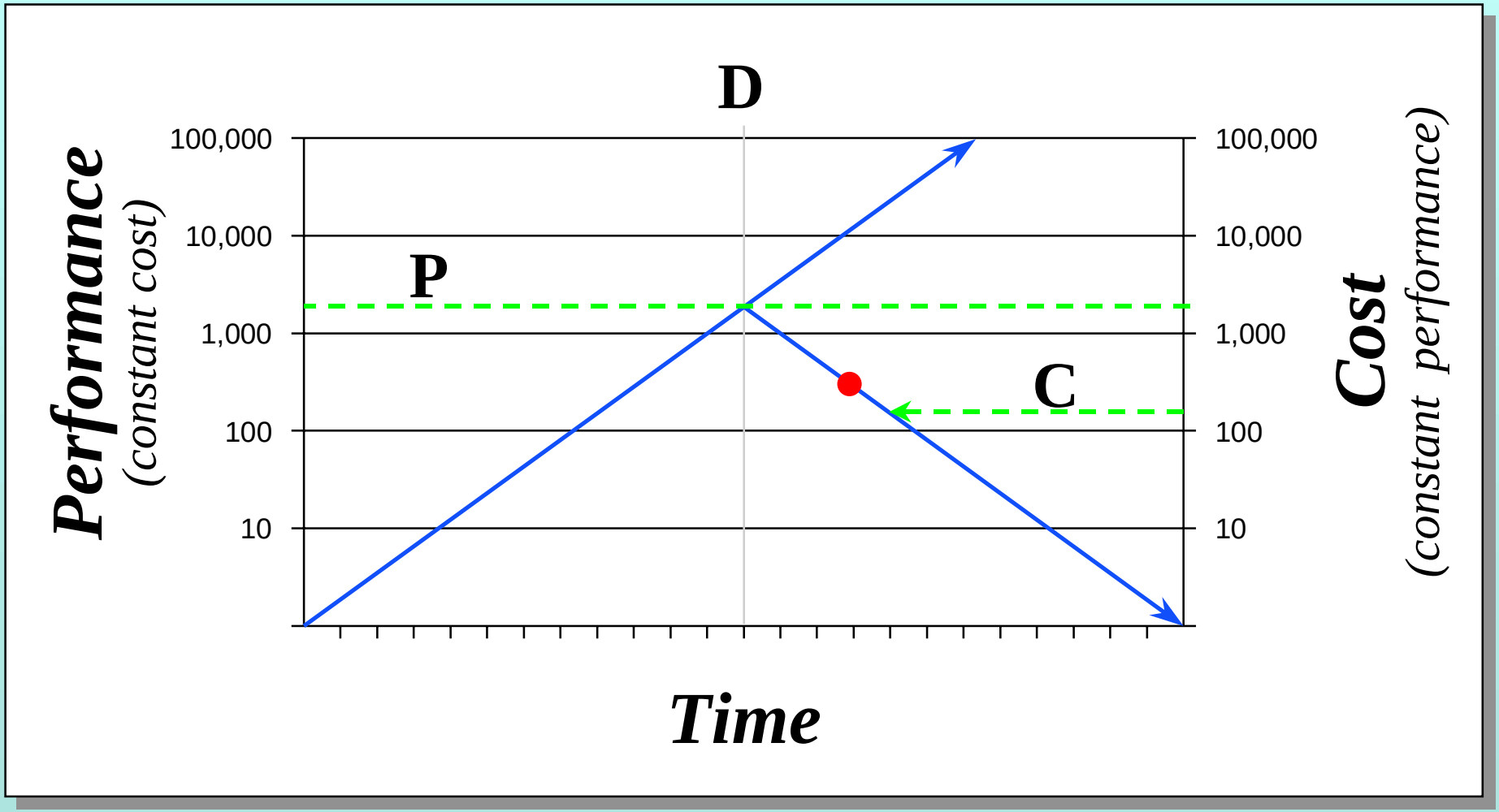
Moore's Law: *The Effect*



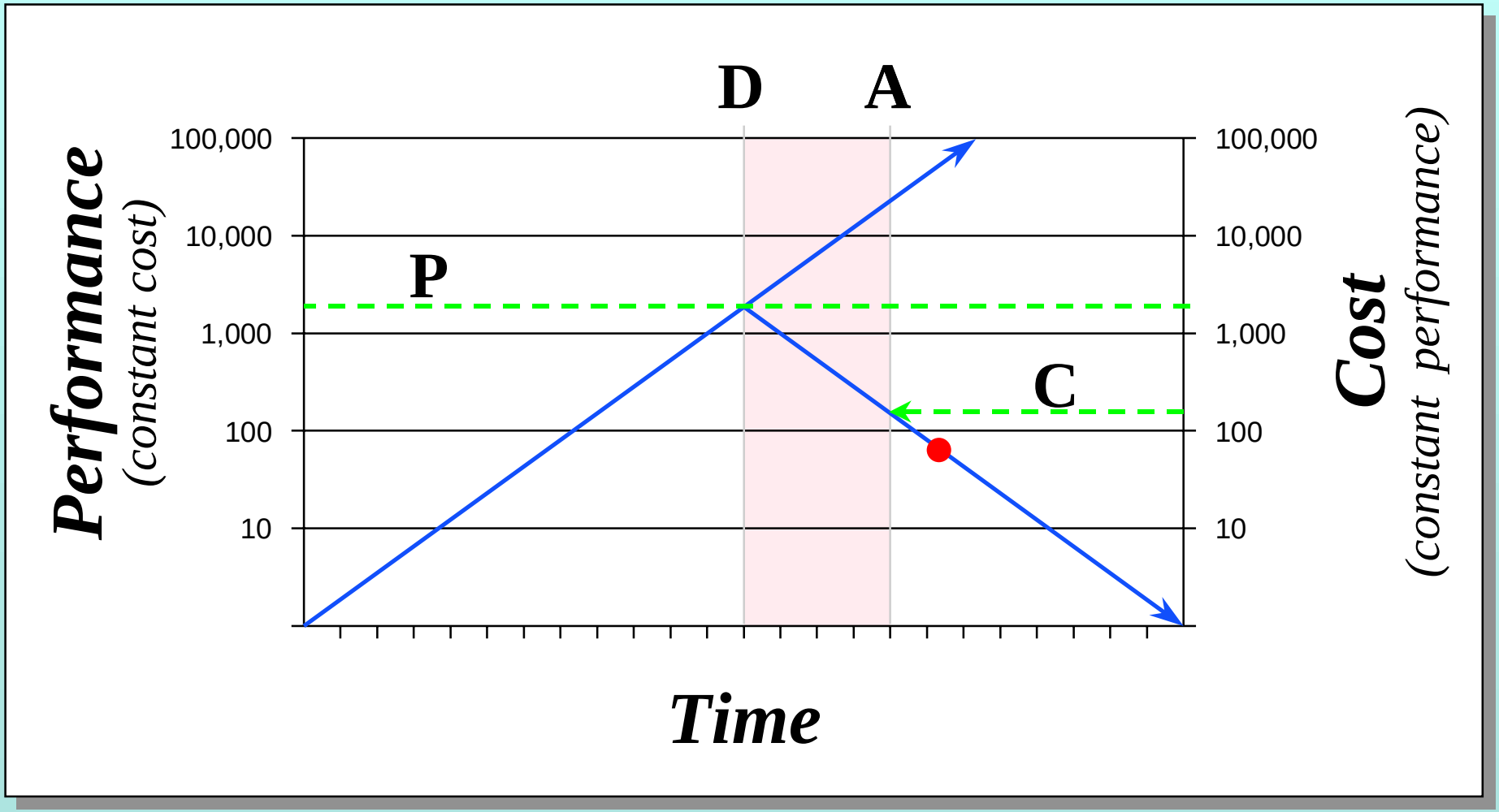
Moore's Law: *The Effect*



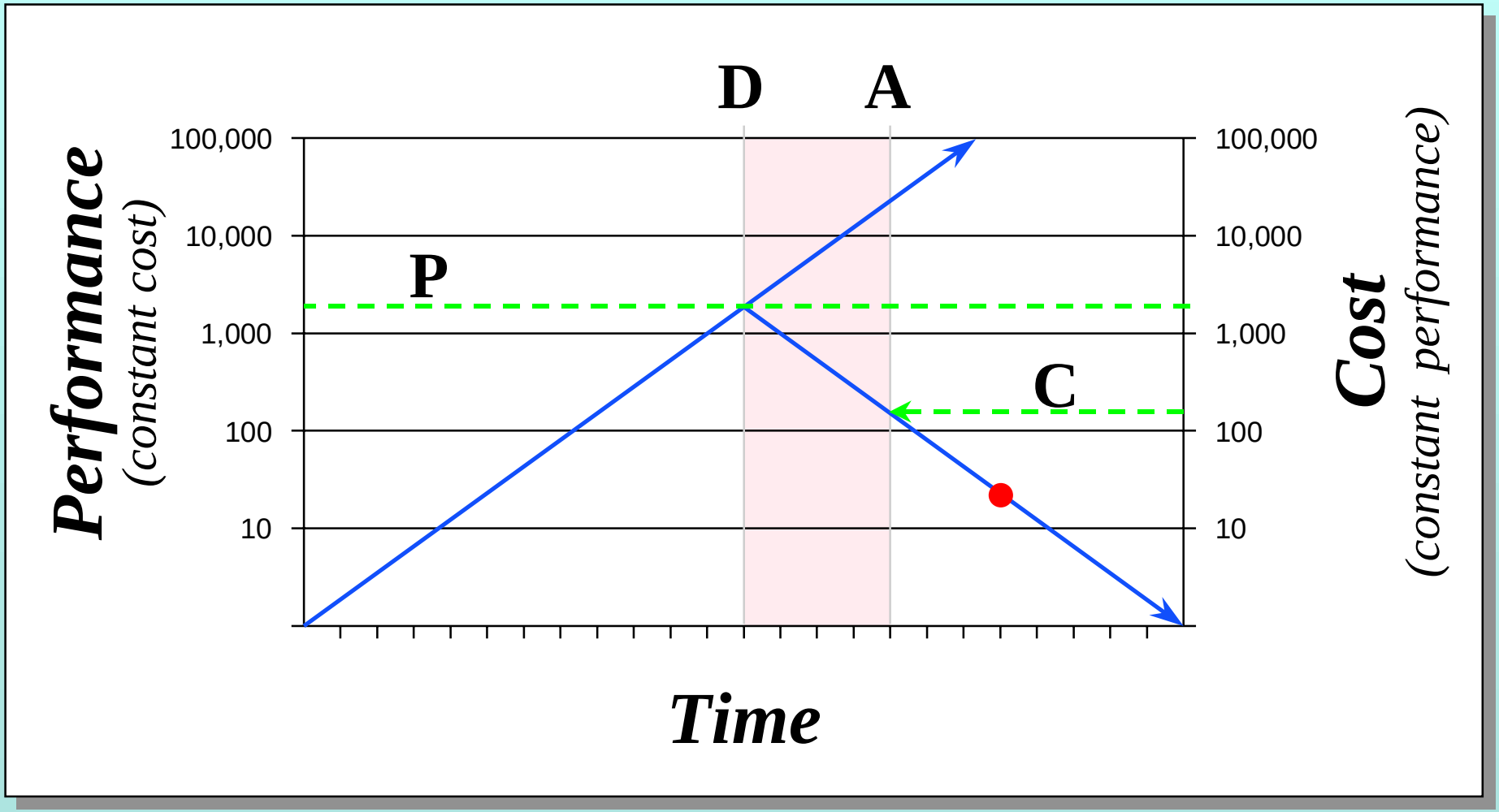
Moore's Law: *The Effect*



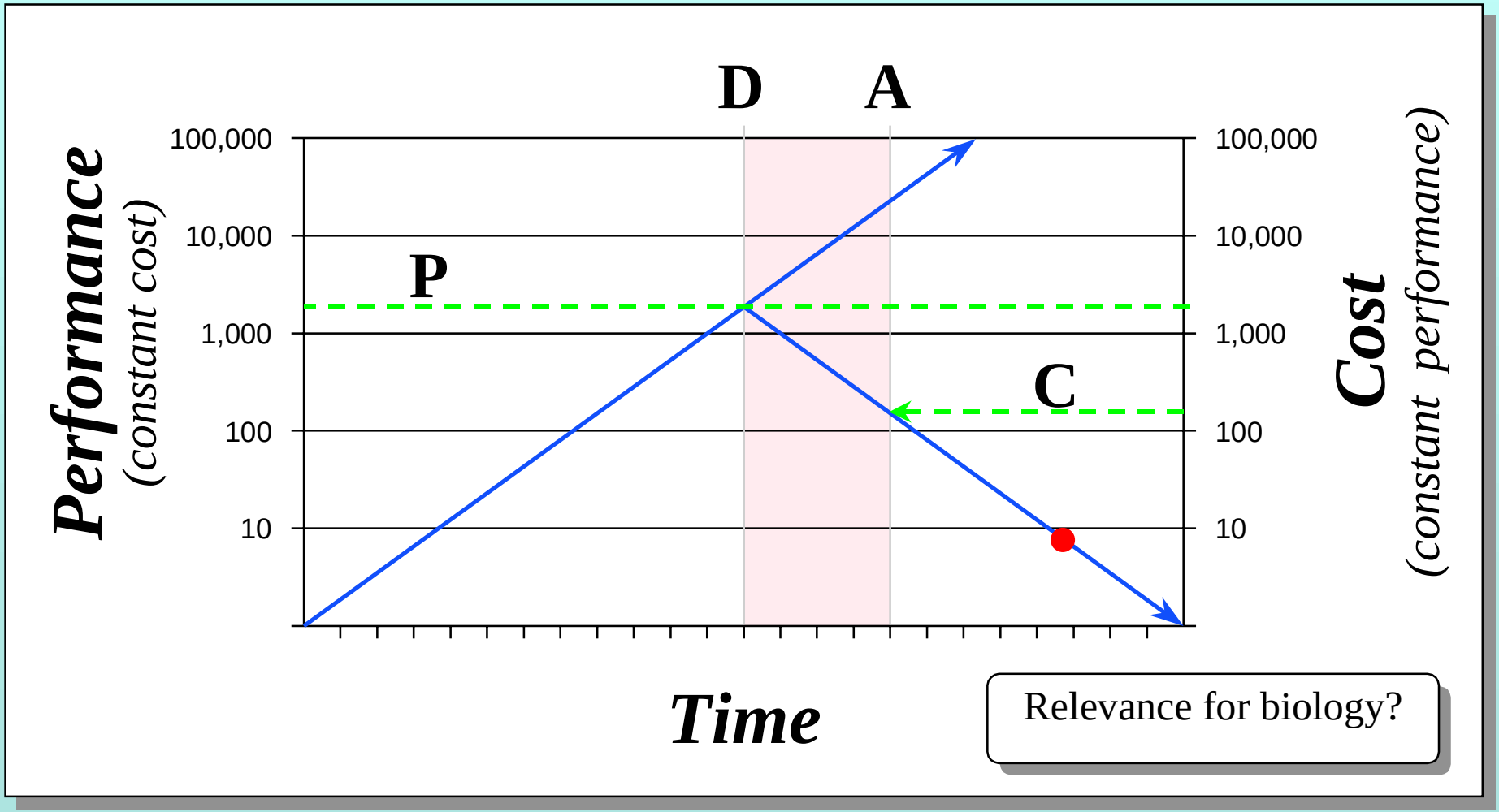
Moore's Law: *The Effect*



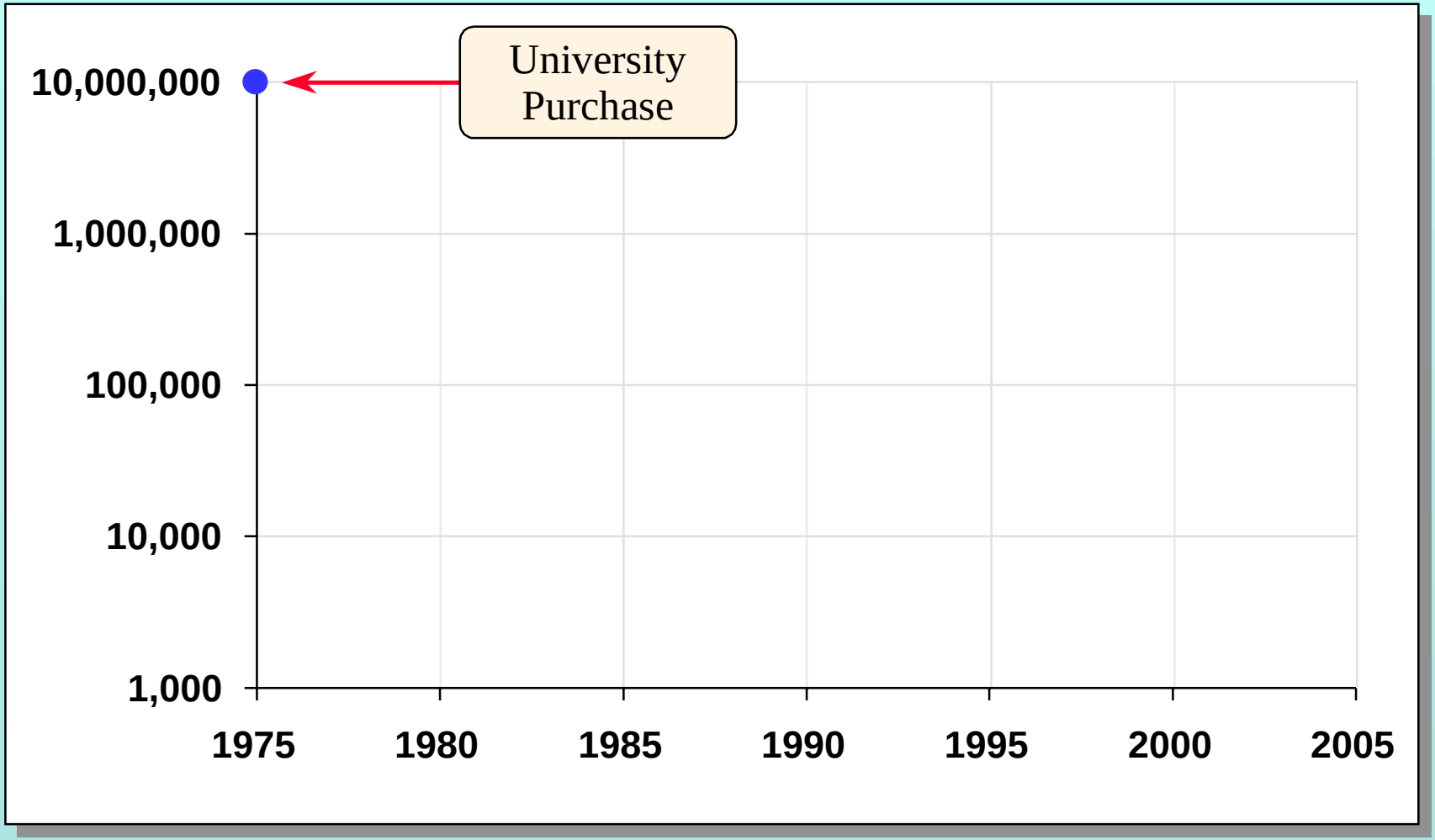
Moore's Law: *The Effect*



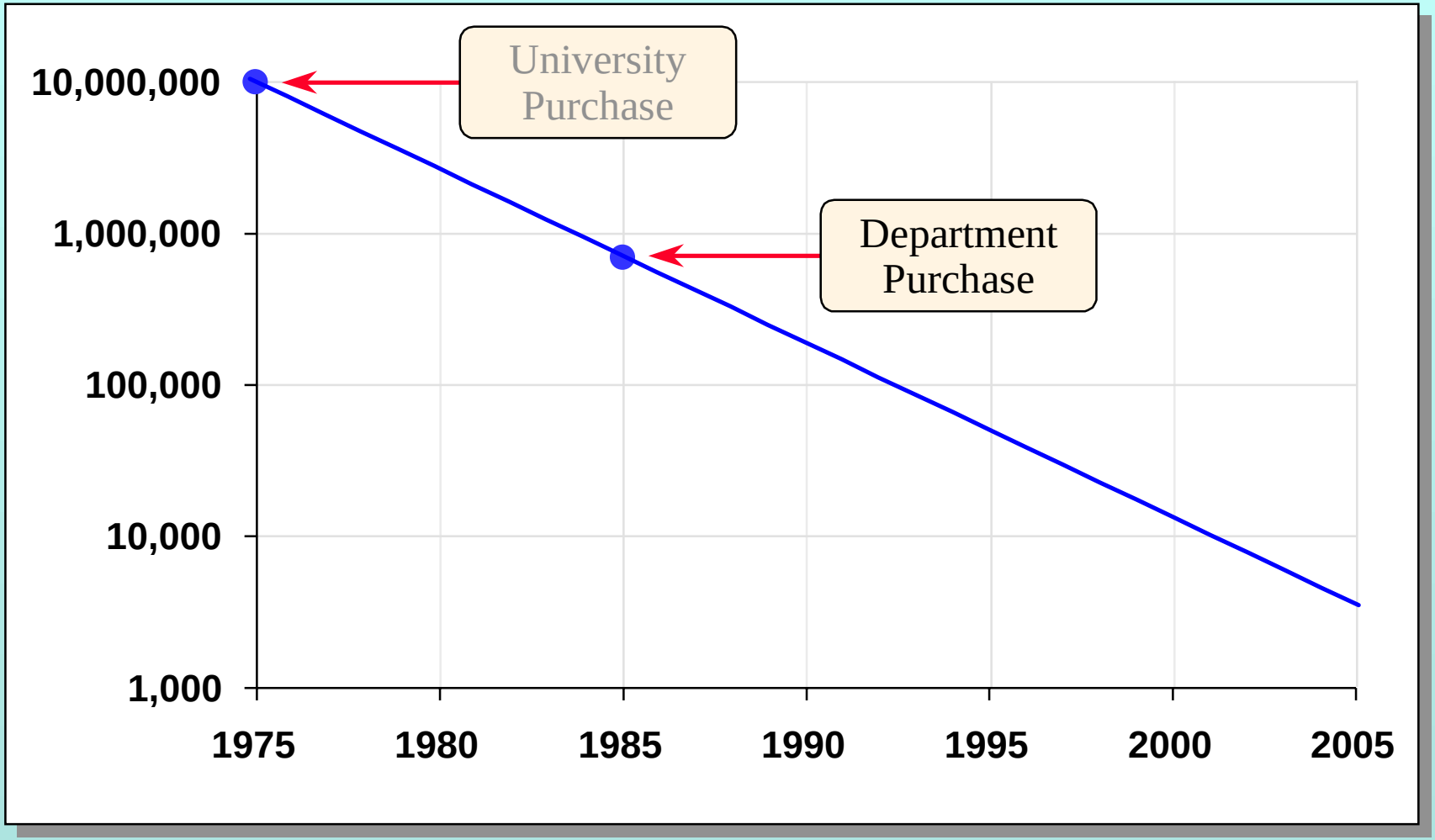
Moore's Law: *The Effect*



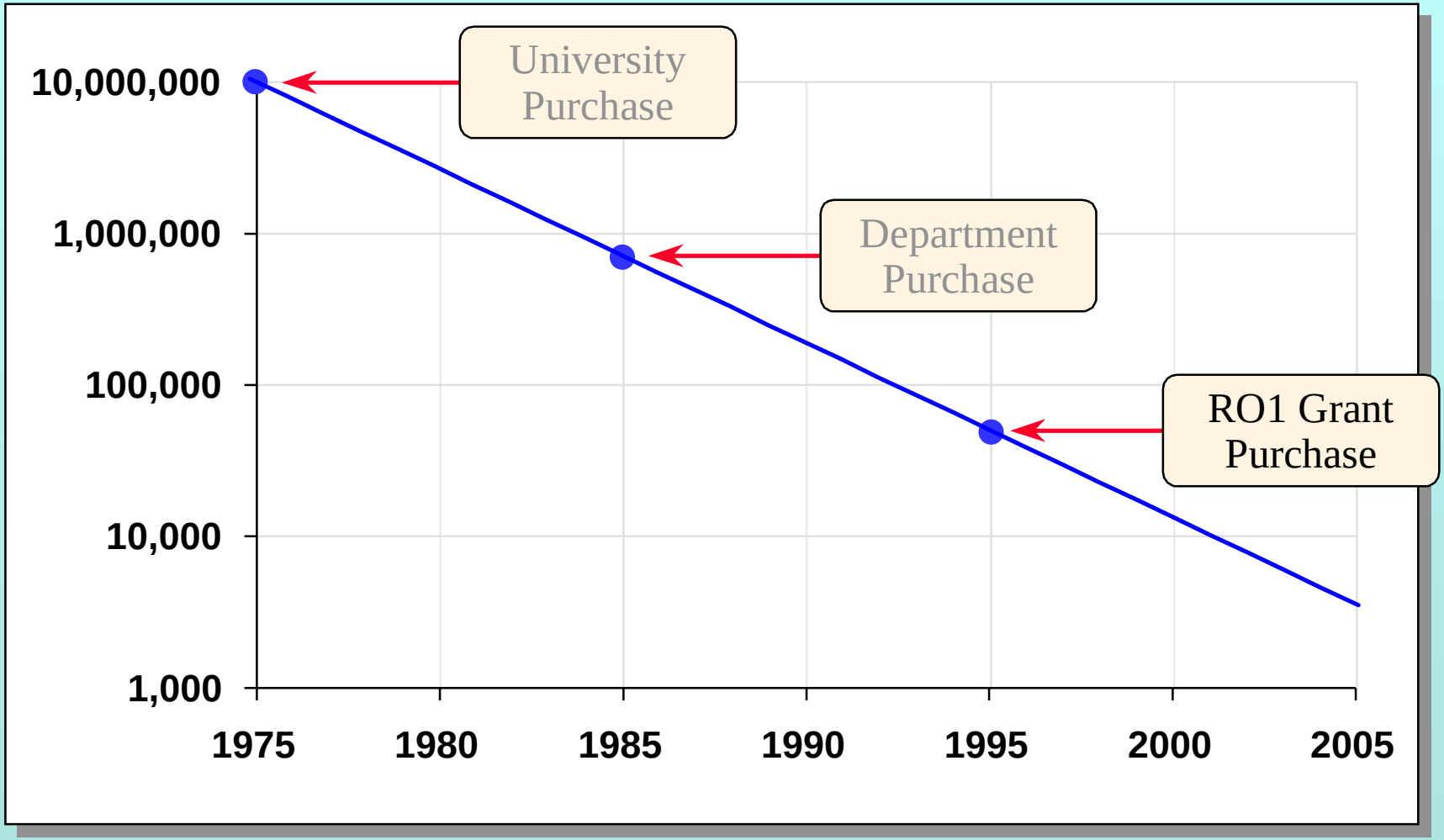
Cost (constant performance)



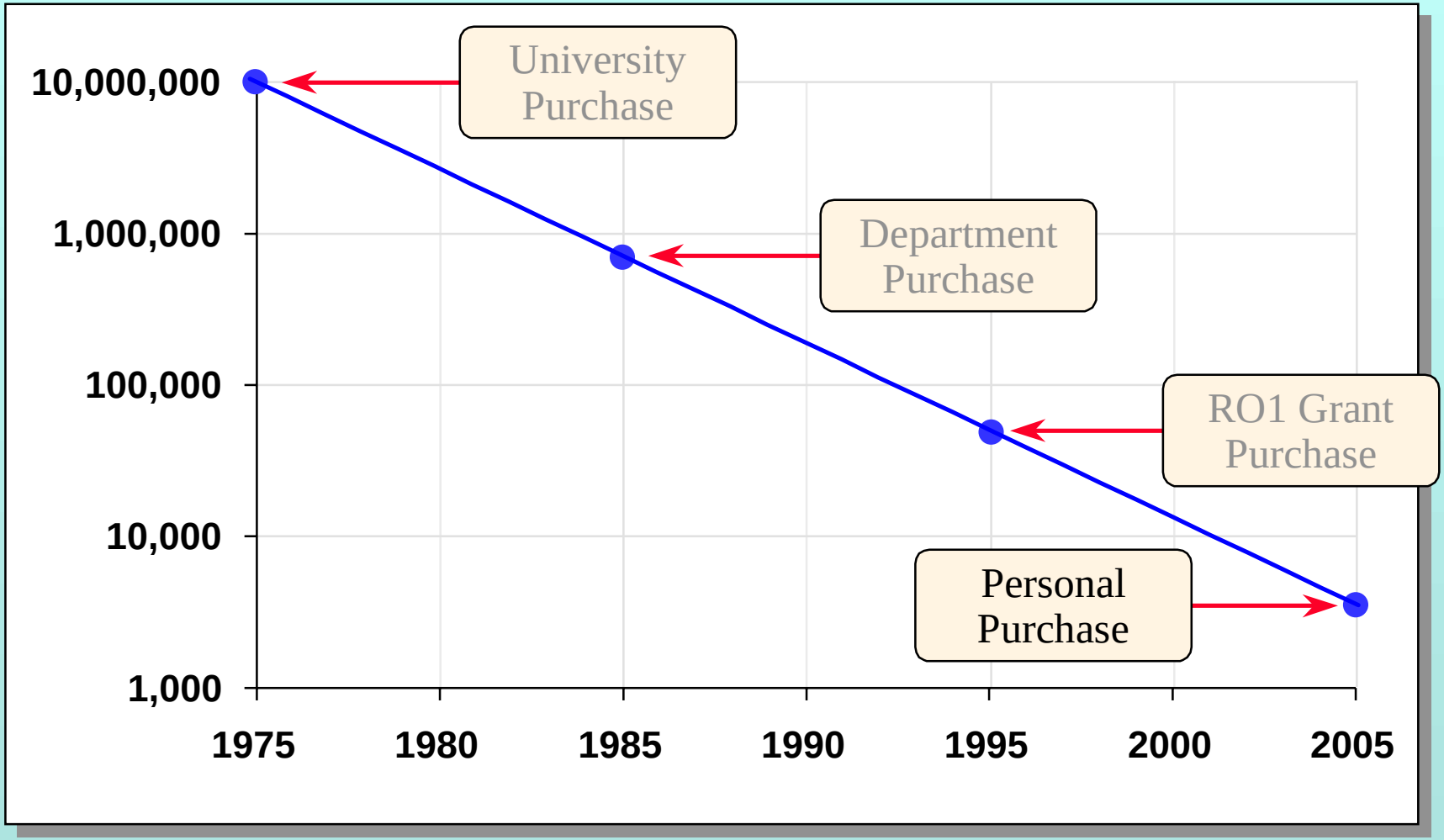
Cost (constant performance)



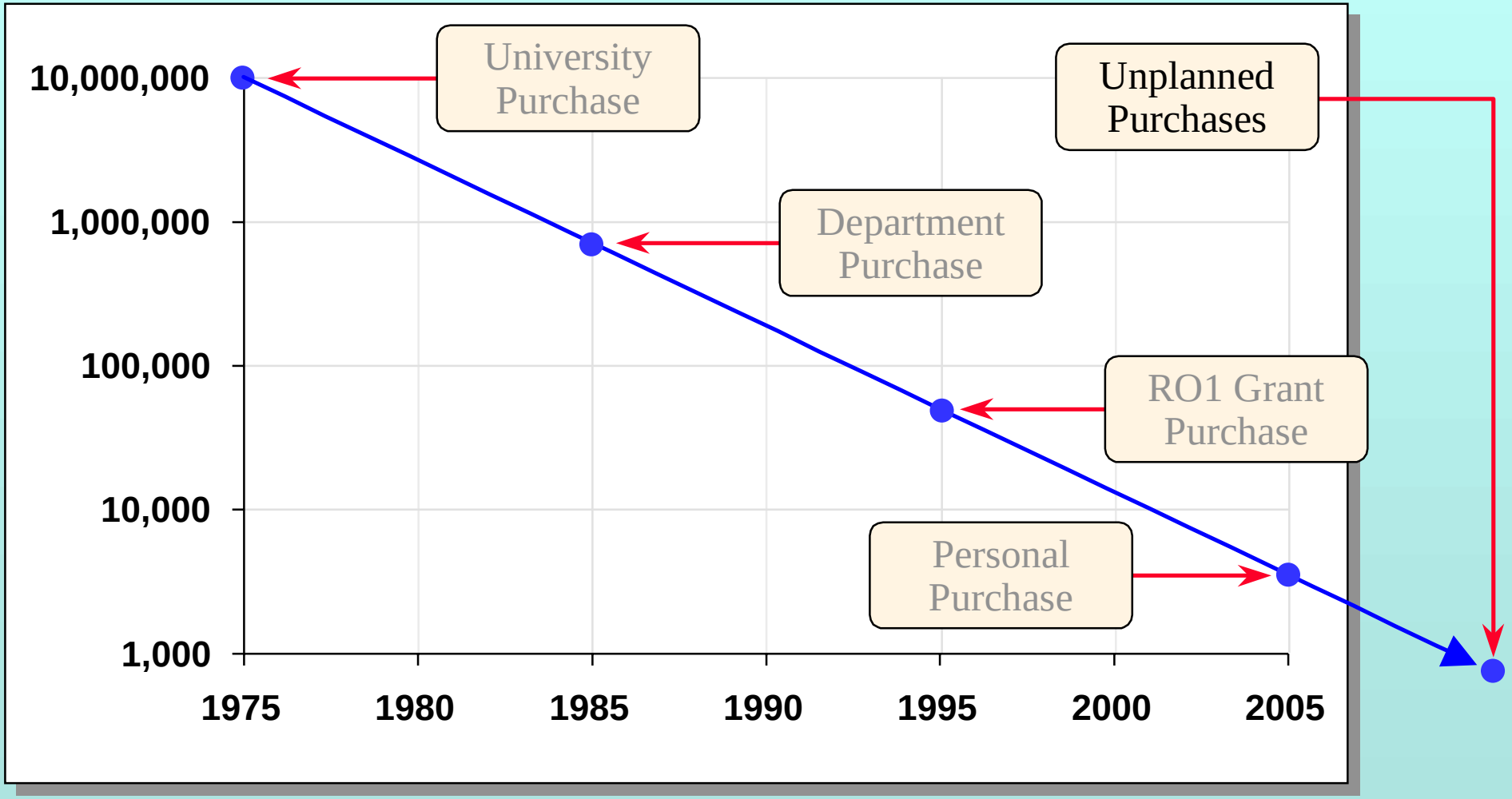
Cost (constant performance)



Cost (constant performance)



Cost (constant performance)



IT-Biology Synergism

IT is Special

Information Technology:

- *affects the performance **and** the management of tasks*

IT is Special

Information Technology:

- *affects the performance **and** the management of tasks*
