Next Steps for Working Scientists: Access to Information

(http://www.esp.org/rjr/codata.pdf)

Robert J. Robbins
Fred Hutchinson Cancer Research Center
1100 Fairview Avenue North, LV-101
Seattle, Washington 98109

rrobbins@fhcrc.org
(206) 667 2920

Abstract

Over the next few years, the relentless exponential effect of Moore's Law will have profound effects upon the use computation in science and technology. By 2005, analytical power previously available only at supercomputer centers will exist on every desktop and the volume of electronic data flow will be enormous. Even now, a current Intel computer delivers more MIPS than the first Cray and GenBank acquires more data every ten weeks than it did in its first ten years.

The information infrastructure needed to support the explosion in scientific computation and scientific data will be substantial. If working scientists are to have adequate access to these resources, significant changes in the way information infrastructure is provided will be required.

Topics

- Moore's Law constantly transforms IT (and everything else).
- Information Technology (IT) has a special relationship with biology.
- Current approaches to supporting bio-information infrastructure seem inadequate for 21st-century biology.
- Without better support, much post-genome-era biology may shift entirely into the private sector.

Moore's Law

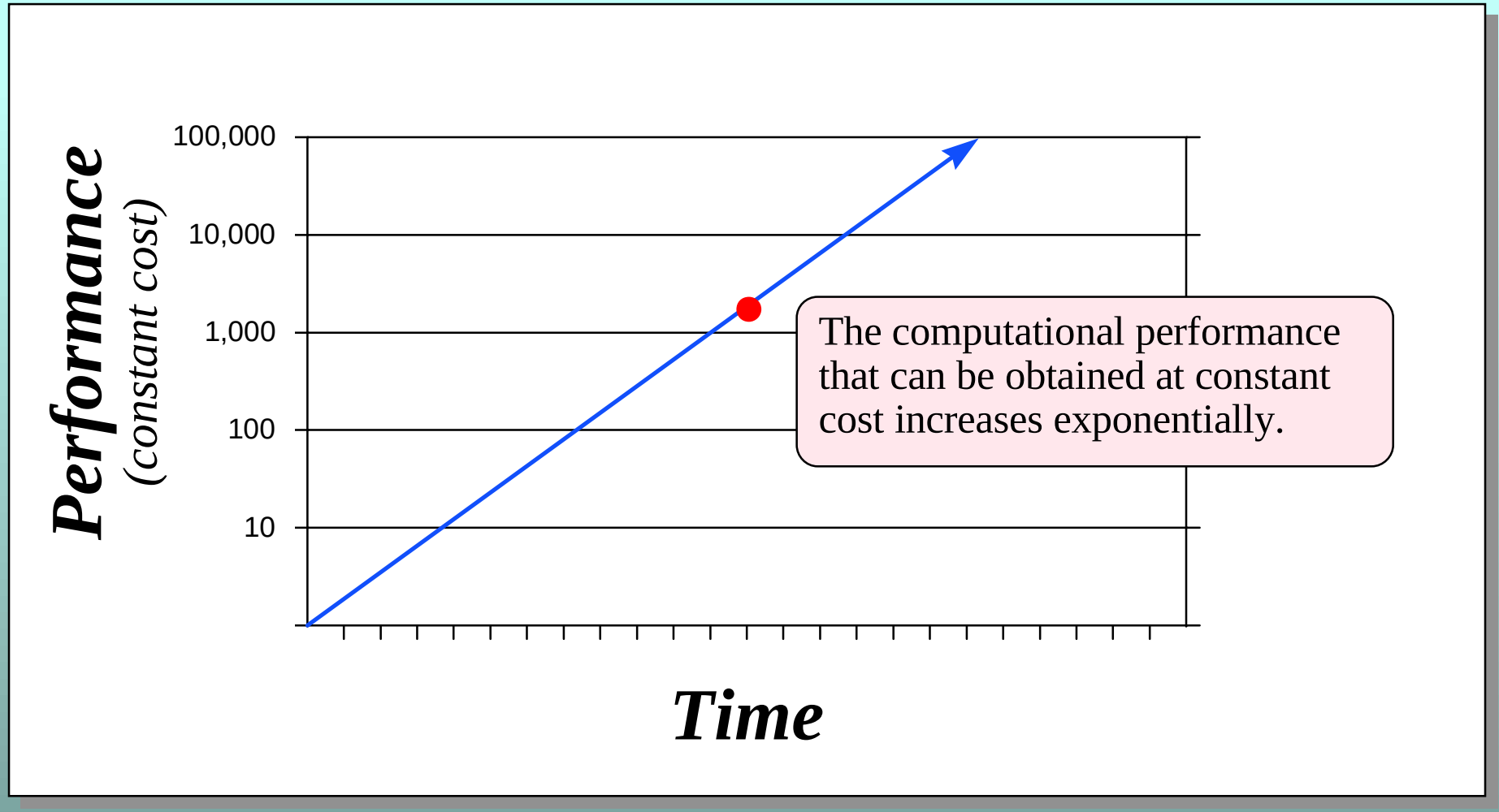
Transforms InfoTech (and everything else)

Moore's Law: *The Statement*

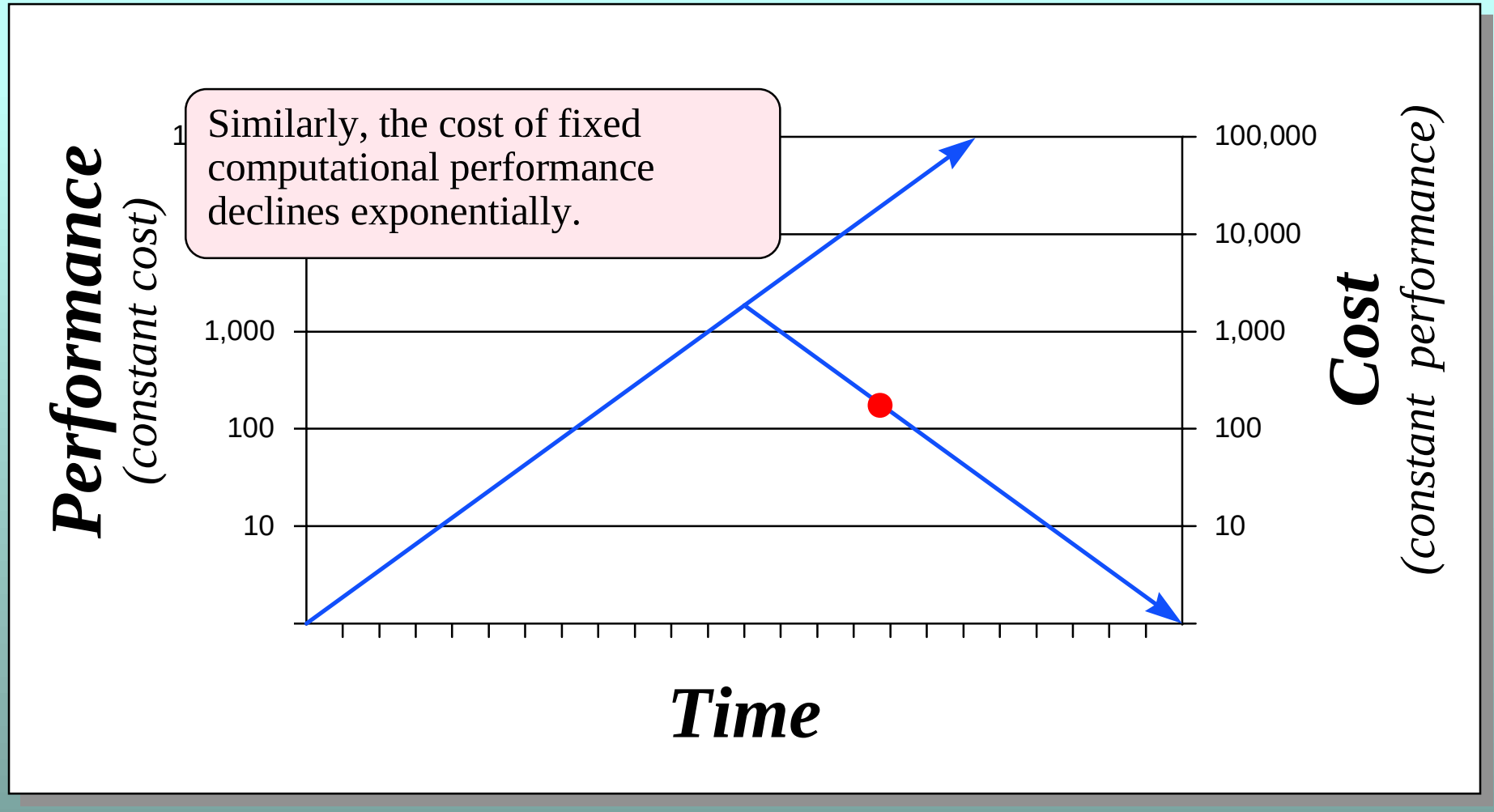
Every eighteen months, the number of transistors that can be placed on a chip doubles.

Gordon Moore, co-founder of Intel...

Moore's Law: *The Effect*



Moore's Law: *The Effect*



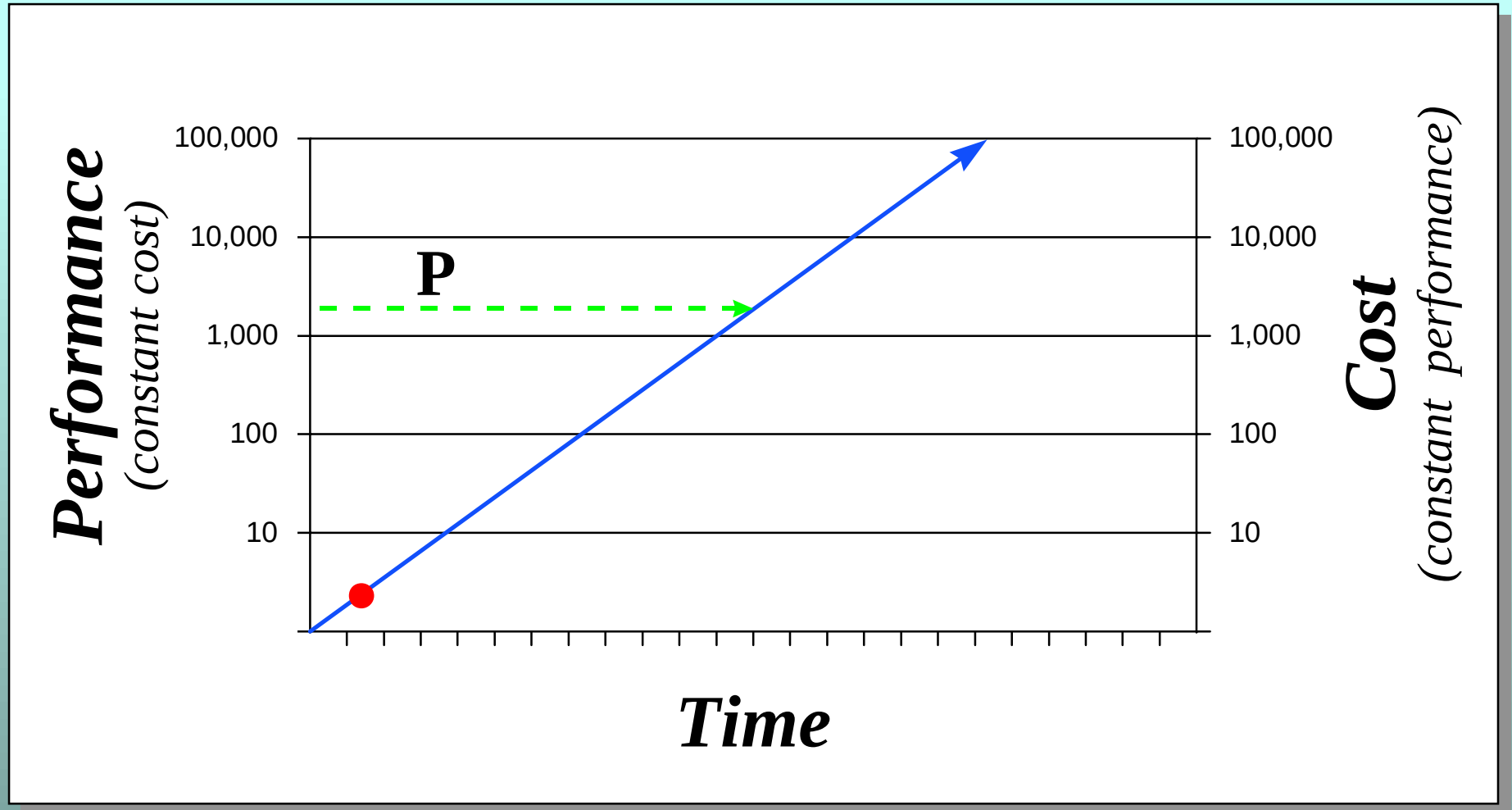
Moore's Law: *The Effect*

Three Phases of Novel IT Applications

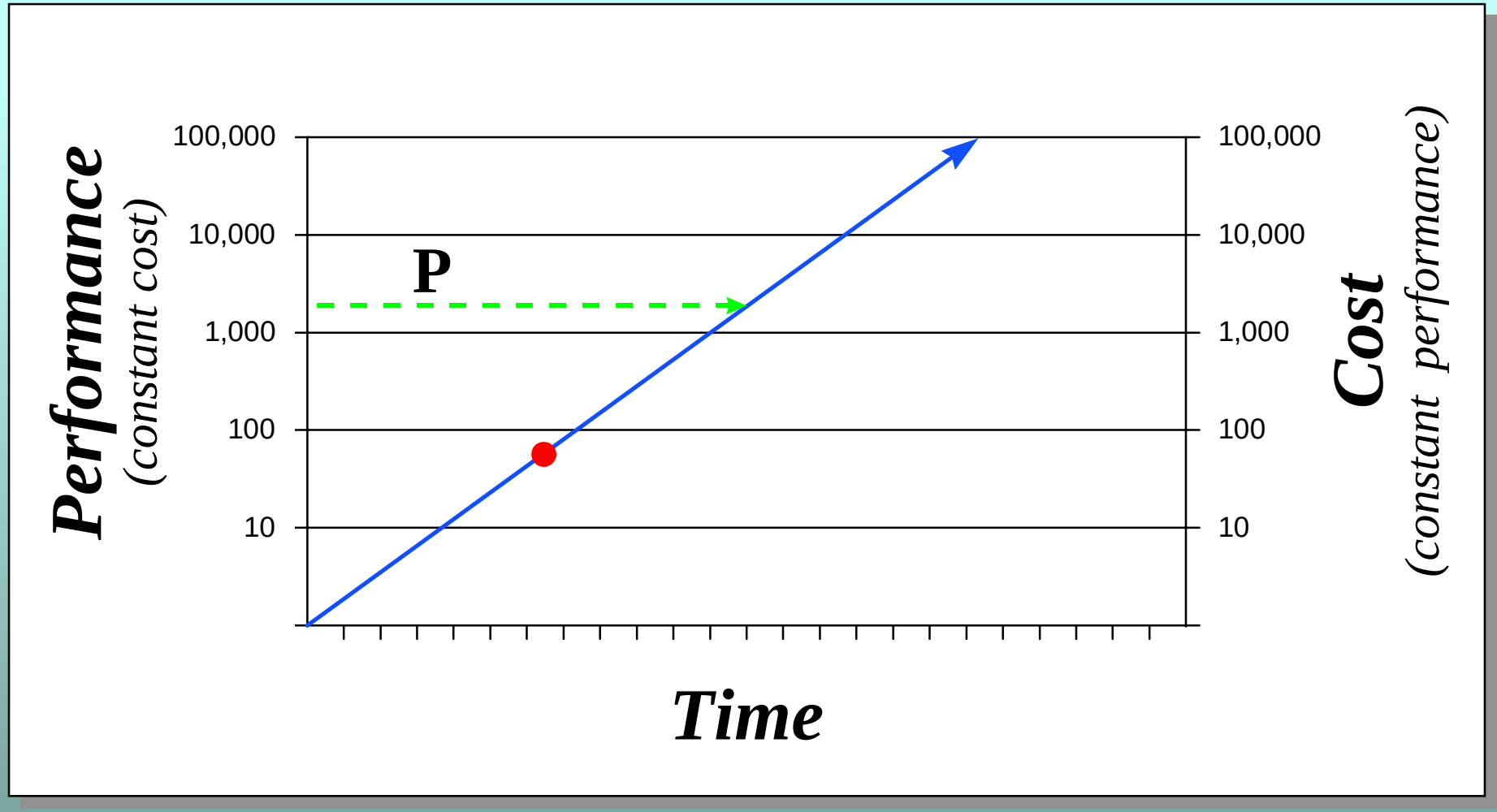
- It's Impossible
- It's Impractical
- It's Overdue

In many fields, those who are overdue with key IT projects have experienced catastrophic losses in competitive advantage.

