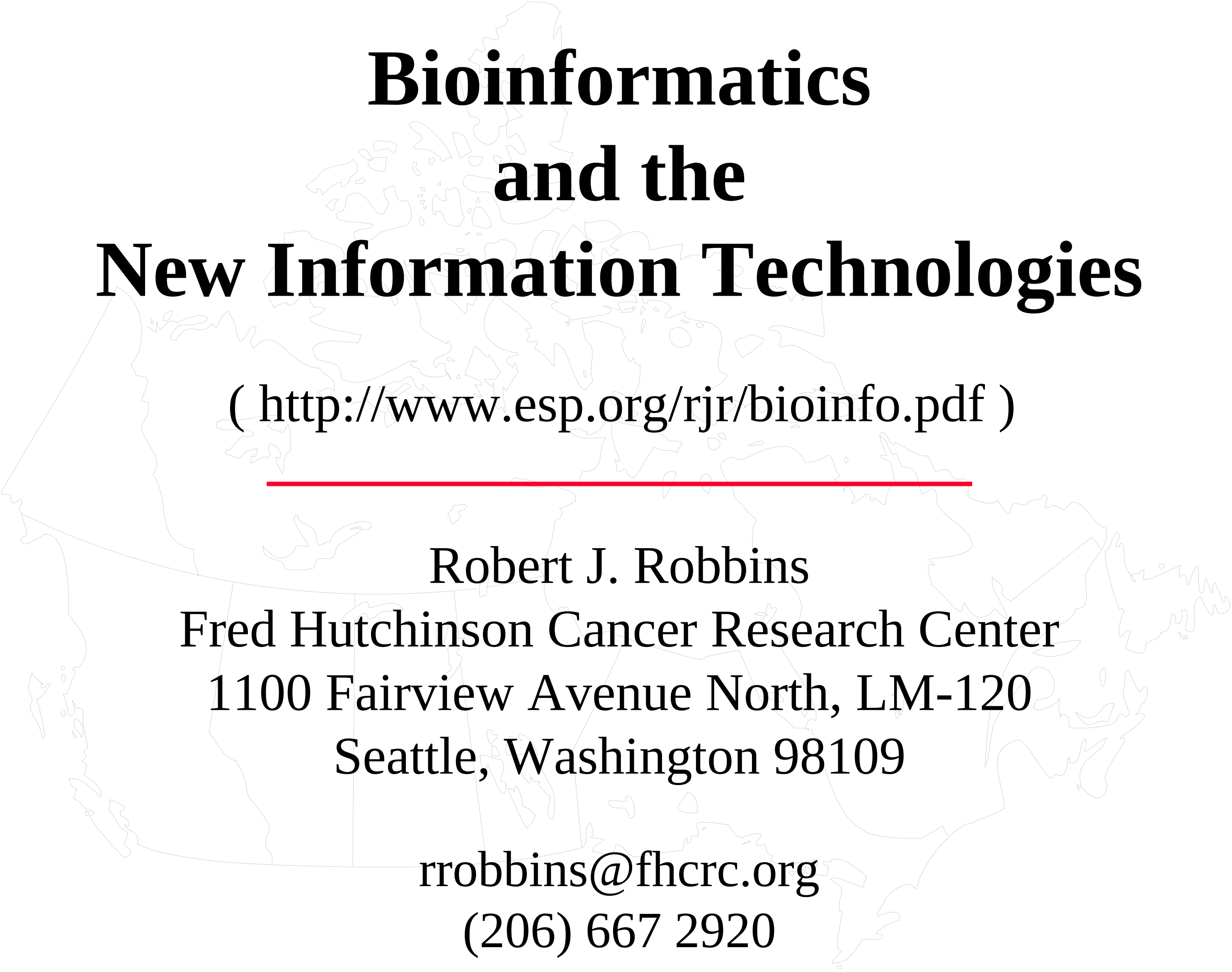


Bioinformatics and the New Information Technologies

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

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

rrobbins@fhcrc.org
(206) 667 2920

Bioinformatics and the New Information Technologies

What's so new?

Computers have been available to biologists for at least 25 years...

rrobbins@fhcrc.org
(206) 667 2920

Manhattan Purchase

7 November 1626

High and Mighty Lords,

Yesterday the ship the *Wapen van Amsterdam* arrived here. It sailed from New Netherland out of the River Mauritius on the 23d of September. They report that our people are in good spirit and live in peace. The women also have borne some children there. They have purchased the Island Manhattes from the Indians for the value of 60 guilders. It is 11,000 morgens in size [about 22,000 acres].

Your High and Mightinesses' obedient,

P Schaghen

Manhattan Purchase

In 1626, representatives of the Dutch West Indies Trading Company purchased all of Manhattan from the local residents for a price generally considered to be equivalent to $24.00.

Good deal for the buyer, bad deal for the seller, yes?

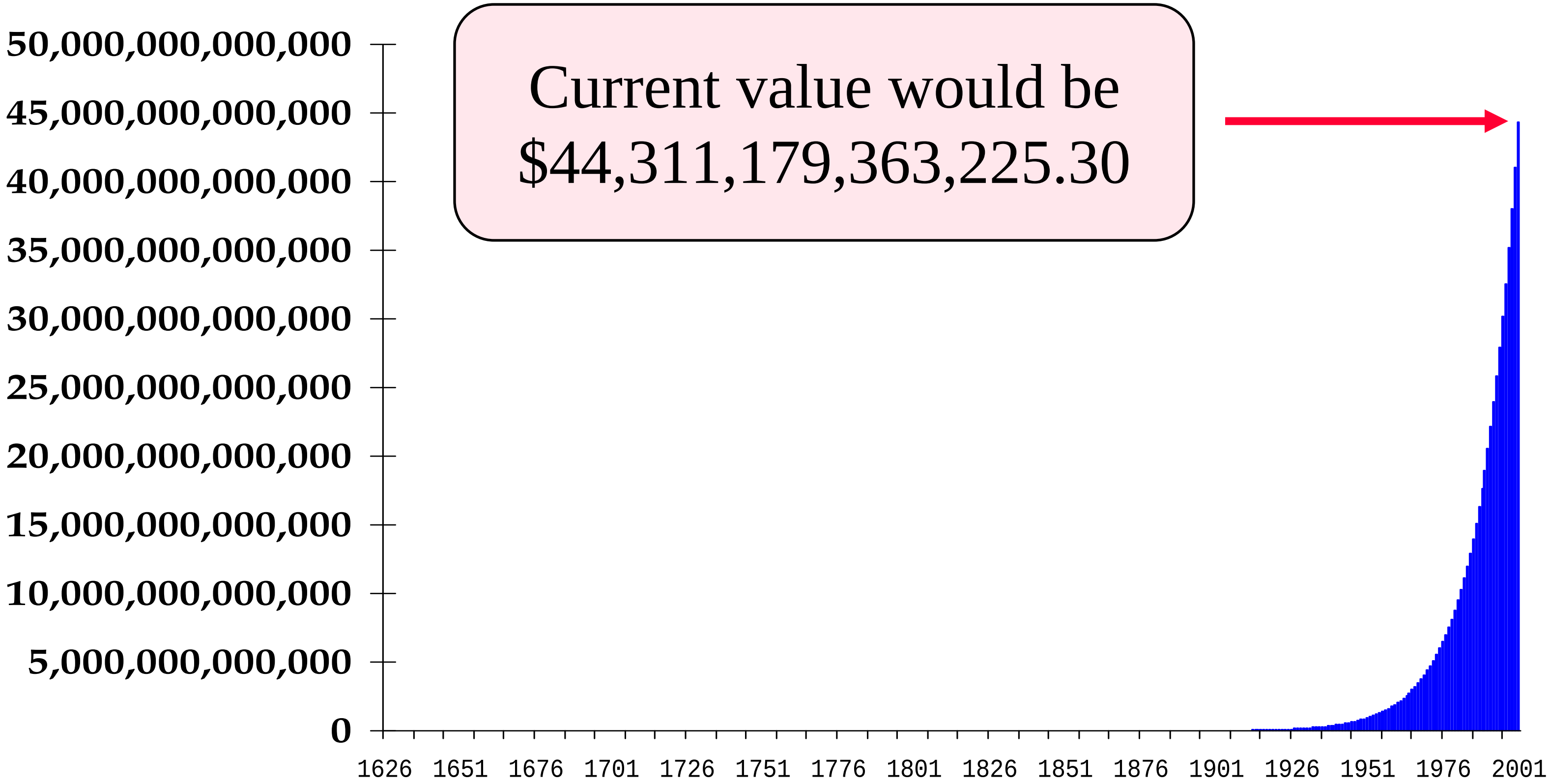
Manhattan Purchase

In 1626, representatives of the Dutch West Indies Trading Company purchased all of Manhattan from the local residents for a price generally considered to be equivalent to $24.00.

Good deal for the buyer, bad deal for the seller, yes?

Suppose the local residents had invested HALF of the $24.00 at 8% interest. What would that be worth now?

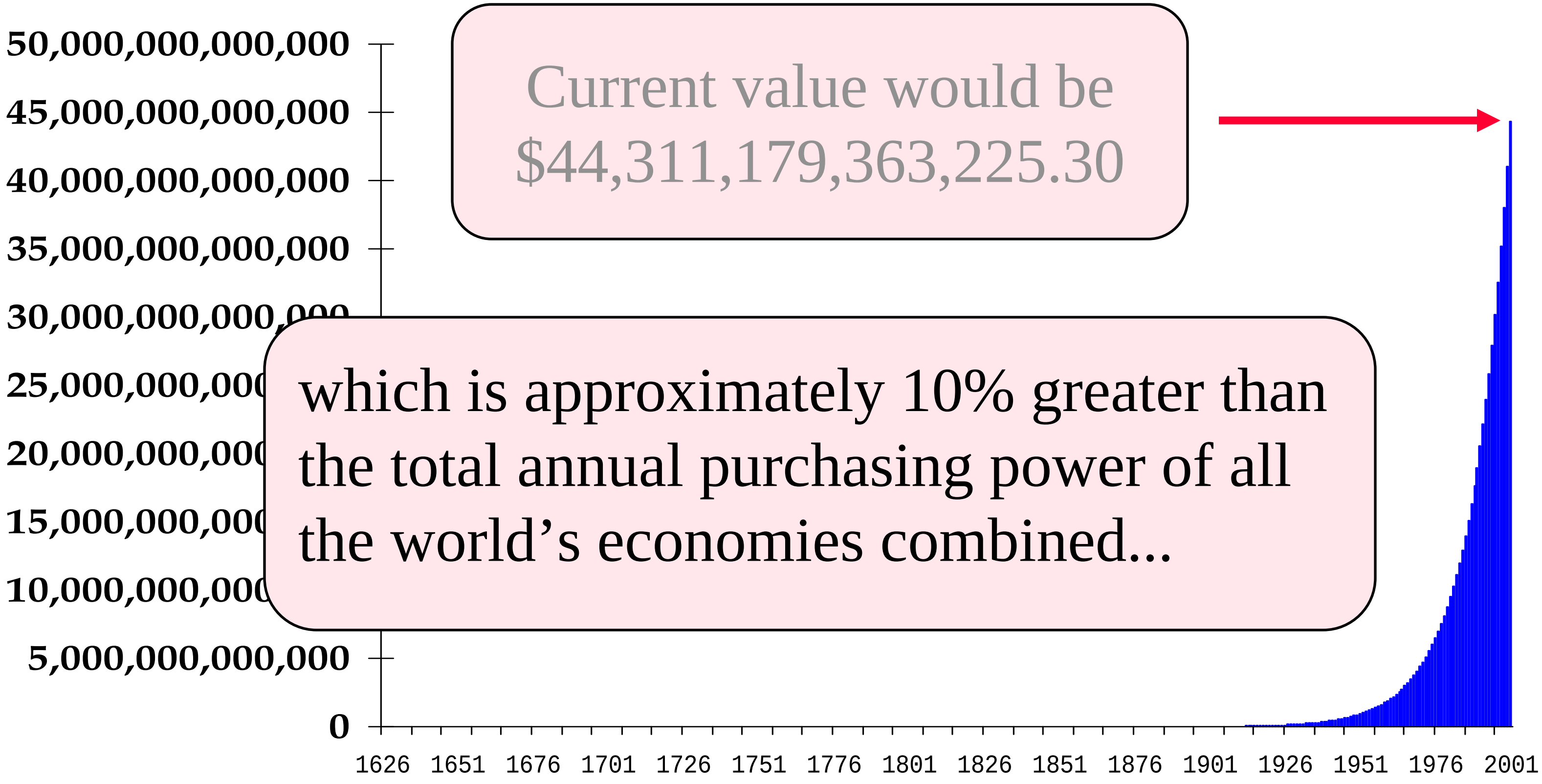
Manhattan Purchase



Current value would be $44,311,179,363,225.30

Twelve dollars, invested at 8% compound interest

Manhattan Purchase

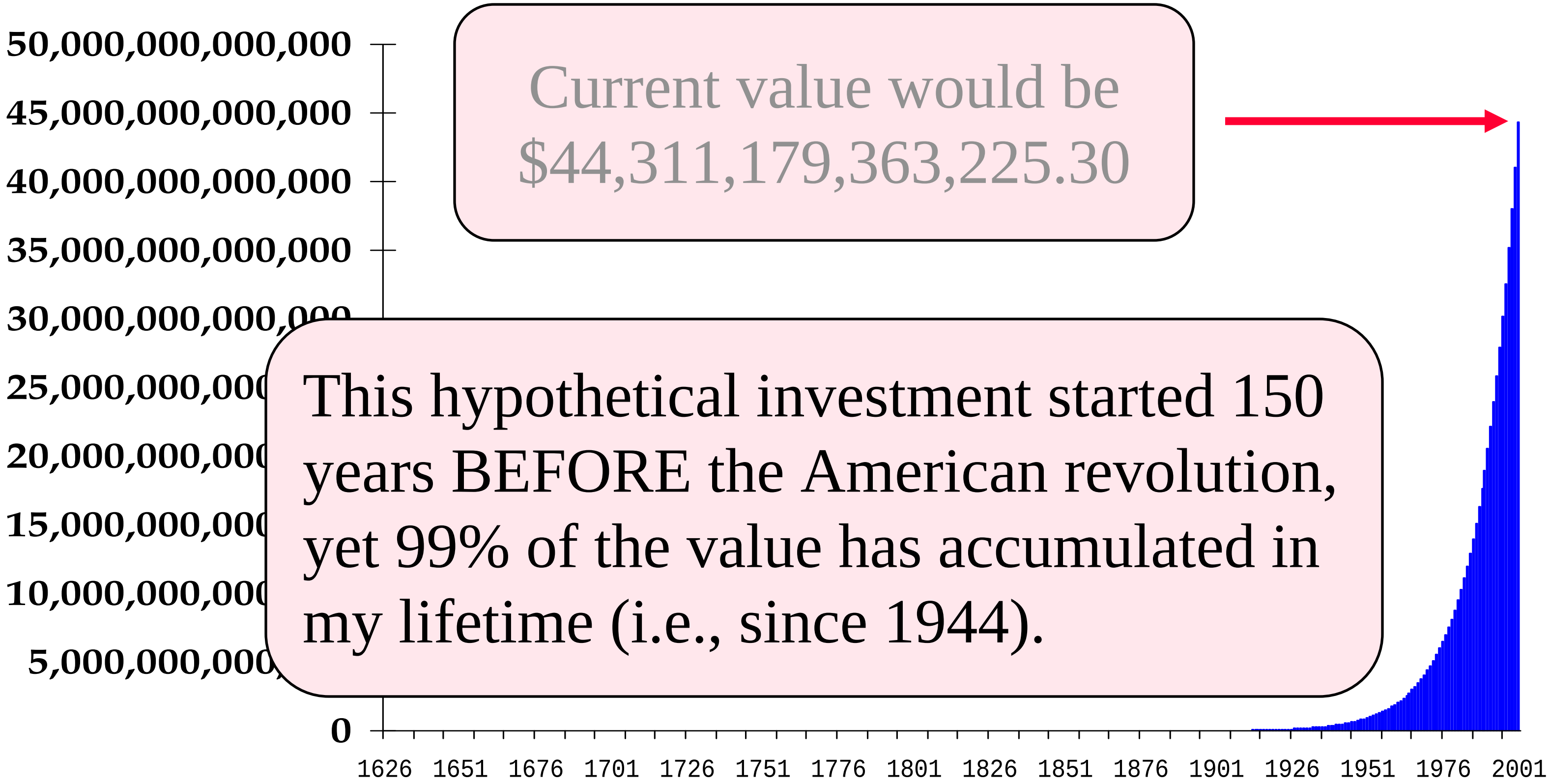


Current value would be $44,311,179,363,225.30

which is approximately 10% greater than the total annual purchasing power of all the world's economies combined...

