Information Technology and 21st Century Biology: Bioinformatics and Beyond

Robert J. Robbins
Fred Hutchinson Cancer Research Center
1100 Fairview Avenue North, J4-300
Seattle, Washington 98109

rrobbins@fhcrc.org
(206) 667 4778

Magic

Magic

To a person from 1890, much current technology would seem like magic.

Magic

To a person from 1890, much current technology would seem like magic.

What technology of 2110 would seem magical to a person from 2000?

Magic

To a person from 1890, much current technology would seem like magic.

What technology of 2110 would seem magical to a person from 2000?

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

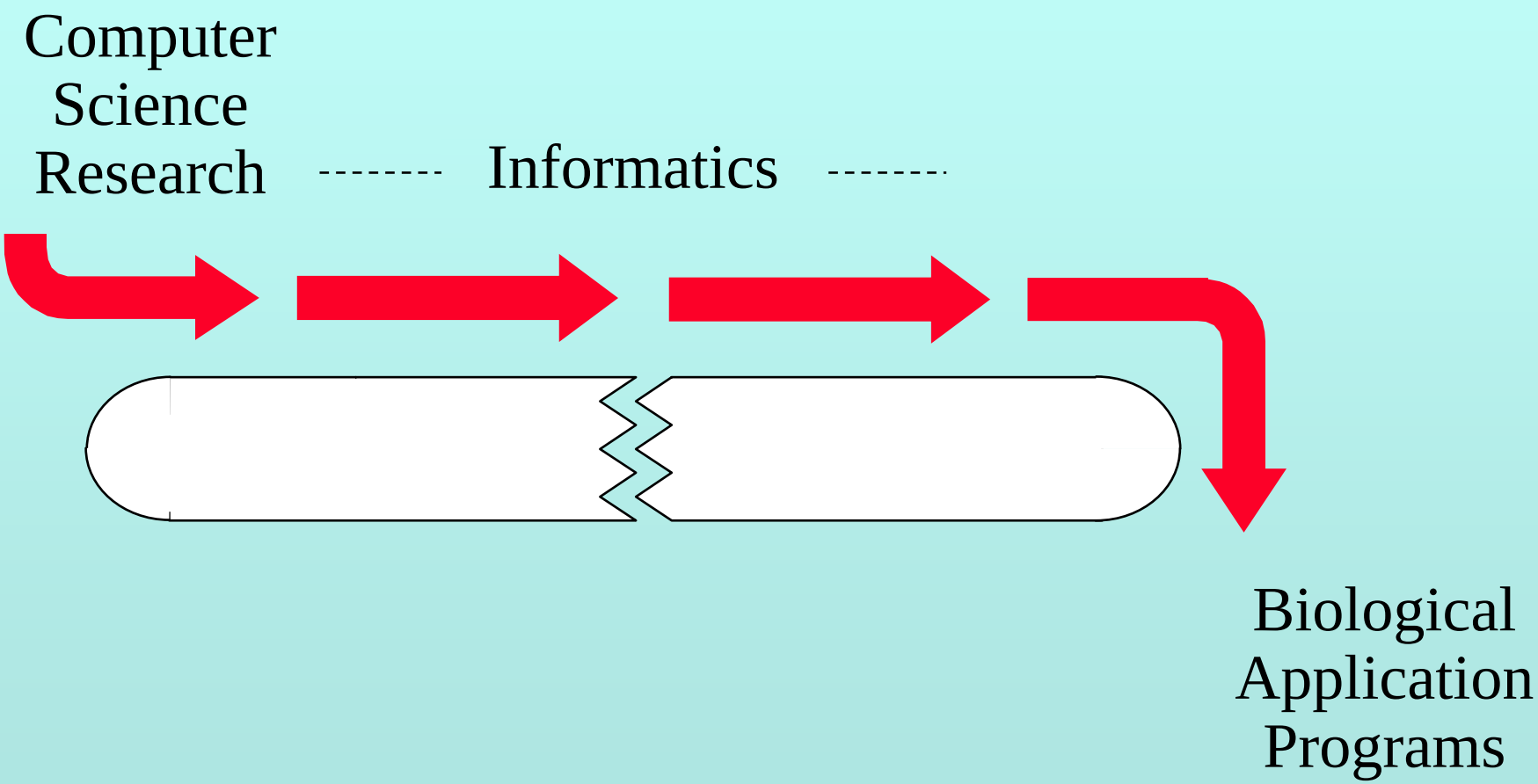
Magic

BioInformatics
will play a critical role in
delivering the magic technology
of the 21st Century

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

What is Informatics

What is Informatics



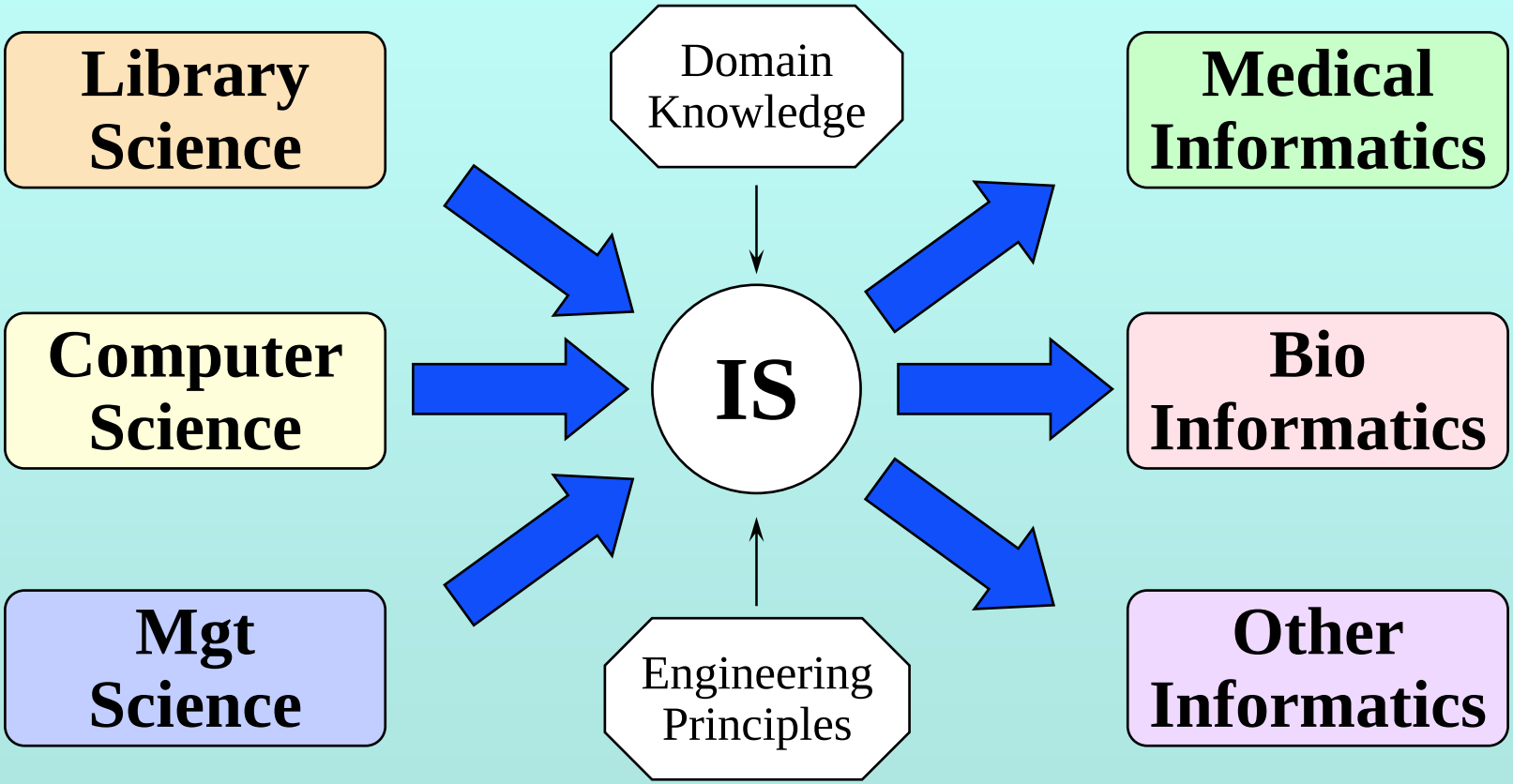
What is Informatics

Informatics combines expertise from:

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

All tempered with an engineering mindset...

What is Informatics



What is Informatics

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.

What is Informatics

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.

Teaching BioInformatics

Informatics teaching at many levels:

- *application usage*
- *routine technical worker*
- *architecture and design*
- *academic researcher*

Teaching BioInformatics

Informatics teaching at many levels:

- *application usage*
- *routine technical worker*
- *architecture and design*
- *academic researcher*

Consider: Multiple programs in statistics at most major universities.

Teaching BioInformatics

In

At all levels, instruction should stress

technical competence

biological understanding

With emphasis on

system architecture

database design

statistics at most major universities.