- *allows the manipulation of huge amounts of highly complex data*

IT is Special

Information Technology:

- *affects the performance **and** the management of tasks*
- *allows the manipulation of huge amounts of highly complex data*
- *is incredibly plastic*

(programming and poetry are both exercises in pure thought)

IT is Special

Information Technology:

- *affects the performance **and** the management of tasks*
- *allows the manipulation of huge amounts of highly complex data*
- *is incredibly plastic*
  (programming and poetry are both exercises in pure thought)
- *improves exponentially* *(Moore's Law)*

Biology is Special

Life is Characterized by:

- *individuality*

Biology is Special

Life is Characterized by:

- *individuality*
- *historicity*

Biology is Special

Life is Characterized by:

- *individuality*
- *historicity*
- *contingency*

Biology is Special

Life is Characterized by:

- *individuality*
- *historicity*
- *contingency*
- *high (digital) information content*

Biology is Special

Life is Characterized by:

- *individuality*
- *historicity*
- *contingency*
- *high (digital) information content*

No law of large numbers...

Biology is Special

Life is Characterized by:

- *individuality*
- *historicity*
- *contingency*
- *high (digital) information content*

No law of large numbers, since every living thing is genuinely unique.

IT-Biology Synergism

- *Physics needs calculus, the method for manipulating information about statistically large numbers of vanishingly small, independent, equivalent things.*

IT-Biology Synergism

- *Physics needs calculus, the method for manipulating information about statistically large 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 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 Schrödinger. 1944. *What is Life*.