Twelve dollars, invested at 8% compound interest

Manhattan Purchase



Current value would be $44,311,179,363,225.30

This hypothetical investment started 150 years BEFORE the American revolution, yet 99% of the value has accumulated in my lifetime (i.e., since 1944).

Twelve dollars, invested at 8% compound interest

Manhattan Purchase

Compound interest can be staggeringly powerful and global-scale phenomena can beggar the imagination.

Manhattan Purchase

Compound interest can be staggeringly powerful and global-scale phenomena can beggar the imagination.

The challenges of biodiversity informatics will be on this scale...

Abstract

The relentless exponential effect of Moore's Law is having profound effects upon the role of 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 desktop computer delivers more MIPS than the first Cray and GenBank acquires more data every week than it did in its first ten years.

The potential information storage capacity of the biosphere is astounding. Efforts to document and comprehend the diversity of the biosphere on a global scale will constitute one of the greatest data-management challenges of all time.

Topics

- Moore’s Law constantly transforms IT (and everything else).

Topics

- Moore’s Law constantly transforms IT (and everything else).
- Information Technology (IT) has a special relationship with biology.

Topics

- Moore’s Law constantly transforms IT (and everything else).
- Information Technology (IT) has a special relationship with biology.
- Bioinformatics will transform 21st-century biology.

Topics

- Moore’s Law constantly transforms IT (and everything else).
- Information Technology (IT) has a special relationship with biology.
- Bioinformatics will transform 21st-century biology.
- Documenting biospheric diversity on a global scale will constitute one of the greatest data-management challenges of all time.

Moore's Law

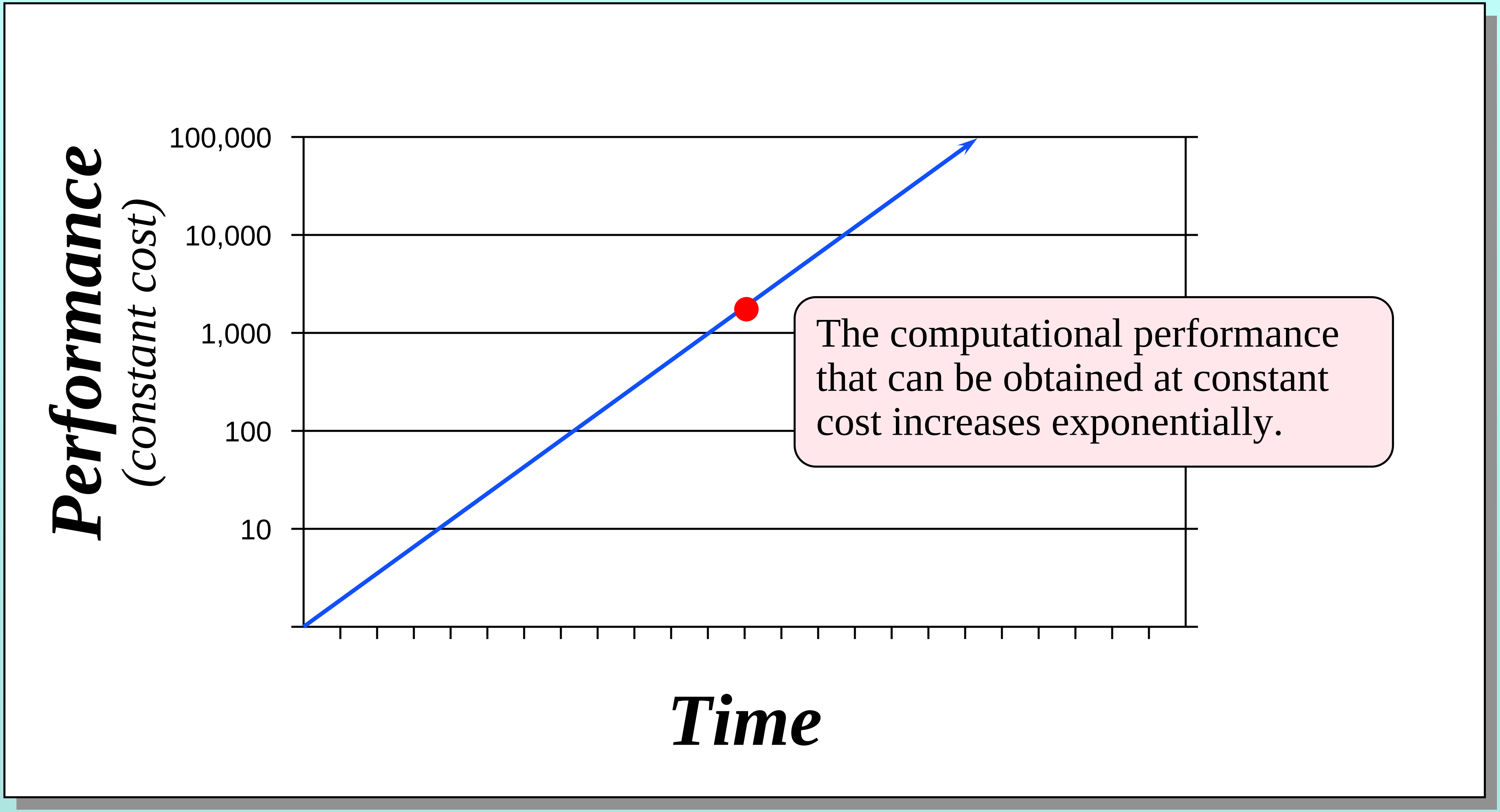
Transforms InfoTech (and everything else)

Moore's Law: *The Statement*

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

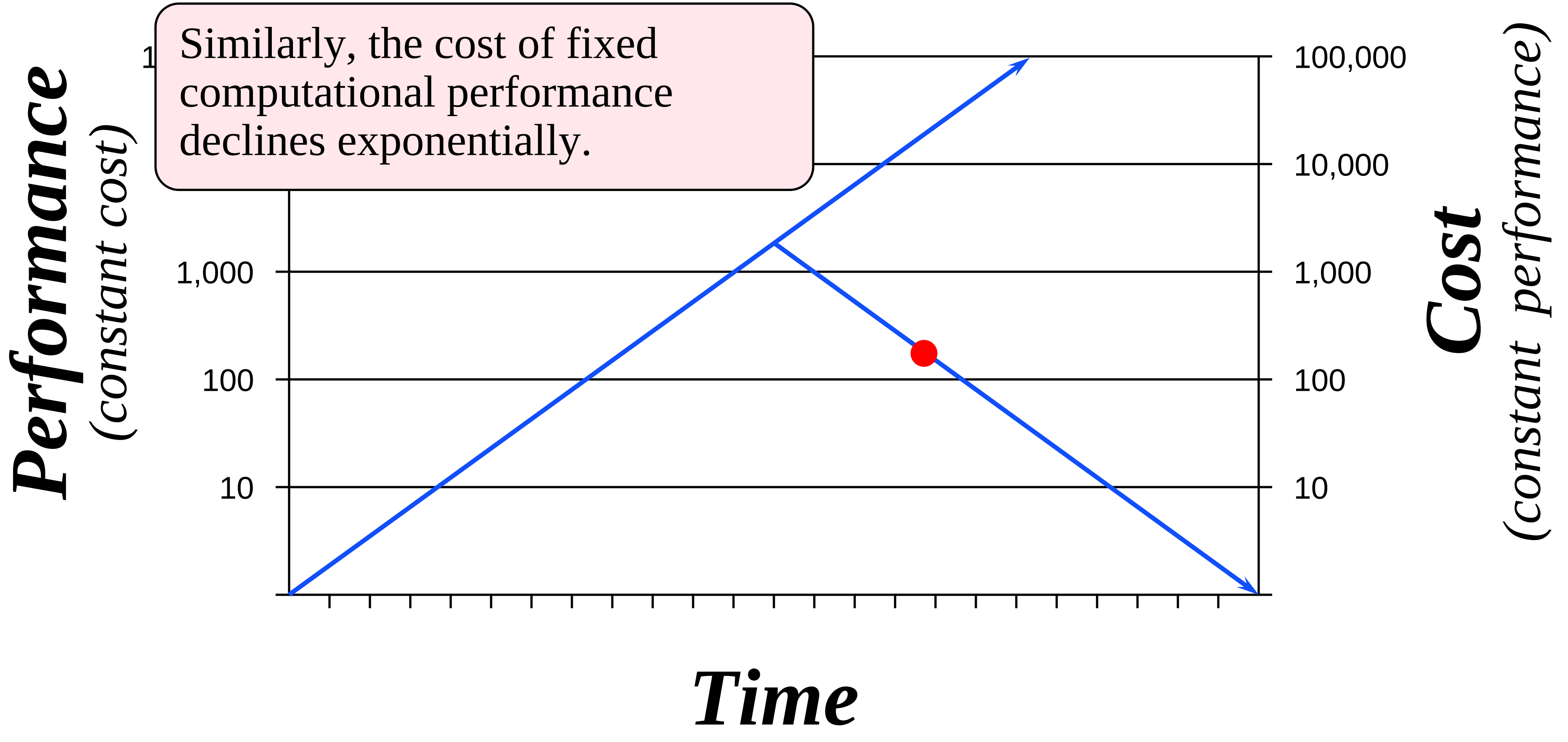
Gordon Moore, co-founder of Intel...

Moore’s Law: *The Effect*



The computational performance that can be obtained at constant cost increases exponentially.

Moore’s Law: *The Effect*



Similarly, the cost of fixed computational performance declines exponentially.

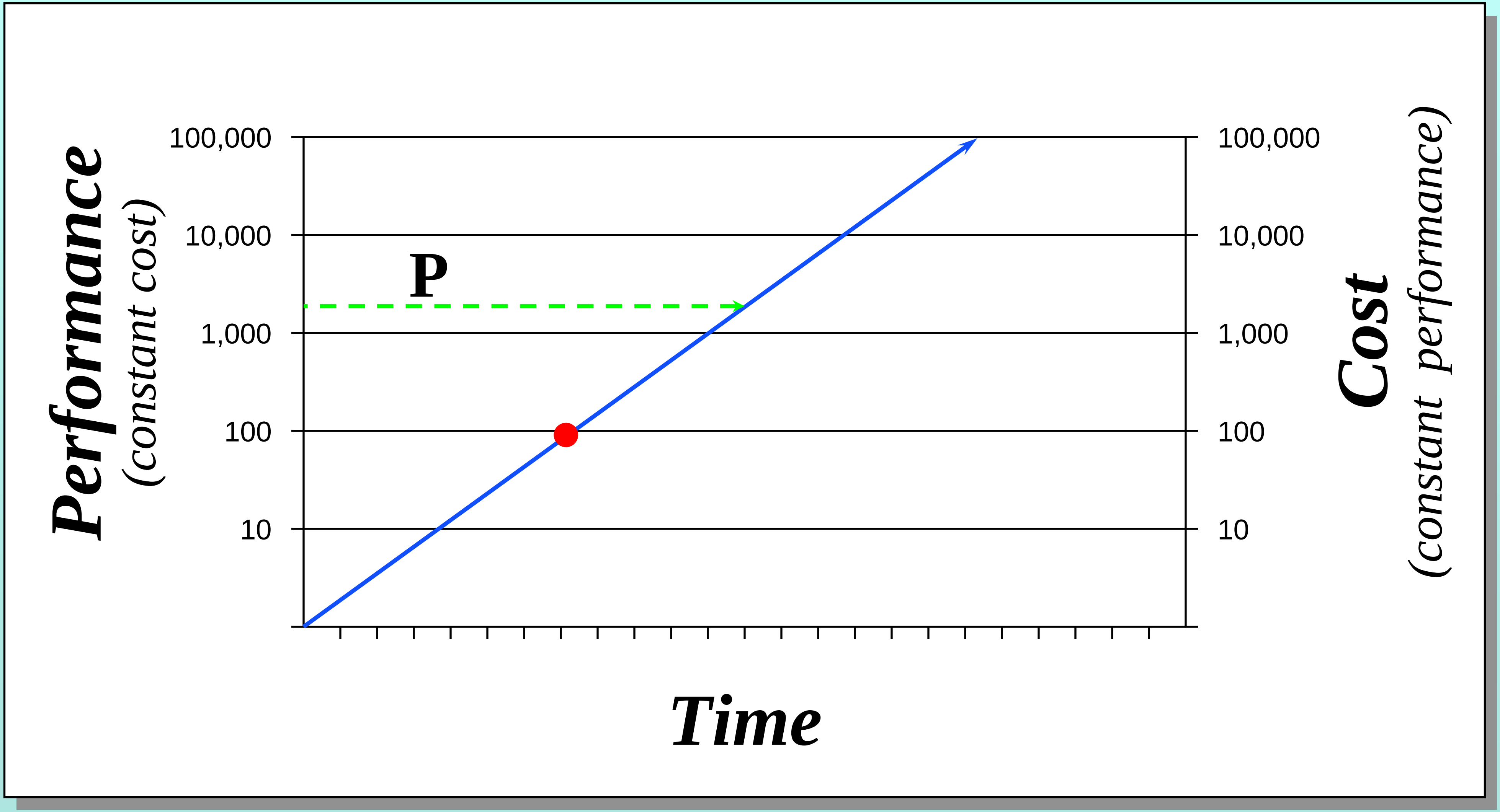
Moore's Law: *The Effect*

Three Phases of Novel IT Applications

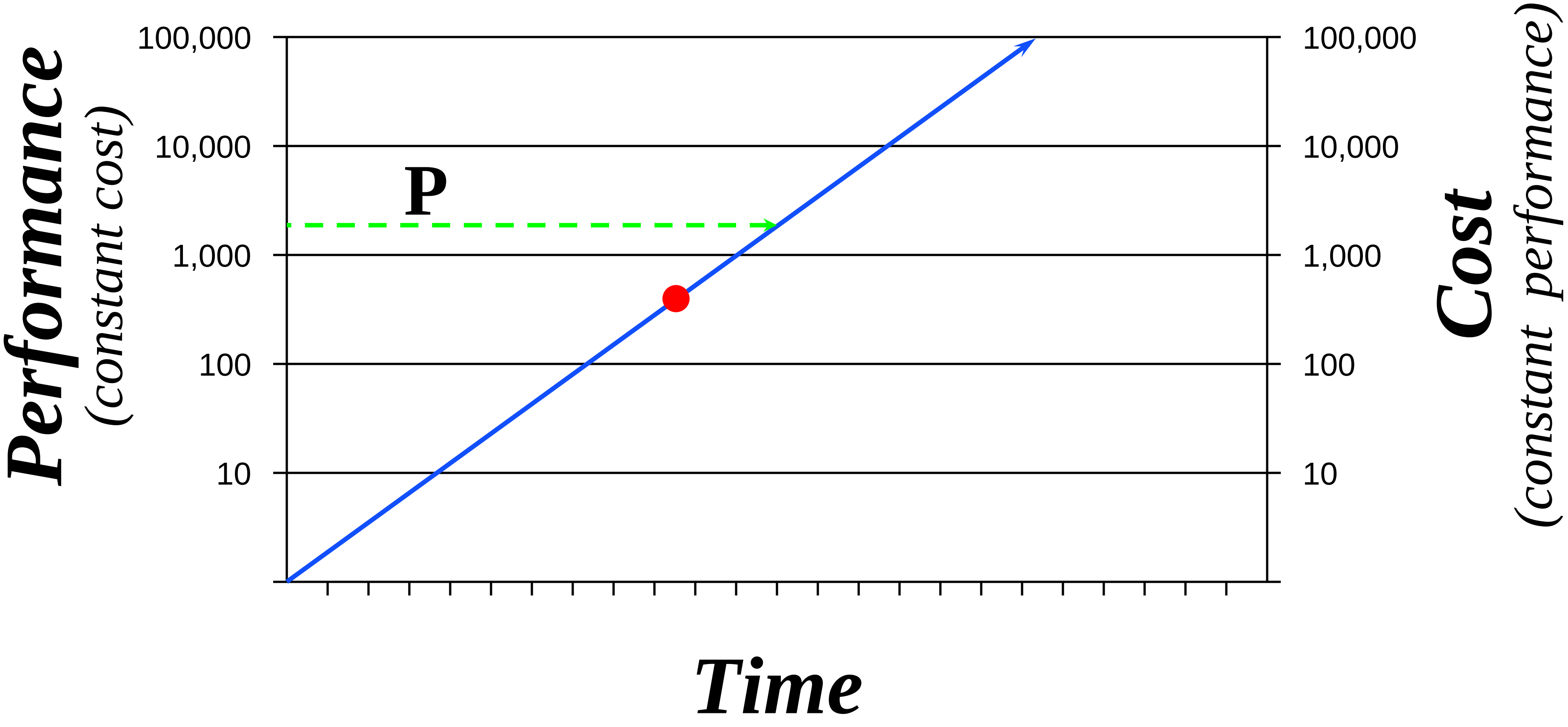
- It's Impossible
- It's Impractical
- It's Overdue