SCPD - Content View - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites

Address http://scpd.stanford.edu/scpd/courses/contentView/bio/ Go Links

Google Search Web AutoFill Options

Stanford Center for Professional Development
STANFORD UNIVERSITY

Jump To... Search

ABOUT SCPD | COURSE PORTFOLIO | PROGRAMS | HOW TO ENROLL | ENROLLED STUDENTS | MEMBERS & PARTNERS | HELP & FAQ

Academic Courses | Professional Ed. Courses | Courselets | Content View

Questions? Contact us

Clark Center

The Clark Center is home to the new Bioengineering Department, a joint endeavor of the School of Engineering and the School of Medicine. Professor Scott Delp (Mechanical Engineering) and Professor Paul Yock (Medicine) co-chair the Bioengineering Department.

Biosciences and Bioengineering

Biosciences and Bioengineering at Stanford

Advances in computer technologies, the completion of the Human Genome Project, and a variety of innovations in engineering disciplines have revolutionized research and industry in many areas: pharmaceuticals, health care, medicine, human biology, informatics, medical devices and more.

SCPD is rapidly expanding the number of Stanford programs and courses available to industry professionals who need to know more about the relationships between technology and biology. Browse our current portfolio of courses and programs in biosciences and bioengineering to find the educational option that best meets your needs.

Degree and Certificate Programs

- Bioinformatics Certificate
- Cardiovascular Bioengineering Certificate NEW
- Computational Genomics Certificate
- Professional Master of Science Degree in Biomedical Informatics program. Applications due by November 30, 2004 to begin the program in Spring 2004-05.
- Data Mining and Applications Certificate

Academic Courses

- BIOC 218 - Computational Molecular Biology

- Free Seminar Sign-up
- Academic Schedule
- Course Advisor for Biomedical Informatics
- Bioengineering Department
- Bioinformatics Online
- Cardiovascular Device Short Course
- Stanford Engineering and Science Institute
- Guide to Bioinformatics Curriculum
- BioX at Stanford
- Stanford Advanced Project Management

Internet

Googling BioInformatics

2,370,000 - bioinformatics

Googling BioInformatics

2,370,000 - bioinformatics

989,000 - bioinformatics college | university

887,000 - bioinformatics grant | grants | support | funding

857,000 - bioinformatics training | education

830,000 - bioinformatics career | careers | job | jobs | employment

702,000 - bioinformatics links

4,560 - "bioinformatics links"

534,000 - bioinformatics journal | journals

51,800 - bioinformatics bellevue | seattle | kirkland | redmond

21,900 - bioinformatics "degree program | programs"

10,700 - bioinformatics “community college"

BioInformatics is Global

"bioinformatics links"

domains found in the first 100 hits:

edu com org se uk pt gr tw be sg ca fi cz no dk ch hk

BioInformatics is Global

"bioinformatics links"

domains found in the first 100 hits:

edu com org se uk pt gr tw be sg ca fi cz no dk ch hk

Still, what's the big deal?

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

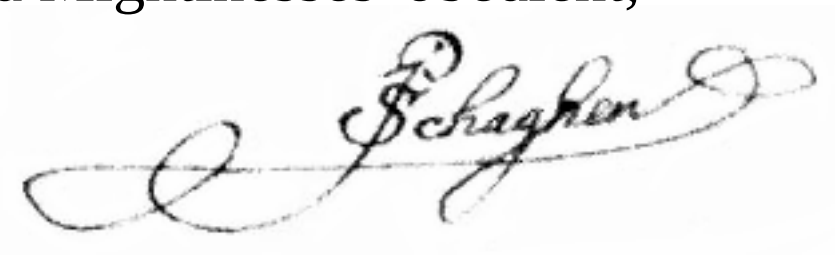
Manhattan Purchase

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,



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?

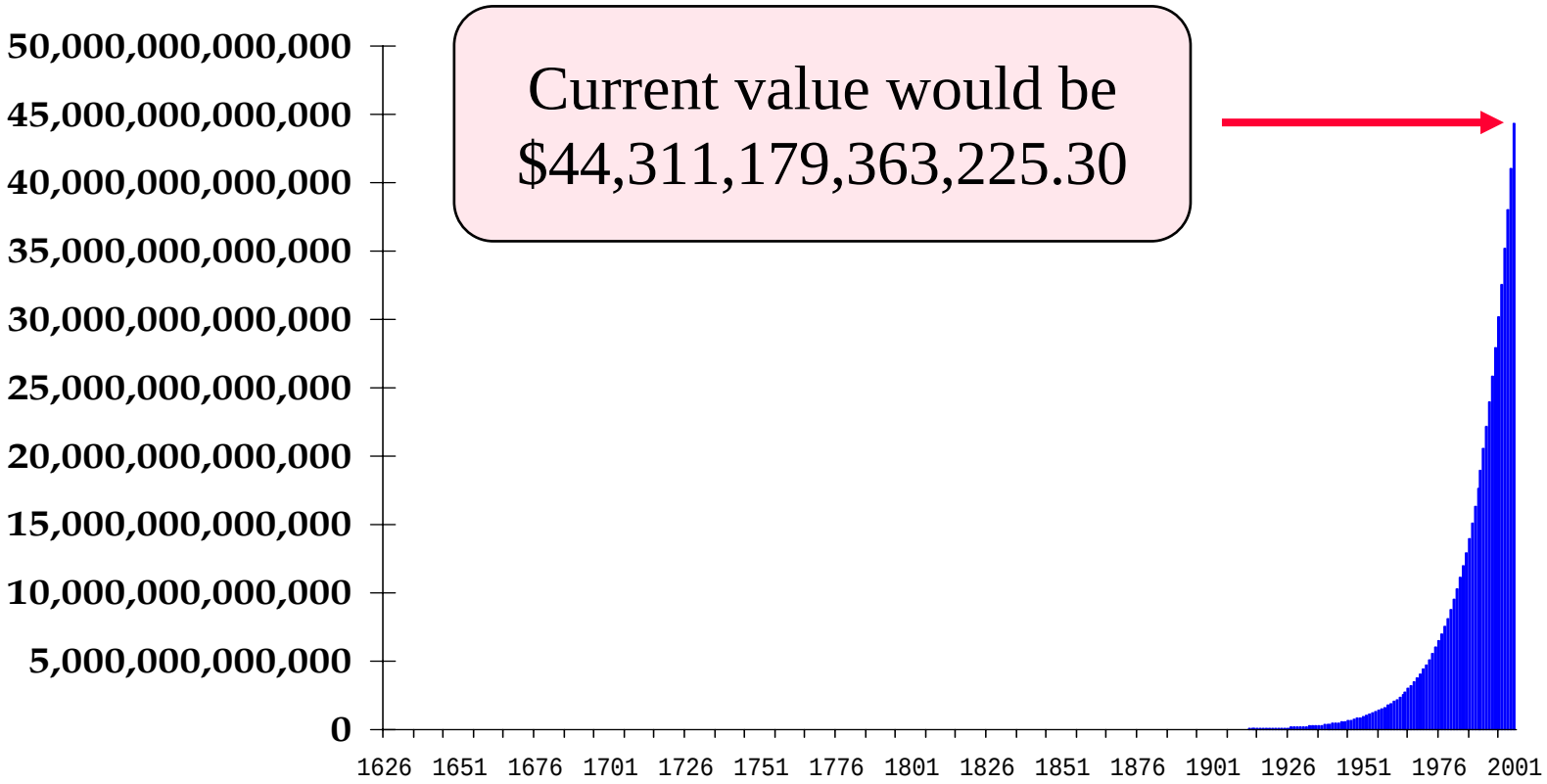
Manhattan Purchase

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?