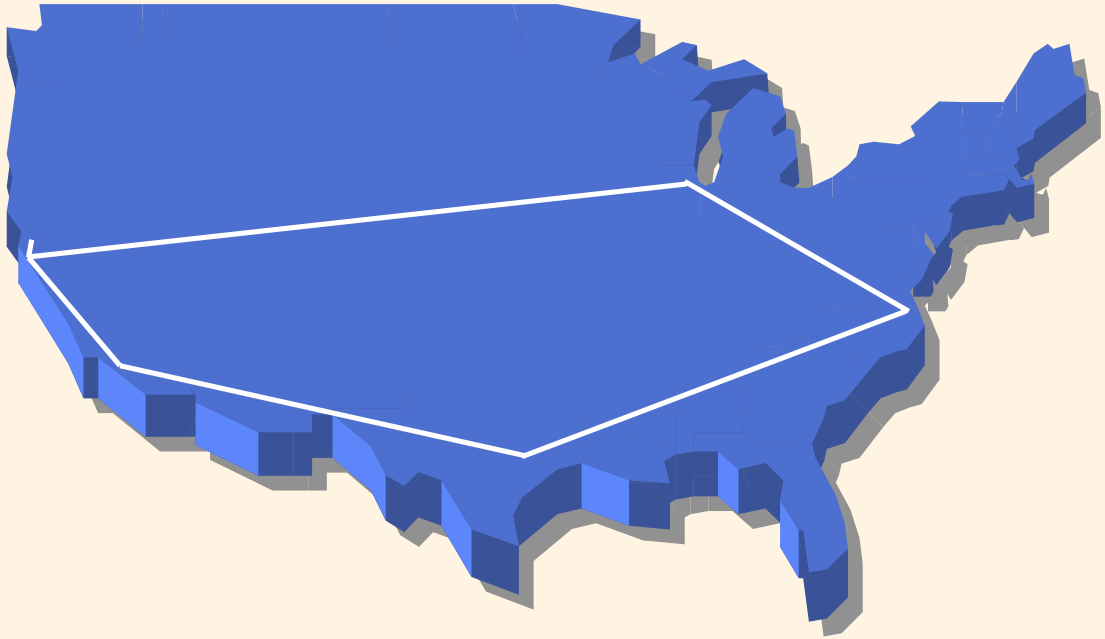
Genetics as Code

[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 Schrödinger. 1944. *What is Life*.

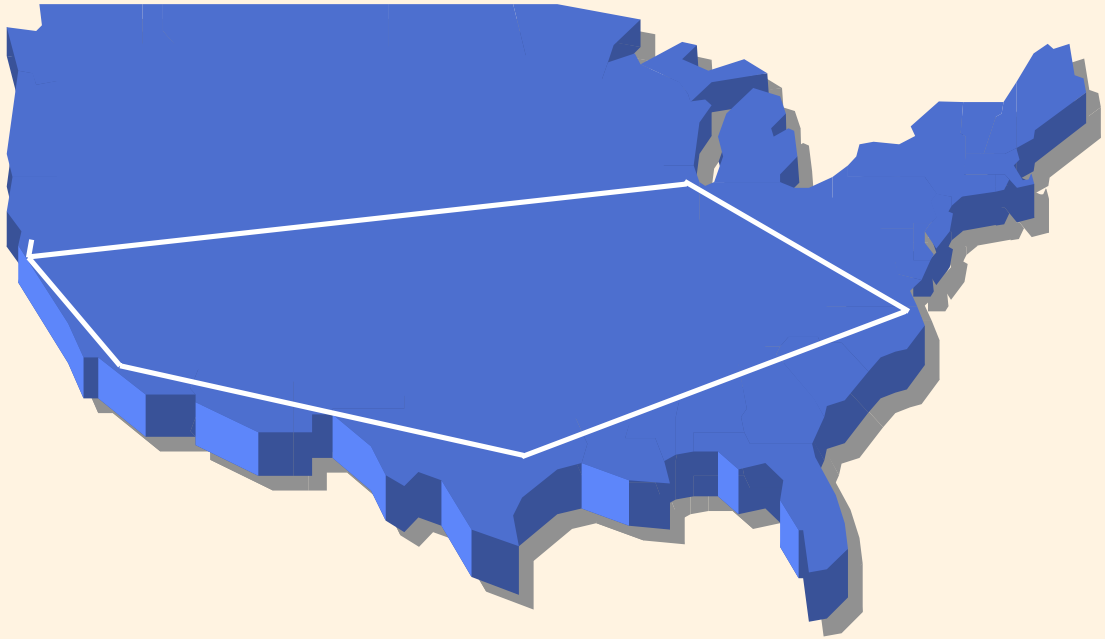
One Human Sequence

We now know that Schrödinger's mysterious human "code-script" consists of 3.3 billion base pairs of DNA.



One Human Sequence

We now know that Schrödinger's mysterious human "code-script" consists of 3.3 billion base pairs of DNA.



Typed in 10-pitch font, one human sequence would stretch for more than 5,000 miles. Digitally formatted, it could be stored on one CD-ROM. Biologically encoded, it fits easily within a single cell.

Bio-digital Information

DNA is a highly efficient digital storage device:

- There is more mass-storage capacity in the DNA of a side of beef than in all the hard drives of all the world's computers.

Bio-digital Information

DNA is a highly efficient digital storage device:

- There is more mass-storage capacity in the DNA of a side of beef than in all the hard drives of all the world's computers.
- Duplicating the mass storage capacity in the DNA of the entire biosphere would require 10^{27} 10 gB hard disks.

Bio-digital Information

DNA is a highly efficient digital storage device:

- There is more mass-storage capacity in the DNA of a side of beef than in all the hard drives of all the world's computers.
- Duplicating the mass storage capacity in the DNA of the entire biosphere would require 10^{27} 10 gB hard disks. That many hard disks would have a volume of 3.9×10^{13} cubic miles.

Bio-digital Information

DNA is a highly efficient digital storage device:

- There is more mass-storage capacity in the DNA of a side of beef than in all the hard drives of all the world's computers.
- Duplicating the mass storage capacity in the DNA of the entire biosphere would require 10^{27} 10 gB hard disks. That many hard disks would have a volume of 3.9×10^{13} cubic miles. The volume of the earth is 1.8×10^{11} cubic miles.

Genomics: 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

Human Genome Project - Goals

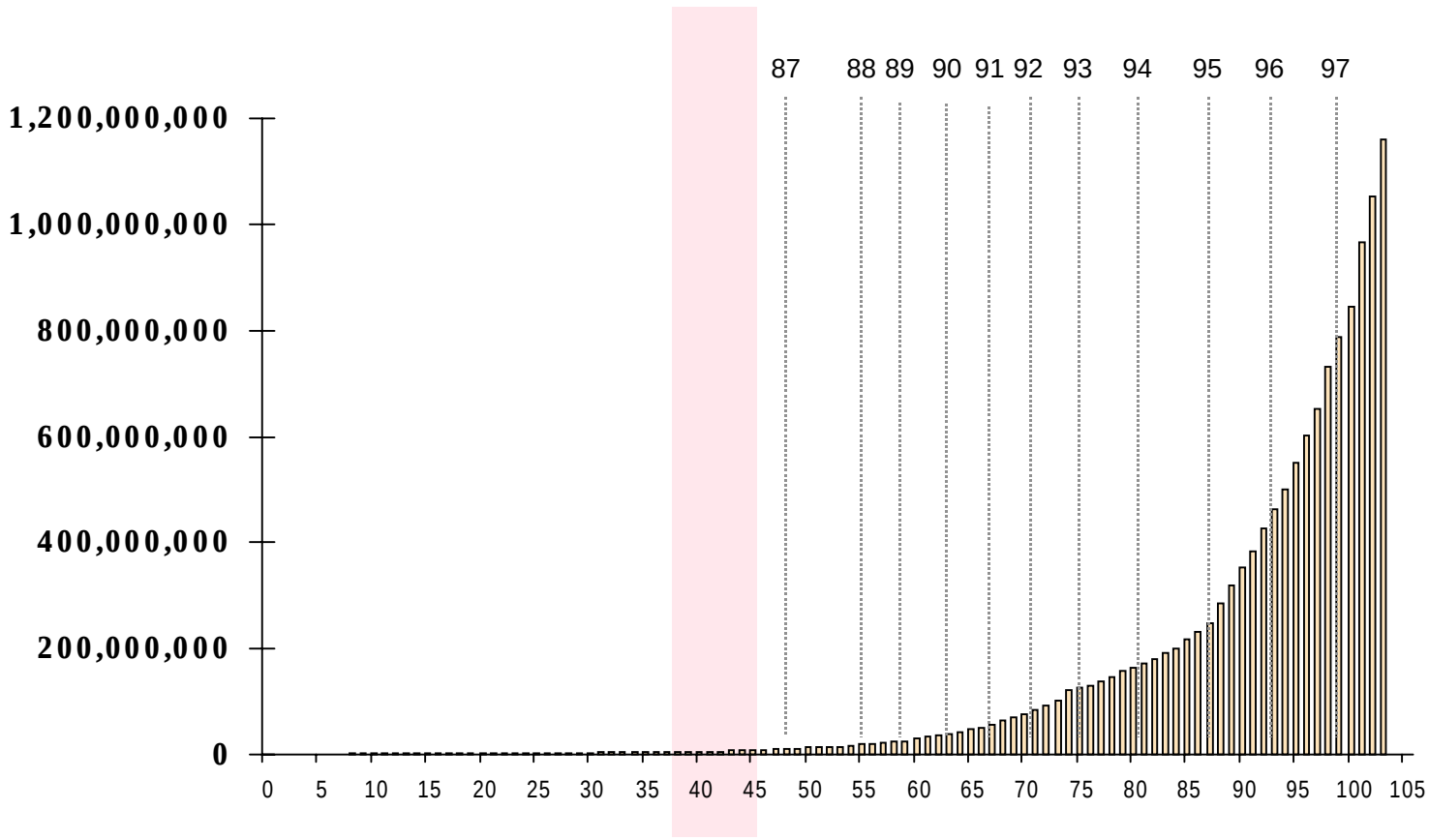
- construction of a high-resolution genetic map of the human genome;
- production of a variety of physical maps of all human chromosomes and of the DNA of selected model organisms;
- determination of the complete sequence of human DNA and of the DNA of selected model organisms;
- development of capabilities 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.*