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

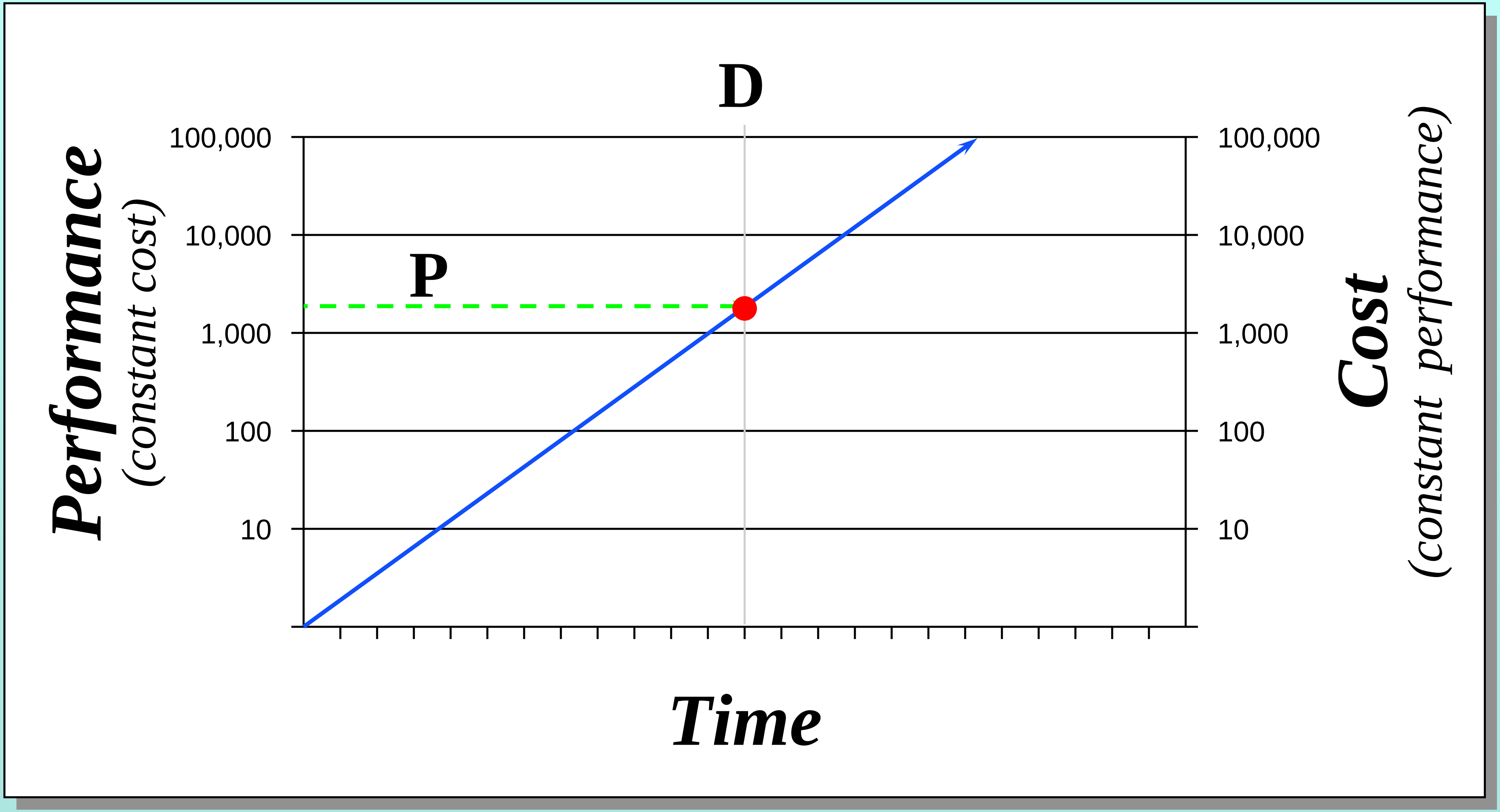
Moore's Law: *The Effect*



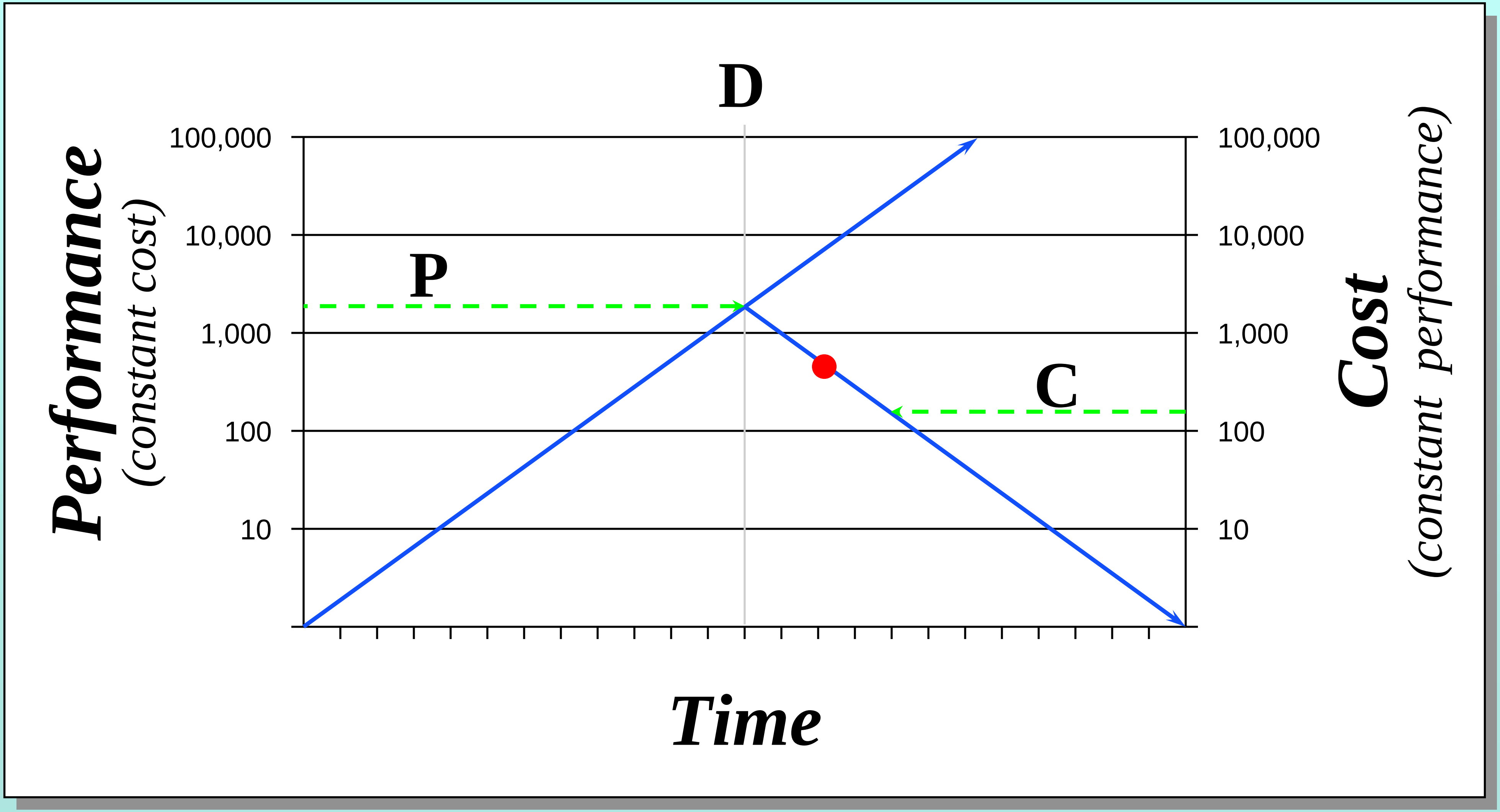
Moore’s Law: *The Effect*



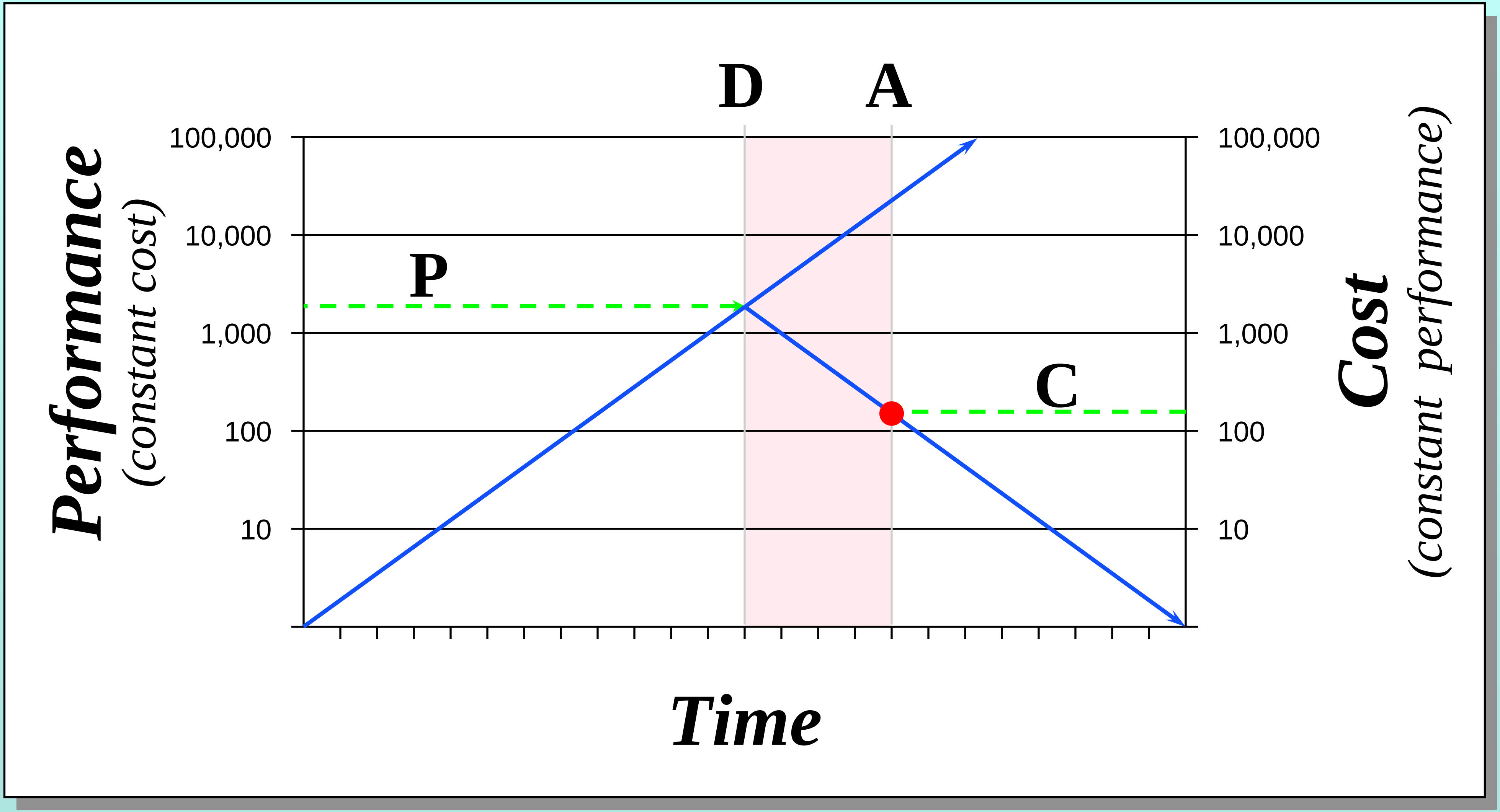
Moore’s Law: *The Effect*



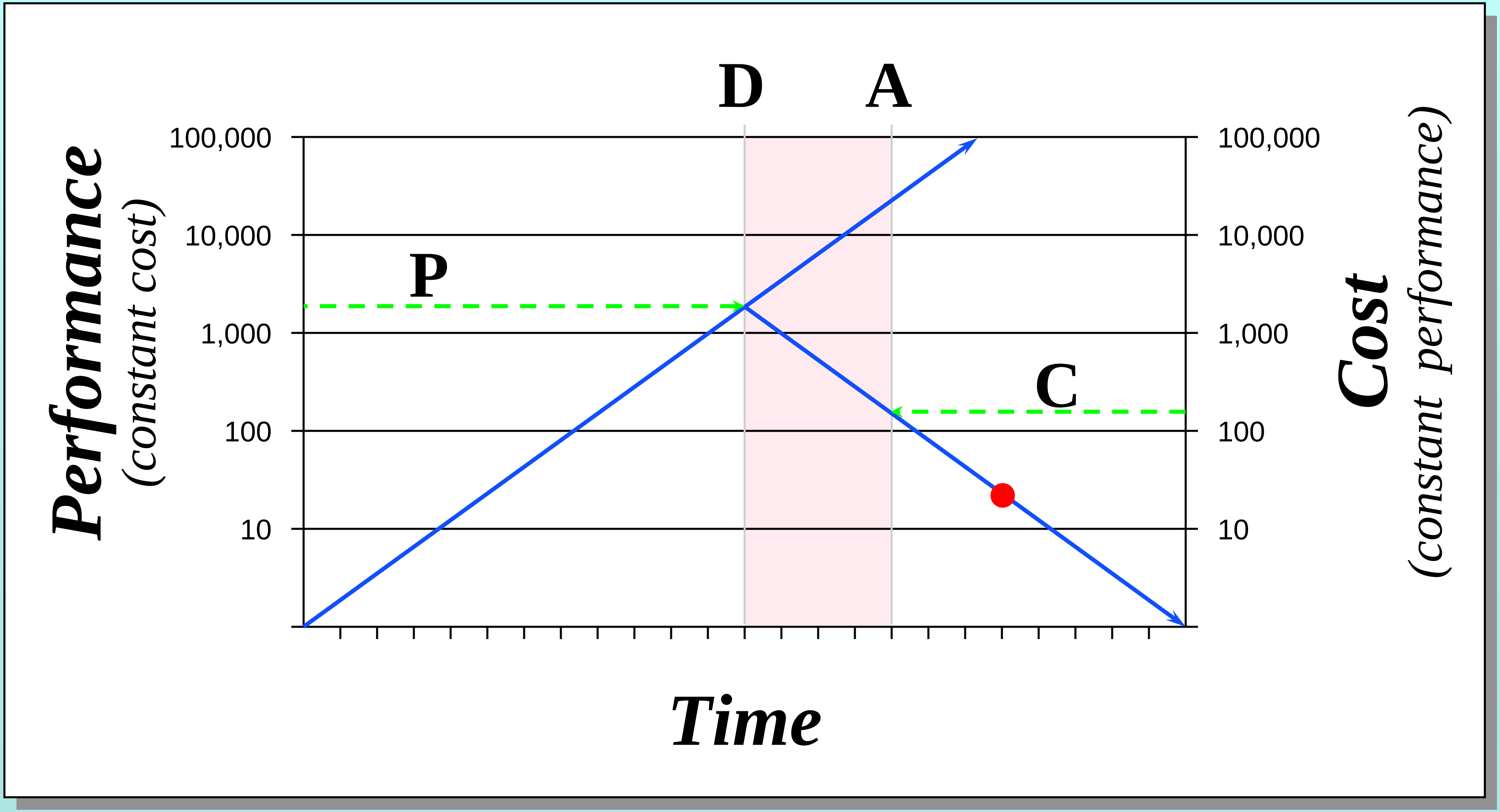