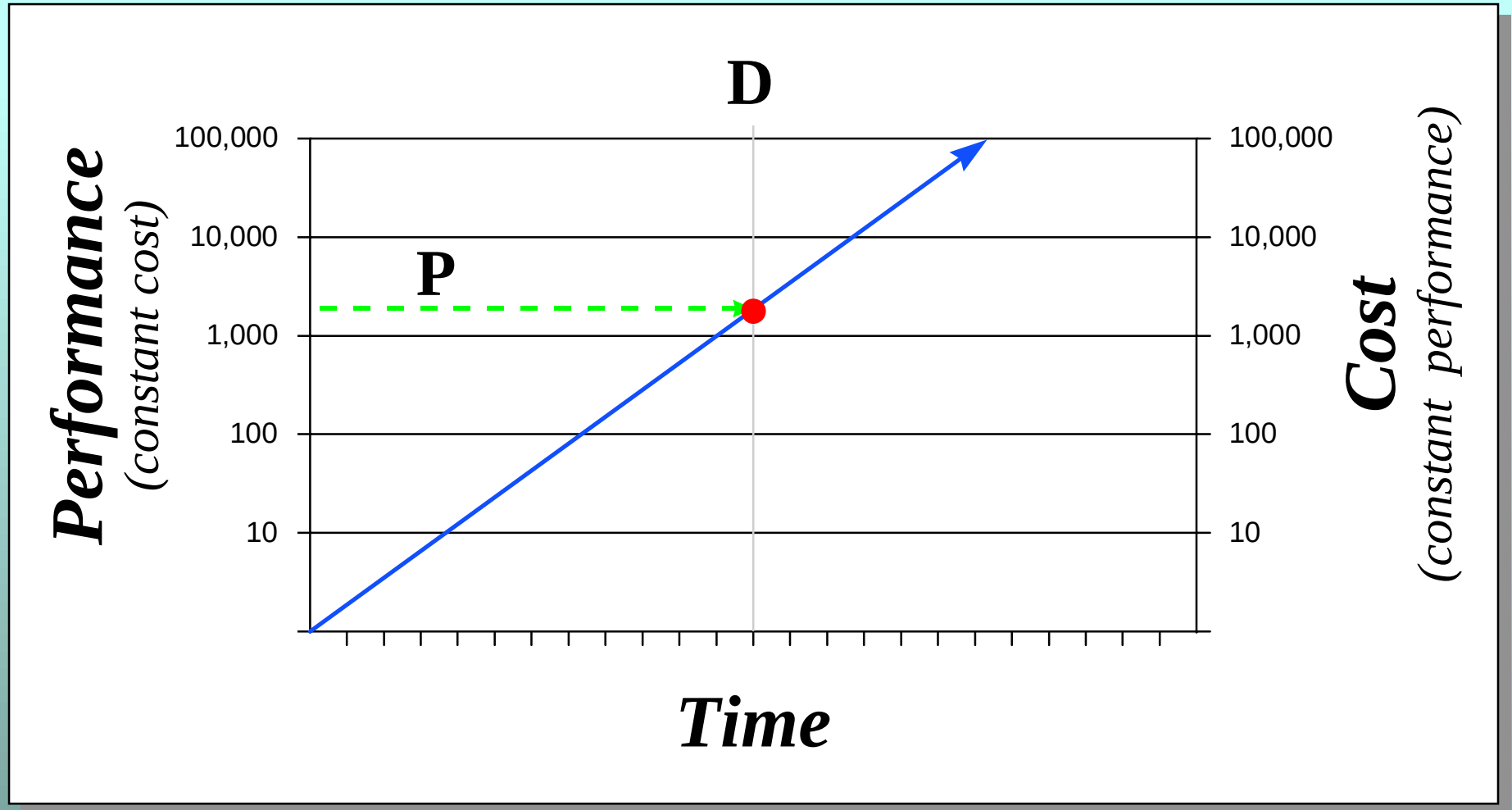
Moore's Law: *The Effect*



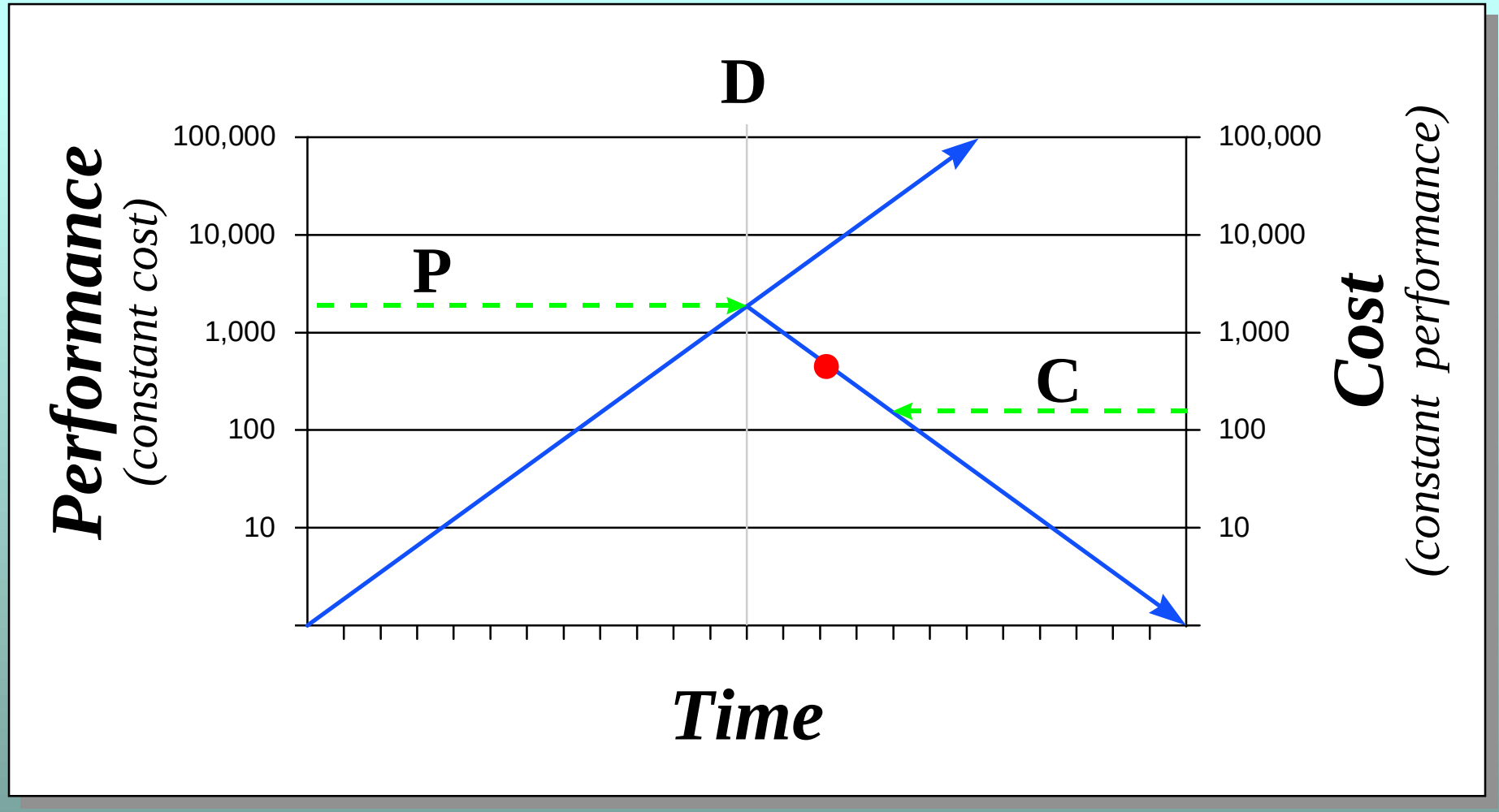
Moore's Law: *The Effect*



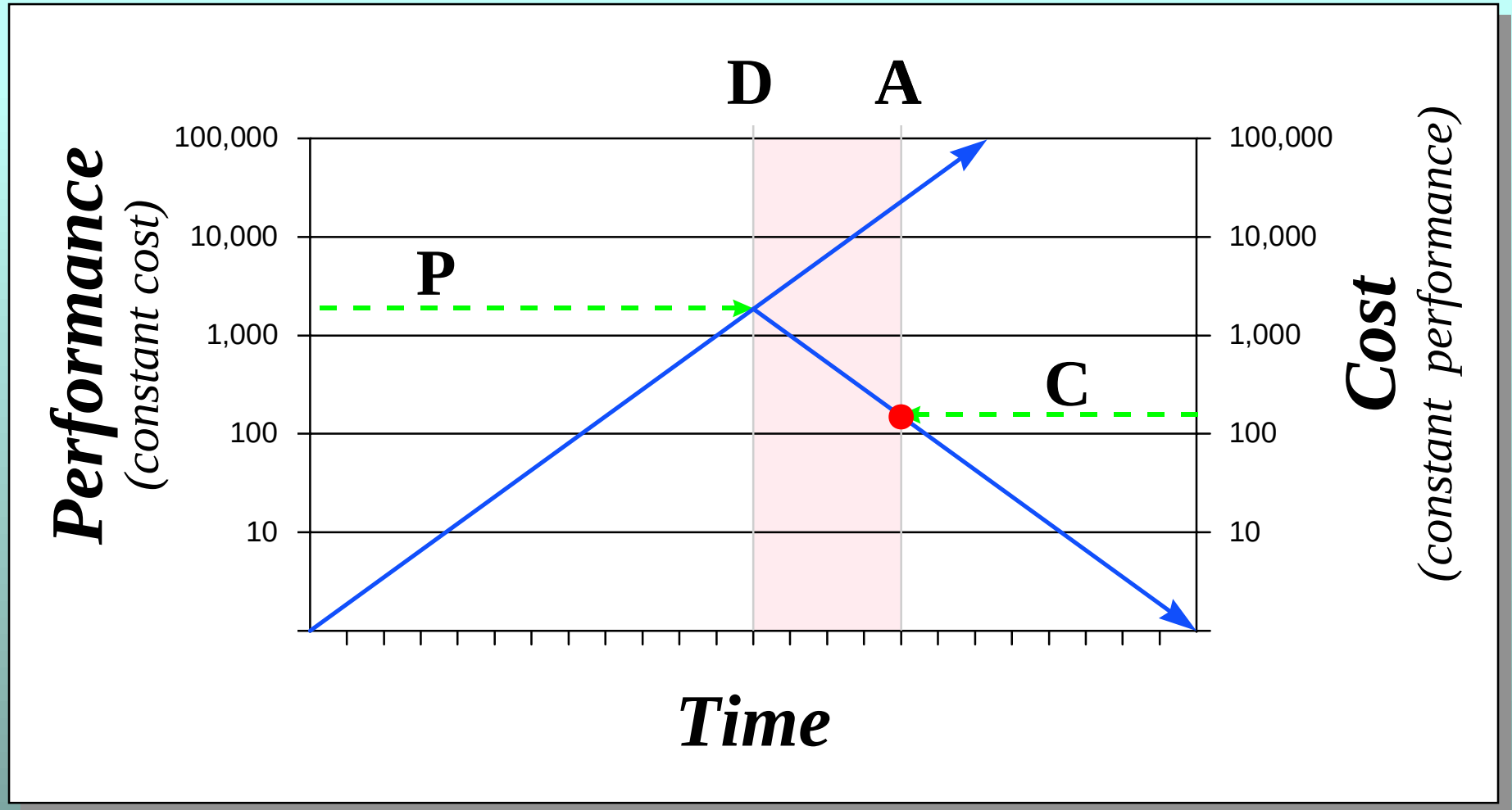
Moore's Law: *The Effect*



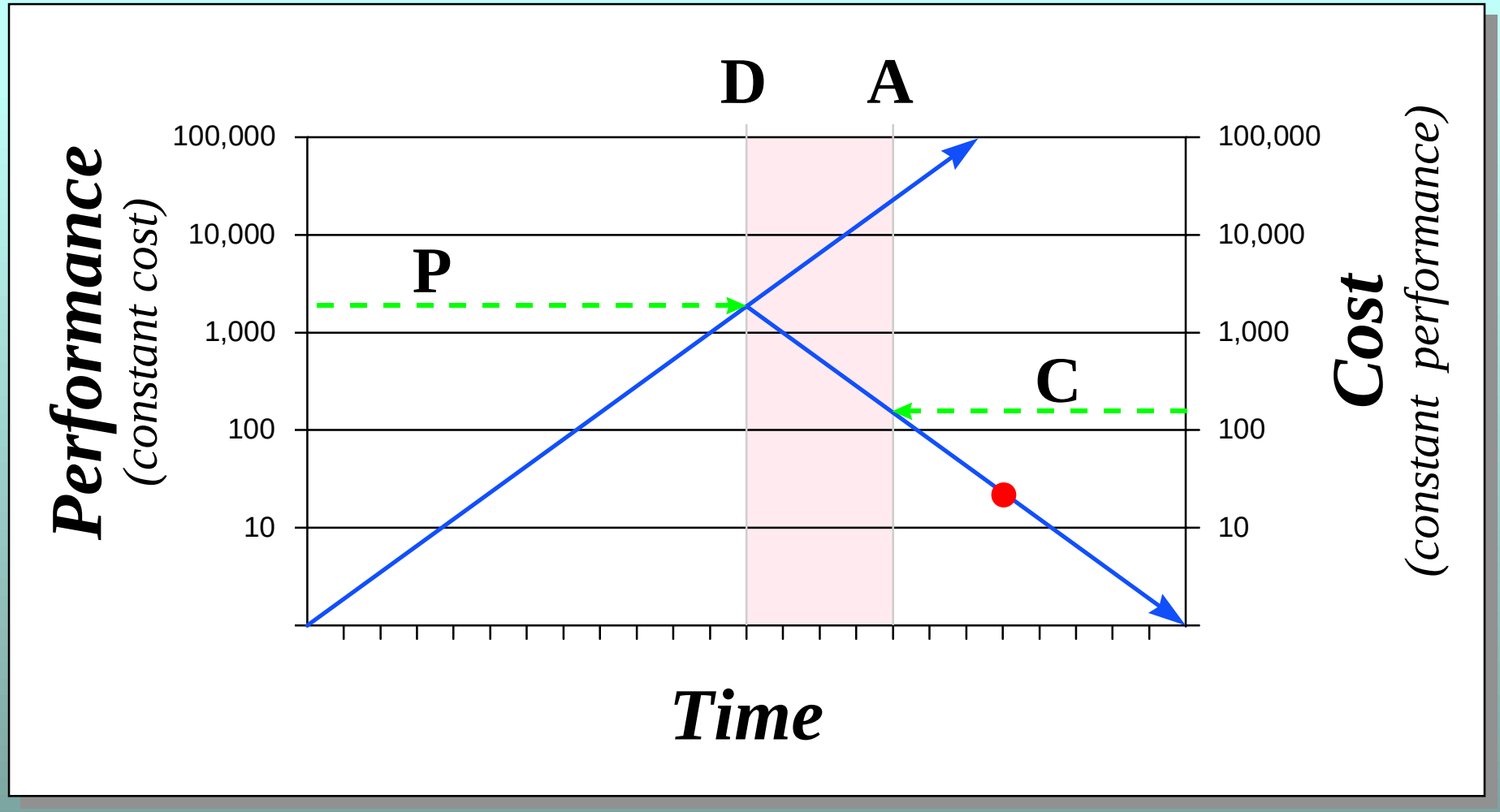
Moore's Law: *The Effect*



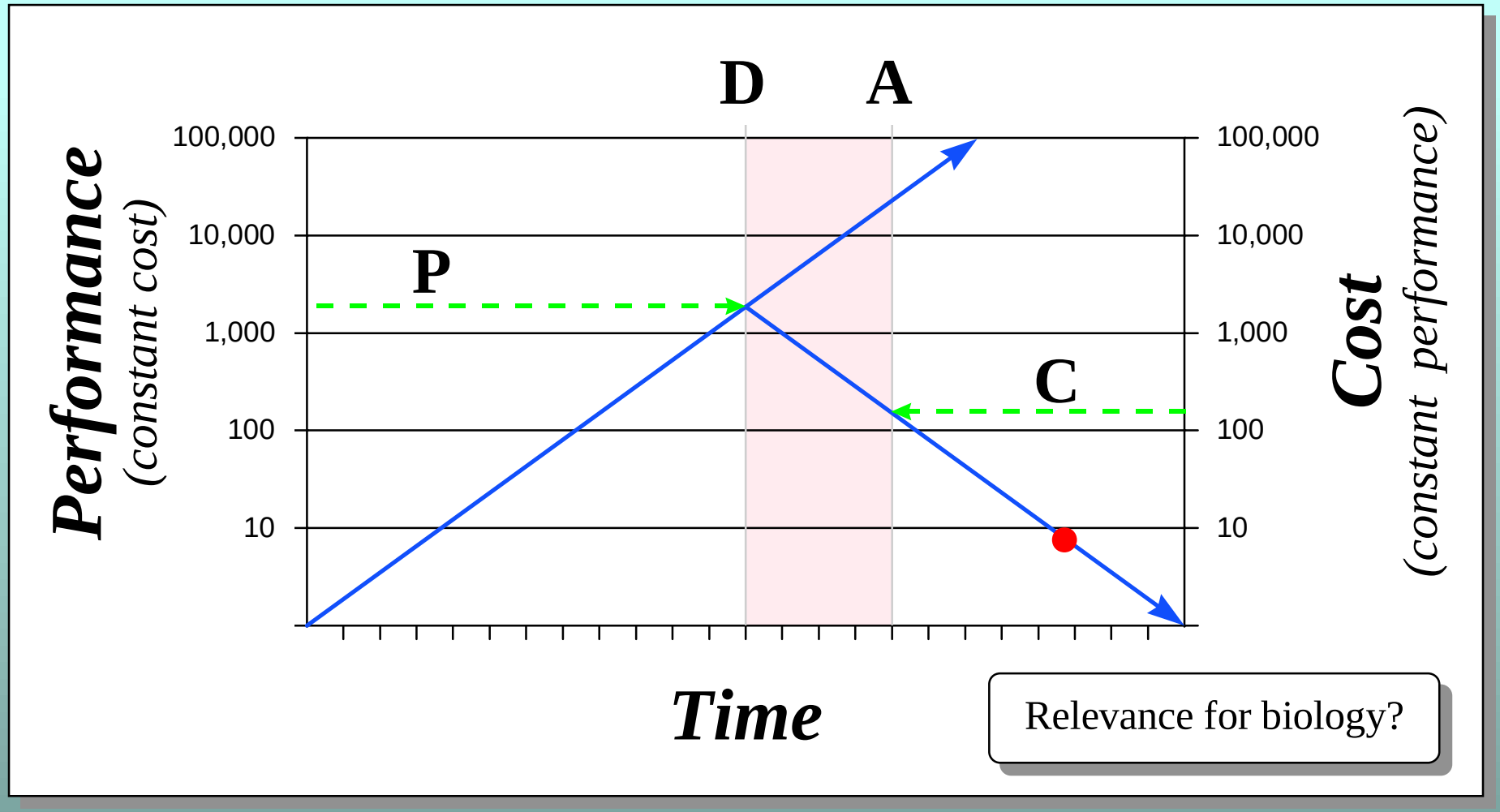
Moore's Law: *The Effect*



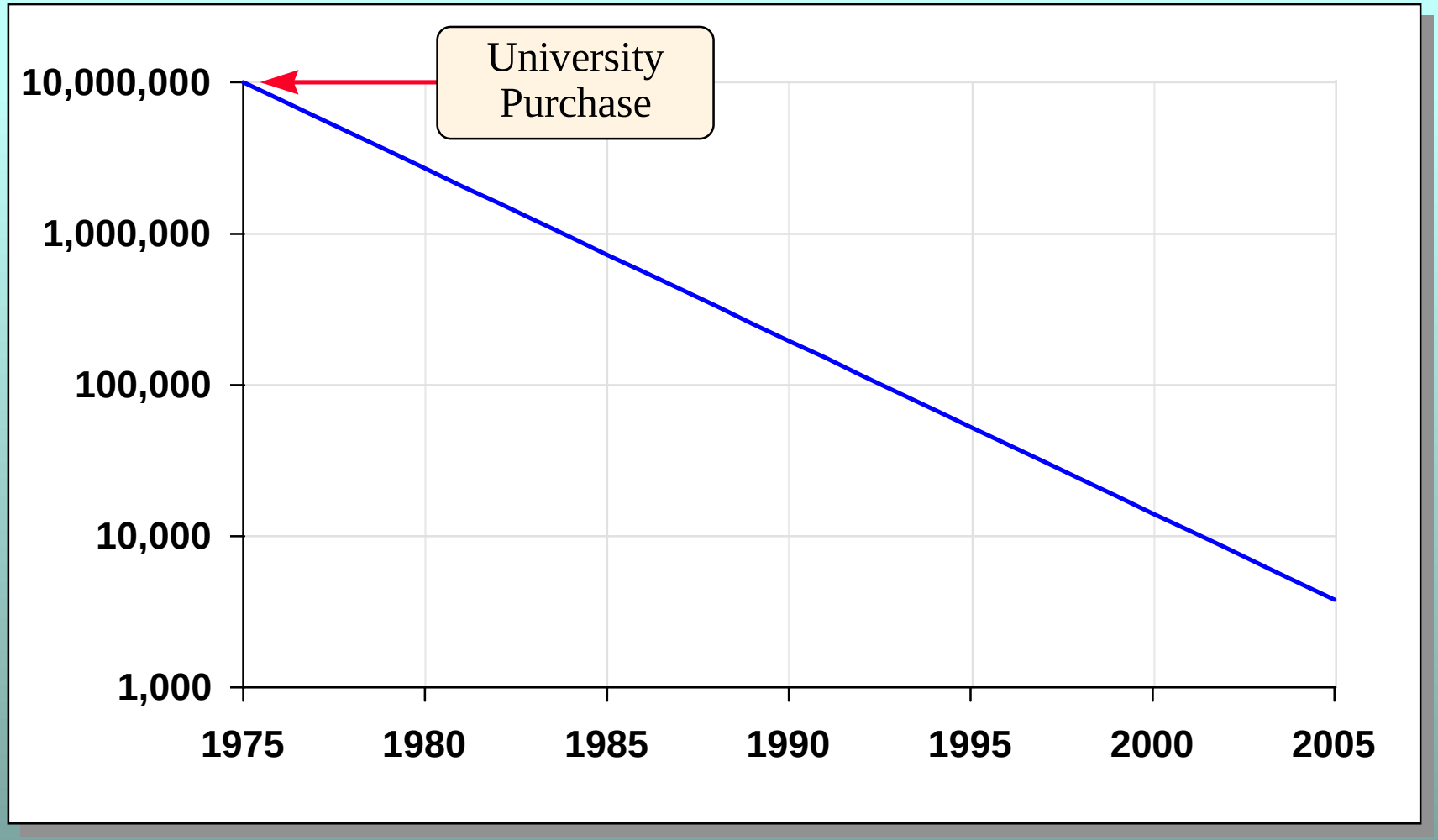
Moore's Law: *The Effect*



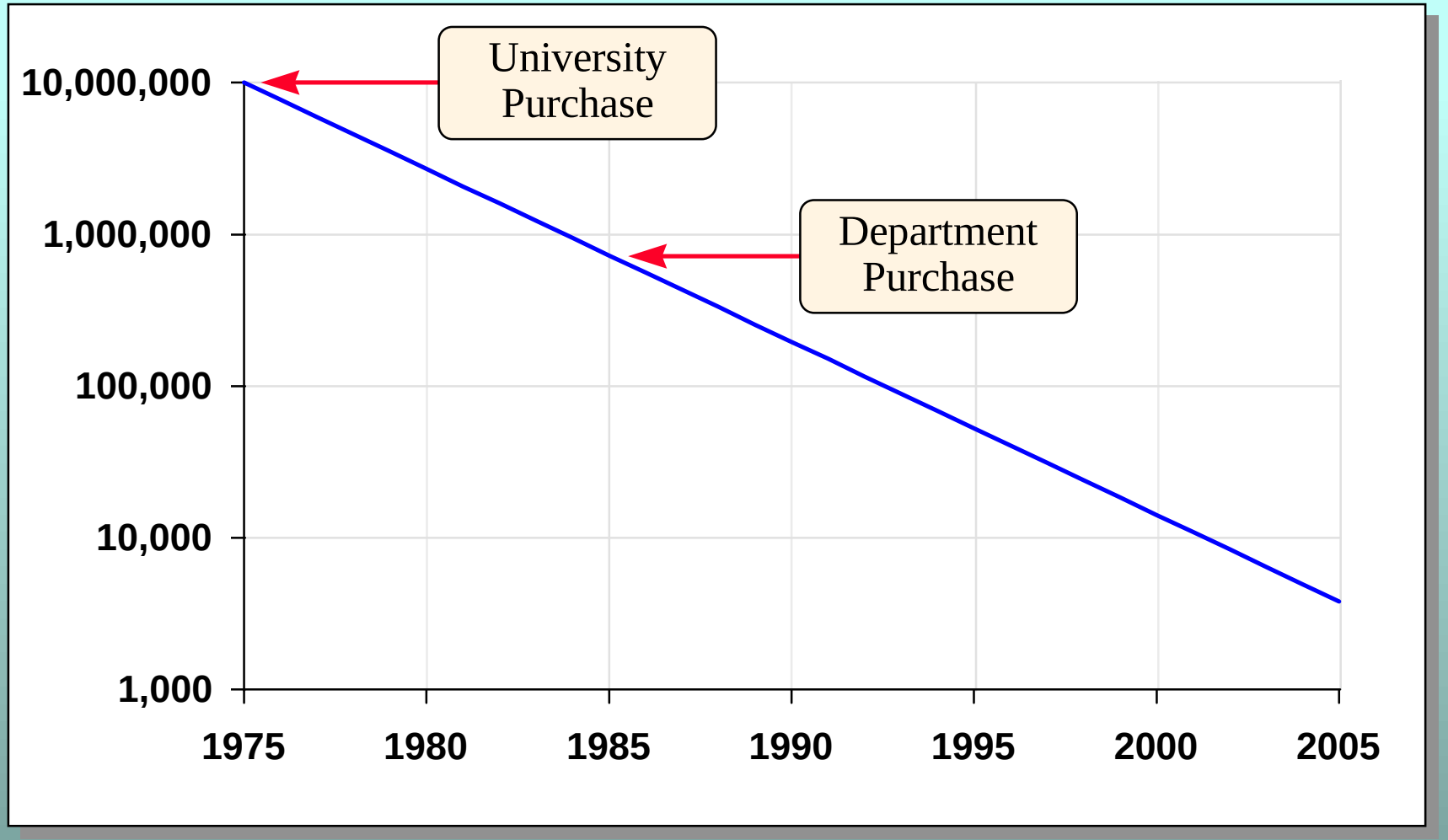
Moore's Law: *The Effect*



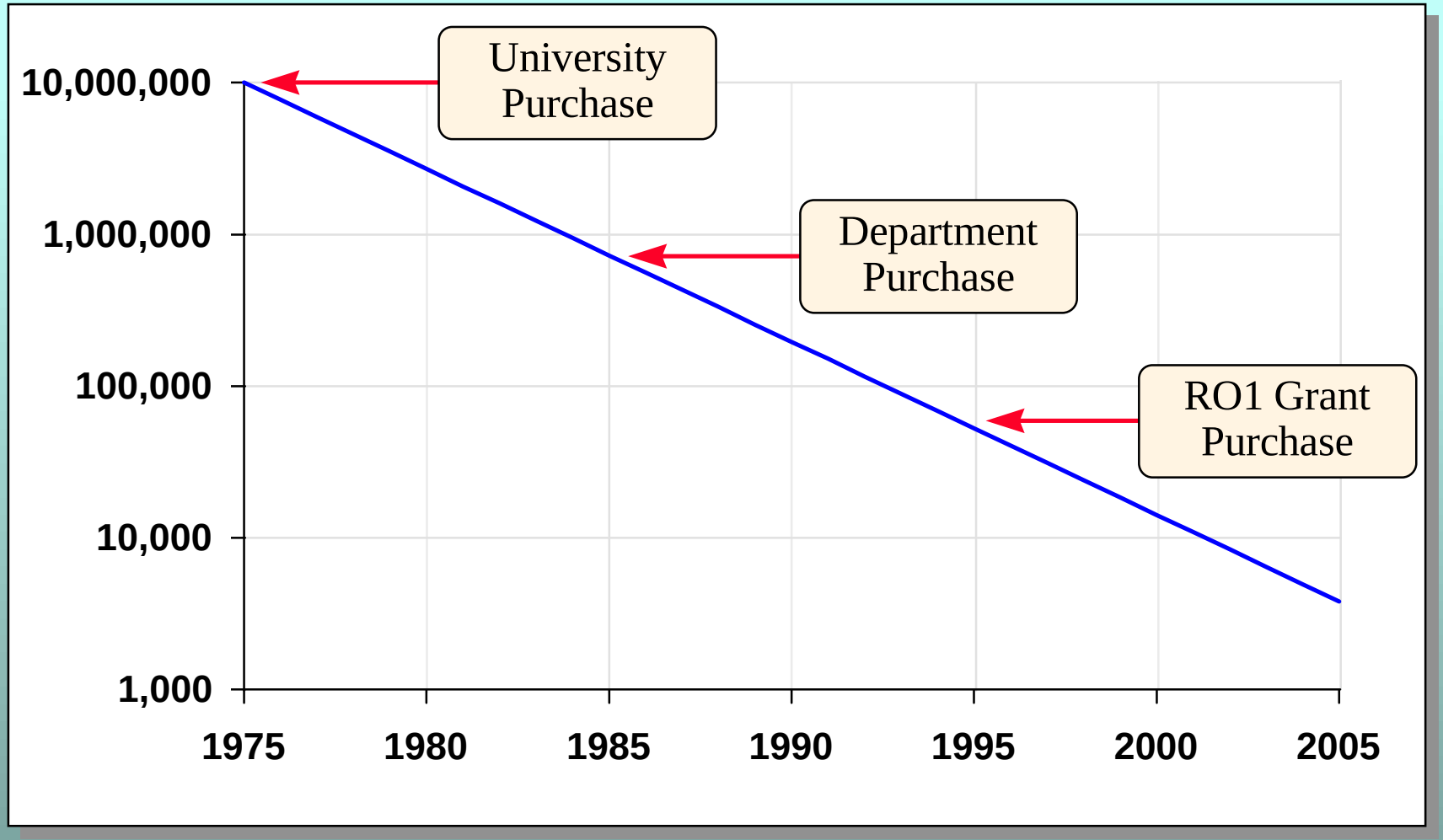
Cost (constant performance)



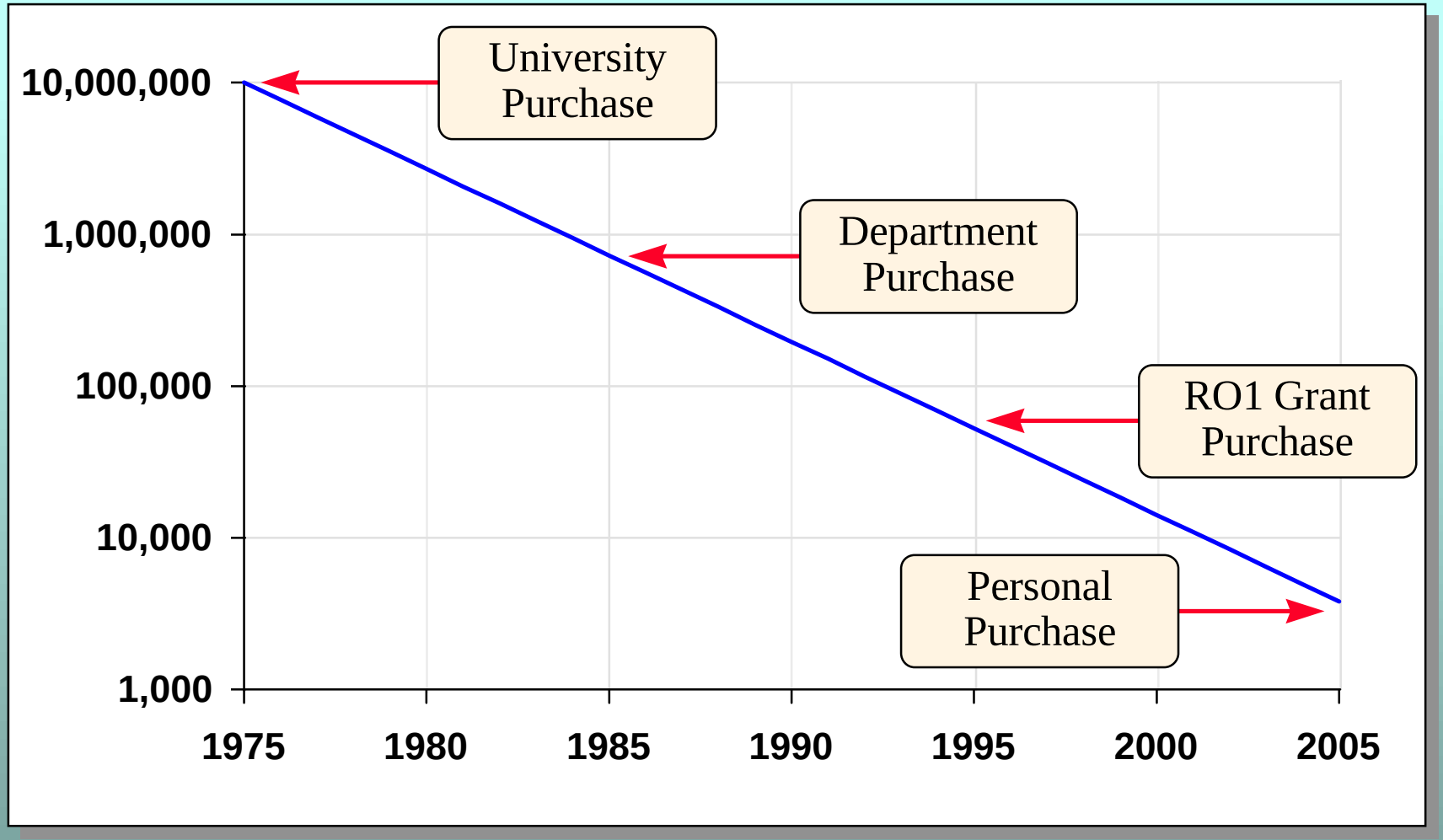
Cost (constant performance)



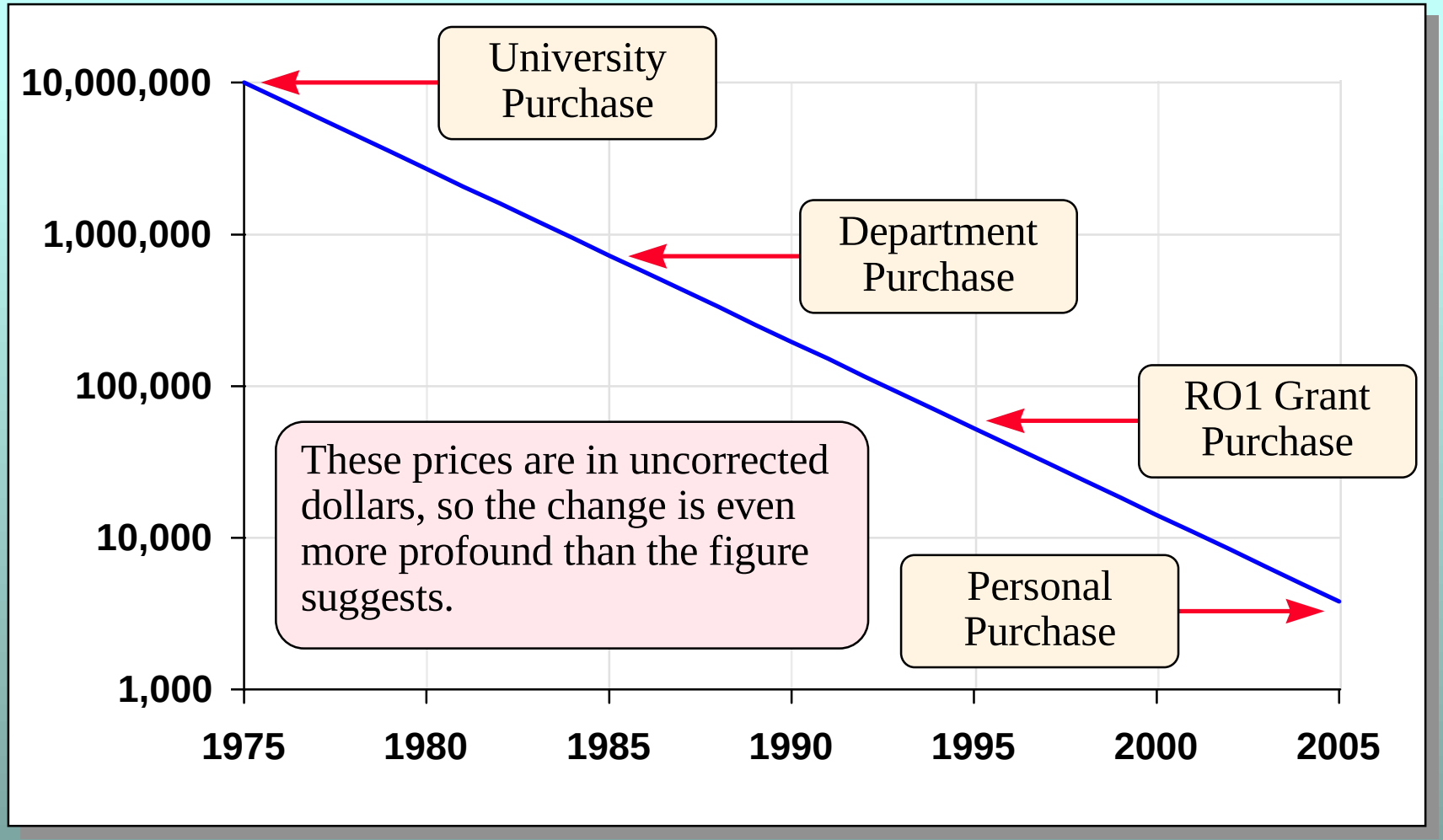
Cost (constant performance)