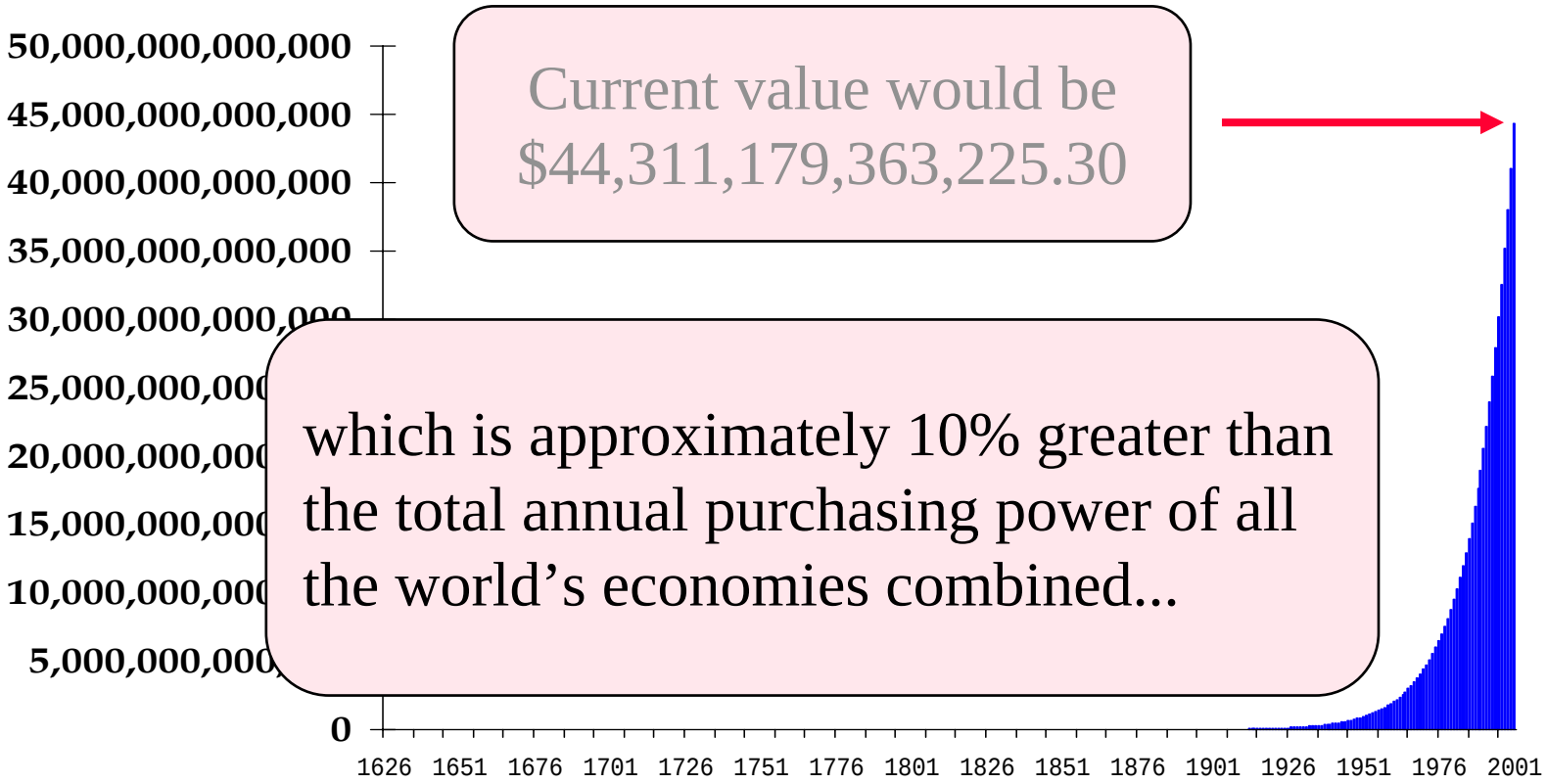
Suppose the local residents had invested HALF of the $24.00 at 8% interest. What would that be worth now?

Manhattan Purchase



Twelve dollars, invested at 8% compound interest

Manhattan Purchase

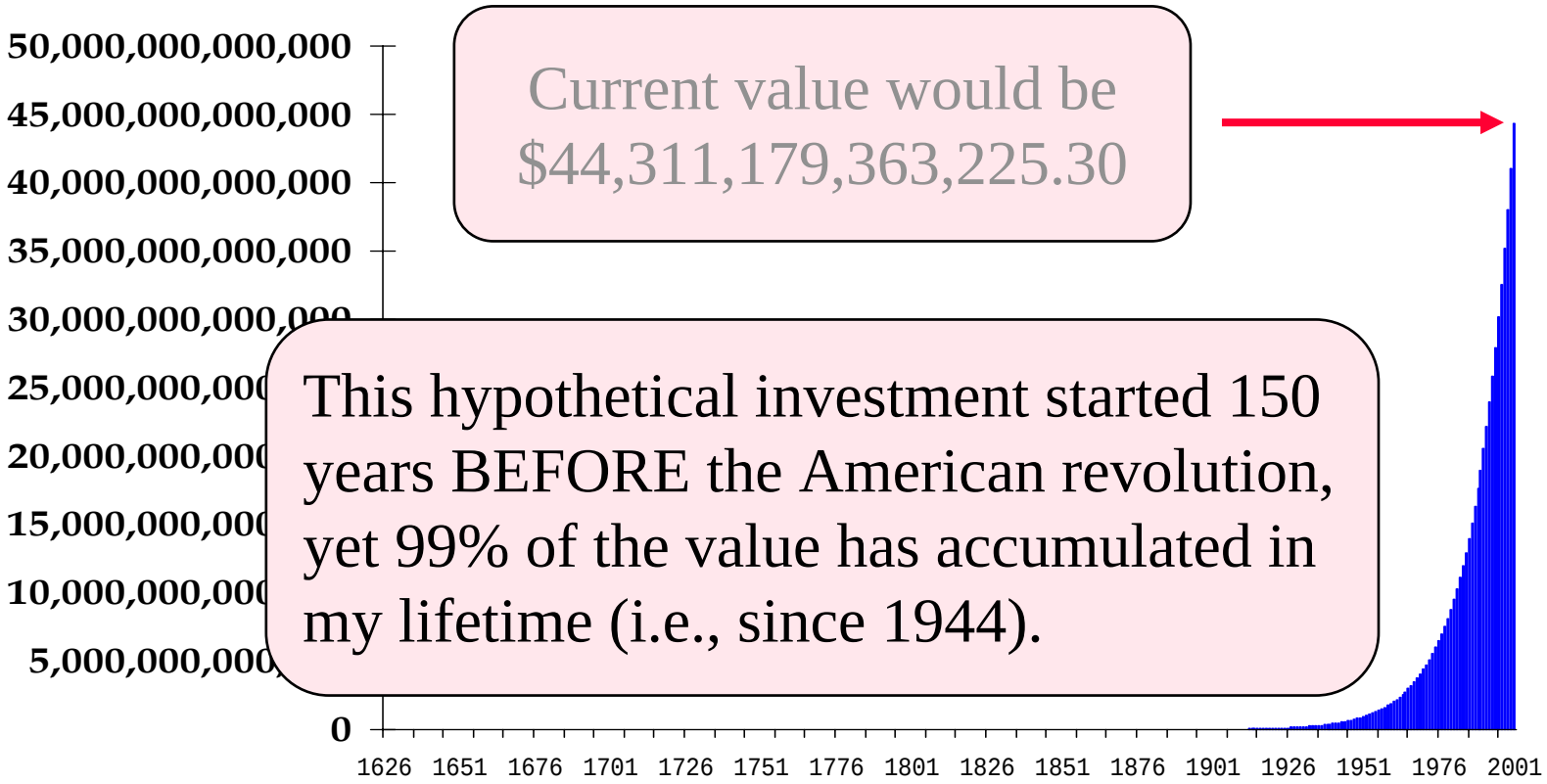


Current value would be $44,311,179,363,225.30

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



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 21st-century bioinformatics will be on this scale...

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

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

Biology is Special

- *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*.

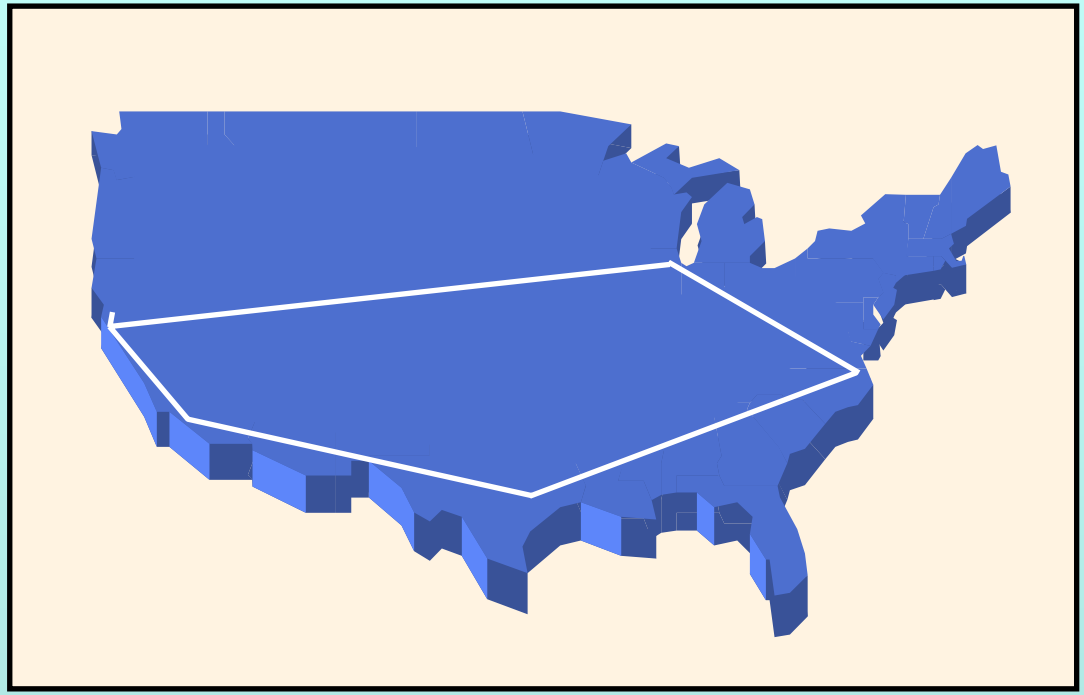
The Digital Basis of Life

[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*.