Moore's Law: *The Effect*



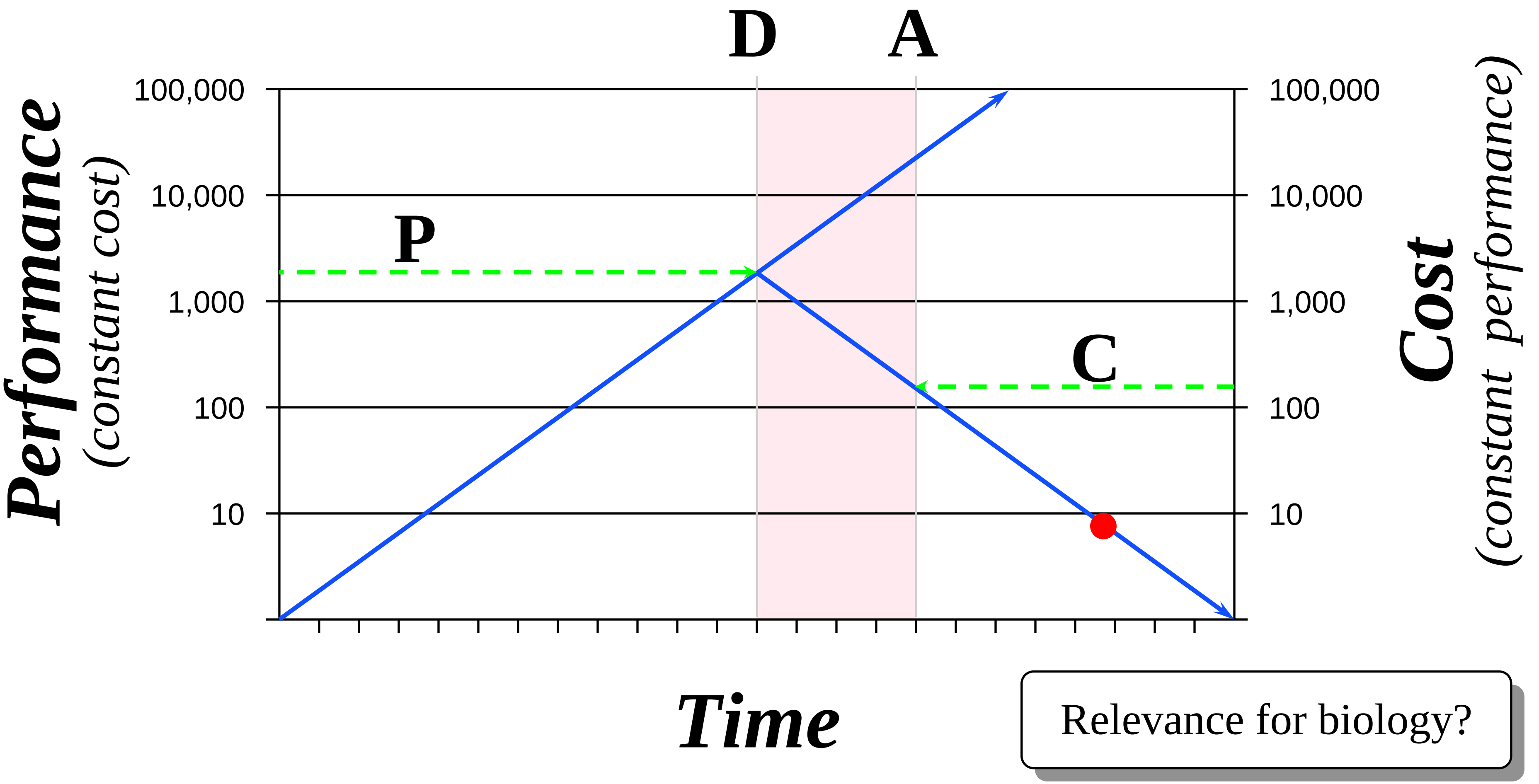
Moore’s Law: *The Effect*



Moore's Law: *The Effect*

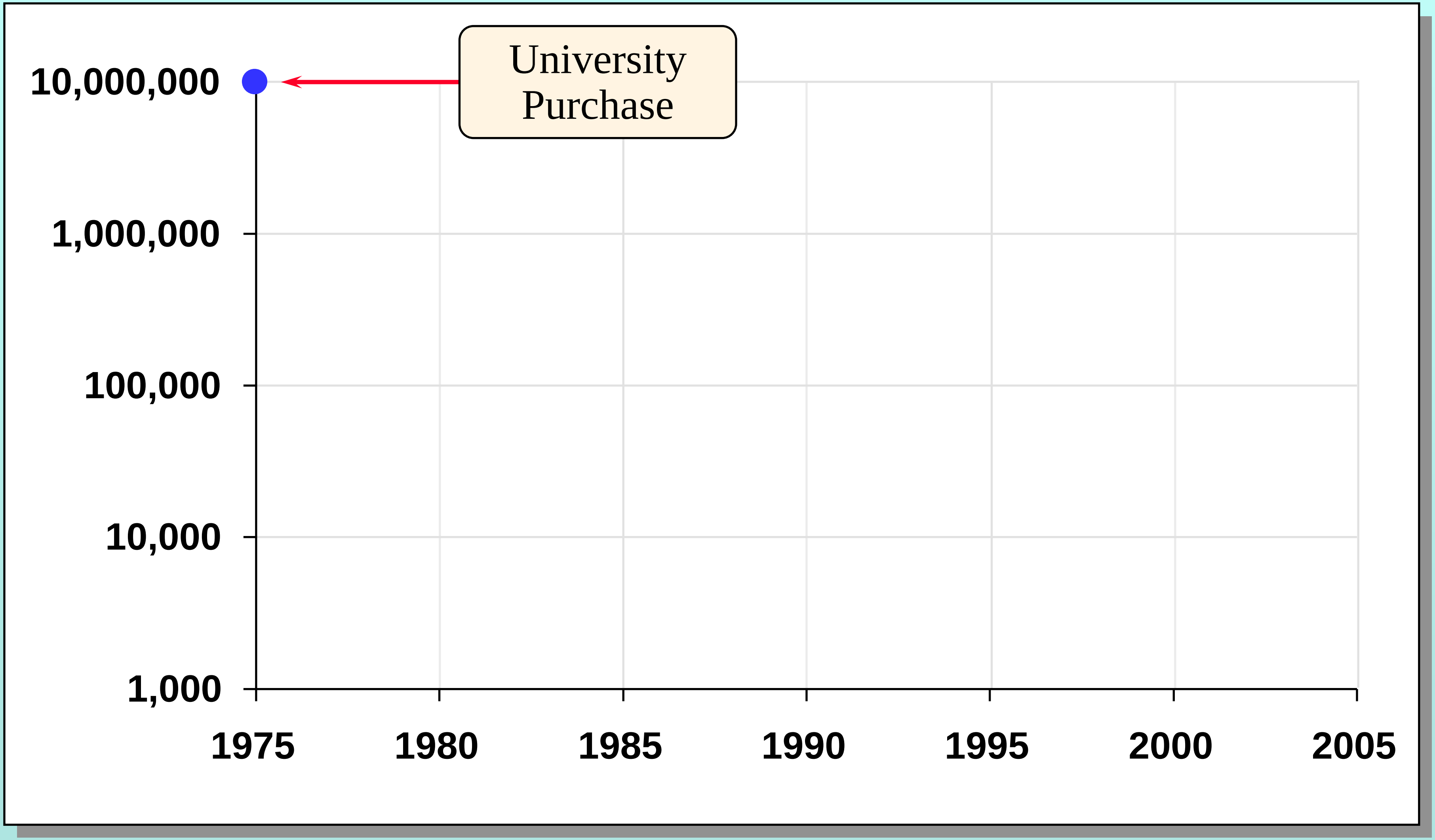


Moore's Law: *The Effect*

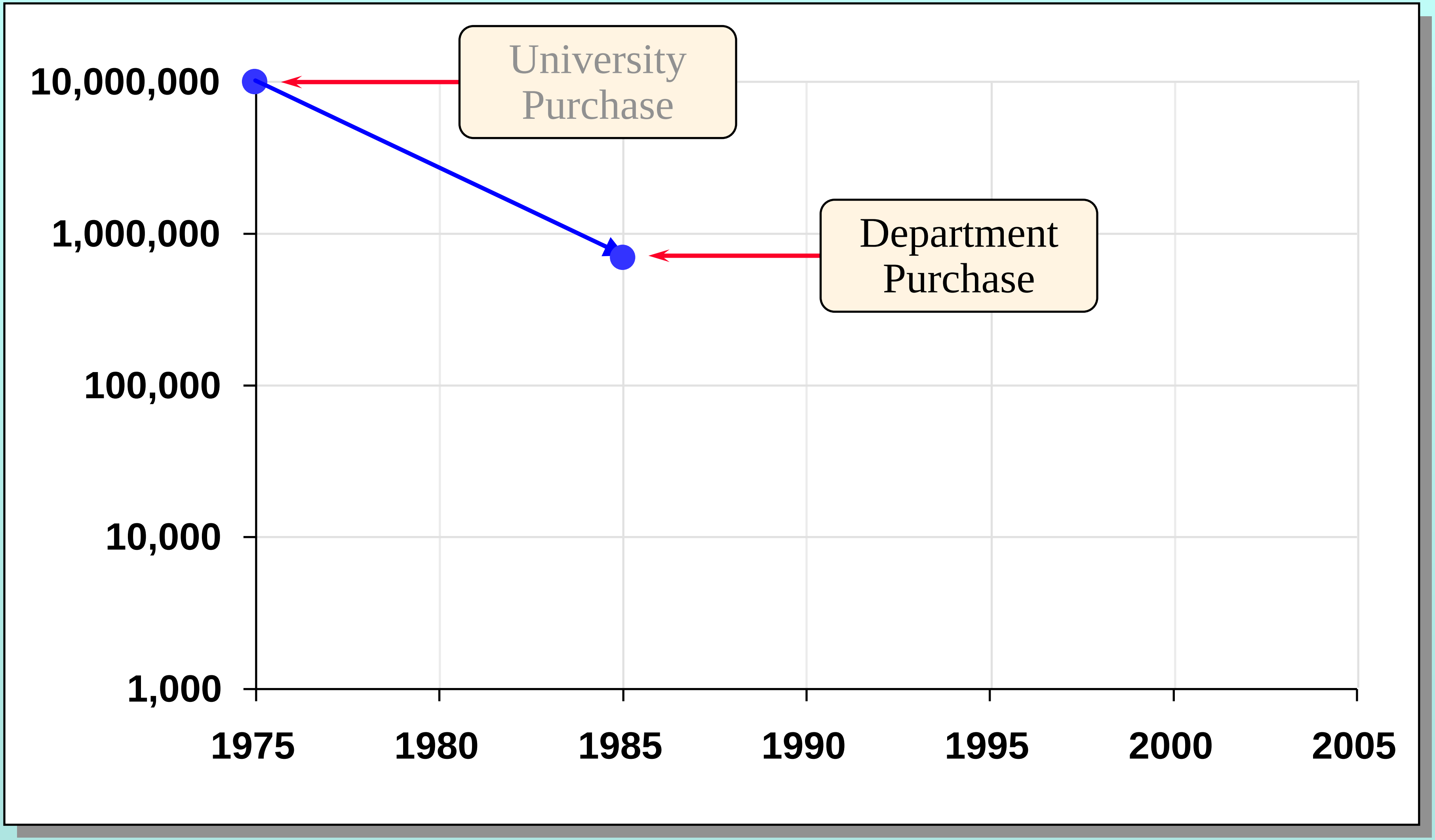


Relevance for biology?

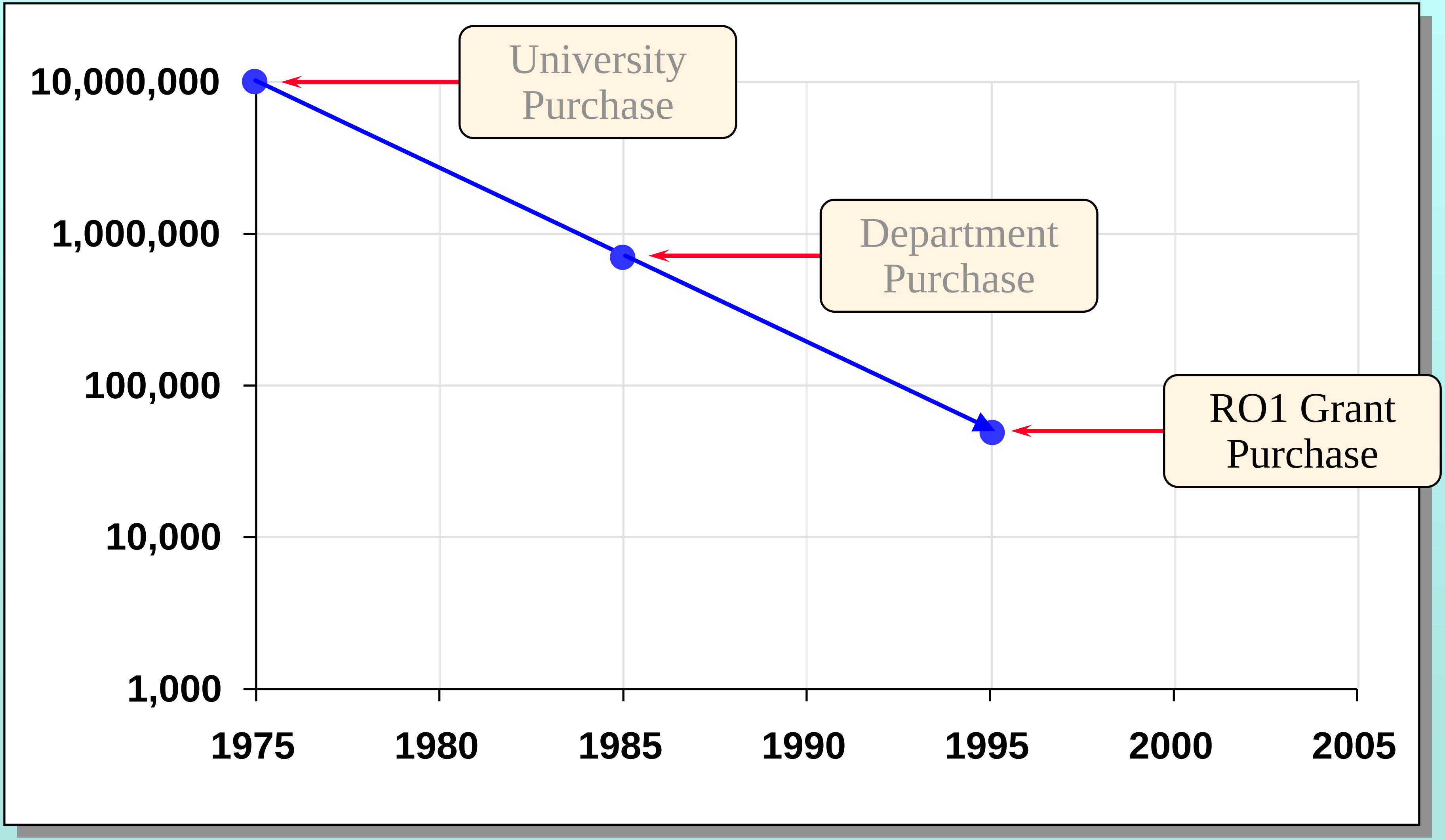
Cost (constant performance)