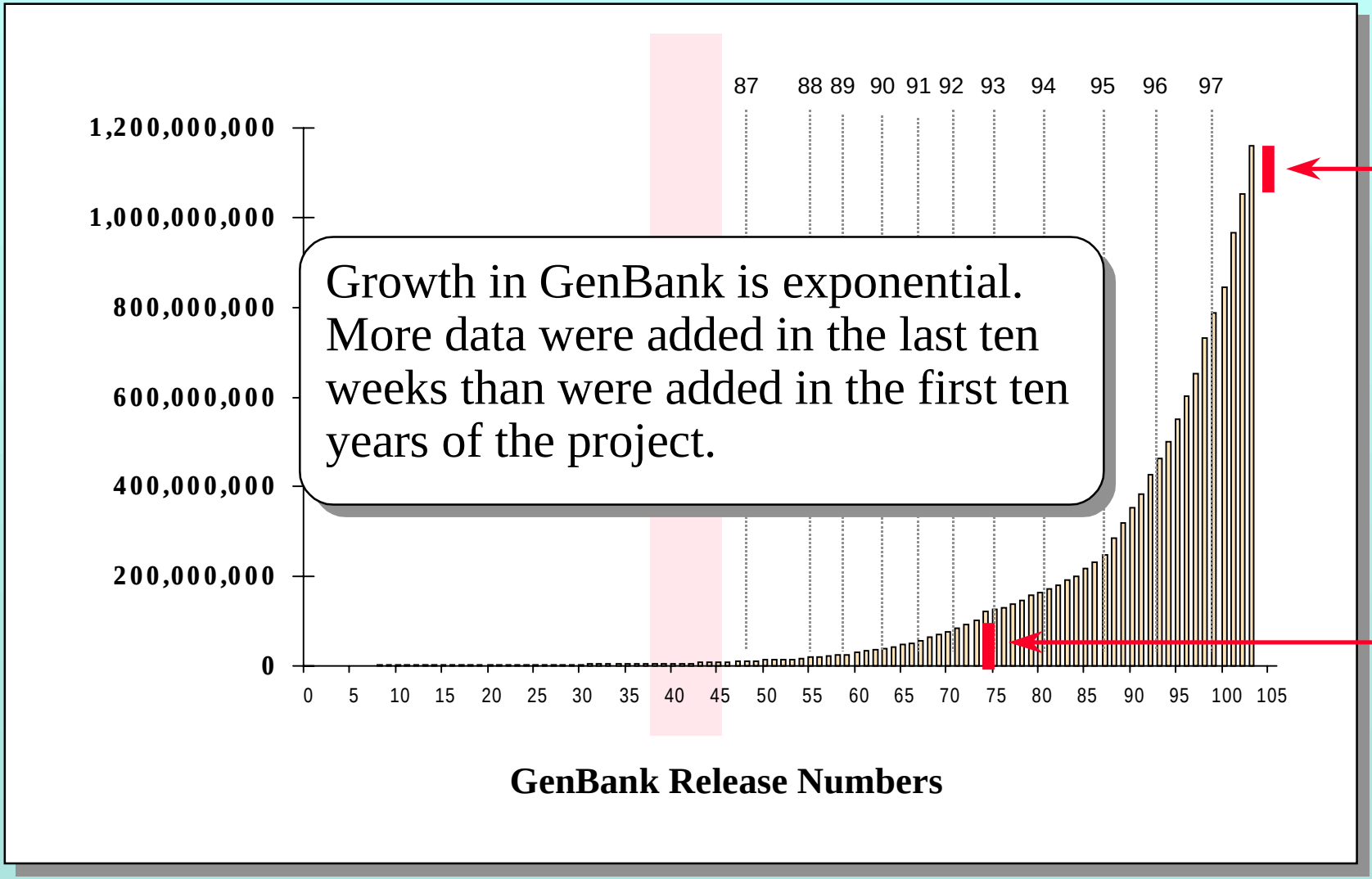
GenBank Totals *(Release 103)*

DIVISION	Entries	Per Cent	Base Pairs	Per Cent
Phage Sequences (PHG)	1,313	0.074%	2,138,810	0.184%
Viral Sequences (VRL)	45,355	2.568%	44,484,848	3.834%
Bacteria (BCT)	38,023	2.153%	88,576,641	7.634%
Plant, Fungal, and Algal Sequences (PLN)	44,553	2.523%	92,259,434	7.951%
Invertebrate Sequences (INV)	29,657	1.679%	105,703,550	9.110%
Rodent Sequences (ROD)	36,967	2.093%	45,437,309	3.916%
Primate Sequences (PRI1–2)	75,587	4.280%	134,944,314	11.630%
Other Mammals (MAM)	12,744	0.722%	12,358,310	1.065%
Other Vertebrate Sequences (VRT)	17,713	1.003%	17,040,159	1.469%
High-Throughput Genome Sequences (HTG)	1,120	0.063%	72,064,395	6.211%
Genome Survey Sequences (GSS)	42,628	2.414%	22,783,326	1.964%
Structural RNA Sequences (RNA)	4,802	0.272%	2,487,397	0.214%
Sequence Tagged Sites Sequences (STS)	52,824	2.991%	18,161,532	1.565%
Patent Sequences (PAT)	87,767	4.970%	27,593,724	2.378%
Synthetic Sequences (SYN)	2,577	0.146%	5,698,945	0.491%
Unannotated Sequences (UNA)	2,480	0.140%	1,933,676	0.167%
EST1-17	**1,269,737**	**71.905%**	**466,634,317**	**40.217%**
TOTALS	**1,765,847**	**100.000%**	**1,160,300,687**	**100.000%**