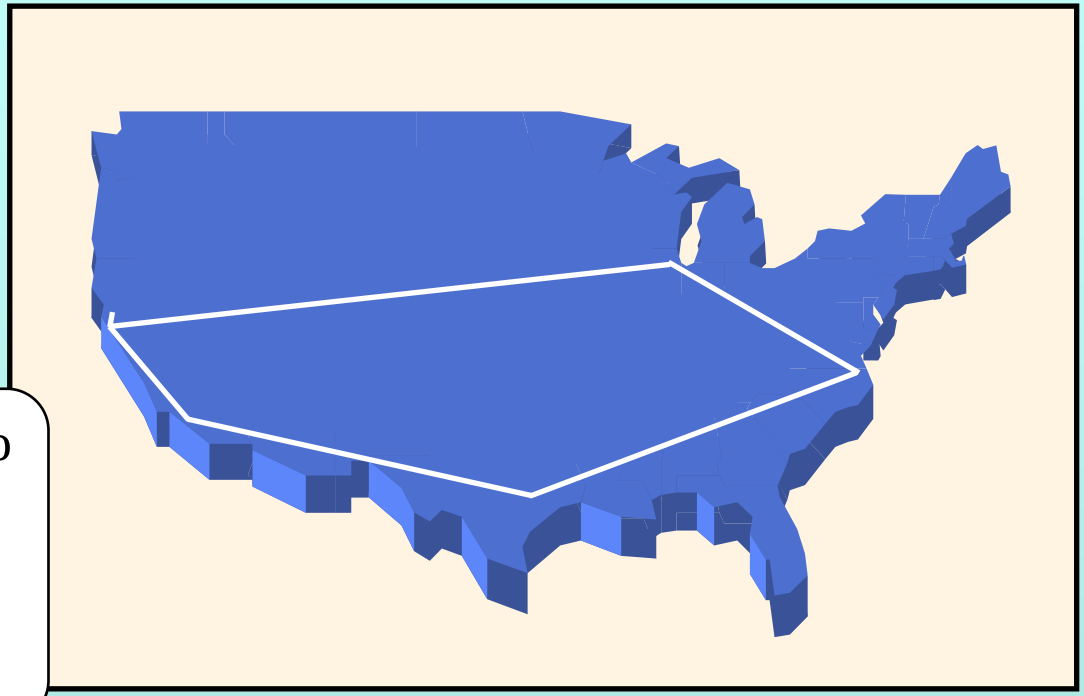
The Digital Basis of Life

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.

The Digital Basis of Life

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.

The Digital Basis of Life

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^{26} 100 gB hard disks.

The Digital Basis of Life

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^{26} 100 gB hard disks. That many hard disks would have a volume of 3.9×10^{12} cubic miles.

The Digital Basis of Life

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^{26} 100 gB hard disks. That many hard disks would have a volume of 3.9×10^{12} cubic miles. The volume of the earth is 1.8×10^{11} cubic miles.

The Digital Basis of Life

DNA is a highly efficient digital storage device:

- There is more mass-storage capacity in the DNA
- ONA
  gB
  a
  e of

the earth is 1.8×10^{11} cubic miles.

Converting ALL of the mass of the inner planets into 100 gB hard drives would not match the digital storage capacity of the biosphere.