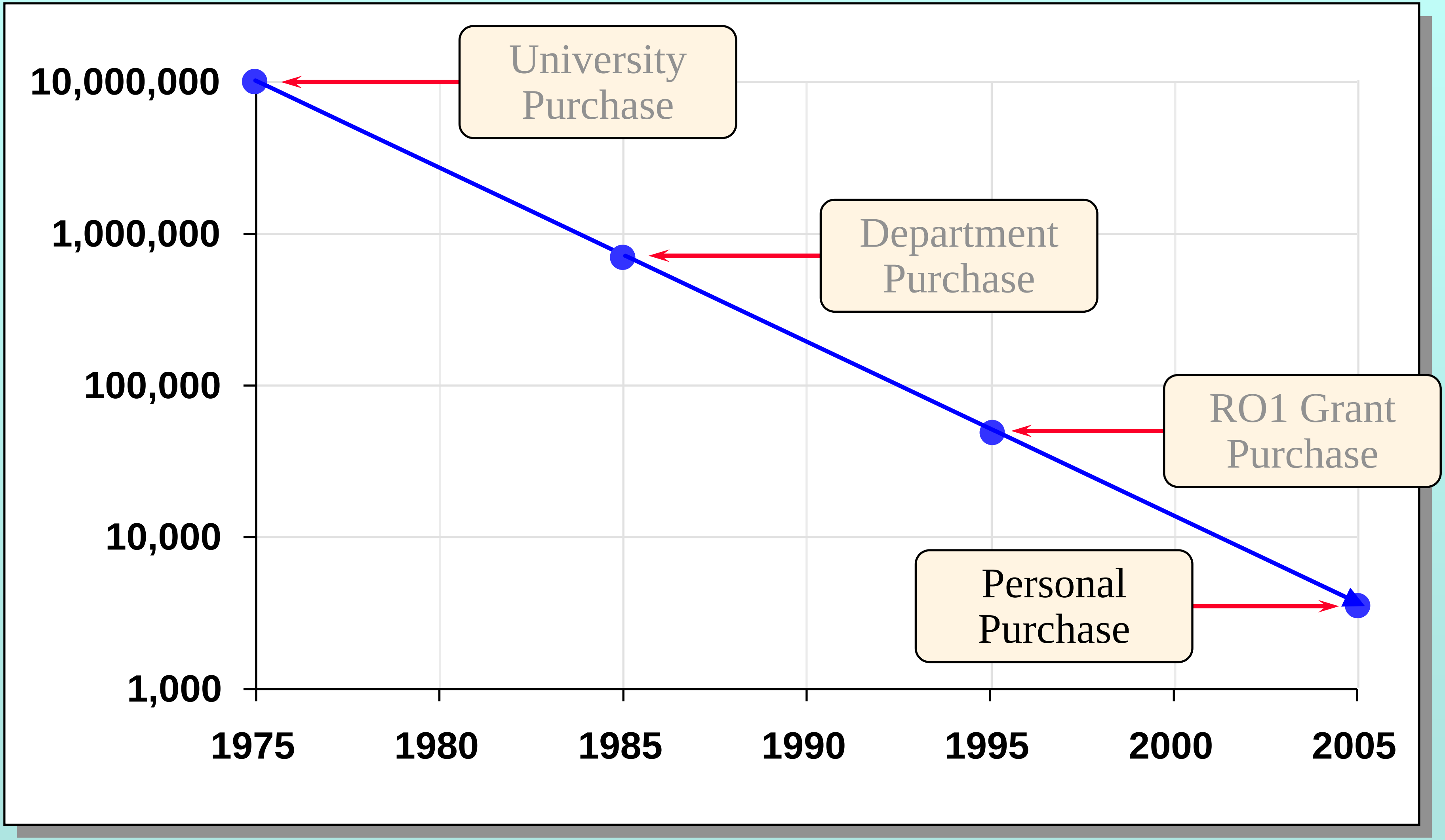
Cost (constant performance)



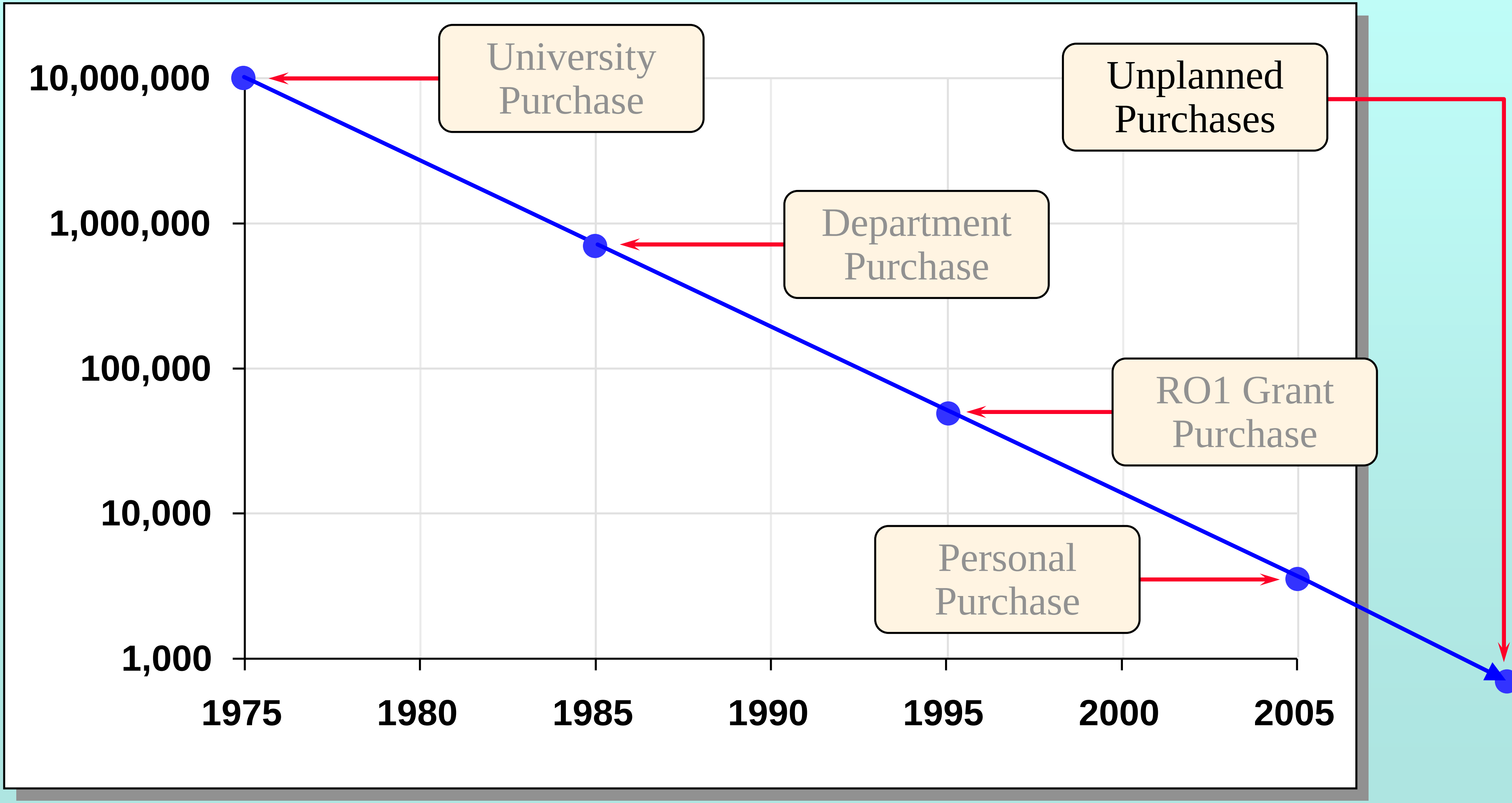
Cost (constant performance)



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.*

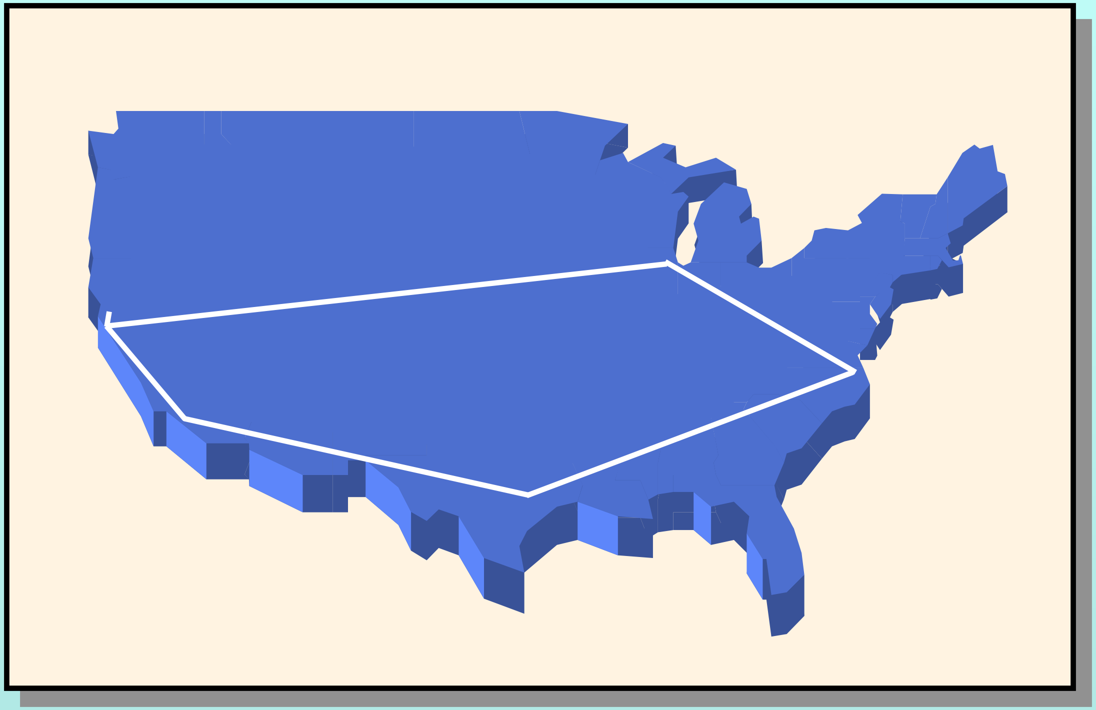
The Digital Basis of Life

[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

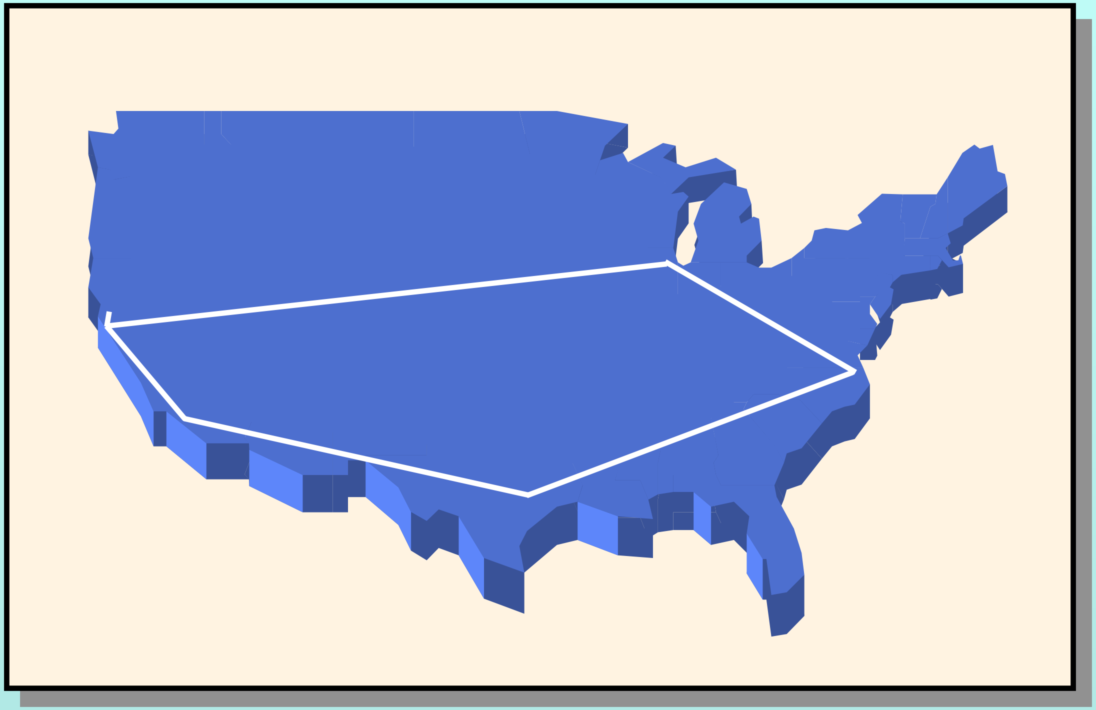
We now know that Schrödinger's mysterious human "code-script" consists of 3.3 billion base pairs of DNA.



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.
- 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.