Cost (constant performance)



Cost (constant performance)



Catching the Wave

Catching the Wave

Fields Transformed by IT:

- *finance & banking*

Catching the Wave

Fields Transformed by IT:

- *finance & banking*
- *travel*

Catching the Wave

Fields Transformed by IT:

- *finance & banking*
- *travel*
- *discount retailing*

Catching the Wave

Fields Transformed by IT:

- *finance & banking*
- *travel*
- *discount retailing*
- *biomedical research ?*

Catching the Wave

Fields Transformed by IT:

- *finance & banking*
- *travel*
- *discount retailing*
- *biomedical research ?*

Why biomedical research? (i) biology is inherently information rich, (ii) appropriately powered computers are now affordable for the research community, and (iii) post-genome biology will thrive on computation.

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...

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*.

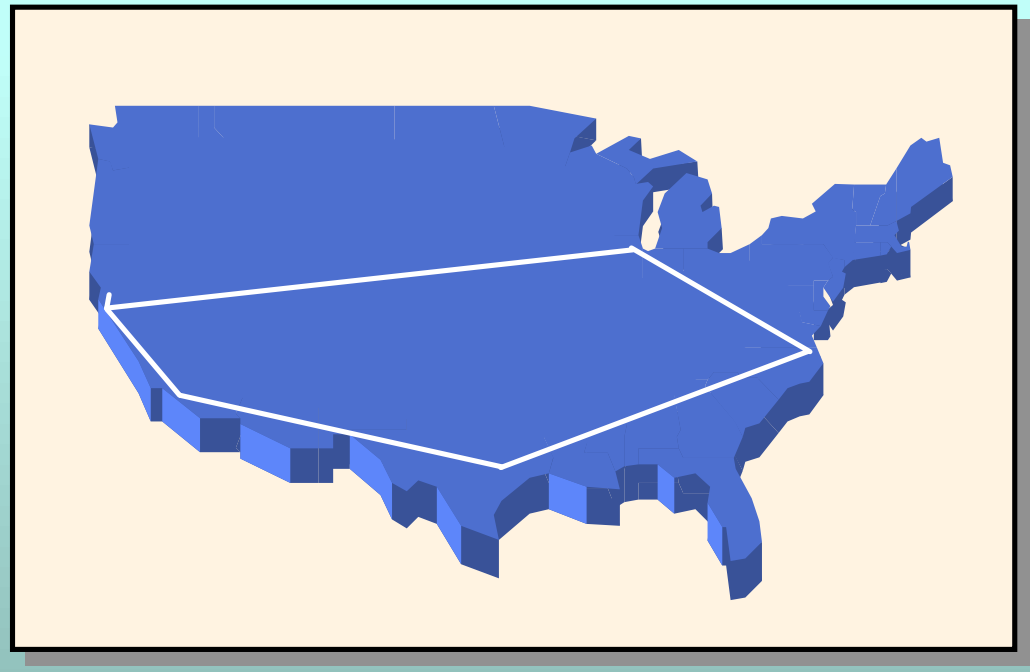
The Digital Basis of Life

[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*.

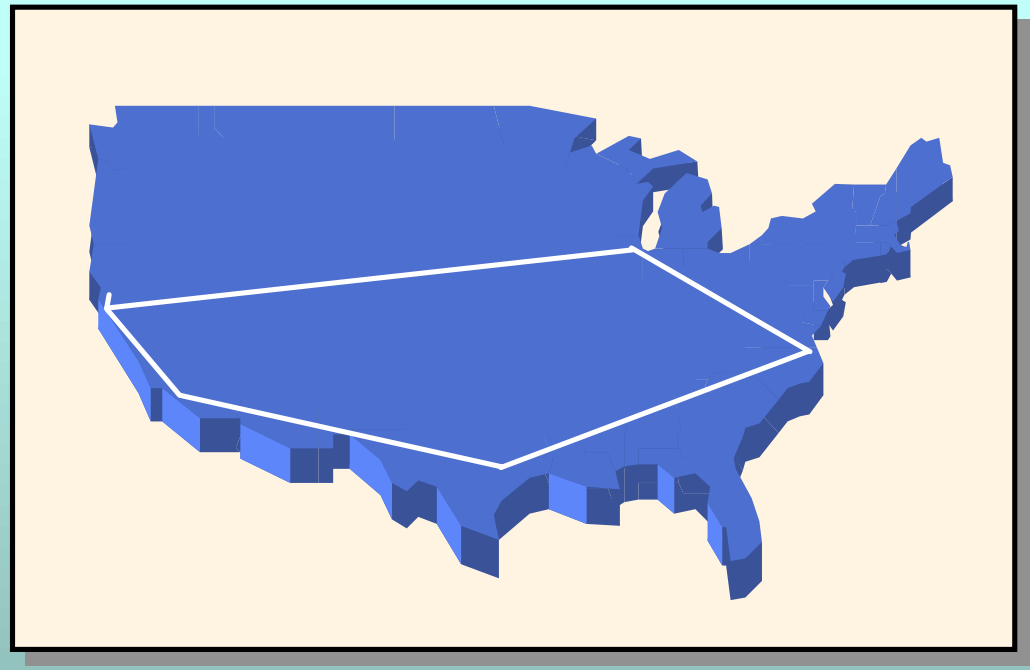
The Digital Basis of Life

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



The Digital Basis of Life

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



Information is passed from parent to child in form that is genuinely, not metaphorically digital. The biological encoding of digital information is incredibly efficient.