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

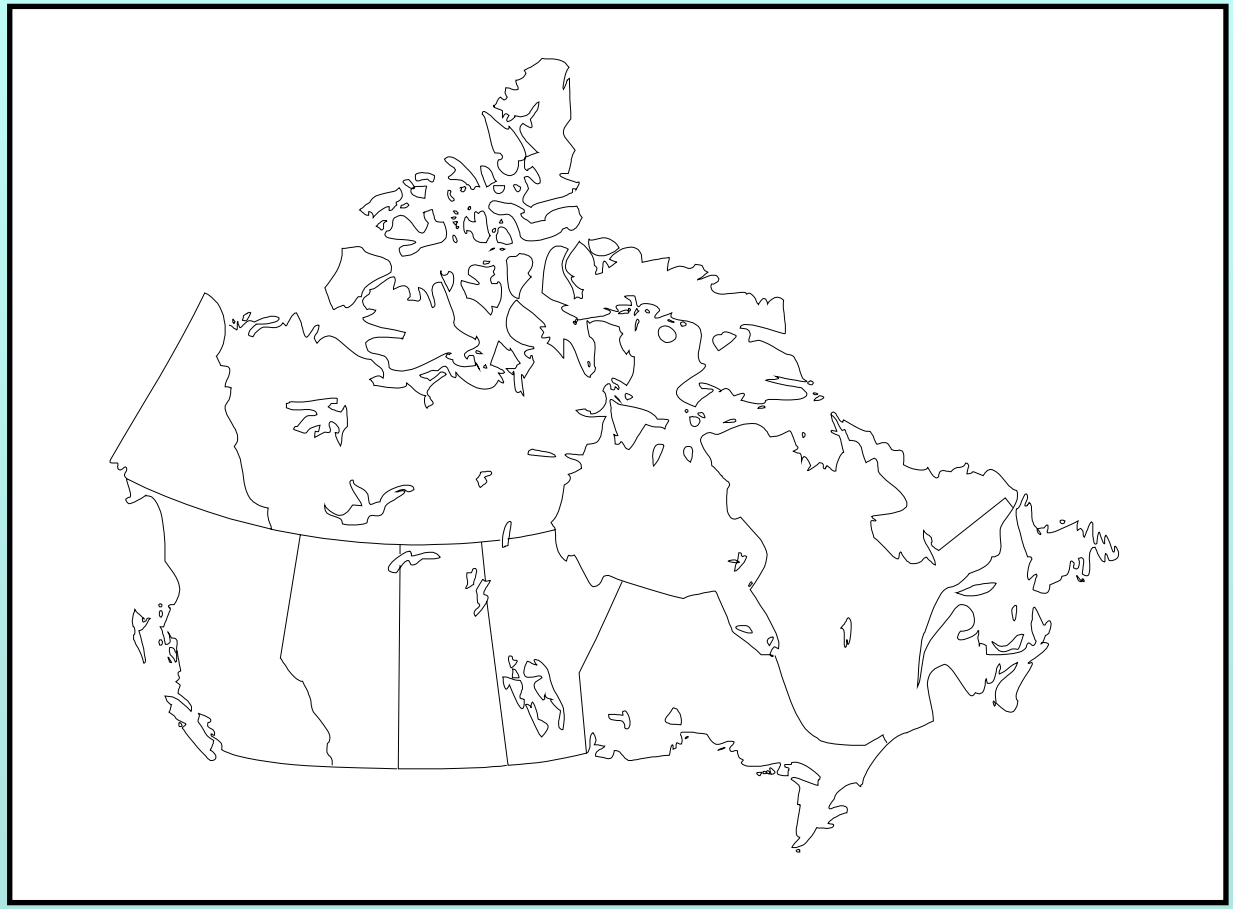
Documenting Biodiversity

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

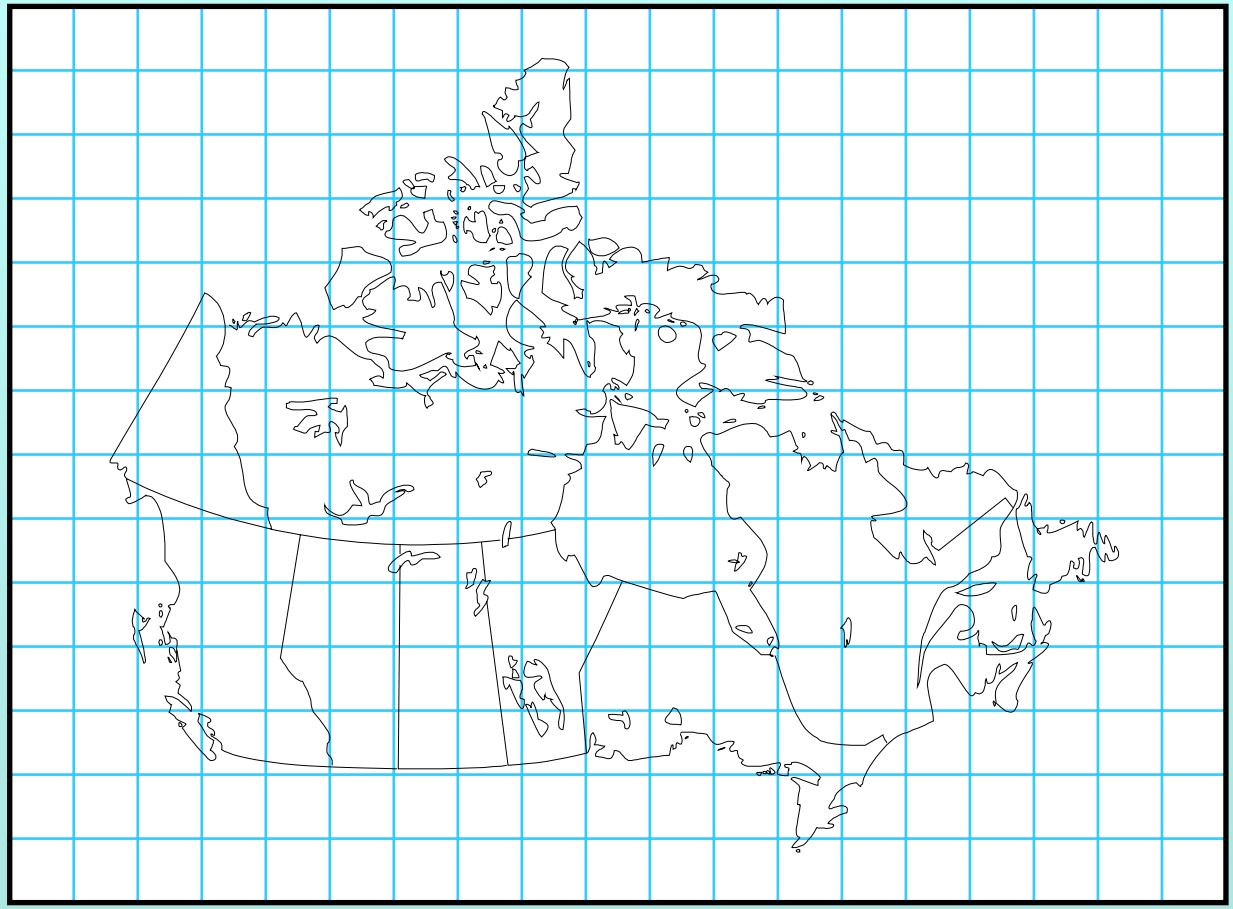
At an elementary level, documenting biodiversity involves tracking species presence/absence per unit of area.



Documenting Biodiversity

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

But at what resolution?

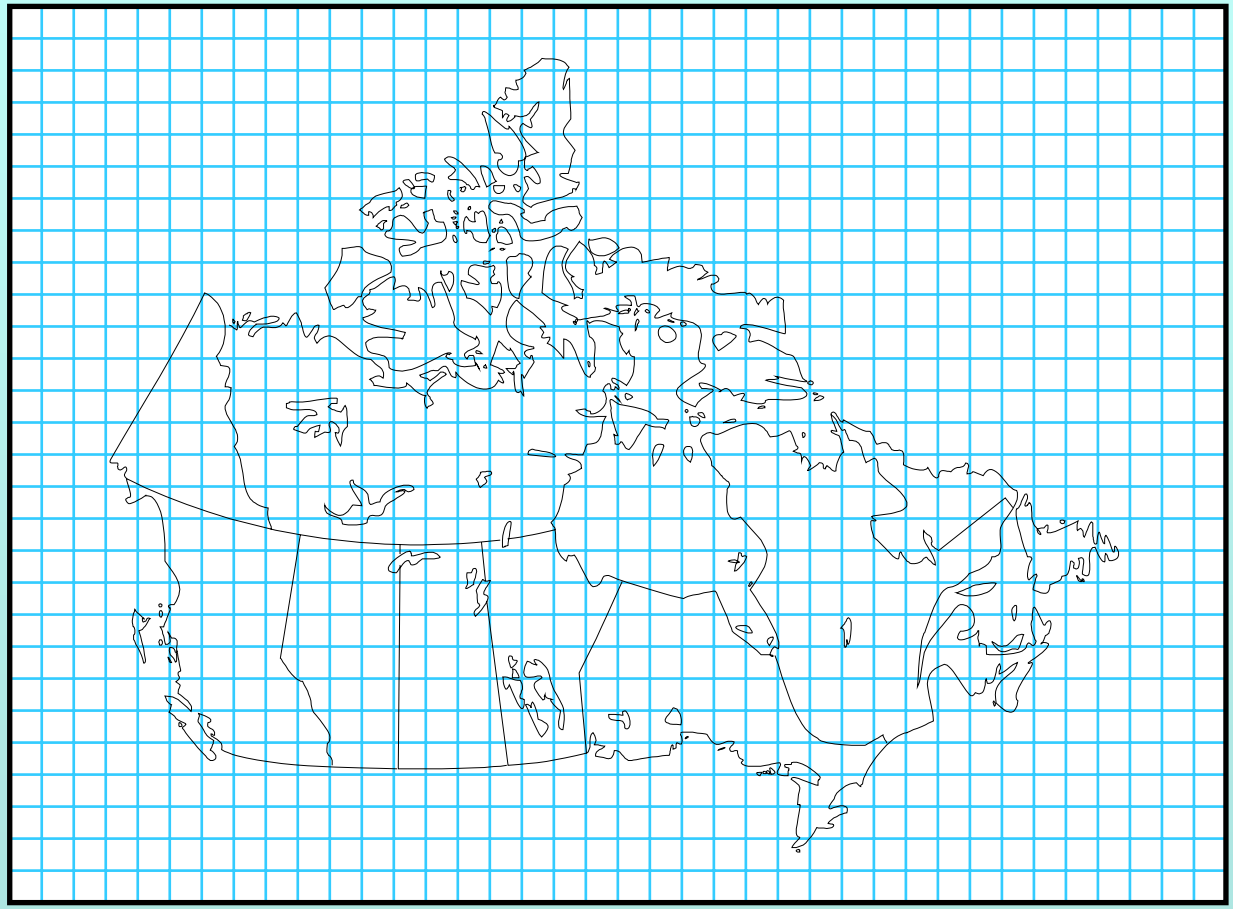


Documenting Biodiversity

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...

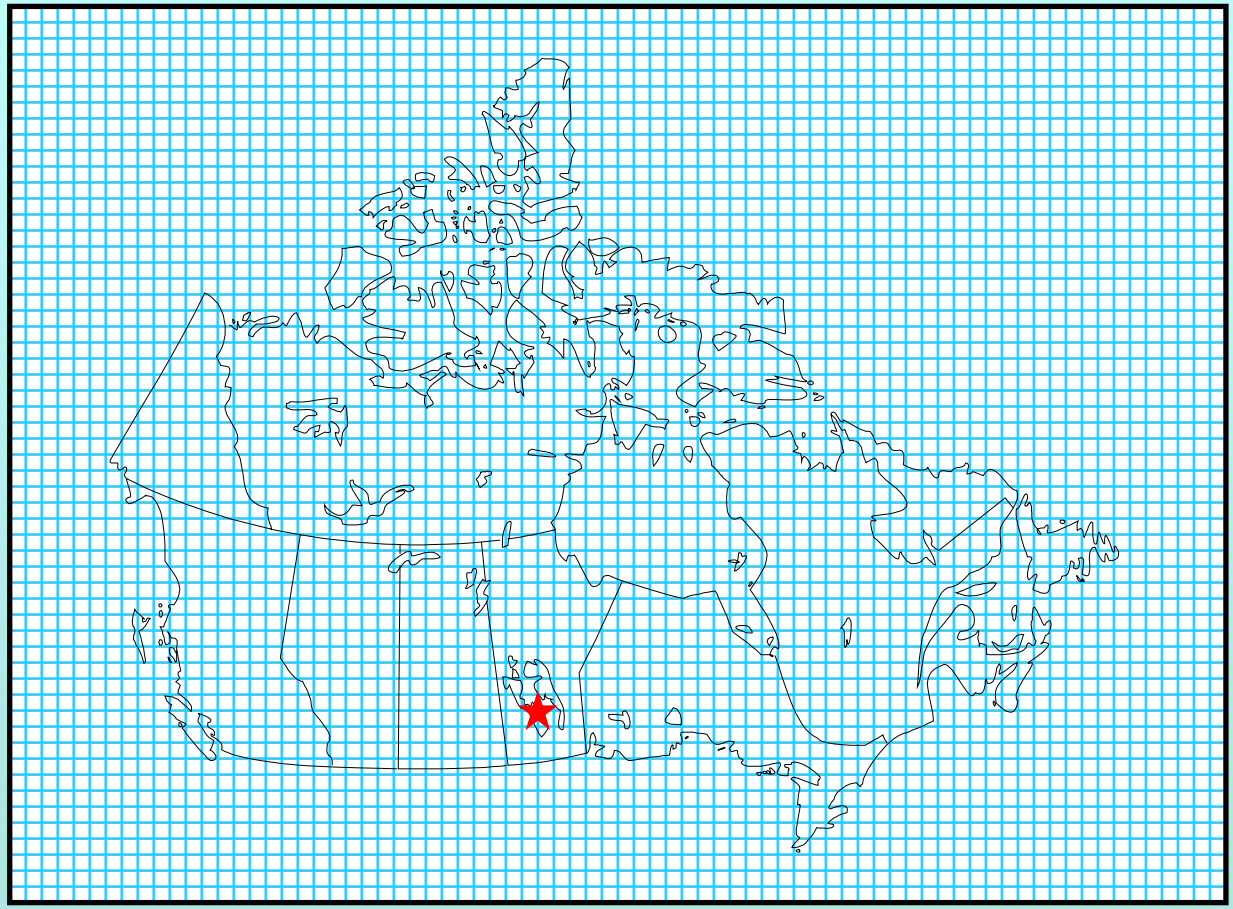


Documenting Biodiversity

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

But at what resolution?