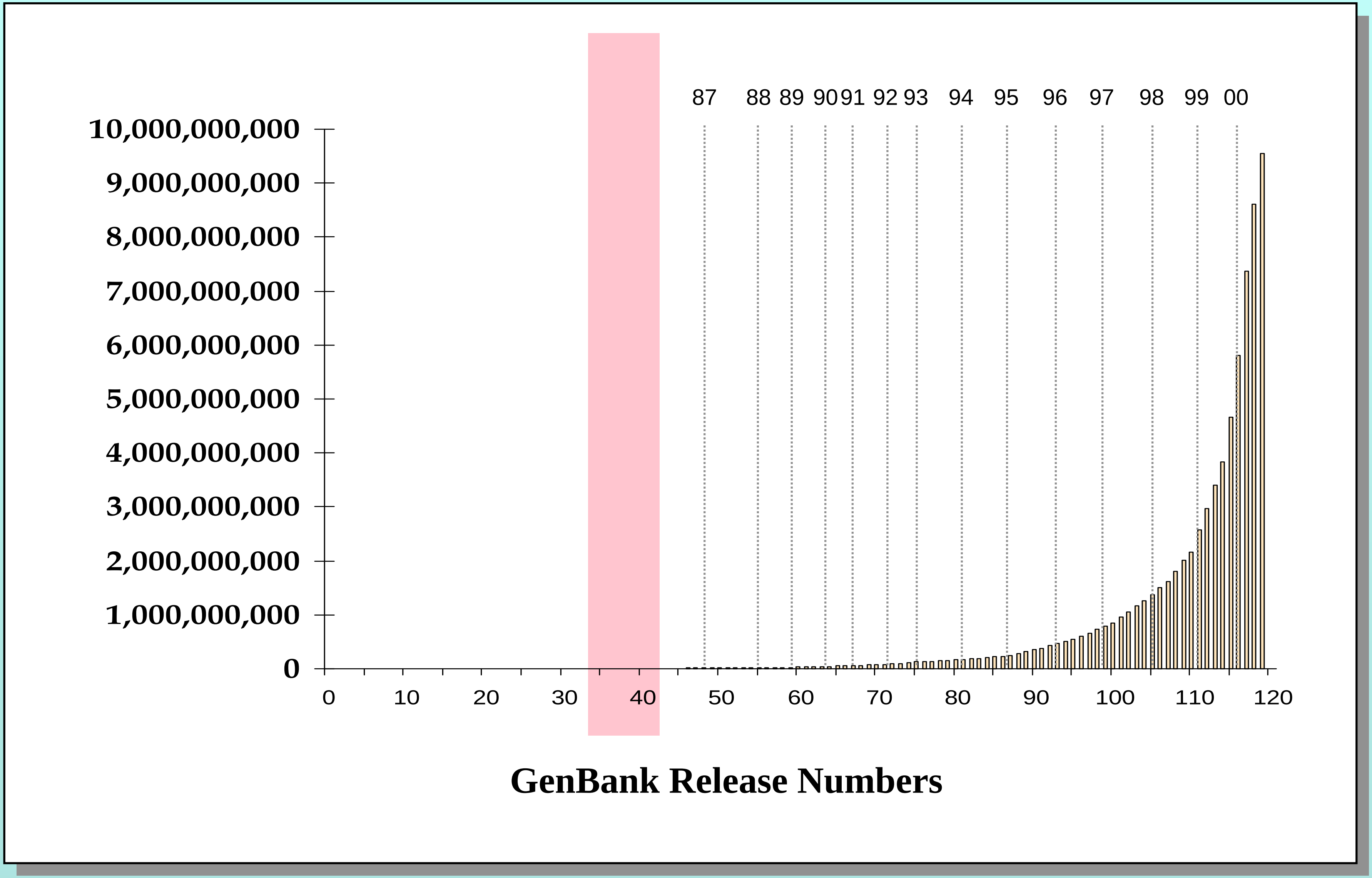
Genomics: An Example

Infrastructure and the HGP

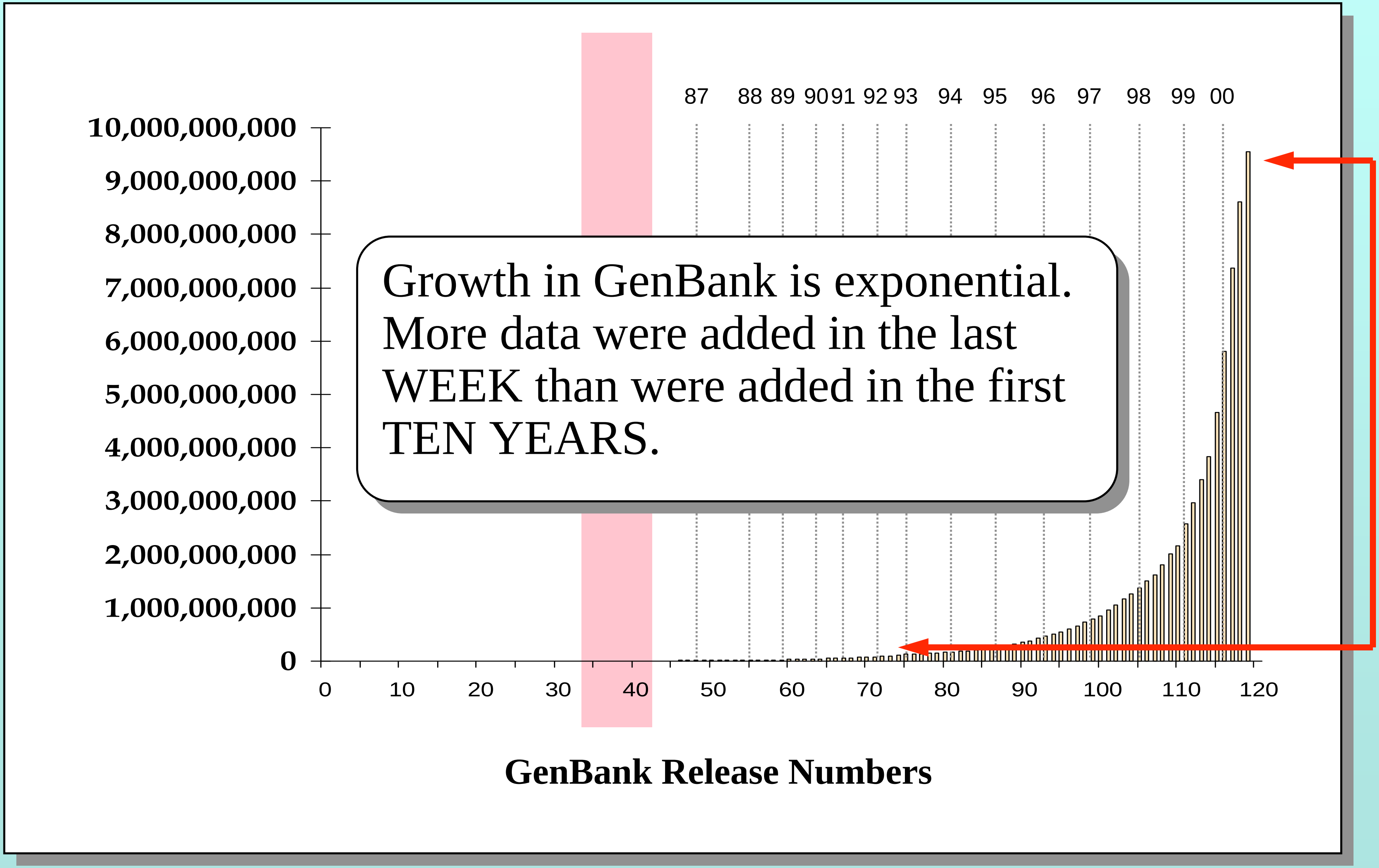
Progress towards all of the [Genome Project] goals will require the establishment of well-funded centralized facilities, including a stock center for the cloned DNA fragments generated in the mapping and sequencing effort and a data center for the computer-based collection and distribution of large amounts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

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

Base Pairs in GenBank

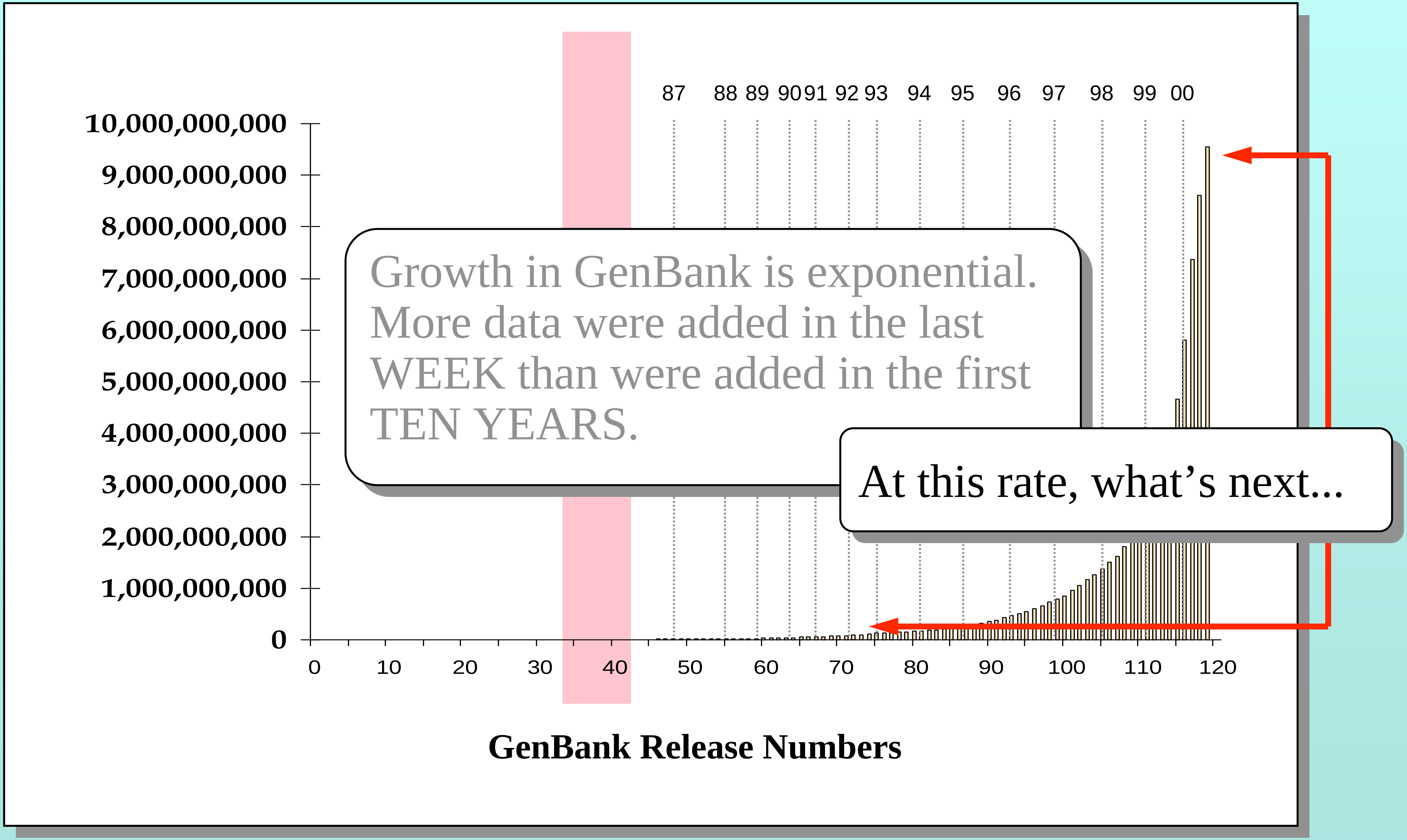


Base Pairs in GenBank



Growth in GenBank is exponential. More data were added in the last WEEK than were added in the first TEN YEARS.

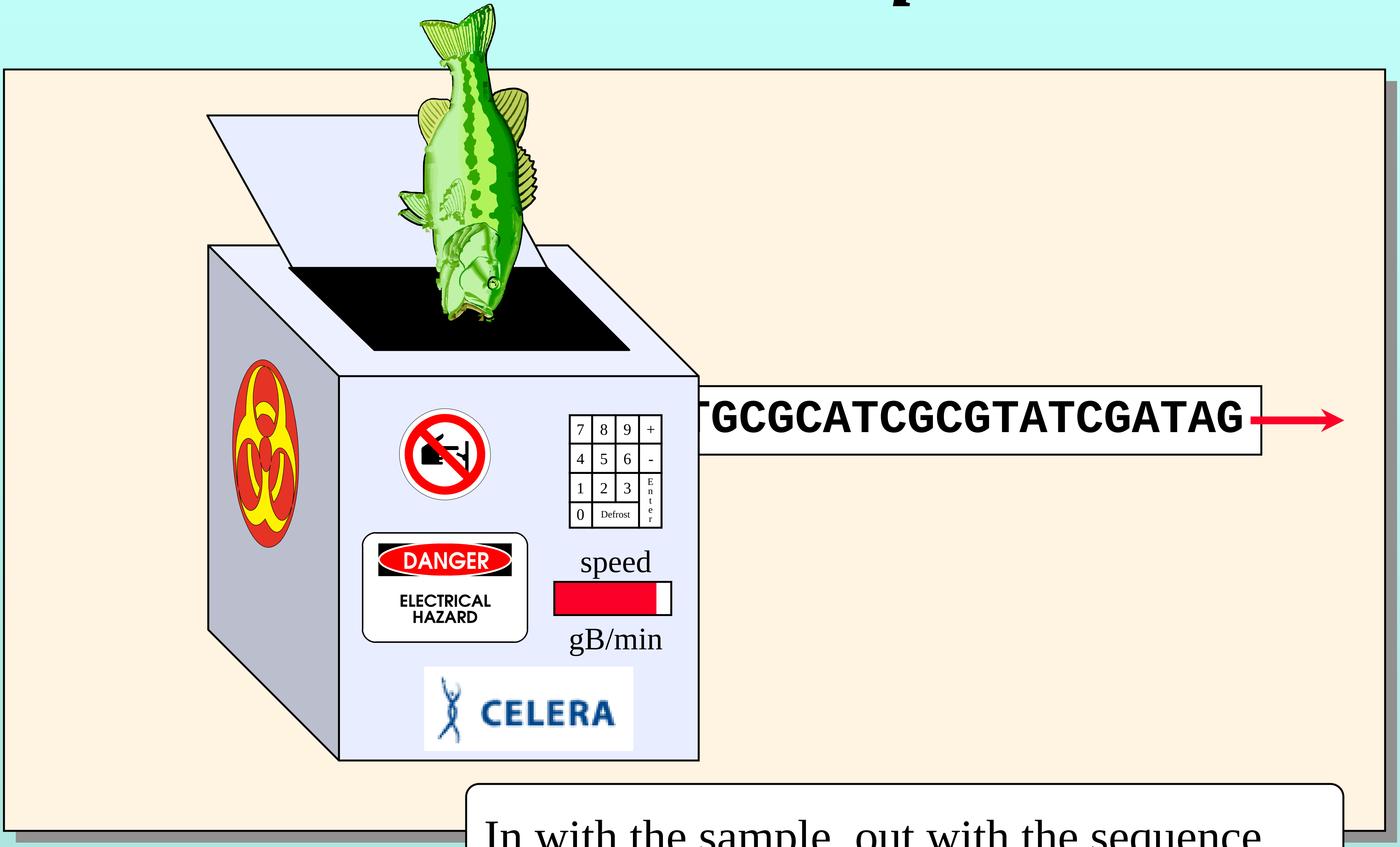
Base Pairs in GenBank



Growth in GenBank is exponential. More data were added in the last WEEK than were added in the first TEN YEARS.

At this rate, what’s next...

Celera *Bass-o-Matic Sequencer*



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...

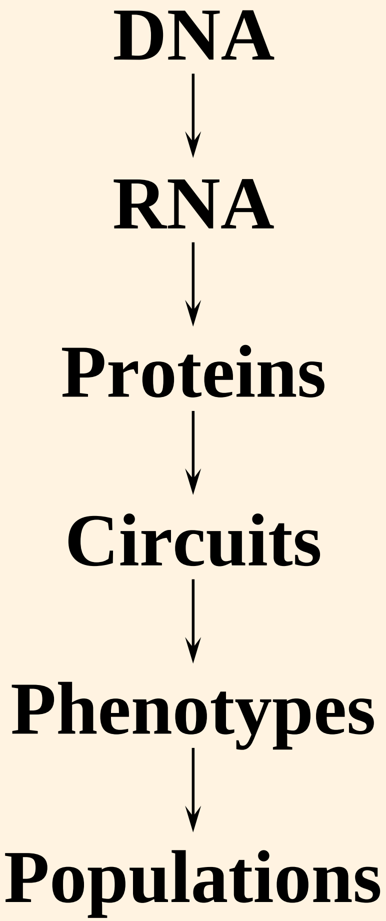
21st Century Biology

The Science

Fundamental Dogma

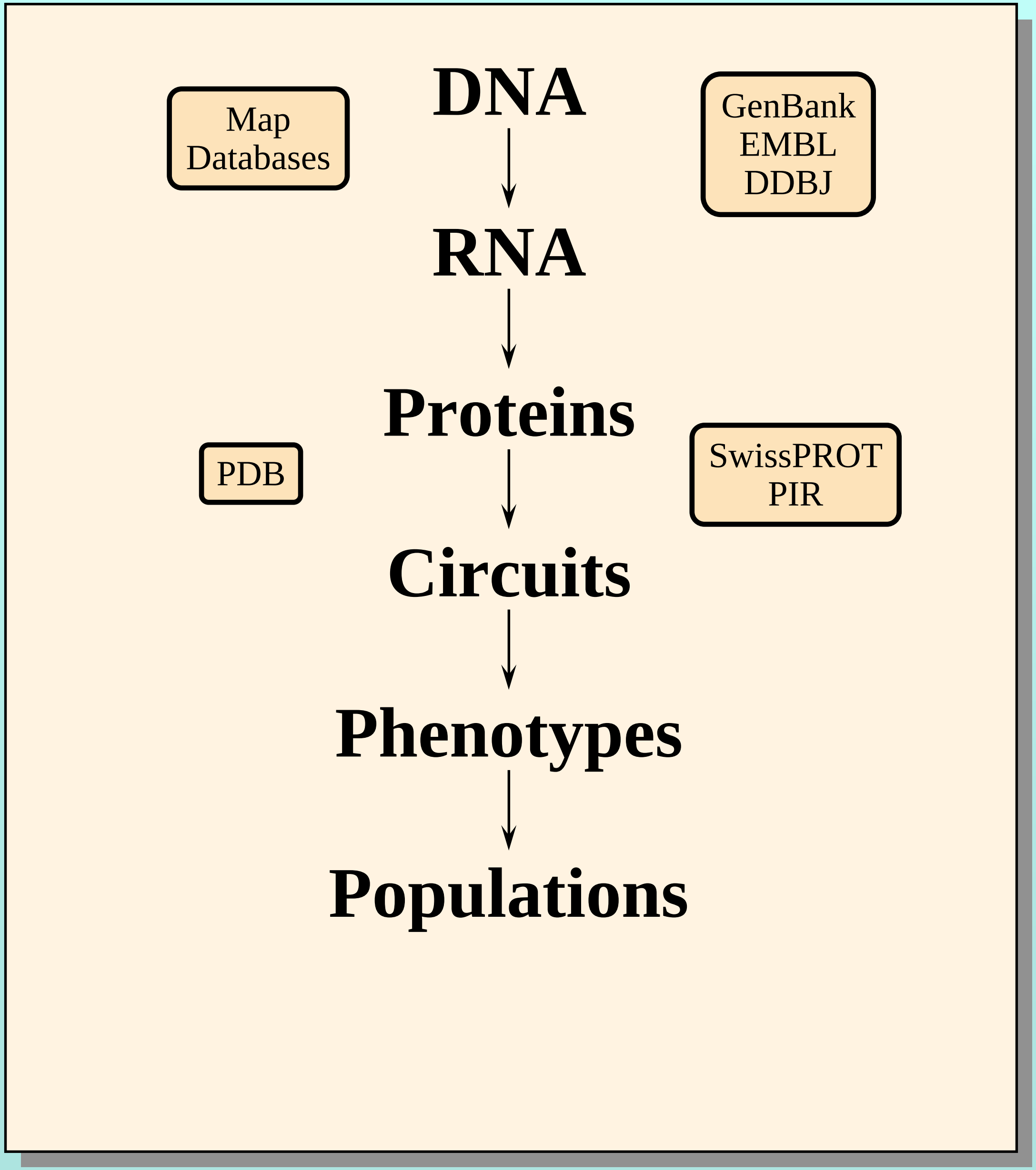
The fundamental dogma of molecular biology is that genes act to create phenotypes through a flow of information from DNA to RNA to proteins, to interactions among proteins (regulatory circuits and metabolic pathways), and ultimately to phenotypes.