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.
- Storing all of the (redundant) information in all of the world's DNA on computer hard disks would require that the entire surface of the Earth be covered to a depth of three miles in Conner 1.0 gB drives.

Genomics: An Example

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.

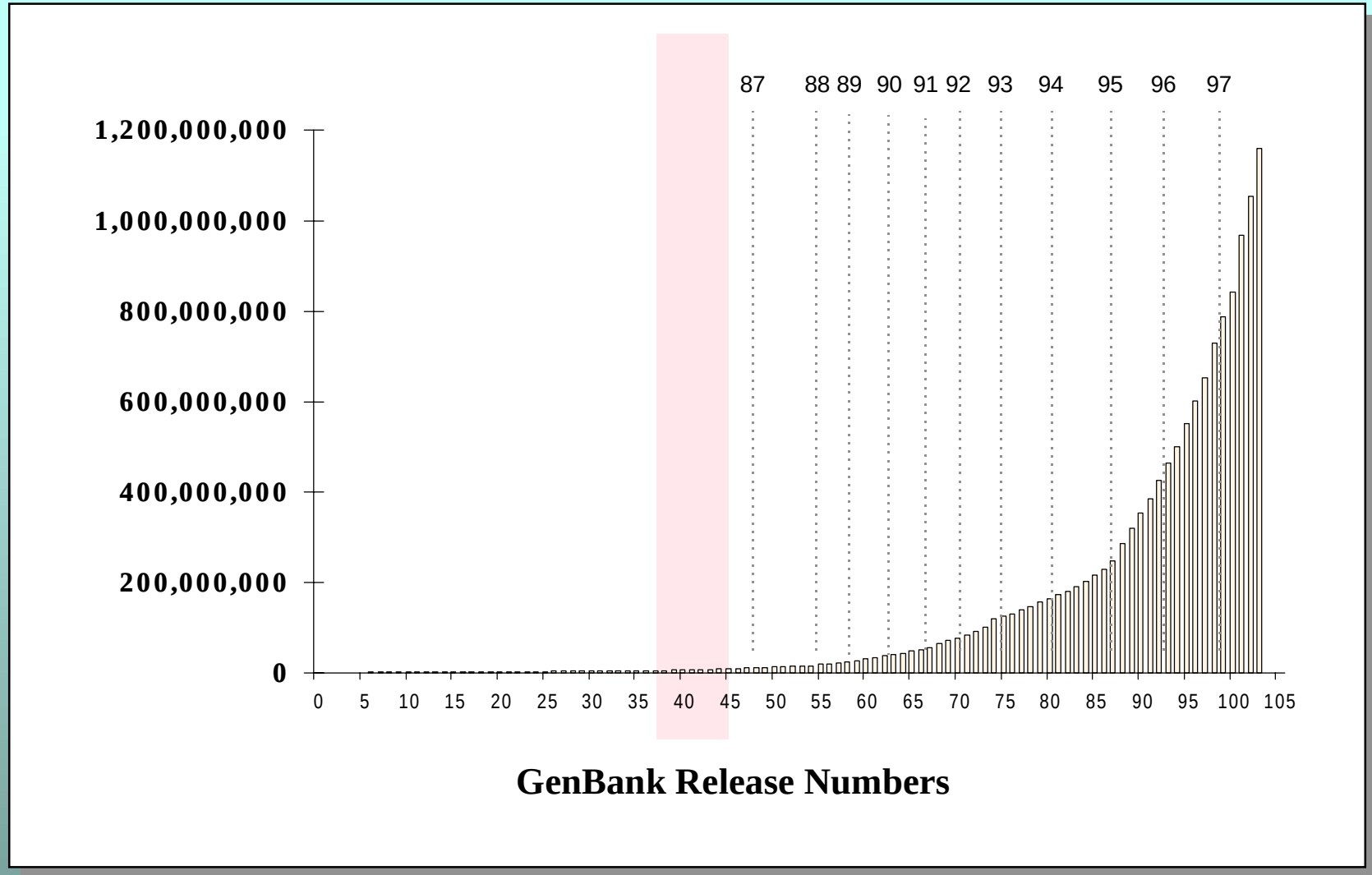
USDOE. 1990. *Understanding Our Genetic Inheritance. The U.S. Human Genome Project: The First Five Years.*

Infrastructure and the HGP

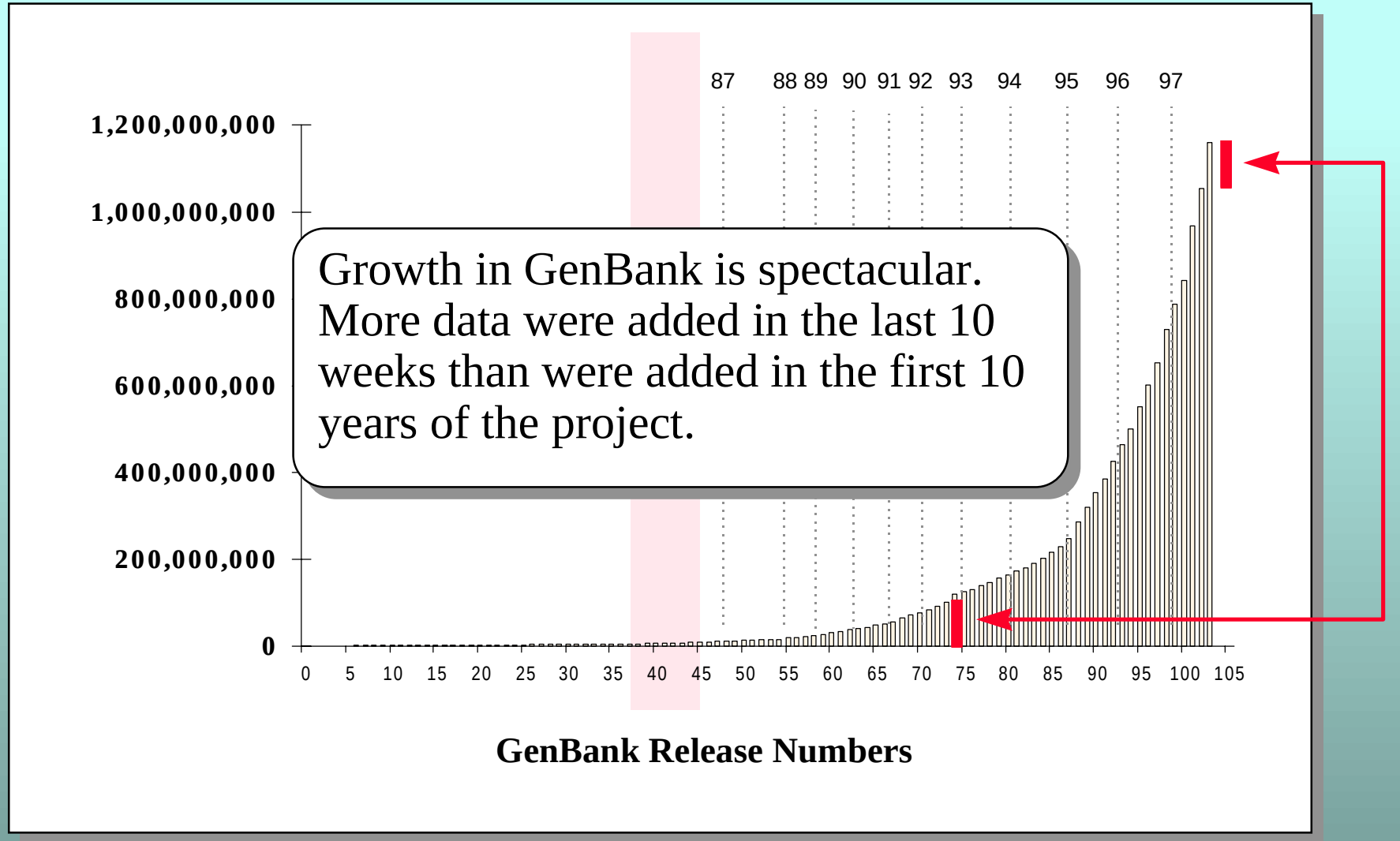
Progress towards all of the [Genome Project] goals will require the establishment of well-funded centralized facilities, including a stock center for the cloned DNA fragments generated in the mapping and sequencing effort and a data center for the computer-based collection and distribution of large amounts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

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

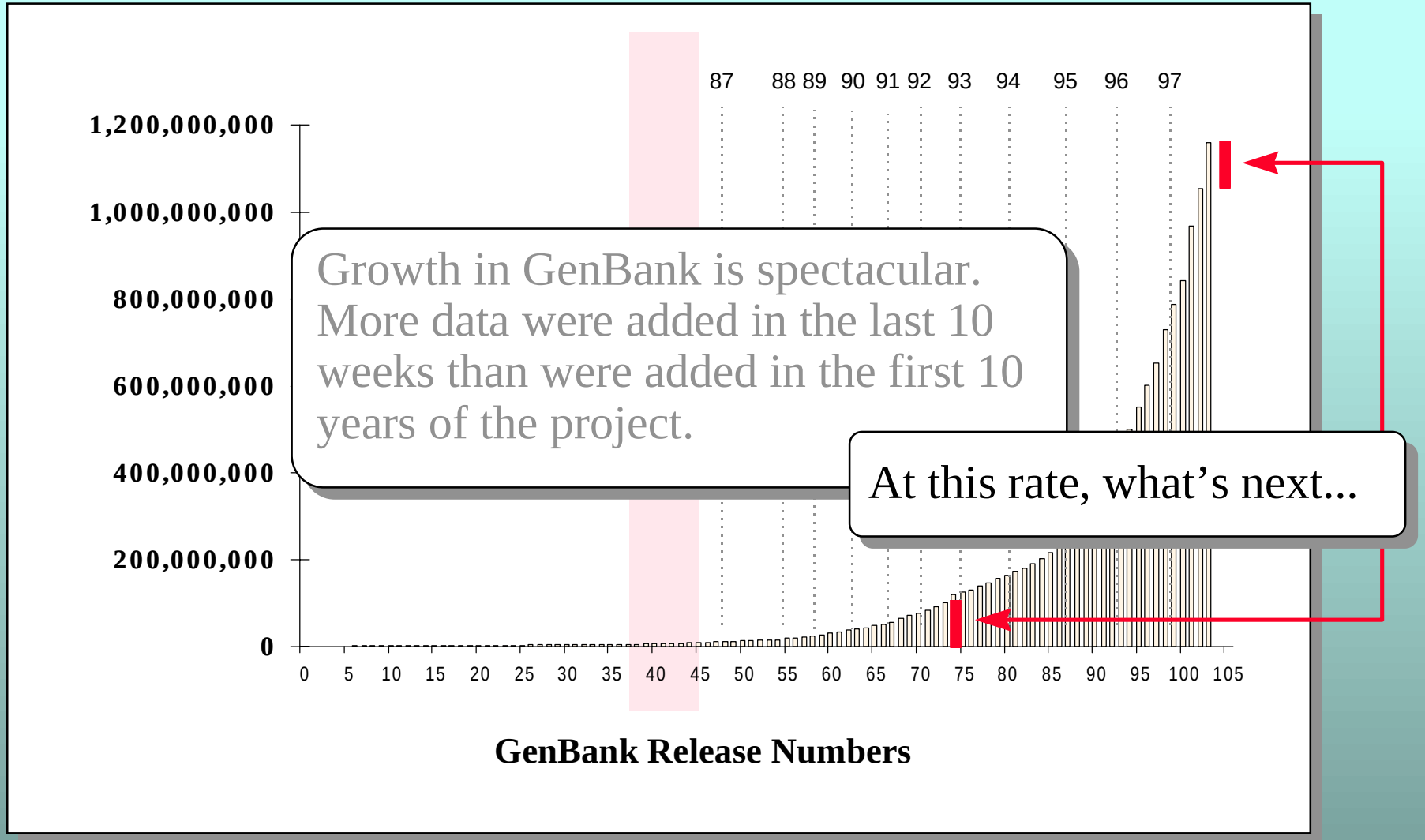
Base Pairs in GenBank



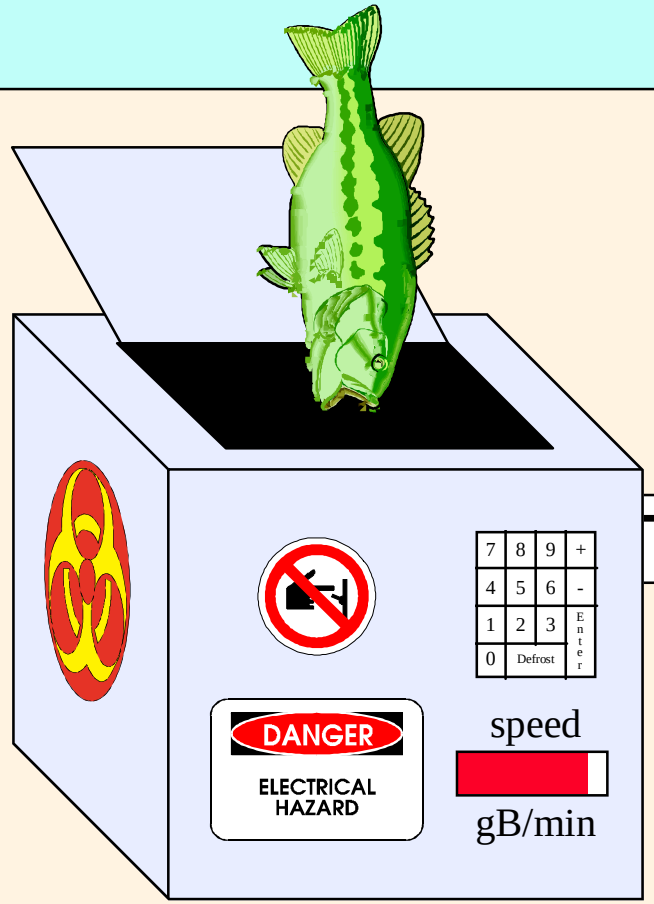
Base Pairs in GenBank



Base Pairs in GenBank



ABI *Bass-o-Matic Sequencer*



TGCGCATCGCGTATCGATAG

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

What's Really Next

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

What's Really Next

The post-genome era 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...

What's Really Next

The post-genome era 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...

The path to understanding will require even more data...