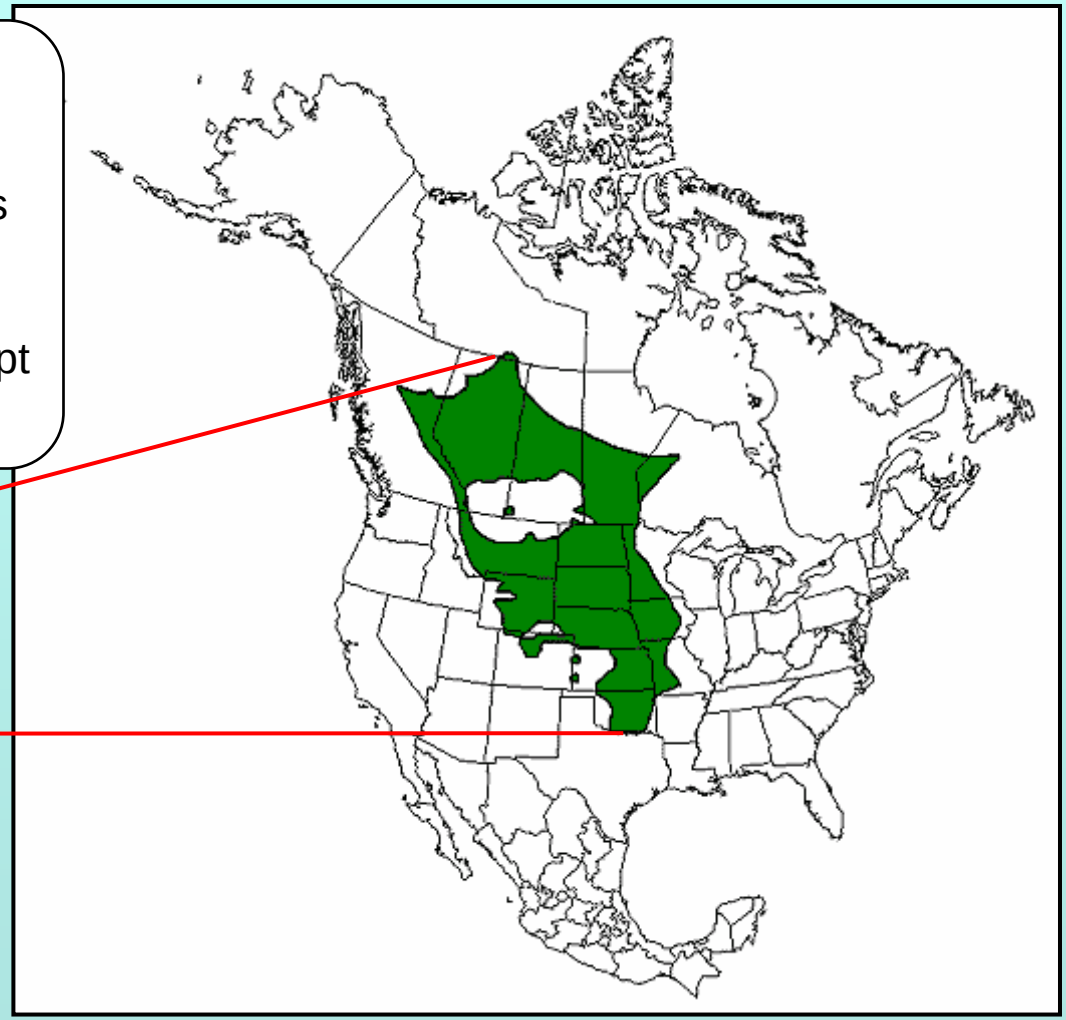
Localized, time-series probability distributions are also needed.



Documenting Biodiversity

It's one thing to say that:

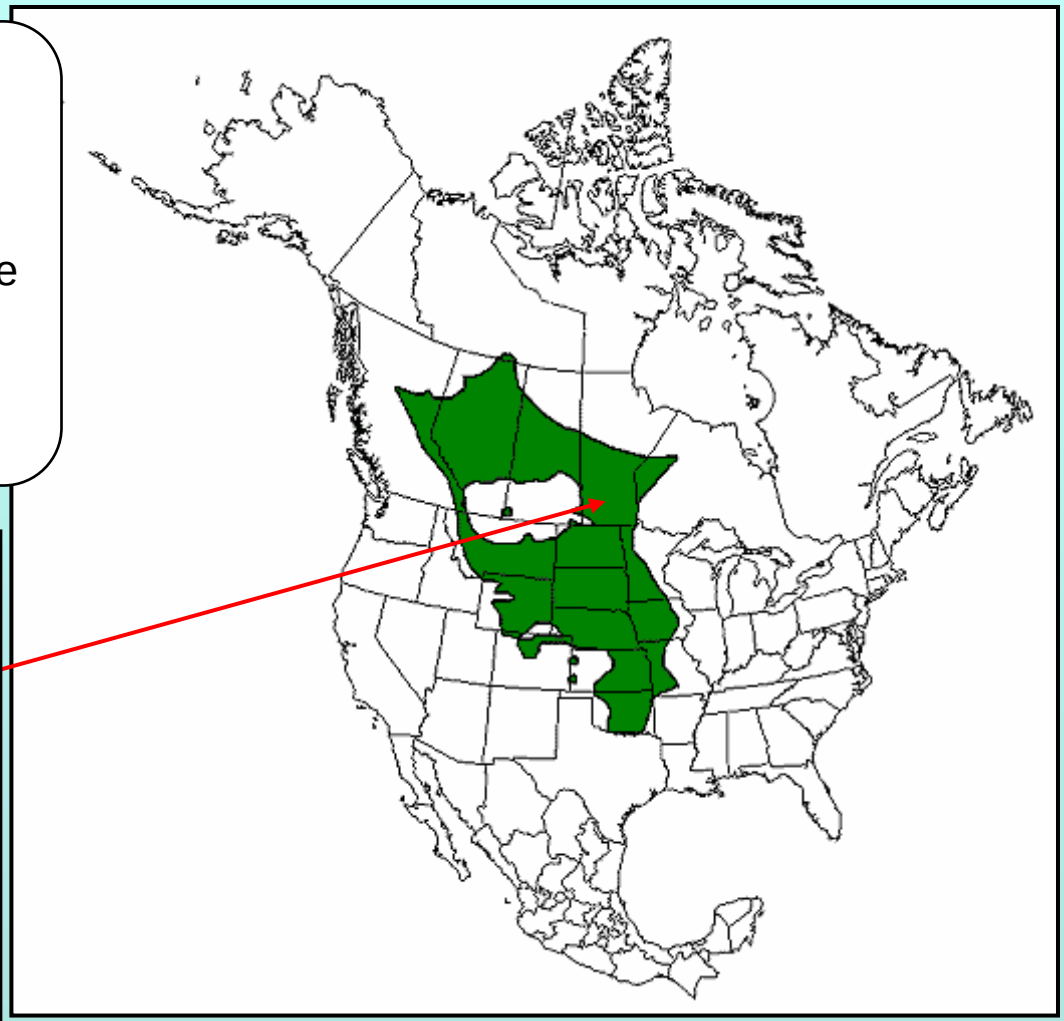
The red-sided garter snake occurs throughout central North America and is found in the southern part of Manitoba up to Flin Flon. It is absent from the extreme southwestern grasslands except for Spruce Woods Provincial Park.



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.

Reality check: budgets

Reality Check

Resource Availability:

- Compared to the recent past, current government spending on biomedical information infrastructure is huge.

Reality Check

Resource Availability:

- Compared to the recent past, current government spending on biomedical information infrastructure is huge.
- Compared to what's needed, current government spending on bio-medical information infrastructure is tiny.

Rhetorical Question

Which is likely to be more complex:

- identifying, documenting, and tracking the whereabouts of **all parcels** in transit in the UPS system at one time

Rhetorical Question

Which is likely to be more complex:

- identifying, documenting, and tracking the whereabouts of **all parcels** in transit in the UPS system at one time
- identifying, documenting, and tracking all data, all materials, and all equipment relevant to all aspects of all publicly funded biomedical research, in all fields and on all topics.

Business Factoids

Five years ago, United Parcel Service:

- used redundant multi-terabyte databases to track all packages in transit
- had 4,000 full-time employees dedicated to IT
- spent one billion dollars per year on IT
- had 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 %

Federal Funding of Biomedical-IT

Appropriate funding level:

- approx. 5-15% of research funding
- *i.e.*, **billions** of dollars per year

Federal Funding of Biomedical-IT

Appropriate funding level:

- approx. 5-15% of research funding
- *i.e.*, **billions** of dollars per year

Seem high?