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



Fundamental Dogma

Although a few databases already exist to distribute molecular information,

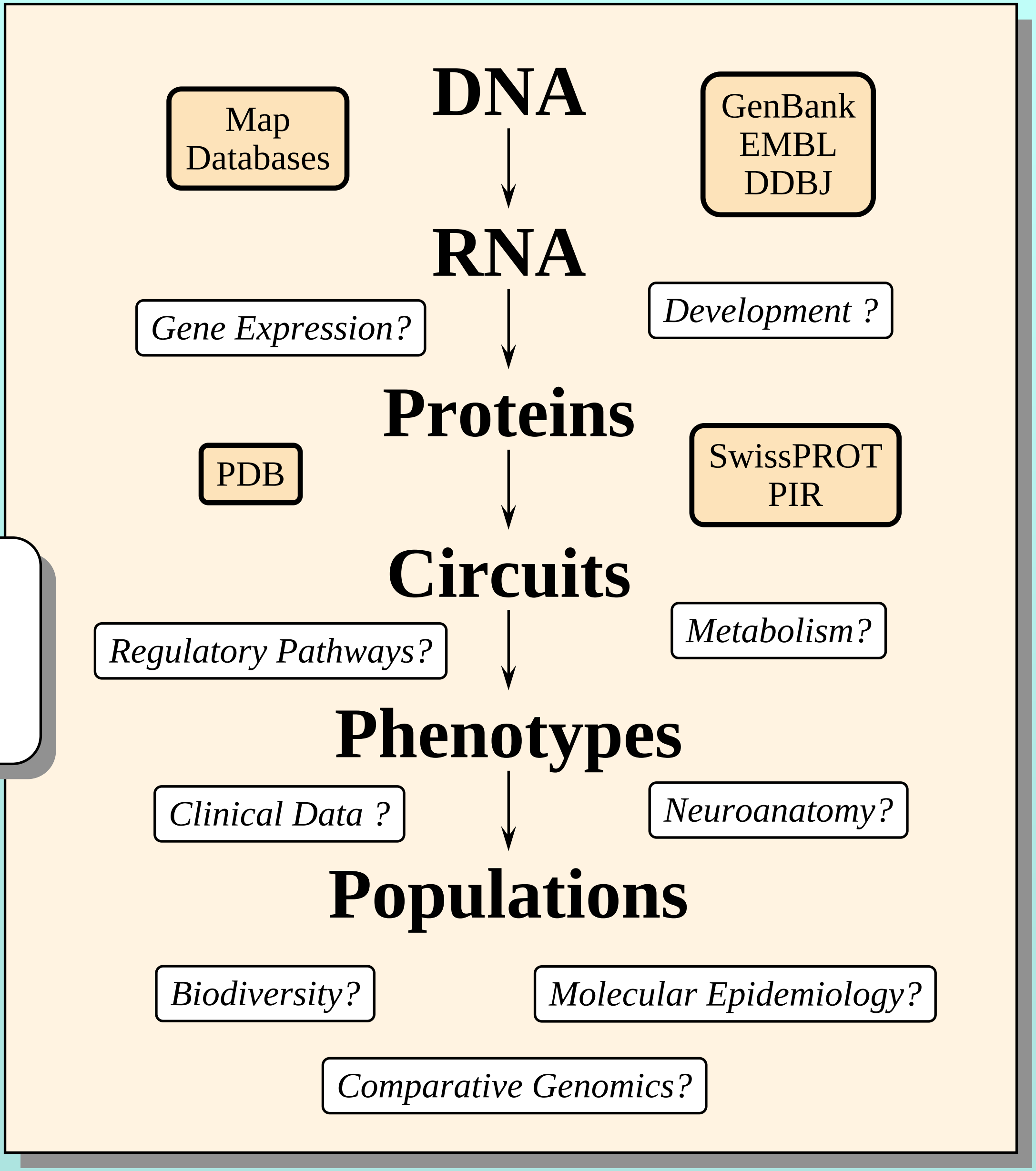


© 2001, Robert J. Robbins

Fundamental Dogma

Although a few databases already exist to distribute molecular information,

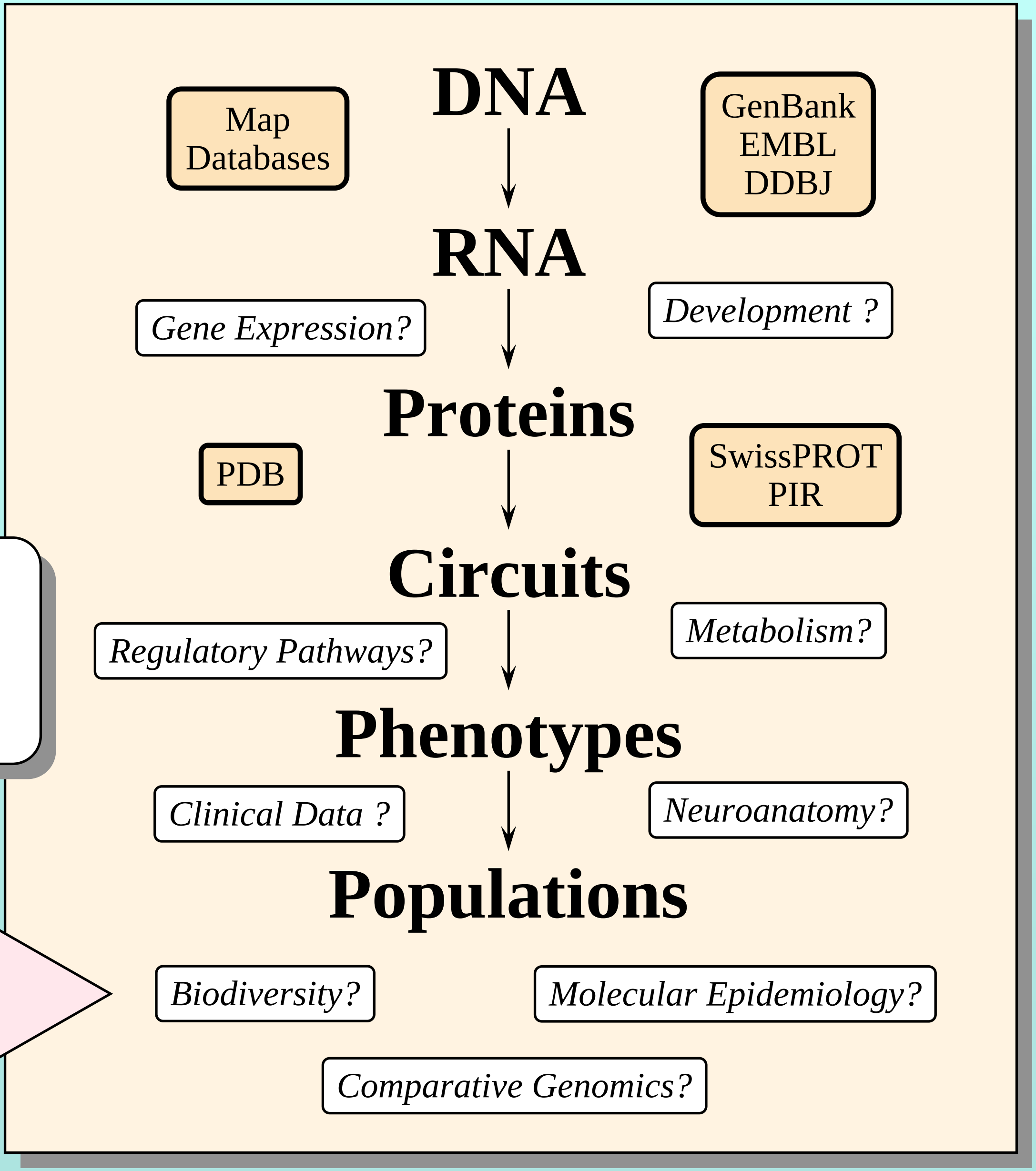
the post-genomic era will need many more to collect, manage, and publish the coming flood of new findings.



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

Post-Genome Era

The Post-Genome Era

Post-genome research involves:

- applying genomic tools and knowledge to more general problems

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

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.

Documenting Global Biodiversity

Documenting Biodiversity

Documenting and comprehending global biodiversity will require access to global data sets on:

- species diversity

Documenting Biodiversity

Documenting and comprehending global biodiversity will require access to global data sets on:

- species diversity
- species distribution and density

Documenting Biodiversity

Documenting and comprehending global biodiversity will require access to global data sets on:

- species diversity
- species distribution and density
- environmental parameters

Documenting Biodiversity

Documenting and comprehending global biodiversity will require access to global data sets on:

- species diversity
- species distribution and density
- environmental parameters
- times series records of biological and environmental data

Documenting Biodiversity

Documenting and comprehending global biodiversity will require access to global data sets on:

- species diversity
- species distribution and density
- environmental parameters
- times series records of biological and environmental data
- genetic diversity within species

Documenting Biodiversity

Documenting and comprehending global biodiversity will require access to global data sets on:

- species diversity
- species distribution and density
- environmental parameters
- times series records of biological and environmental data
- genetic diversity within species
- individual differences in gene expression

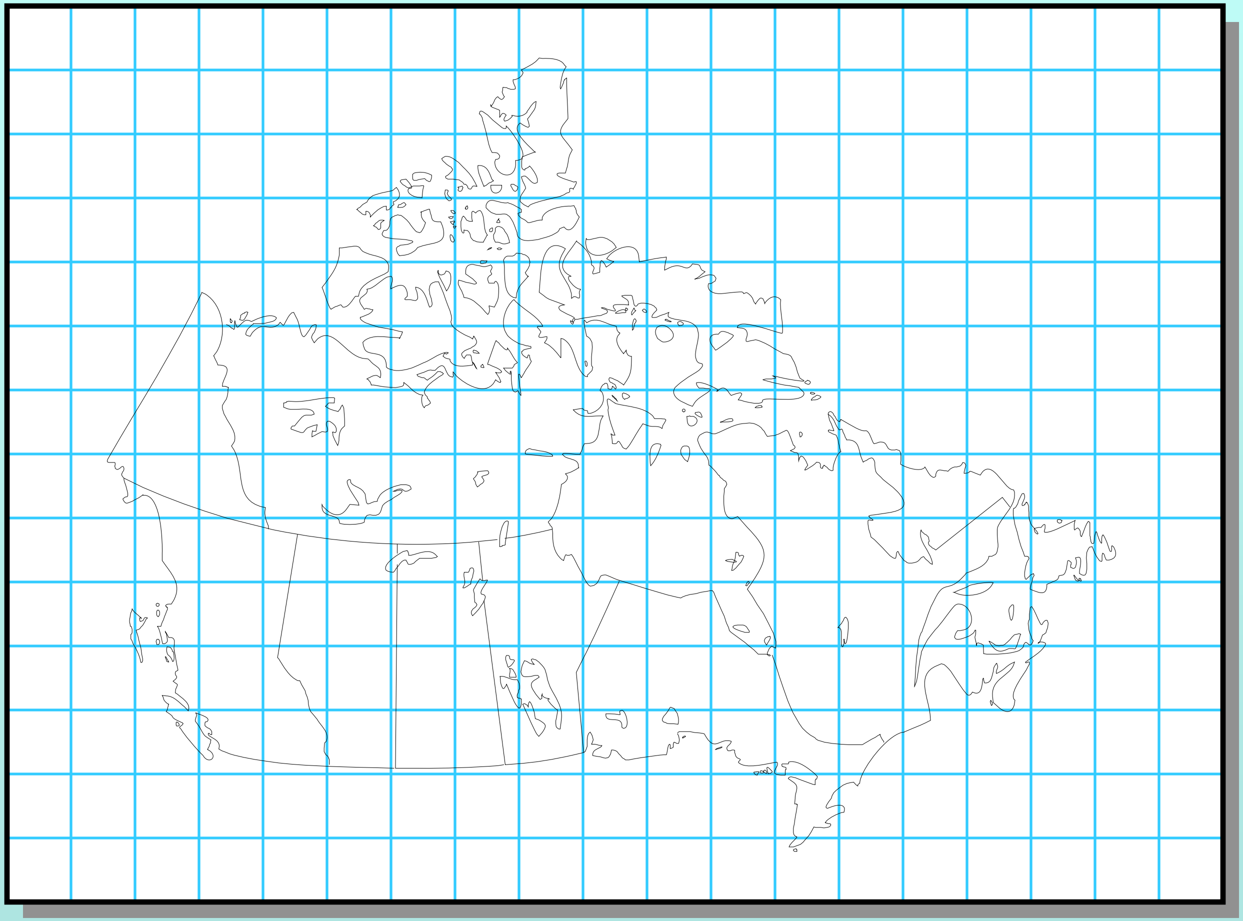
Documenting Biodiversity

At an elementary level, documenting biodiversity involves tracking species presence/absence per unit of area.

Documenting Biodiversity

At an elementary level, documenting biodiversity involves tracking species presence/absence per unit of area.

But at what resolution?

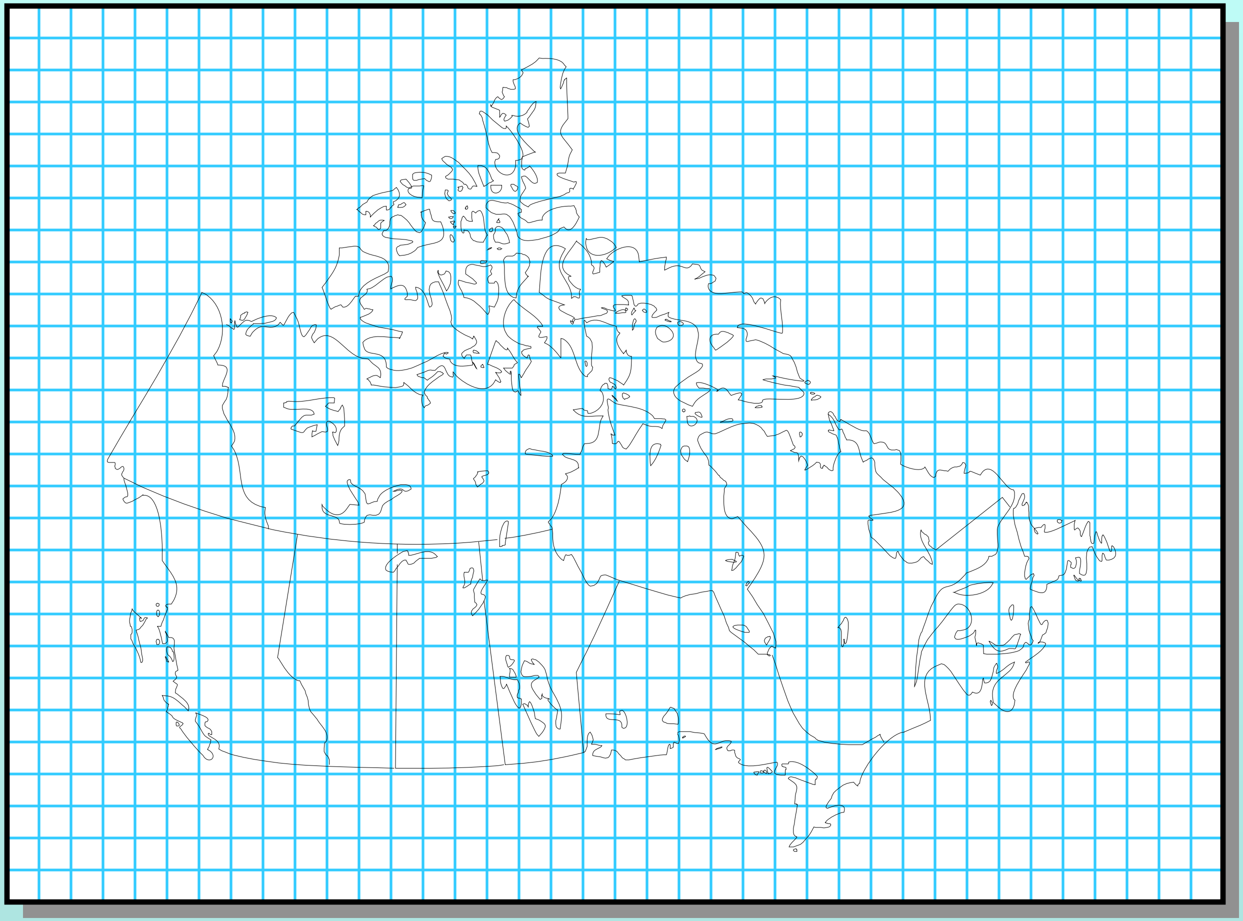


Documenting Biodiversity

At an elementary level, documenting biodiversity involves tracking species presence/absence per unit of area.