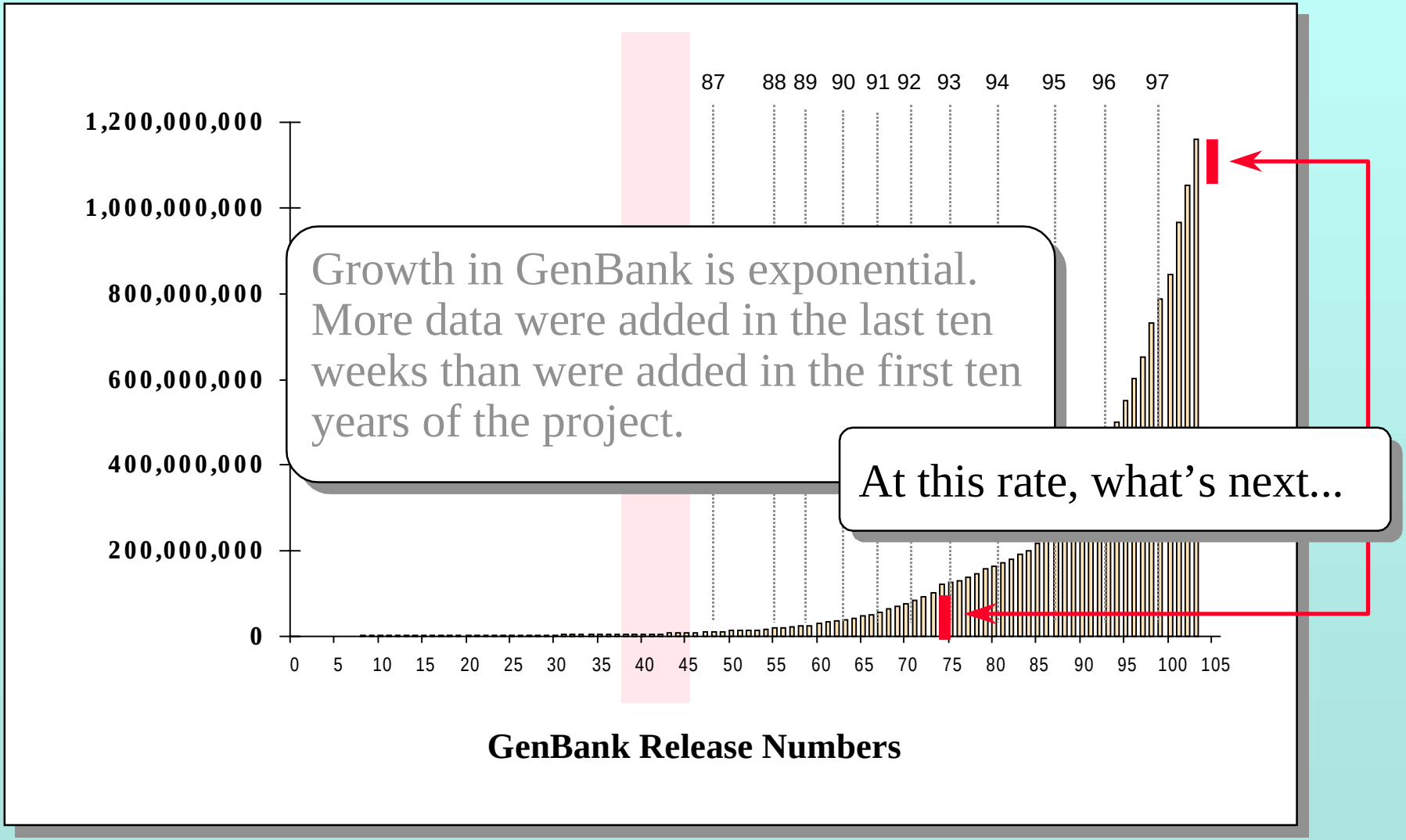
Base Pairs in GenBank



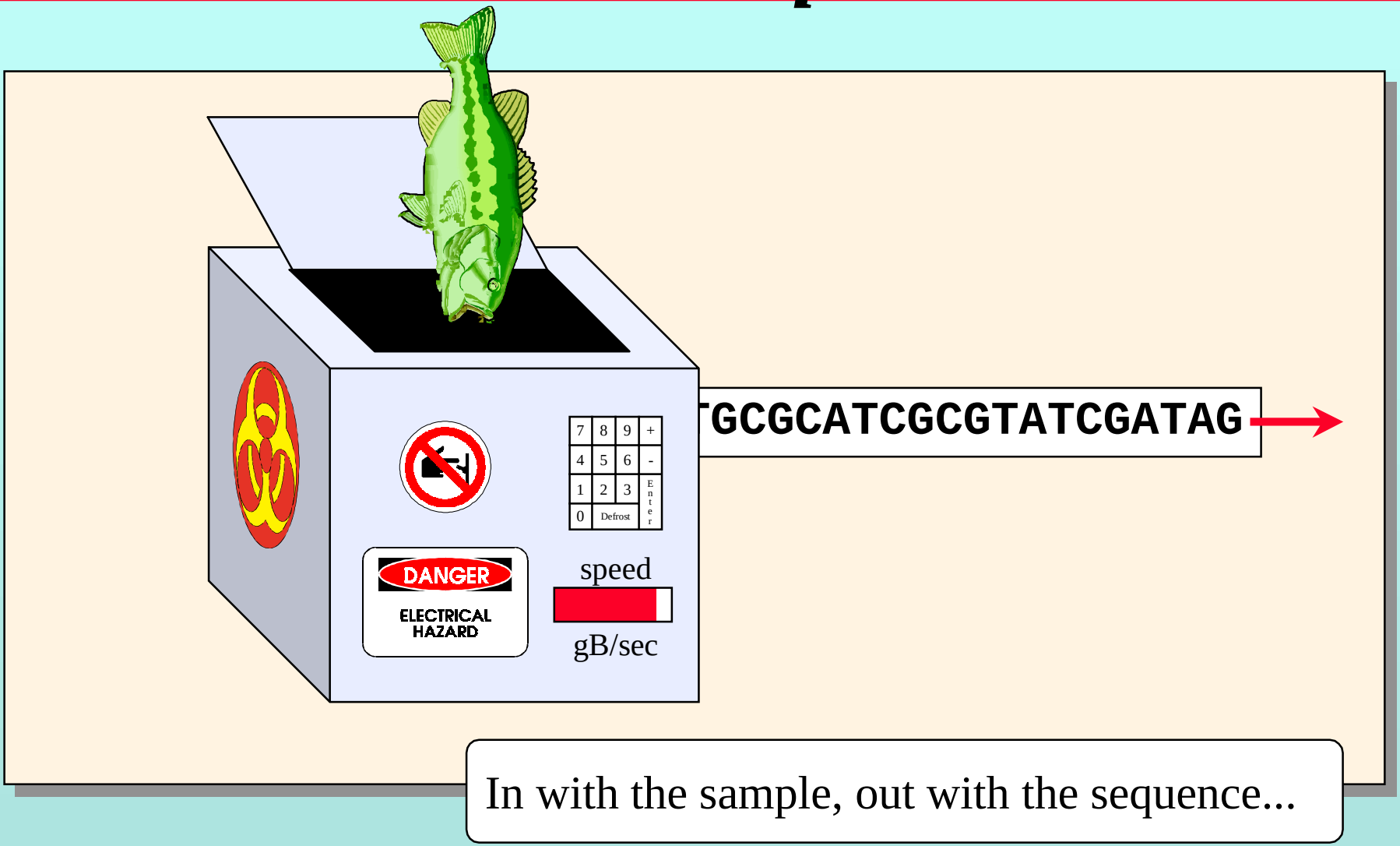
Base Pairs in GenBank



Base Pairs in GenBank



ABI *Bass-o-Matic Sequencer*



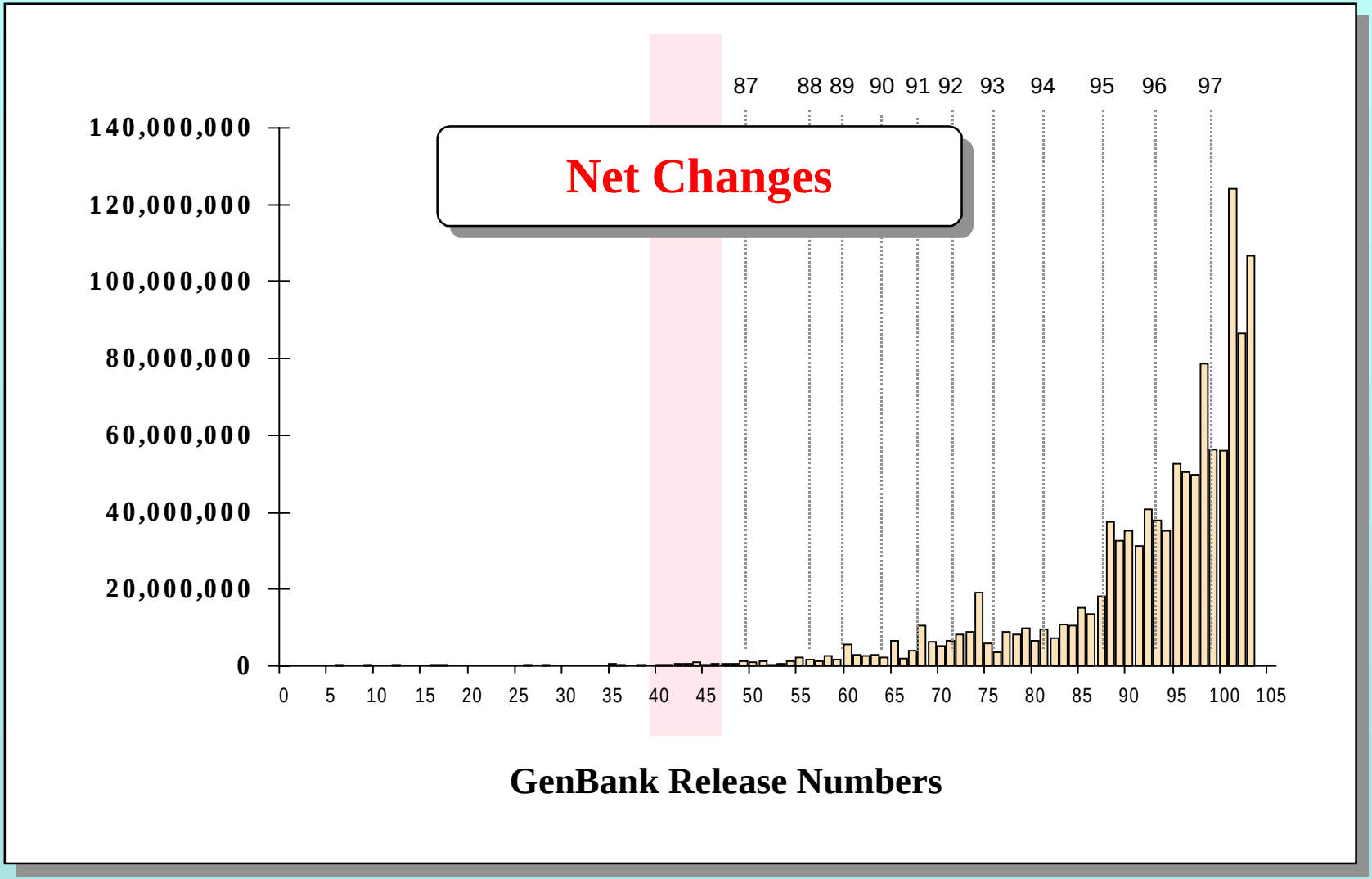
In with the sample, out with the sequence...

What's Really Next

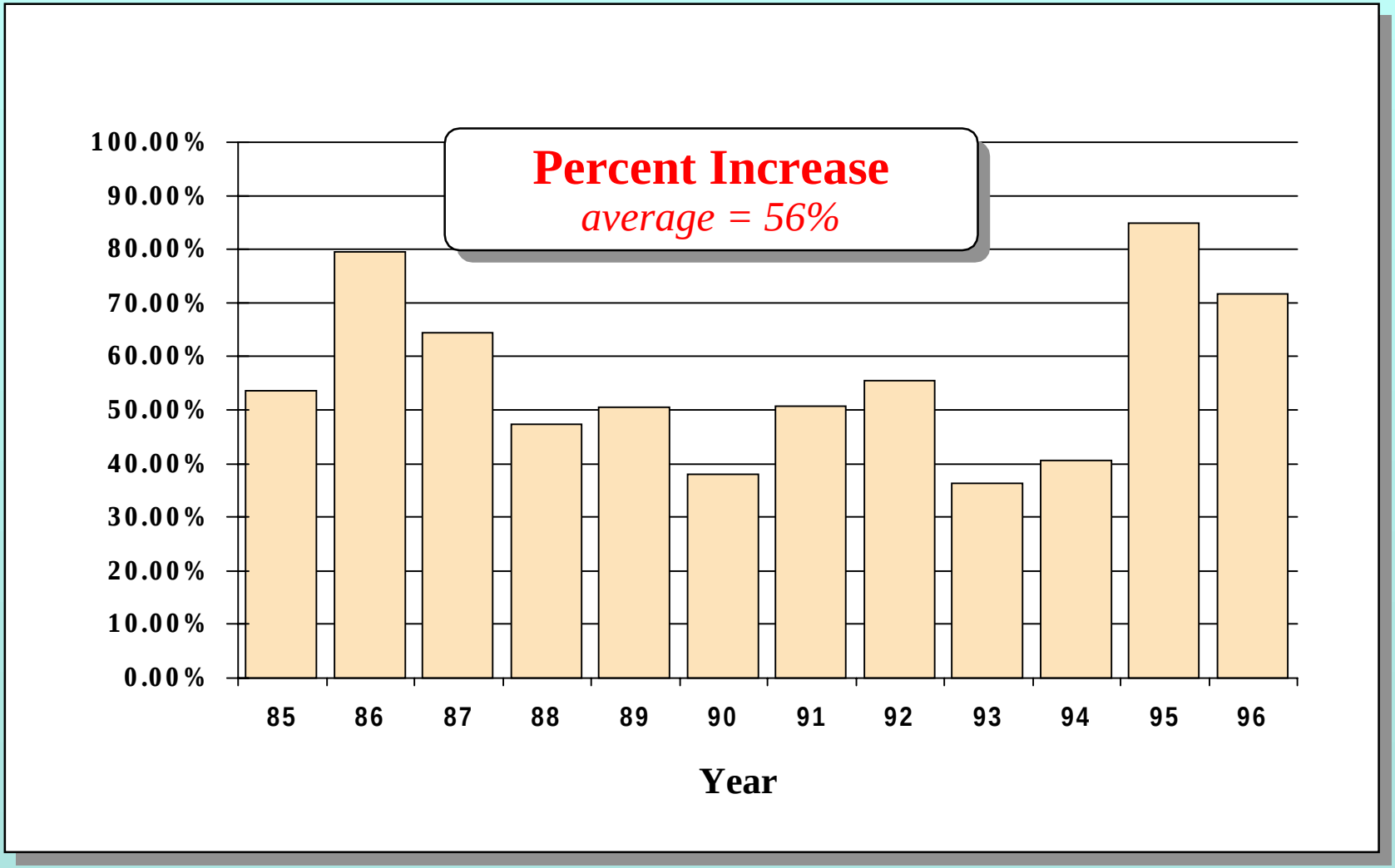
The post-genome era in biological research will take for granted ready access to huge amounts of genomic data.

The challenge will be ***understanding*** those data and using the understanding to solve real-world problems...