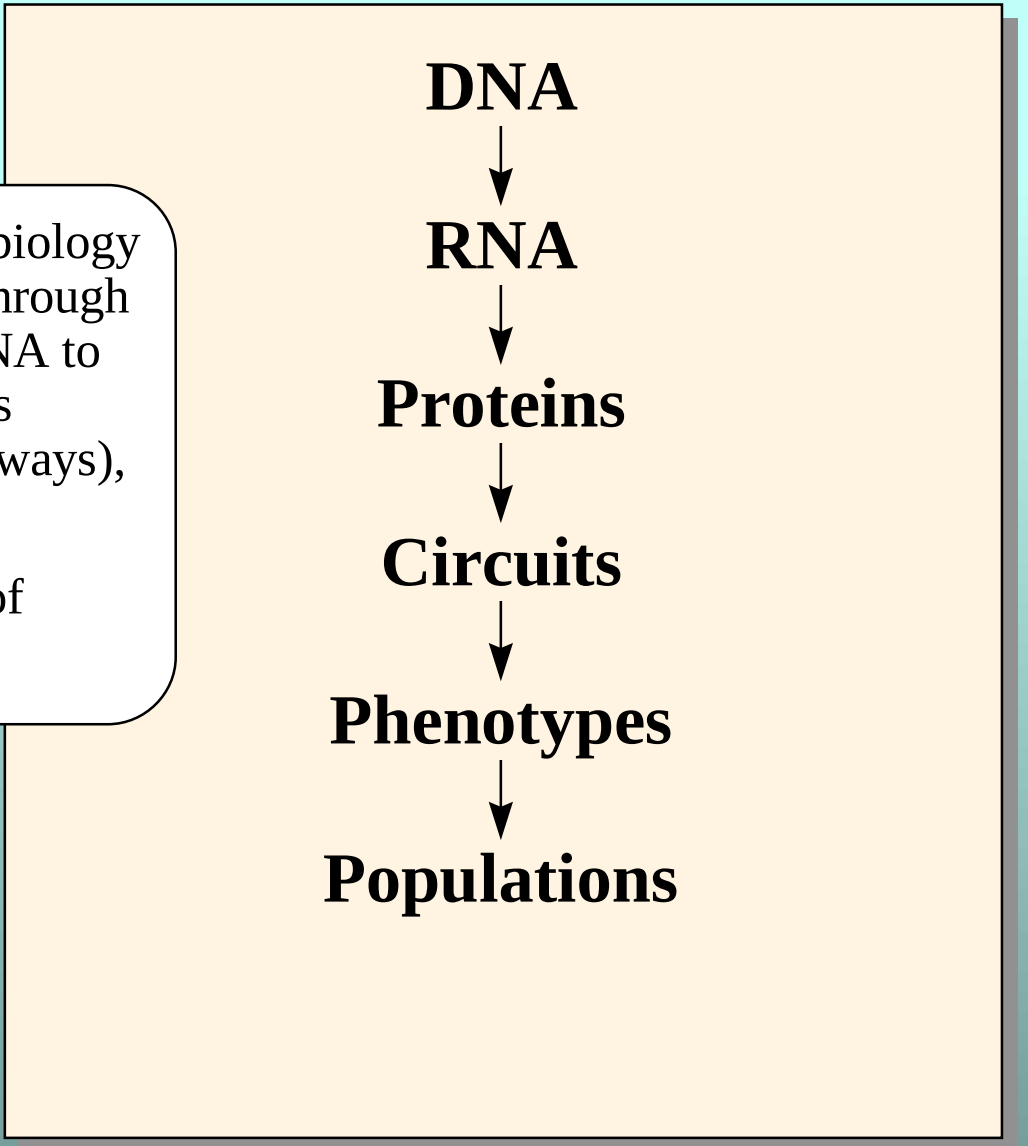
21st Century Biology

The Science

Fundamental Dogma

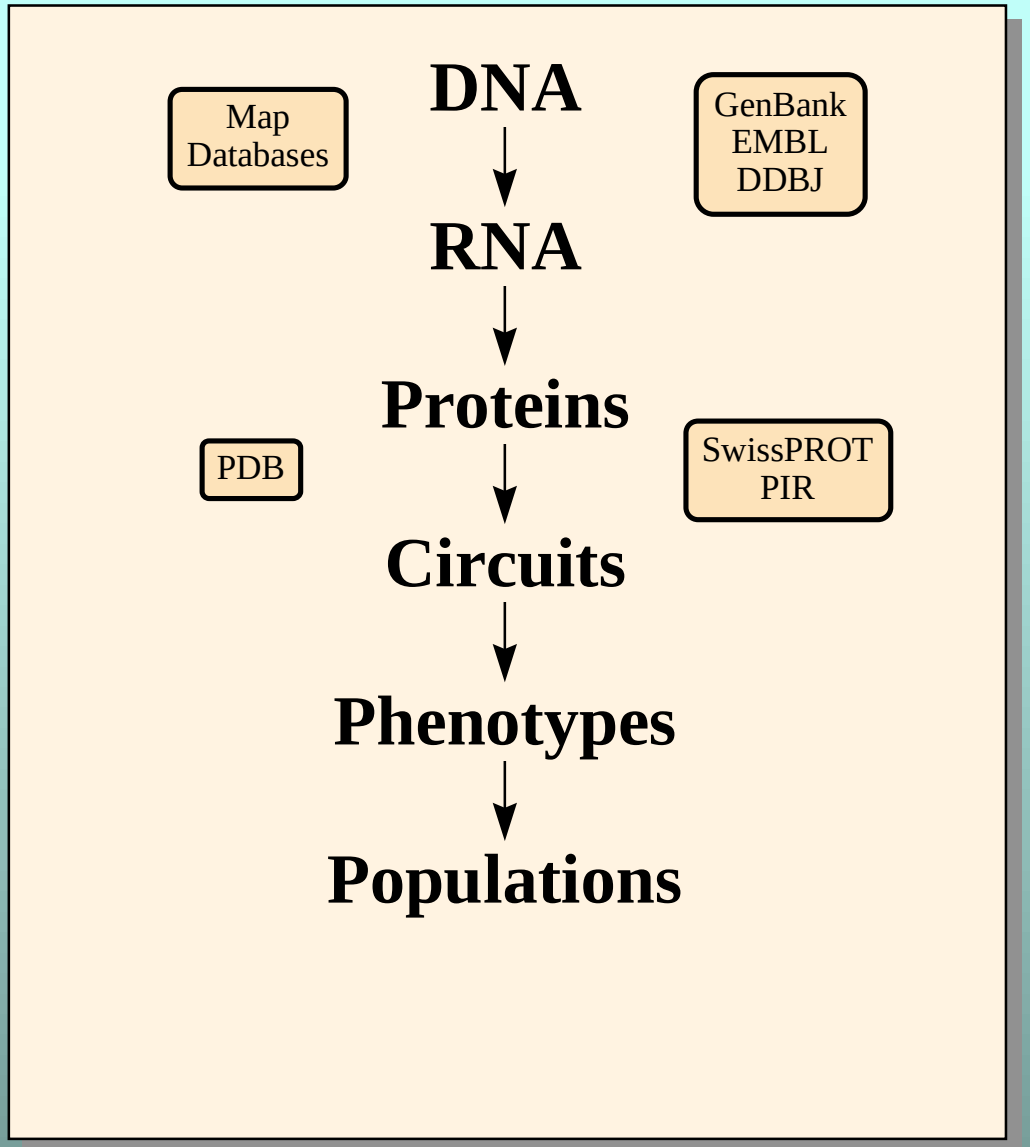
The fundamental dogma of molecular biology is that genes act to create phenotypes through a flow of information from DNA to RNA to proteins, to interactions among proteins (regulatory circuits and metabolic pathways), and ultimately to phenotypes.

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



Fundamental Dogma

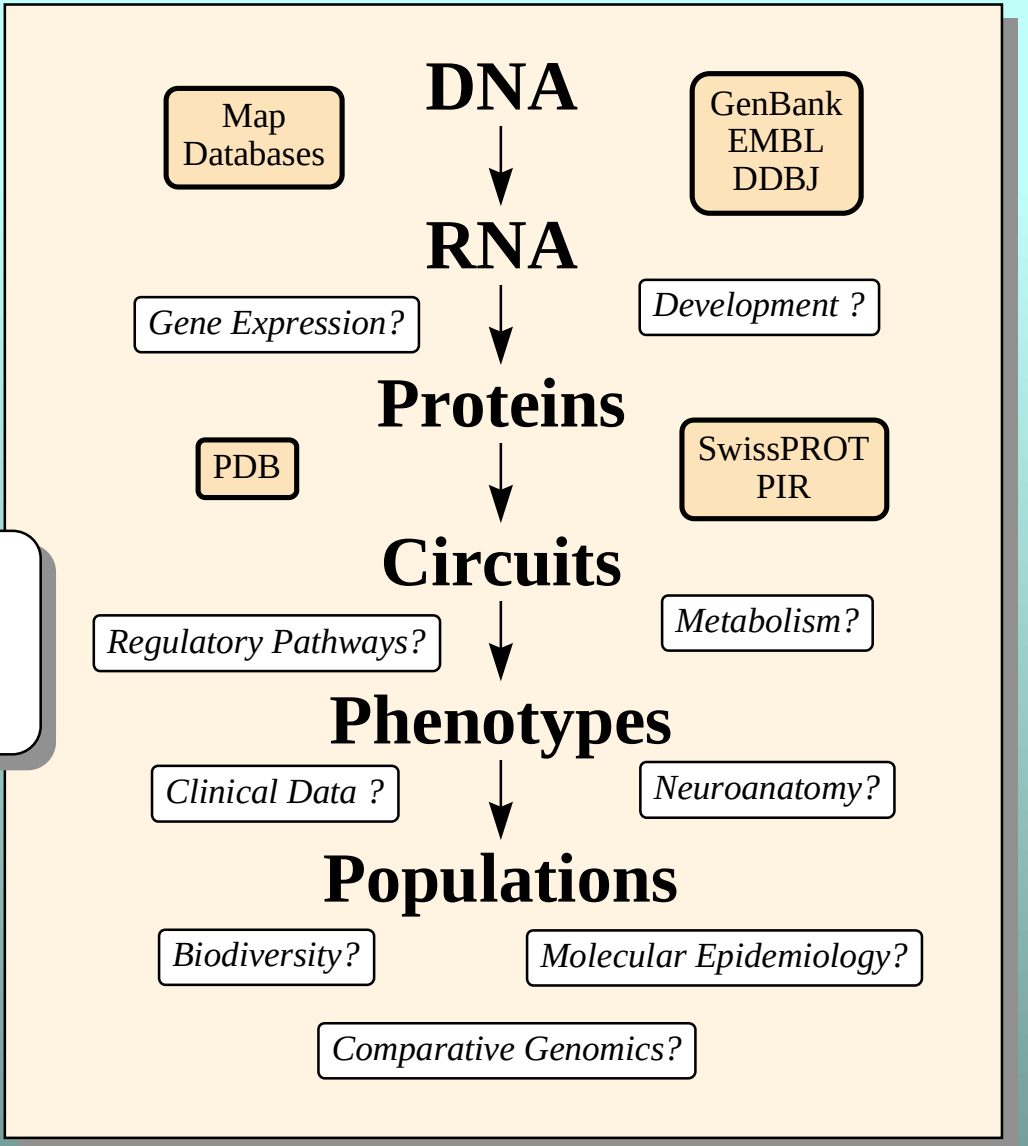
Although a few databases already exist to distribute molecular information,



Fundamental Dogma

Although a few databases already exist to distribute molecular information,

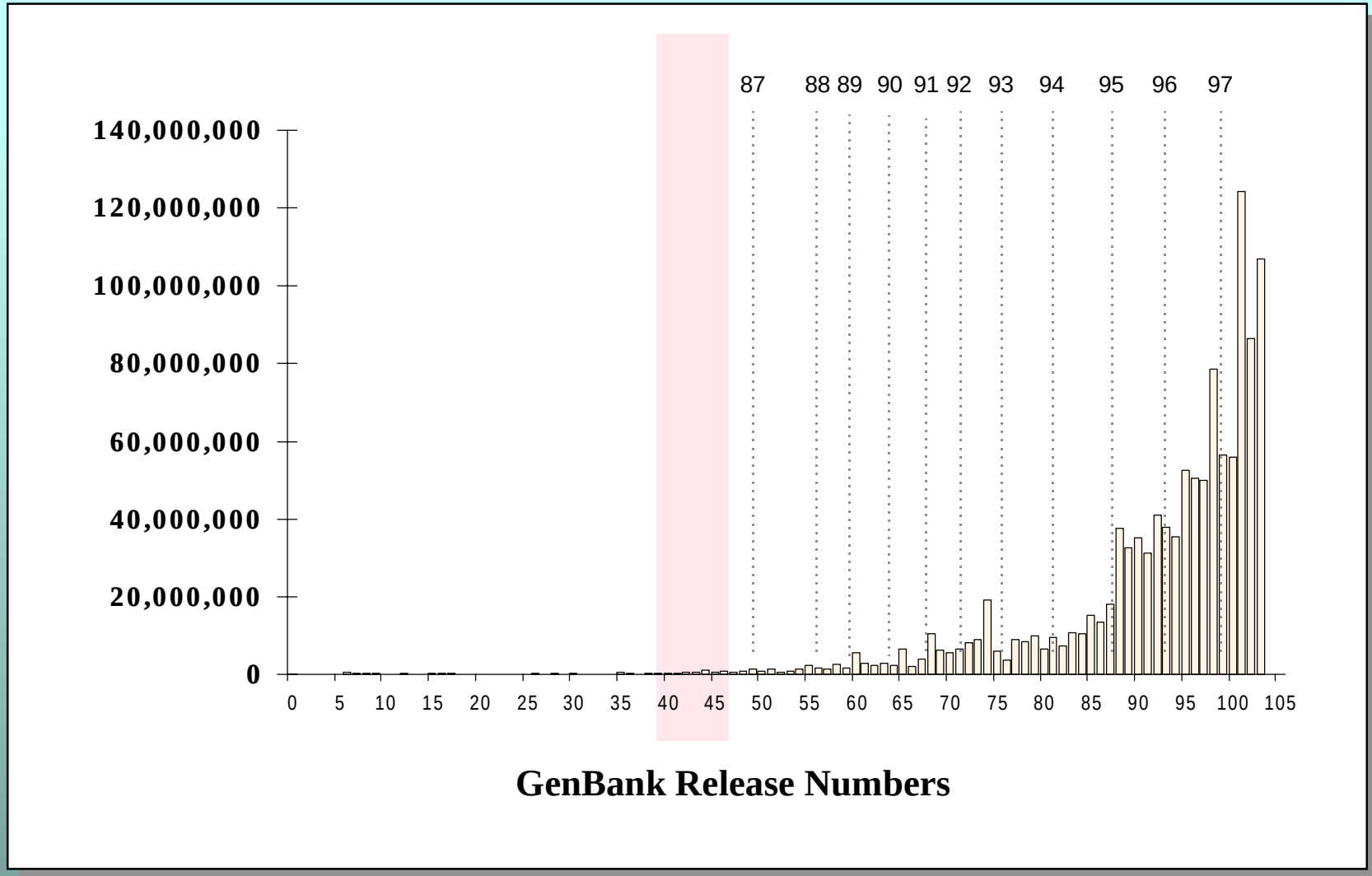
the post-genomic era will need many more to collect, manage, and publish the coming flood of new findings.