But at what resolution?

With every increase in resolution, the data set grows exponentially...

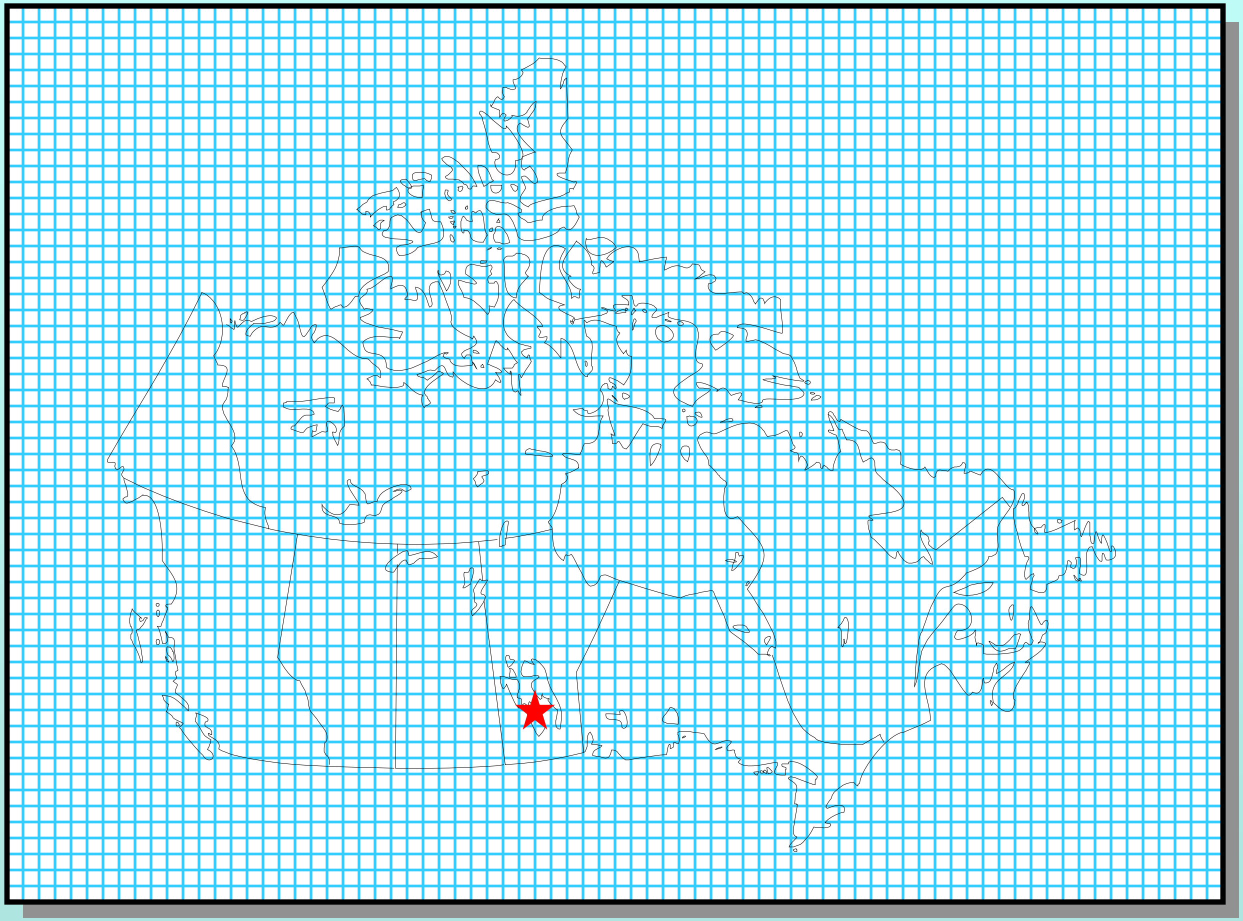


Documenting Biodiversity

At an elementary level, documenting biodiversity involves tracking species presence/absence per unit of area.

But at what resolution?

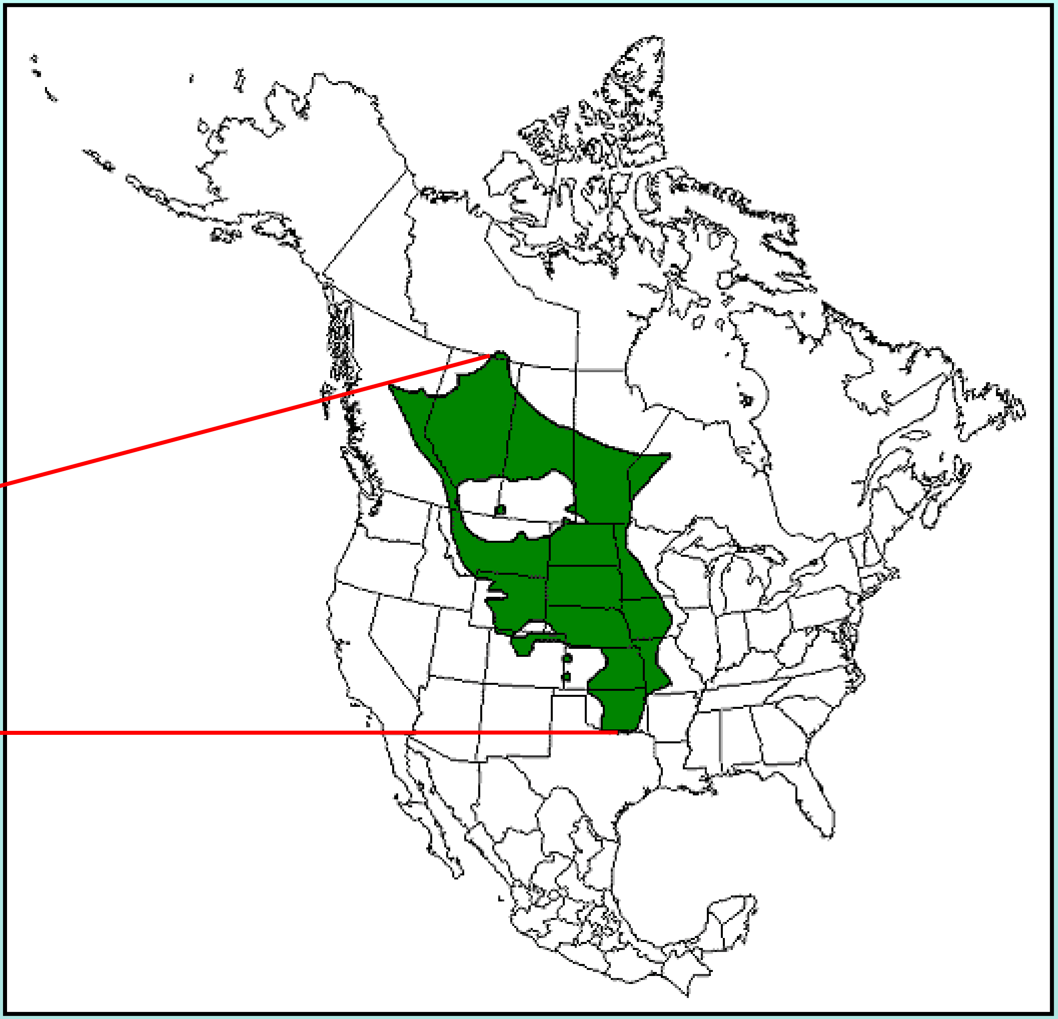
Localized, time-series probability distributions are also needed.



Documenting Biodiversity

It’s one thing to say that:

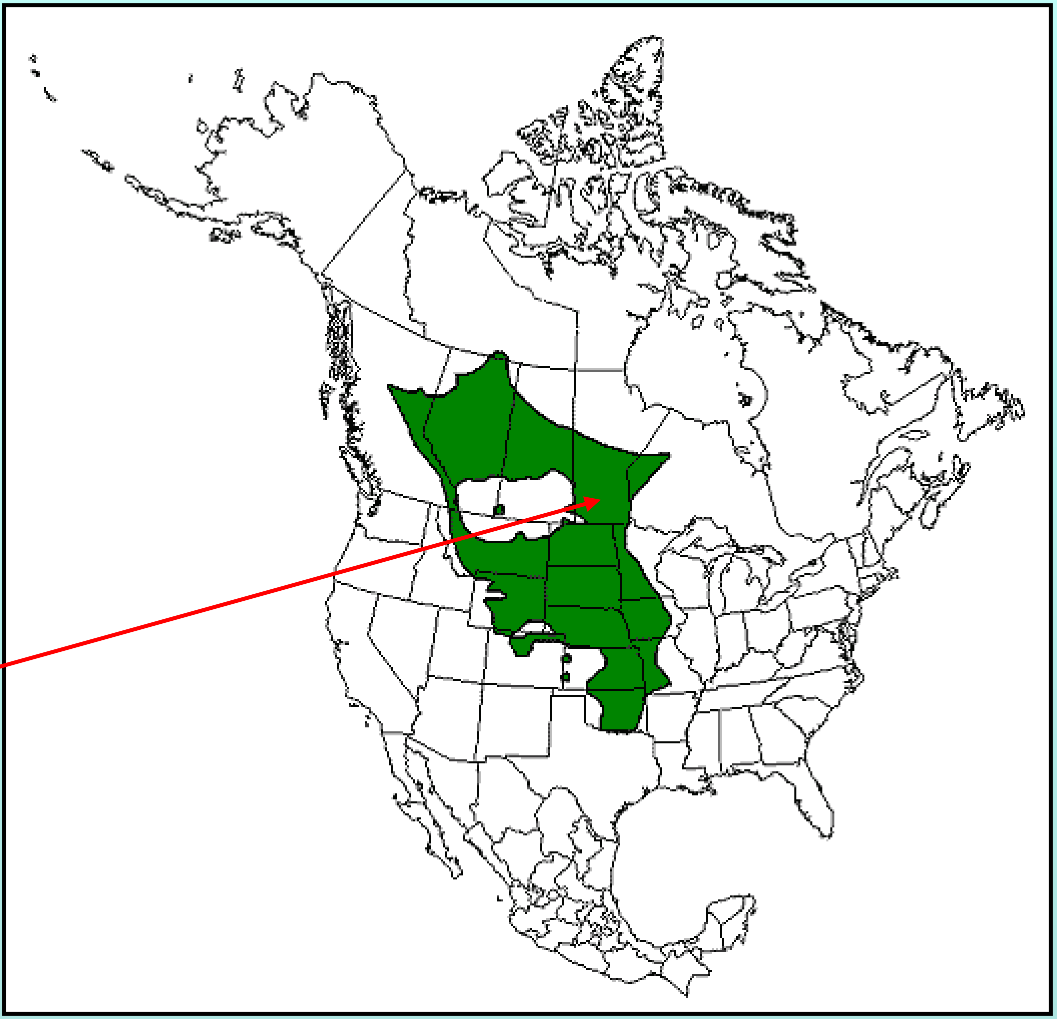
The red-sided garter snake occurs throughout central North America and is found in the southern part of Manitoba up to Flin Flon. It is absent from the extreme southwestern grasslands except for Spruce Woods Provincial Park.



Documenting Biodiversity

It’s another to note:

Every fall and spring, more than 65,000 red-sided garter snakes congregate at local over-wintering dens in the Narcisse Wildlife Management Area. This results in the most locally dense concentration of snakes in the world.



Documenting Biodiversity (km scale)

- Surface of the Earth = 10^9 km^2.

Documenting Biodiversity (km scale)

- Surface of the Earth = 10^9 km^2.
- Representing the distribution of one million species, requires a two-dimensional distribution grid with 10^{15} cells.

Documenting Biodiversity (km scale)

- Surface of the Earth = 10^9 km^2.
- Representing the distribution of one million species, requires a two-dimensional distribution grid with 10^{15} cells.
- Adding the third dimension, at one-km scale, to document the diversity in a bio-film that is, say, ten kilometers thick, will require a three-dimensional grid with 10^{16} cells.

Documenting Biodiversity (km scale)

- Surface of the Earth = 10^9 km^2.
- Representing the distribution of one million species, requires a two-dimensional distribution grid with 10^{15} cells.
- Adding the third dimension, at one-km scale, to document the diversity in a bio-film that is, say, ten kilometers thick, will require a three-dimensional grid with 10^{16} cells.
- Adding time-series probability distributions will increase the complexity substantially.

Documenting Biodiversity (km scale)

- Surface of the Earth = 10^9 km^2.
- Representing the distribution of one million species, requires a two-dimensional distribution grid with 10^{15} cells.
- Adding the third dimension, at one-km scale, to document the diversity in a bio-film that is, say, ten kilometers thick, will require a three-dimensional grid with 10^{16} cells.
- Adding time-series probability distributions will increase the complexity substantially.

But this only tells us where things are in cubic kilometers...

Documenting Biodiversity (meter scale)

- Surface of the Earth = 10^{15} m^2.

Documenting Biodiversity (meter scale)

- Surface of the Earth = 10^{15} m^2.
- Representing the distribution of one million species, requires a two-dimensional distribution grid with 10^{21} cells.

Documenting Biodiversity (meter scale)

- Surface of the Earth = 10^{15} m^2.
- Representing the distribution of one million species, requires a two-dimensional distribution grid with 10^{21} cells.
- Adding the third dimension, at one-meter scale, to document the diversity in a bio-film that is, say, ten kilometers thick, will require a three-dimensional grid with 10^{26} cells.

Documenting Biodiversity (meter scale)

- Surface of the Earth = 10^{15} m^2.
- Representing the distribution of one million species, requires a two-dimensional distribution grid with 10^{21} cells.
- Adding the third dimension, at one-meter scale, to document the diversity in a bio-film that is, say, ten kilometers thick, will require a three-dimensional grid with 10^{26} cells.
- Adding time-series probability distributions will increase the complexity substantially.

Documenting Biodiversity (meter scale)

- Surface of the Earth = 10^{15} m^2.
- Representing the distribution of one million species, requires a two-dimensional distribution grid with 10^{21} cells.
- Adding the third dimension, at one-meter scale, to document the diversity in a bio-film that is, say, ten kilometers thick, will require a three-dimensional grid with 10^{26} cells.
- Adding time-series probability distributions will increase the complexity substantially.

This only tracks species diversity, not genetic diversity...

Documenting Genetic Biodiversity

- What is a reasonable estimate of the per-subject cost to document the economically significant genetic biodiversity in a single population of a single species? Let’s assume an island population with 250,000 individuals.

Documenting Genetic Biodiversity

- What is a reasonable estimate of the per-subject cost to document the economically significant genetic biodiversity in a single population of a single species? Let’s assume an island population with 250,000 individuals.

$1.00 per individual ?

Documenting Genetic Biodiversity

- What is a reasonable estimate of the per-subject cost to document the economically significant genetic biodiversity in a single population of a single species? Let's assume an island population with 250,000 individuals.

  $1.00 per individual ?

  $10.00 per individual ?

Documenting Genetic Biodiversity

- What is a reasonable estimate of the per-subject cost to document the economically significant genetic biodiversity in a single population of a single species? Let’s assume an island population with 250,000 individuals.

$1.00 per individual ?

$10.00 per individual ?

$100.00 per individual ?

Documenting Genetic Biodiversity

- What is a reasonable estimate of the per-subject cost to document the economically significant genetic biodiversity in a single population of a single species? Let’s assume an island population with 250,000 individuals.

  $1.00 per individual ?

  $10.00 per individual ?

  $100.00 per individual ?

- How about $2,000 per individual, for a total of $500,000,000?

Documenting Genetic Biodiversity

- What is a reasonable estimate of the per-subject cost to document the economically significant genetic biodiversity in a single population of a single species? Let’s assume an island population with 250,000 individuals.

  $1.00 per individual ?

  $10.00 per individual ?

  $100.00 per individual ?

- How about $2,000 per individual, for a total of $500,000,000?

That’s the likely budget for DeCODE Genetics’ efforts to characterize the human population of Iceland.

Documenting Biodiversity

Documenting and comprehending biospheric diversity on a global scale will be one of the greatest data-management challenges of all time.

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 %

Slides:

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