What percent of institutional operating budgets goes to other mature infrastructure?

Federal Funding of Biomedical-IT

Warning:

Until more resources become available, finding true SOLUTIONS to biomedical-IT problems will be impossible.

What percent of institutional operating budgets goes to other mature infrastructure?

Reality check:

Inadequate technology

&

“light’s better” solutions

Caution from the Past

Scientific Database Management

Final Report

edited by

James C. French, Anita K. Jones, and John L. Pfalz

Report of the Invitational NSF Workshop on
Scientific Database Management
12–13 March 1990
Charlottesville, Virginia
Anita K. Jones, Chairperson
Technical Report 90-21
August 1990

Caution from the Past

U Va Tech Reports:

- CS-90-21

  J.C. French, A.K. Jones and J.L. Pfaltz, Scientific Database Management (Final Report), August 1990.

  ftp://ftp.cs.virginia.edu/pub/techreports/CS-90-21.ps.Z

- CS-90-22

  J.C. French, A.K. Jones and J.L. Pfaltz, Scientific Database Management (Panel Reports and Supporting Material), August 1990

  ftp://ftp.cs.virginia.edu/pub/techreports/CS-90-22.ps.Z

Caution from the Past

Two major conclusions:

- The single unifying cry of the workshop is that existing data models are inadequate for science data needs. (p. 6)

Caution from the Past

Two major conclusions:

- The single unifying cry of the workshop is that existing data models are inadequate for science data needs. (p. 6)
- The data source dimension (e.g., single or multi-source), which is not generally mentioned in the database literature, may present the most fundamental challenge. (p. 3)

Database Problems

Topics

- Database problems
  - Scientific data are not standard business data.
  - Better formal data models are required.
  - Schema flexibility is essential.
  - More complex logic is needed.

Database I

Basics

Relational Databases

Business Databases:

- FACTS
- REAL OBJECTS
- CLOSED UNIVERSE
- DEDUCTIVE REASONING
- CENTRALLY OPERATED

Relational Databases

Business Databases:	Scientific Databases:
• FACTS	• OBSERVATIONS
• REAL OBJECTS	• HYPOTHETICAL OBJECTS
• CLOSED UNIVERSE	• OPEN UNIVERSE
• DEDUCTIVE REASONING	• INDUCTIVE REASONING
• CENTRALLY OPERATED	• TOTALLY DECENTRALIZED

Relational Databases

Facts:	Observations:
• SOLID • STABLE • GLOBALLY CONSISTENT • STAND ALONE	• SOFT • CONSTANTLY CHANGING • MUTUALLY INCONSISTENT • REQUIRE REFERENCES

Relational Databases

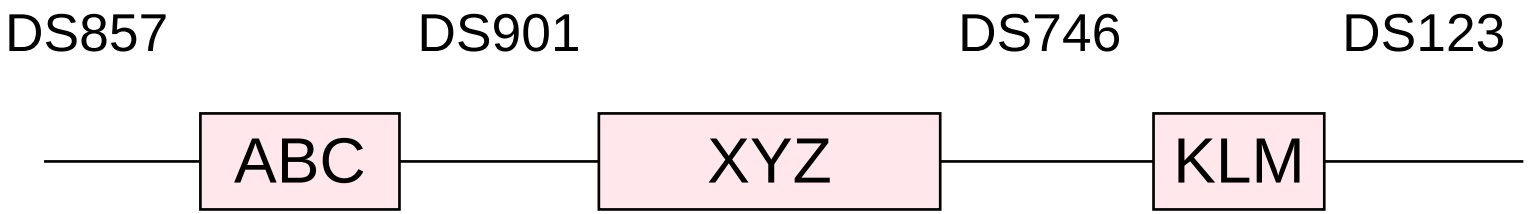
Real Objects:

- CONCRETE
- STABLE (or known instability)
- IMMUTABLE (more or less)

Hypothetical Objects:

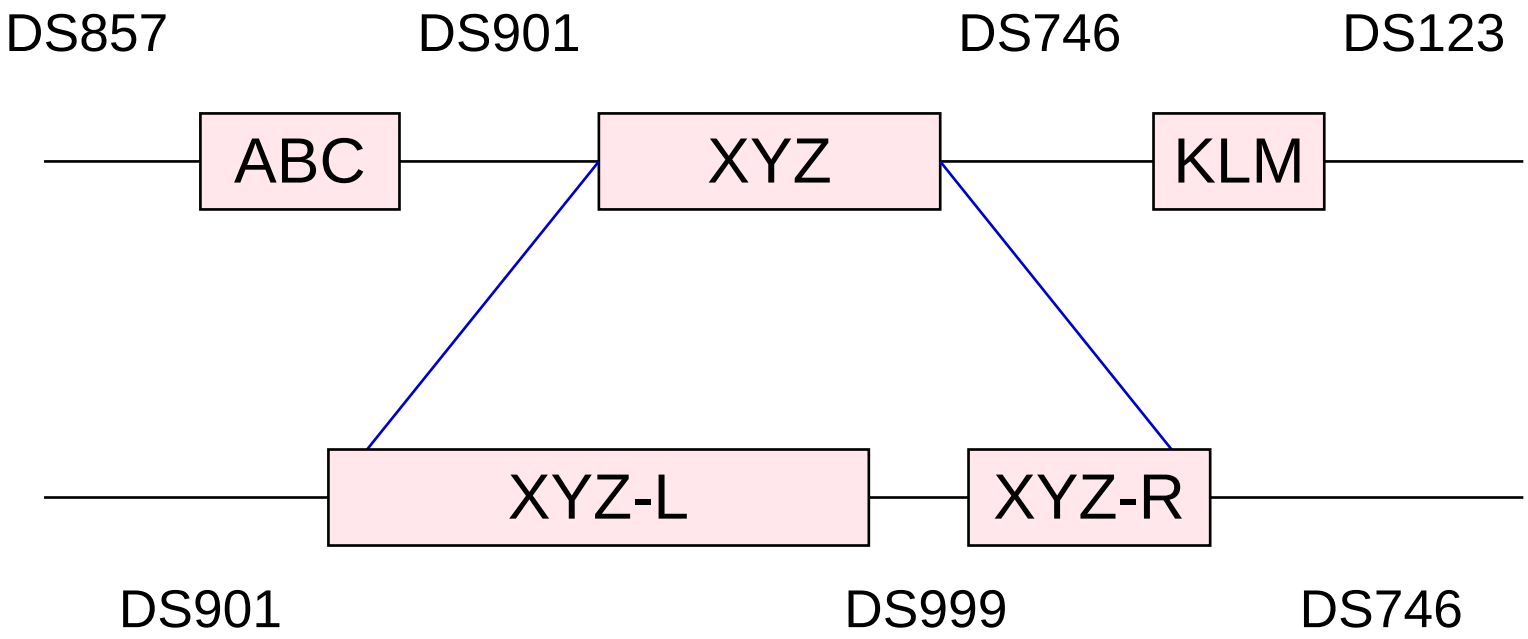
- INSUBSTANTIAL
- UNSTABLE
- HIGHLY MUTABLE (lumping and splitting)

GDB Example:



In principle, the completed genome should consist of alternating coding regions (genes) and non-coding regions (D-segs). Each map object (gene or D-seg) is an individual object, with a primary key and with foreign keys pointing to it.

GDB Example:



But while the genome is being completed, the HYPOTHETICAL genes and D-segs may undergo lumping or splitting, creating challenges for the maintenance of referential integrity.

GDB Example:

DS857 DS901 DS746 DS123

Reality is not negotiable:

Databases must either evolve to track changes in our scientific concepts, or become irrelevant

But while the genome is being completed, the HYPOTHETICAL genes and D-segs may undergo lumping or splitting, creating challenges for the maintenance of referential integrity.

Relational Databases

Closed Universe:	**Open Universe:**
Who, of the registrants for this meeting, came to the meeting?	

Relational Databases

Closed Universe:	Open Universe:
Who, of the registrants for this meeting, came to the meeting? Who, of the registrants for this meeting, did not come to the meeting?	

Relational Databases

Closed Universe:	Open Universe:
Who, of the registrants for this meeting, came to the meeting? Who, of the registrants for this meeting, did not come to the meeting?	Who else did not come to the meeting?