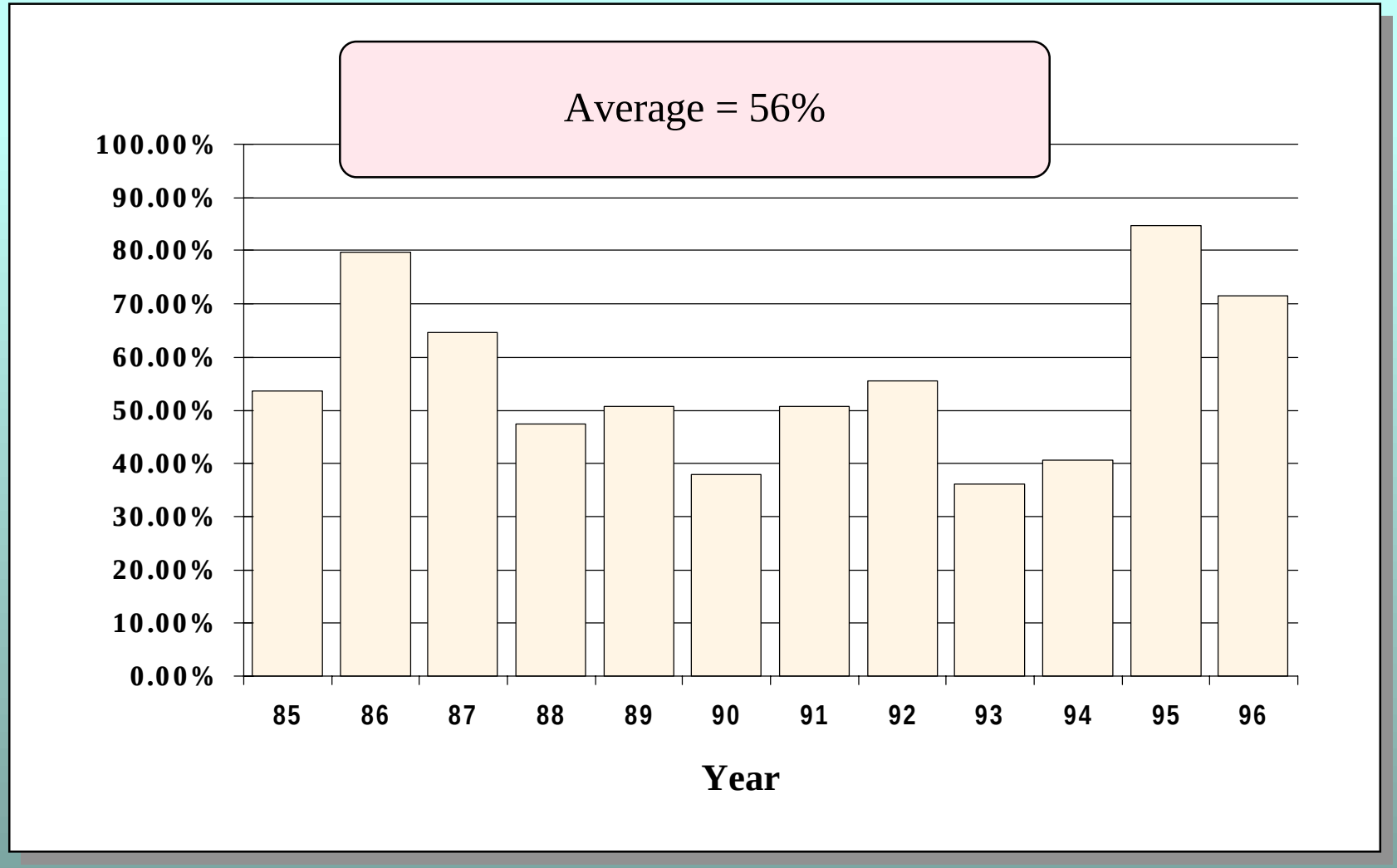
21st Century Biology

Data Volume

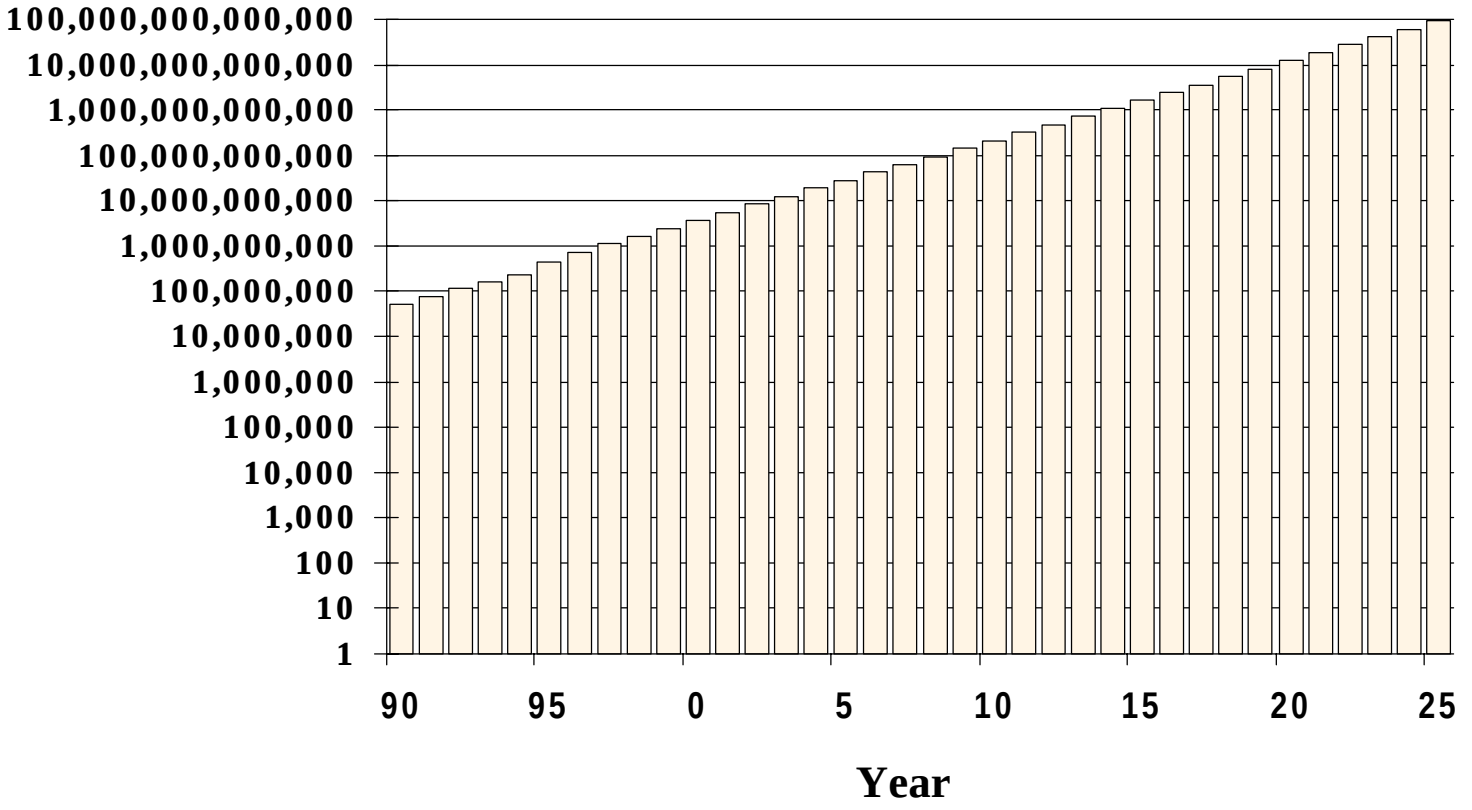
Base Pairs in GenBank *(changes)*



Base Pairs in GenBank *(Percent Increase)*

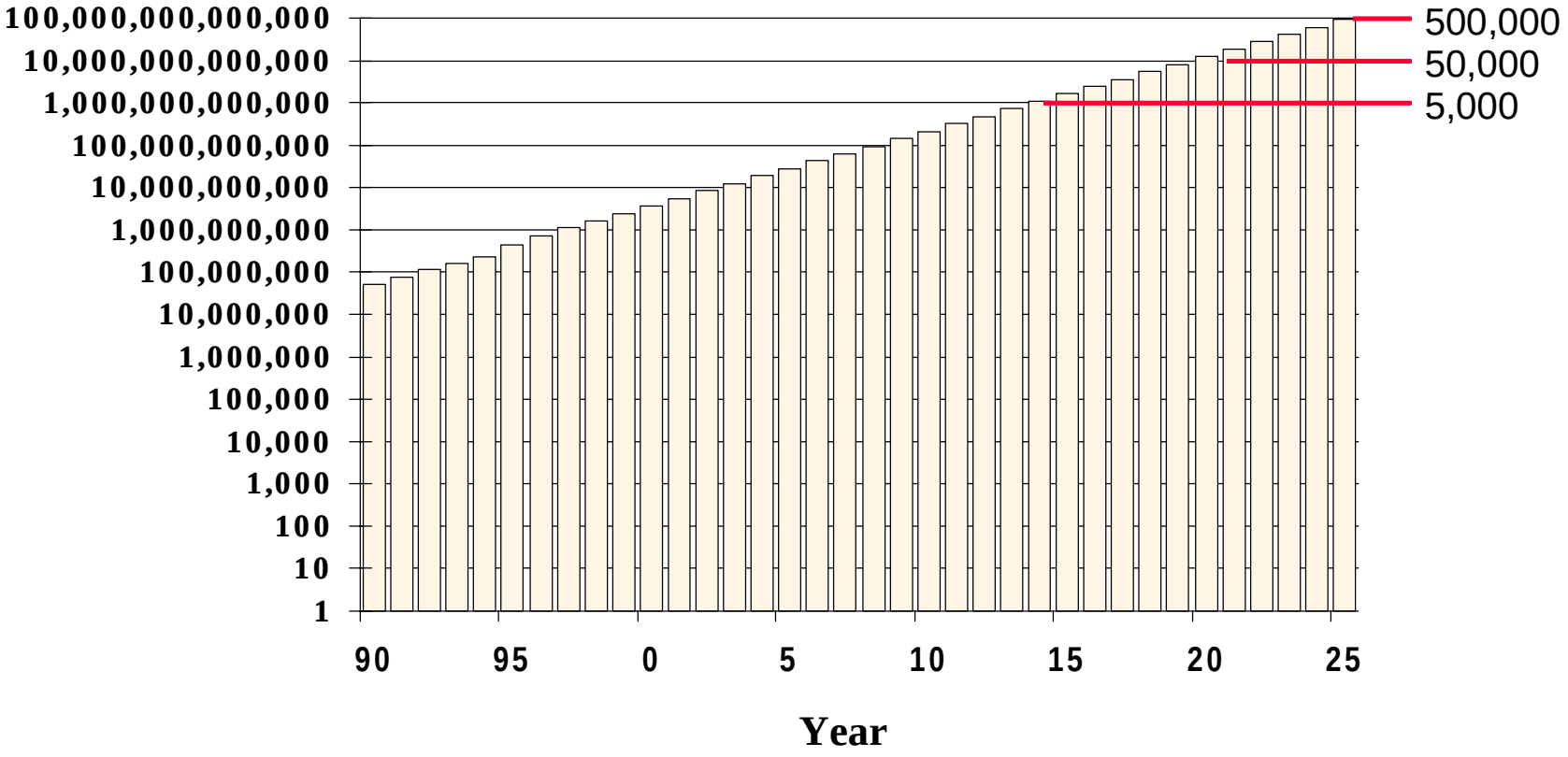


Projected Base Pairs



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

Projected Base Pairs



Ridiculous growth numbers, indicated as number of base pairs per individual medical record in the US.

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.

Funding for Bio-Information Infrastructure

Call for Change

Among the many new tools that are or will be needed (for 21st-century biology), some of those having the highest priority are:

- bioinformatics
- computational biology
- functional imaging tools using biosensors and biomarkers
- transformation and transient expression technologies
- nanotechnologies

Impact of Emerging Technologies on the Biological Sciences: Report of a Workshop. NSF-supported workshop, held 26-27 June 1995, Washington, DC.

The Problem

- IT will play a central role in 21st Century, post-genome-era biology.

The Problem

- IT will play a central role in 21st Century, post-genome-era biology.
- IT moves at “Internet Speed” and responds rapidly to market forces.

The Problem

- IT will play a central role in 21st Century, post-genome-era biology.
- IT moves at “Internet Speed” and responds rapidly to market forces.
- Current levels of support for public bio-information infrastructure are too low.

The Problem

- IT will play a central role in 21st Century, post-genome-era biology.
- IT moves at “Internet Speed” and responds rapidly to market forces.
- Current levels of support for public bio-information infrastructure are too low.
- Compared to internet speed, federal grant-funding decision processes are ponderously slow and inefficient.

Federal Funding of Bio-Databases

The challenges:

- providing adequate funding levels

Federal Funding of Bio-Databases

The challenges:

- providing adequate funding levels
- making timely, efficient decisions

IT Budgets

A Reality Check

Rhetorical Question

Which is likely to be more complex:

- identifying, documenting, and tracking the whereabouts of all parcels in transit in the US at one time, or...

Rhetorical Question

Which is likely to be more complex:

- identifying, documenting, and tracking the whereabouts of all parcels in transit in the US at one time, or...
- identifying, documenting, and analyzing the structure and function of all individual genes in all economically significant organisms; then analyzing all significant gene-gene and gene-environment interactions in those organisms and their environments.

Business Factoids

United Parcel Service:

- uses two redundant 3 Terabyte (yes, 3000 GB) databases to track all packages in transit.

Business Factoids

United Parcel Service:

- uses two redundant 3 Terabyte (yes, 3000 GB) databases to track all packages in transit.
- has 4,000 full-time employees dedicated to IT.

Business Factoids

United Parcel Service:

- uses two redundant 3 Terabyte (yes, 3000 GB) databases to track all packages in transit.
- has 4,000 full-time employees dedicated to IT.
- spends one billion dollars per year on IT.

Business Factoids

United Parcel Service:

- uses two redundant 3 Terabyte (yes, 3000 GB) databases to track all packages in transit.
- has 4,000 full-time employees dedicated to IT.
- spends one billion dollars per year on IT.
- has an income of 1.1 billion dollars (against revenues of 22.4 billion dollars).

Business Comparisons

Company	Revenues	IT Budget	Pct
Chase-Manhattan	16,431,000,000	1,800,000,000	10.95 %
AMR Corporation	17,753,000,000	1,368,000,000	7.71 %
Nation's Bank	17,509,000,000	1,130,000,000	6.45 %
Sprint	14,235,000,000	873,000,000	6.13 %
IBM	75,947,000,000	4,400,000,000	5.79 %
MCI	18,500,000,000	1,000,000,000	5.41 %
Microsoft	11,360,000,000	510,000,000	4.49 %
United Parcel	22,400,000,000	1,000,000,000	4.46 %
Bristol-Myers Squibb	15,065,000,000	440,000,000	2.92 %
Pfizer	11,306,000,000	300,000,000	2.65 %
Pacific Gas & Electric	10,000,000,000	250,000,000	2.50 %
Wal-Mart	104,859,000,000	550,000,000	0.52 %
K-Mart	31,437,000,000	130,000,000	0.41 %

Bio IT Support

A Modest Proposal

Level of Support

Appropriate funding level:

- approx. 5-10% of research funding
- *i.e.*, 1 - 2 **billion** dollars per year

Source of estimate:

- Experience of IT-transformed industries.
- Current support for IT-rich biological research.