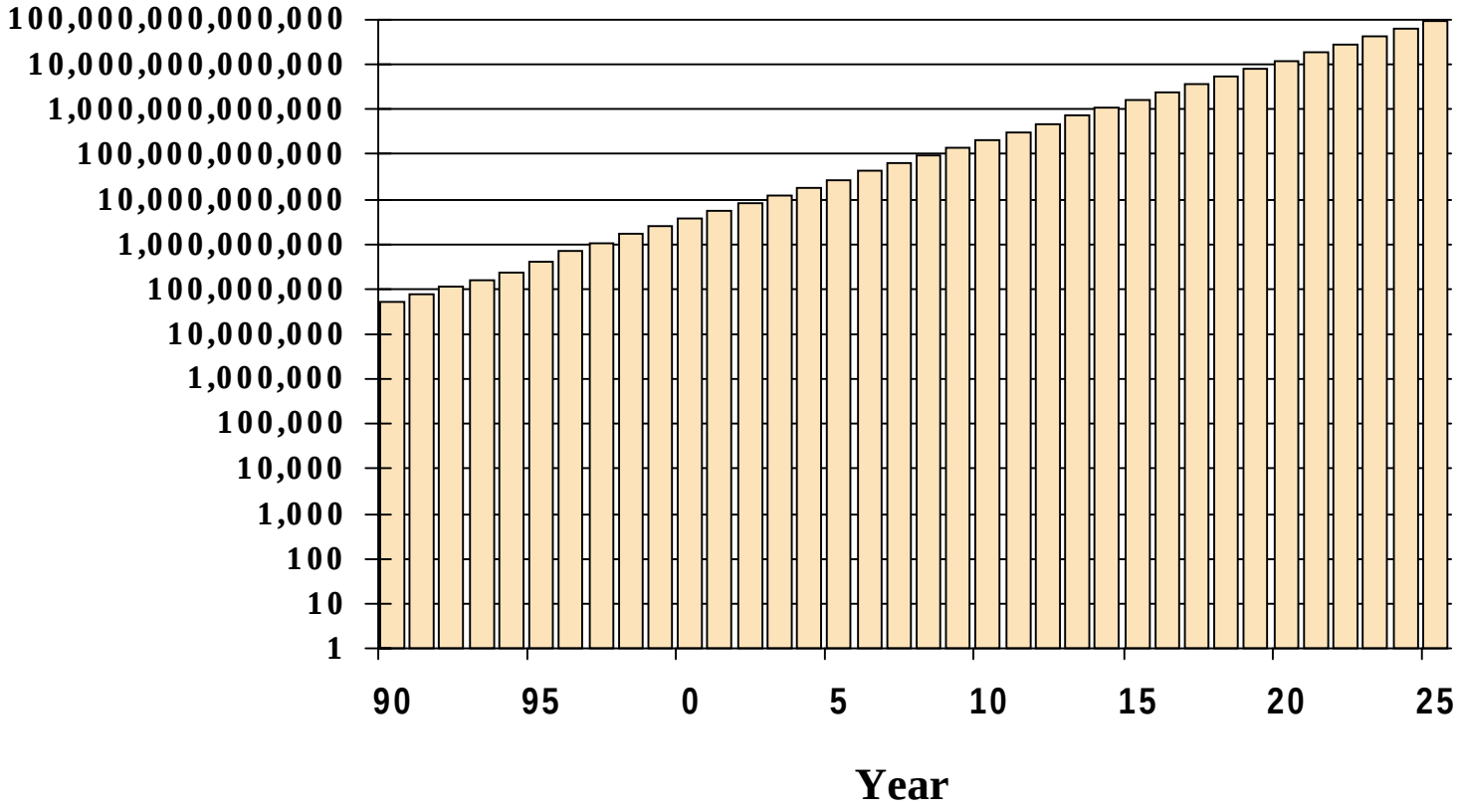
Base Pairs in GenBank



Base Pairs in GenBank *(Percent Increase)*

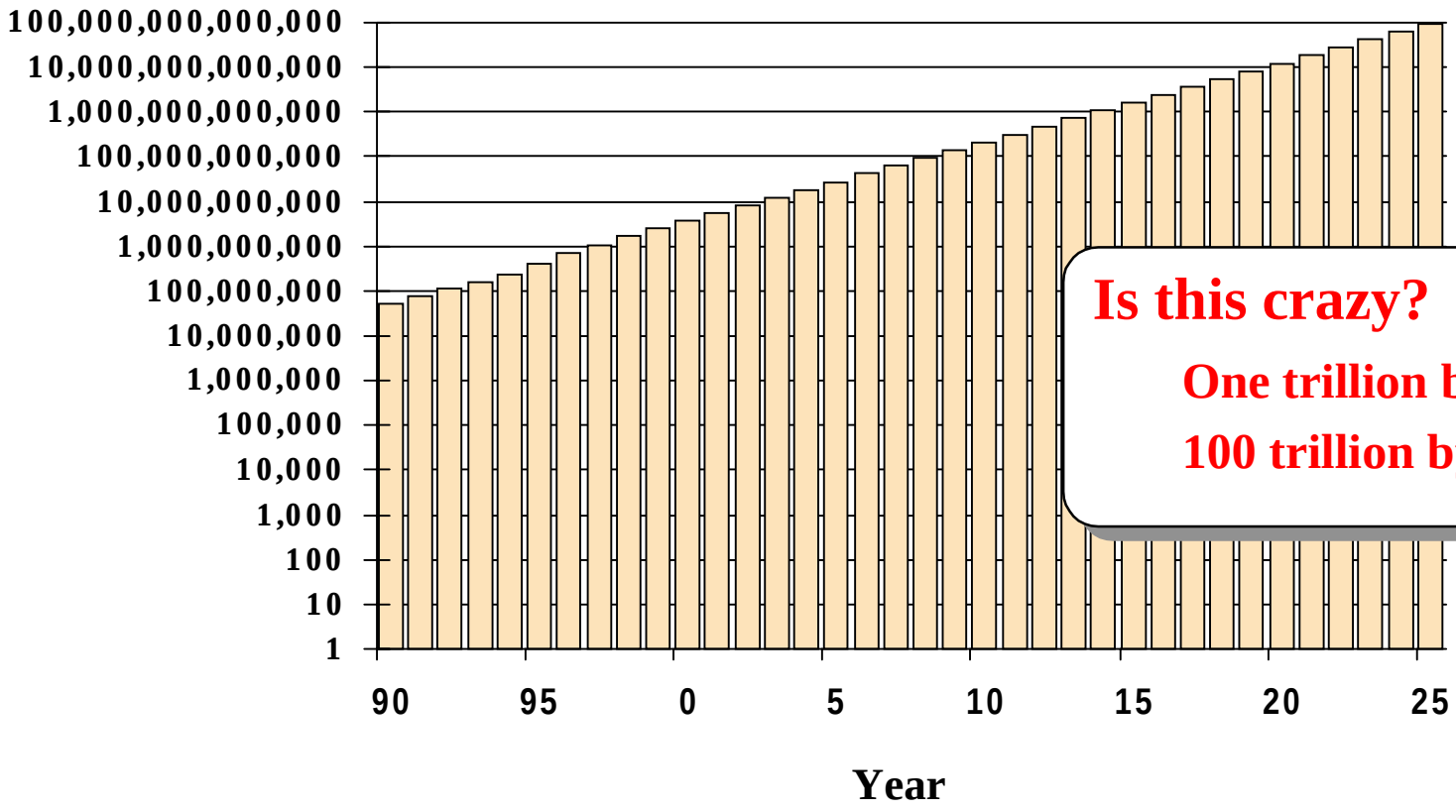


Projected Base Pairs



Assumed annual growth rate: 50%
(less than current rate)

Projected Base Pairs



Is this crazy?

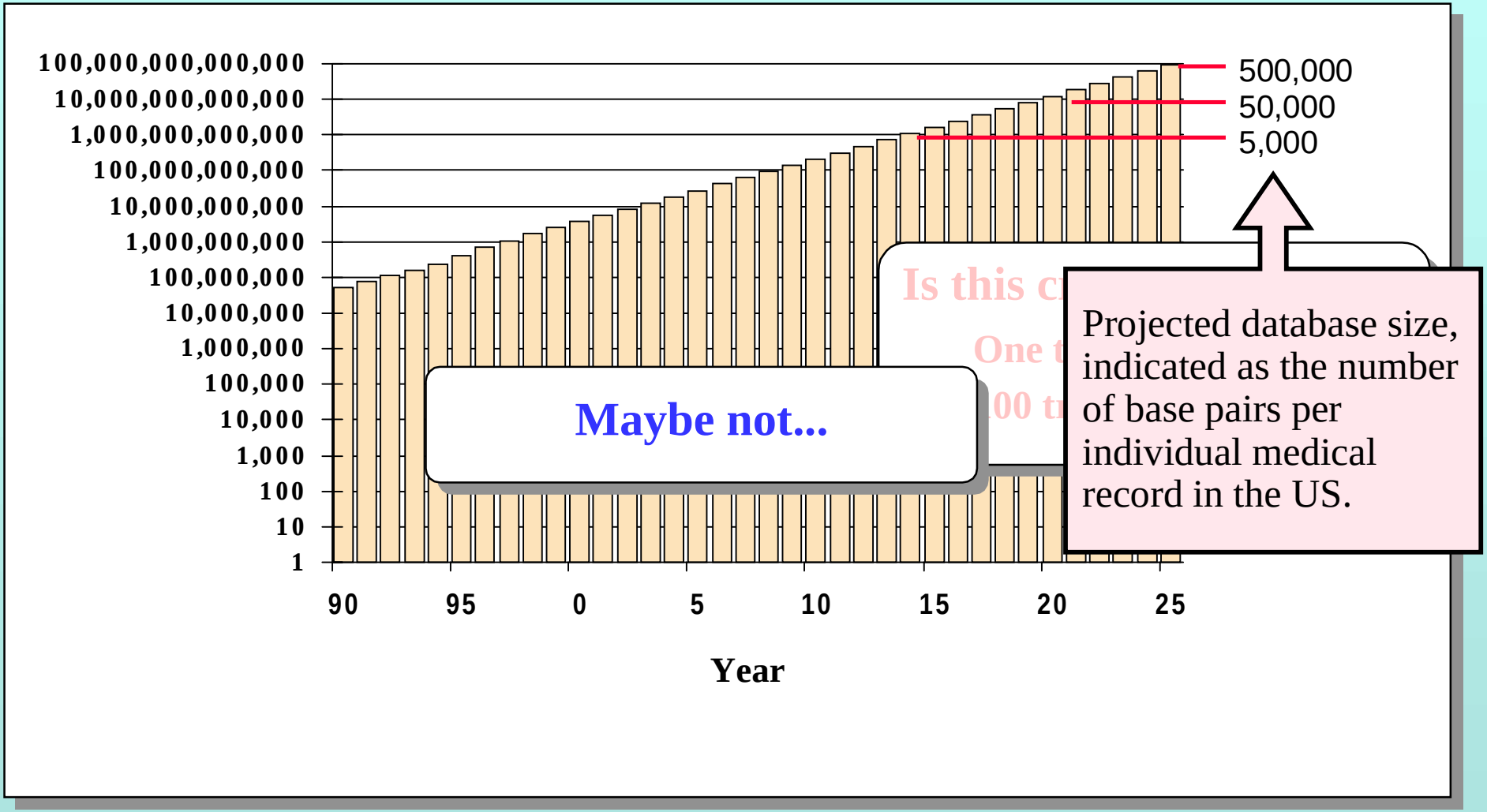
One trillion bp by 2015

100 trillion by 2025

Assumed annual growth rate: 50%

(less than current rate)

Projected Base Pairs



21st Century Biology

Post-Genome Era

The Post-Genome Era

Post-genome research involves:

- applying genomic tools and knowledge to more general problems
- asking new questions, tractable only to genomic or post-genomic analysis
- moving beyond the structural genomics of the human genome project and into the functional genomics of the post-genome era

The Post-Genome Era

Suggested definition:

- functional genomics = biology

The Post-Genome Era

An early analysis:

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

Paradigm Shift in Biology

To use [the] 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.

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

Case of Microbiology

< 5,000	known and described bacteria
5,000,000	base pairs per genome
25,000,000,000	TOTAL base pairs

If a full, annotated sequence were available for all known bacteria, the practice of microbiology would match Gilbert's prediction.