Process of Support

Possible solutions:

- increase the direct support of federal service organizations providing information infrastructure (*e.g.*, NCBI).

Process of Support

Possible solutions:

- increase the direct support of federal service organizations providing information infrastructure (*e.g.*, NCBI).
- reduce supply-side support for investigator-initiated, grant-funded public database projects.

Process of Support

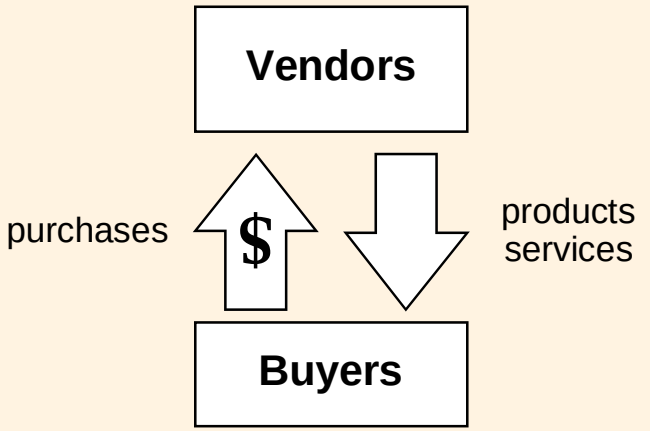
Possible solutions:

- increase the direct support of federal service organizations providing information infrastructure (*e.g.*, NCBI).
- reduce supply-side support for investigator-initiated, grant-funded public database projects.
- increase demand-side support for market-provided biomedical information resources.

Market Forces

In a simple market economy, vendors try to anticipate the needs of buyers and offer products and services to meet those needs.

Real users decide whether or not to buy a product or service, depending upon whether or not it meets a real need at a reasonable price.



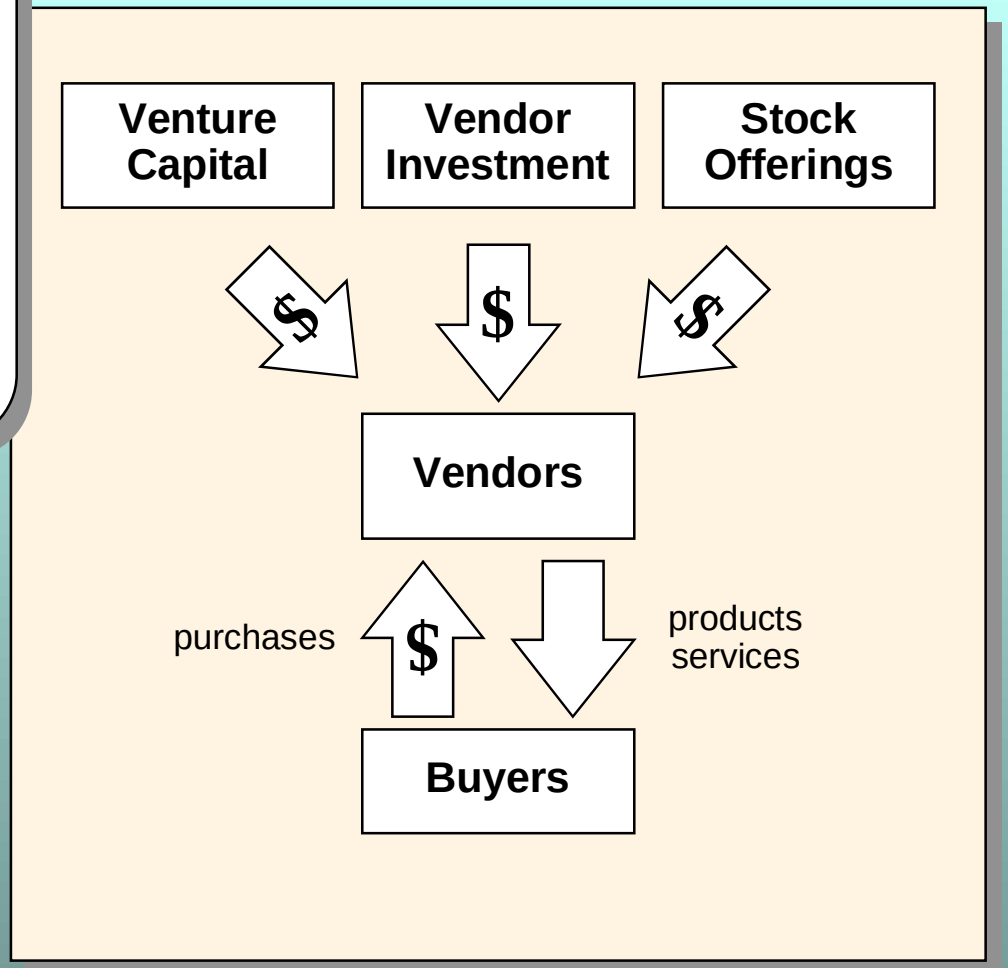
Business 101 Insight:

Successful vendors target a niche and excel at meeting the needs of that niche.

Market Forces

Funding to initiate the development of products and services come from investors, not from buyers.

Investors decide whether or not to provide start-up funding based upon the estimated ability of the vendor to create products and services that will meet real needs at competitive prices.

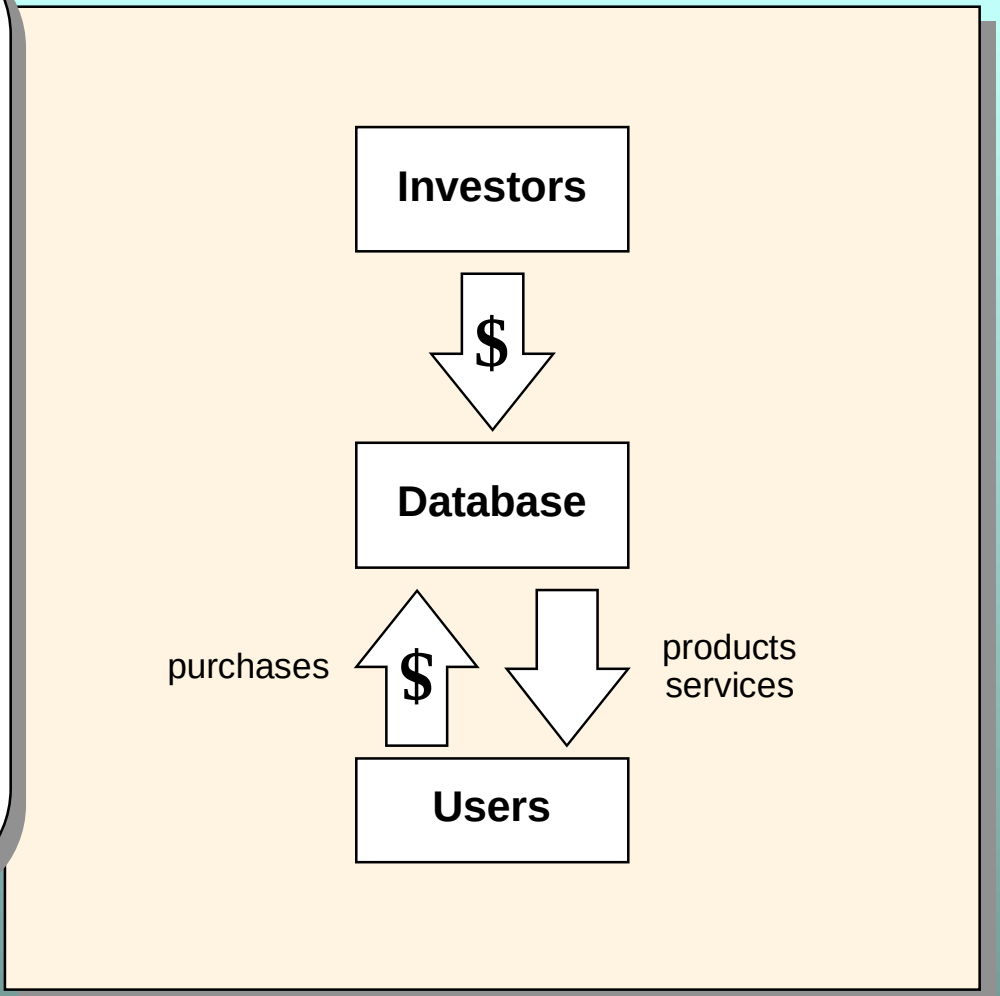


Federal Funding

If biological databases were driven by market forces, individual users would choose what services they need and individual database providers would choose what services to make available.

Investors would provide start-up money on the likelihood of successful products and services being developed.

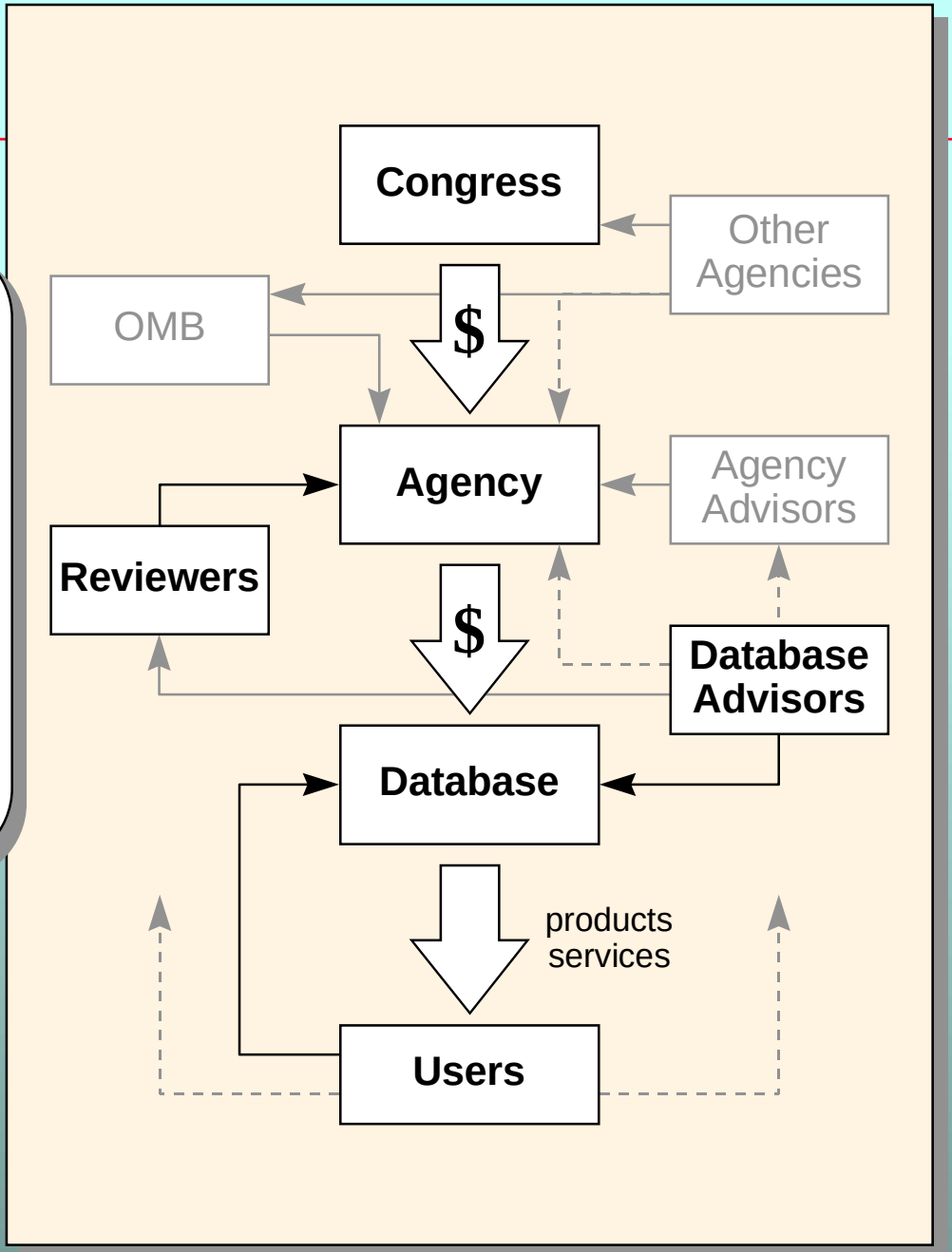
Ultimate success would depend on meeting the needs of real users. Decisions could be made rapidly, in response to changing needs and emerging opportunities.



Federal Funding

Instead, funding decisions for grant-supported biological databases can follow a ponderously slow course, with almost no opportunity for real-time input from real users.

Even with the best of intentions at all levels, this process is slow, inefficient, risk-averse, and non-responsive to the real and changing needs of users.



Federal Funding of Bio-Databases

Creating market forces:

- stop supporting the supply side of biodatabases through slow, inefficient processes.

Federal Funding of Bio-Databases

Creating market forces:

- stop supporting the supply side of biodatabases through slow, inefficient processes.
- start supporting the demand side through fast, efficient processes.

Federal Funding of Bio-Databases

Creating market forces:

- stop supporting the supply side of biodatabases through slow, inefficient processes.
- start supporting the demand side through fast, efficient processes.
- provide guaranteed supplementary funding, redeemable only for access to bio-databases.

Federal Funding of Bio-Databases

Creating market forces:

- stop supporting the supply side of biodatabases through slow, inefficient processes.
- start supporting the demand side through fast, efficient processes.
- provide guaranteed supplementary funding, redeemable only for access to bio-databases.
- data stamps

Federal Funding of Bio-Databases

Creating market forces:

- stop supporting the supply side of biodatabases through slow, inefficient processes.
- start supporting the demand side through fast, efficient processes.
- provide guaranteed supplementary funding, redeemable only for access to bio-databases.
- data stamps, AKA ***food (for-thought) stamps*** **?!**

Food (for thought) Stamps

Funding Agencies could:

- provide a 10% supplement to **every** research grant in the form of “stamps” redeemable only at database providers.
- allow the “stamps” to be transferable among scientists, so that a market for them could emerge.
- provide funding only after the stamps have been redeemed at a database provider.

Food (for thought) Stamps

Problems:

- how to estimate the amount of FFT stamps that would actually be redeemed (and thus the required budget set-aside).
- how to identify “approved” database providers.
- how to initiate the FFT system.
- etc etc

Food (for thought) Stamps

Alternatives (if no solution emerges):

- increasingly inefficient research activities (abject failure will occur when it becomes simpler to repeat research than to obtain prior results).
- loss of access to bio-databases for public-sector research.
- movement of majority of “important” biological research into the private sector.
- loss of American pre-eminence (if other countries solve the problems first).

Slides:

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