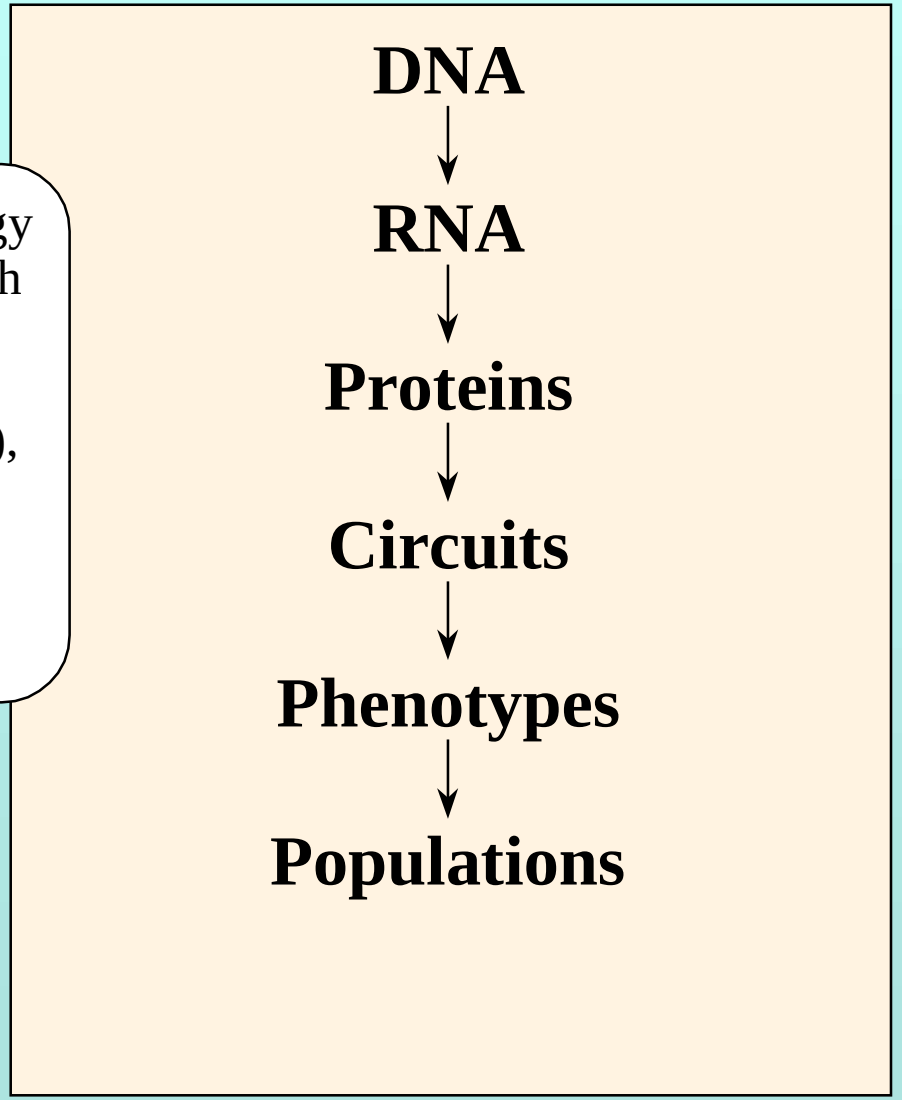
21st Century Biology

The Science

Fundamental Dogma

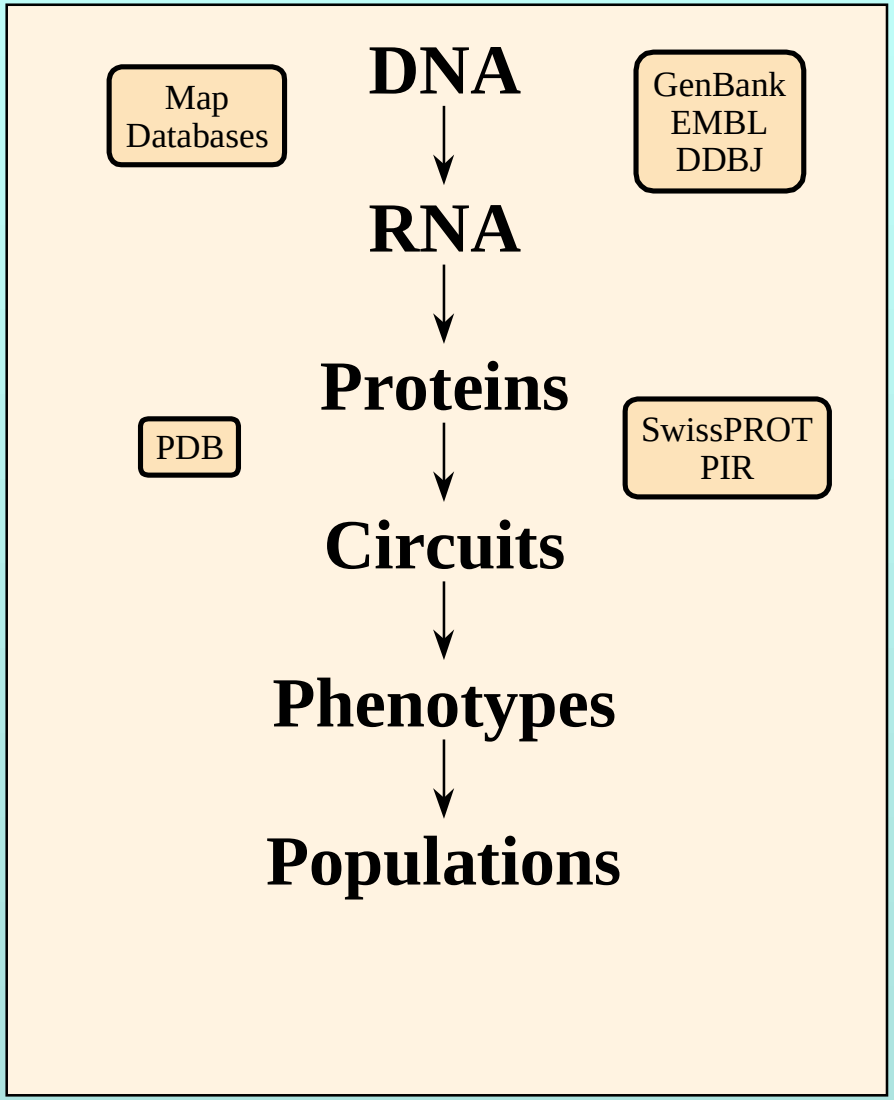
The fundamental dogma of molecular biology is that genes act to create phenotypes through a flow of information from DNA to RNA to proteins, to interactions among proteins (regulatory circuits and metabolic pathways), and ultimately to phenotypes.

Collections of individual phenotypes, of course, constitute a population.



Fundamental Dogma

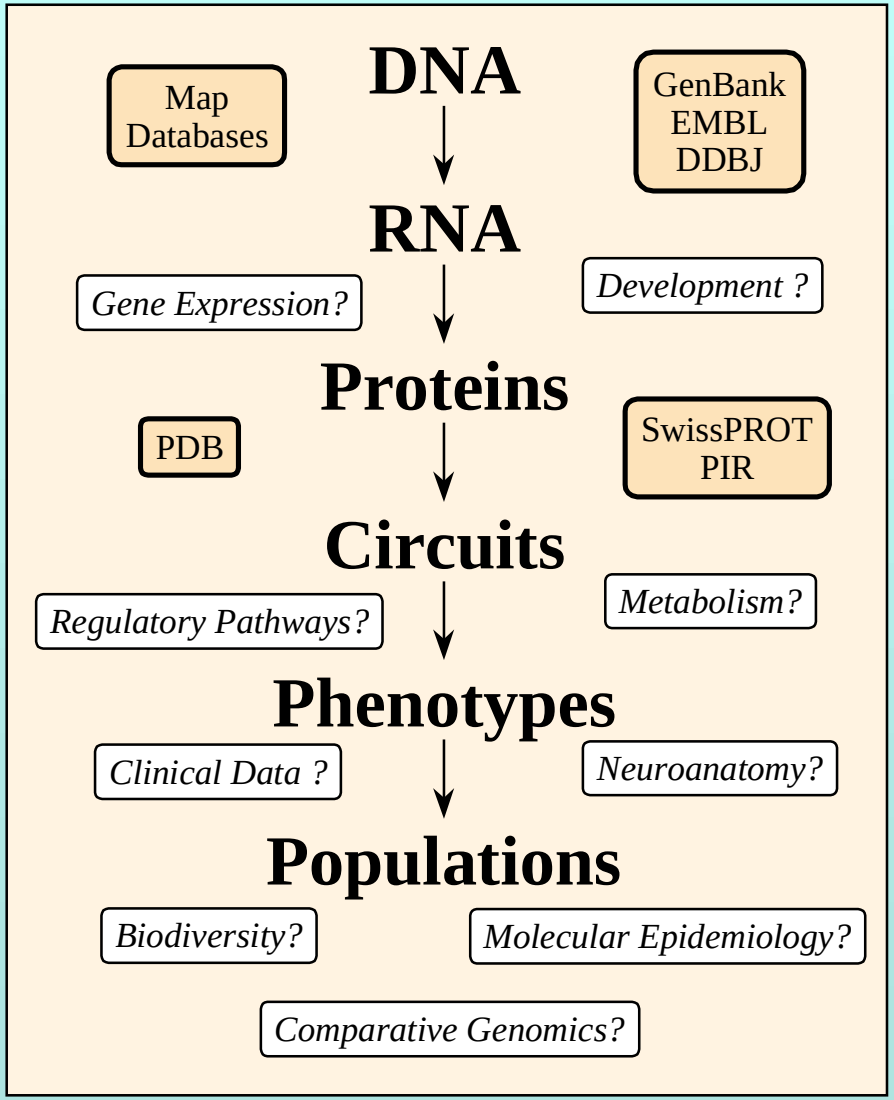
Although a few databases already exist to distribute molecular information,



Fundamental Dogma

Although a few databases already exist to distribute molecular information,

the post-genomic era will need many more to collect, manage, and publish the coming flood of new findings.



21st Century Biology

The People

Human Resources Issues

- Reduction in need for non-IT staff

Human Resources Issues

- Reduction in need for non-IT staff
- Increase in need for IT staff, especially "information engineers"

Human Resources Issues

- Reduction in need for non-IT staff
- Increase in need for IT staff, especially "information engineers"

In modern biology, a general trend is to convert expert work into staff work and finally into computation. New expertise is required to design, carry out, and interpret continuing work.

Human Resources Issues

Elbert Branscomb: "You must recognize that some day you may need as many computer scientists as biologists in your labs."

Human Resources Issues

Elbert Branscomb: "You must recognize that some day you may need as many computer scientists as biologists in your labs."

Craig Venter: "At TIGR, we already have twice as many computer scientists on our staff."

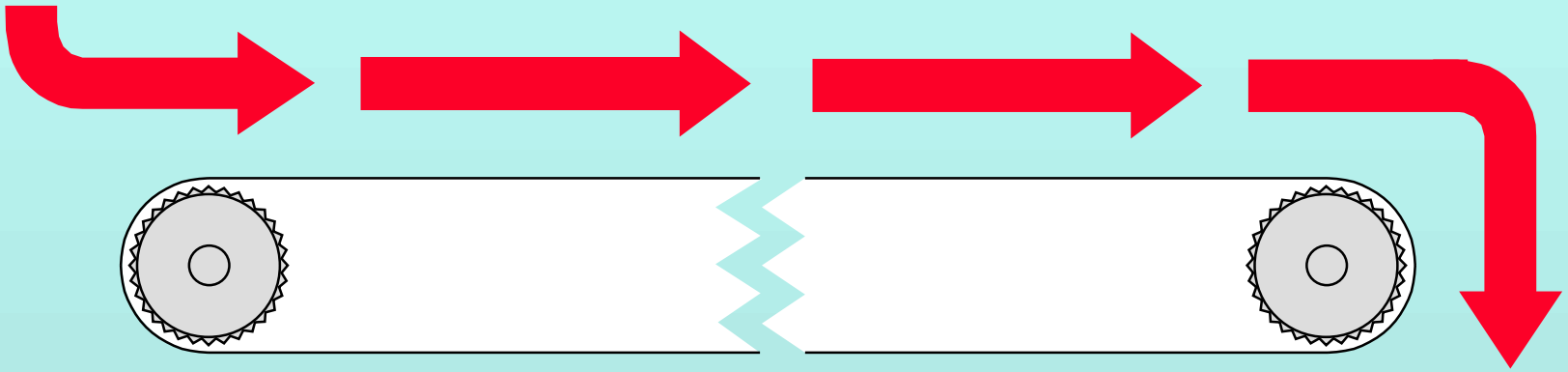
Exchange at DOE workshop on high-throughput sequencing.

New Discipline of Informatics

What is Informatics?

Computer Science Research

Informatics



Biological Application Programs

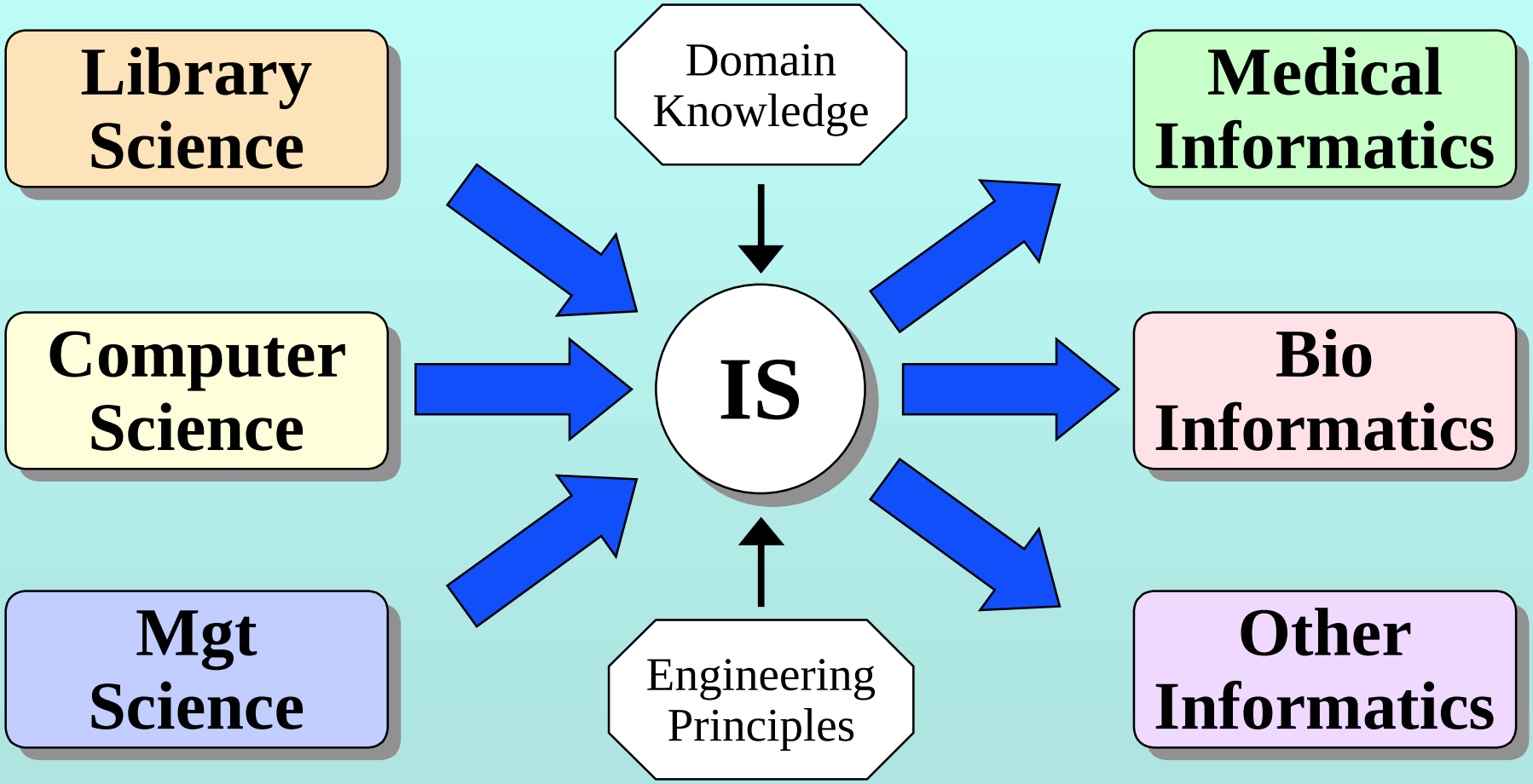
What is Informatics?

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.

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.

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.

The Future is Now

The Future is Now

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.

The Future is Now

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.

Computers also allow us to see ecosystems in action, to see genes being expressed, to see inside living organisms, to see populations in flux, to see ...

Slides:

http://www.esp.org/rjr/aaas2000.pdf