Relational Databases

Deductive Reasoning:	Inductive Reasoning:
• DETERMINISTIC • WELL ESTABLISHED ALGORITHMS (formal logic)	• PROBABALISTIC • METHODS STILL DEBATED (almost at the metaphysical level)

Database II

Data Models

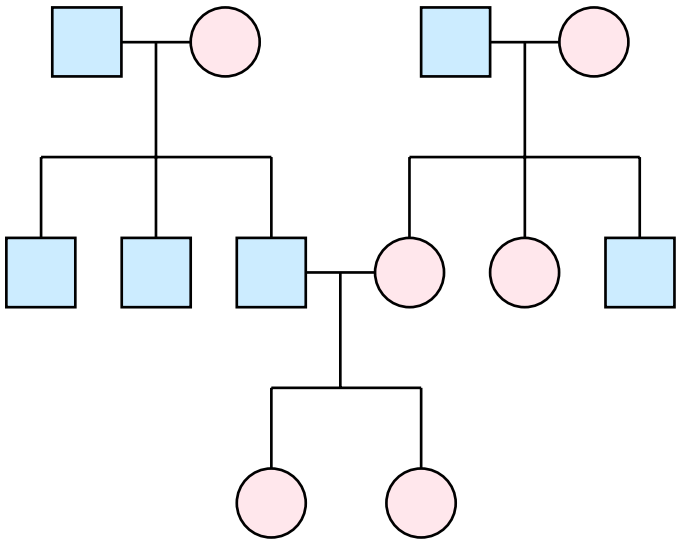
Data-model Challenges

Many bio-data problems involve:

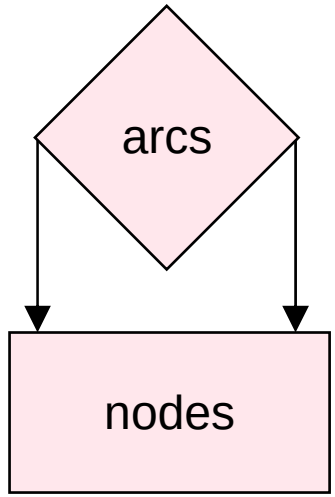
- Graphs: pedigrees, taxonomies, partial orderings, etc…
- Repeat time series observations, with inconsistent results
- Provisional conclusions
- Universal linking tables

Graph Challenges

Pedigree

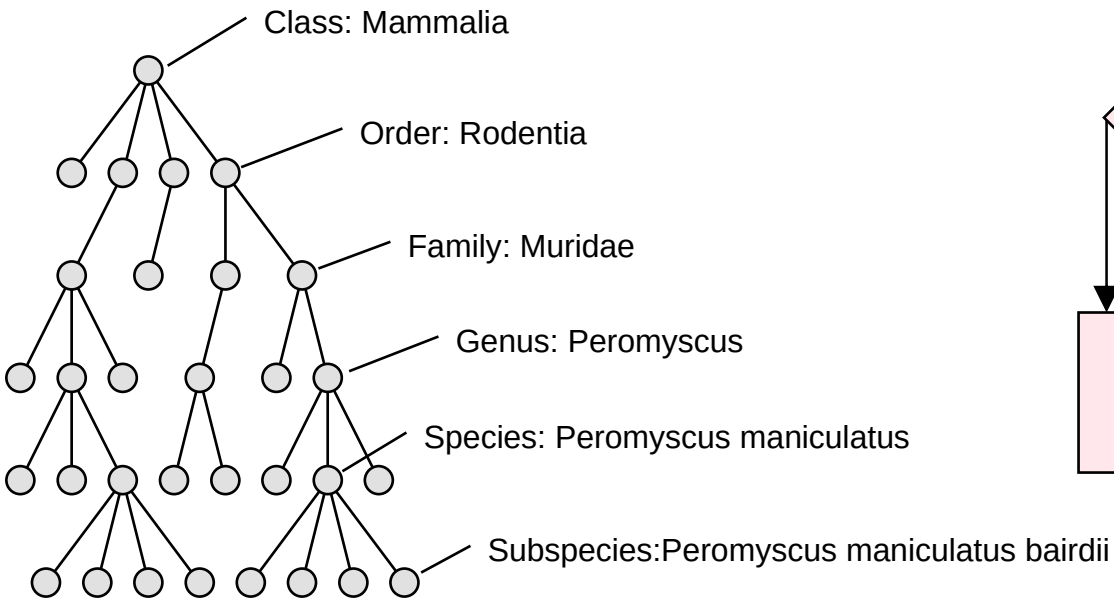


Relational Representation

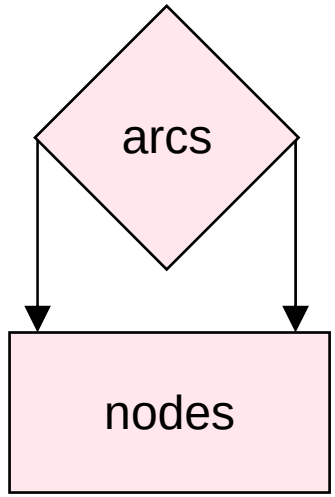


Graph Challenges

Classification Hierarchy



Relational Representation



Graph Challenges

Classification Hierarchy

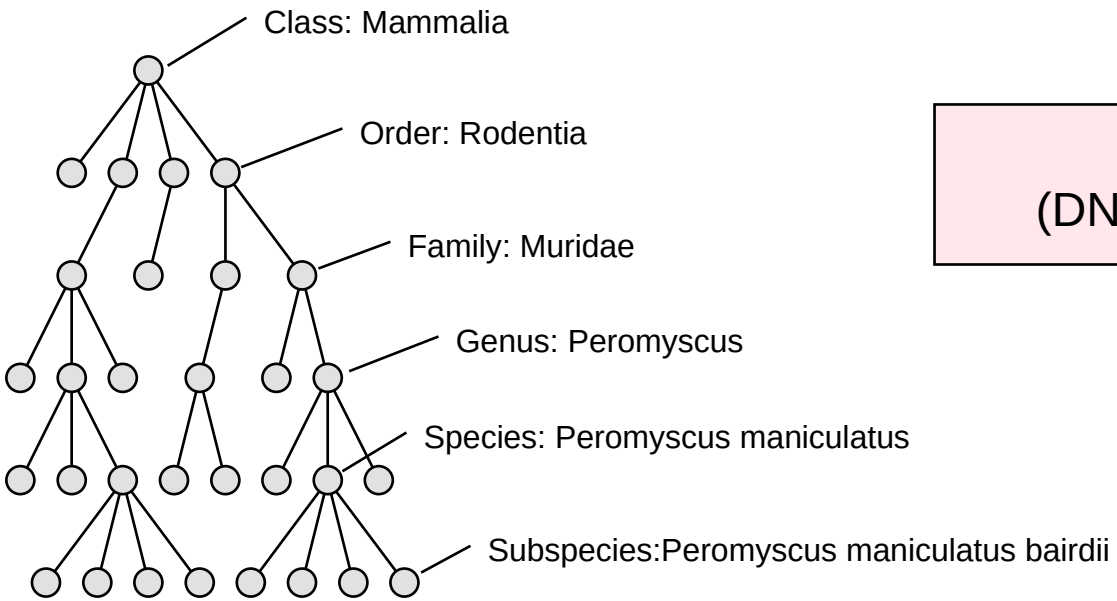
Relational Representation

Graph solutions needed:

It would be nice if database products included a CREATE GRAPH operator, including the ability to declare constraints to be maintained (e.g., directed, acyclic, connected, tree, etc)

Classification Challenges

Classification Hierarchy

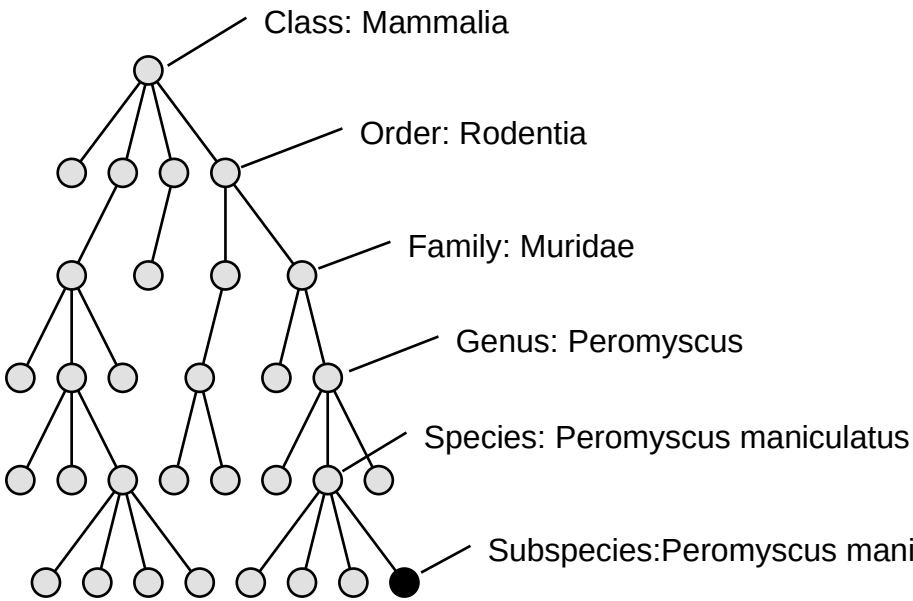


Data Objects to be Classified

Data object
(DNA sequences?)

Classification Challenges

Classification Hierarchy



Data Objects to be Classified

Classified as:

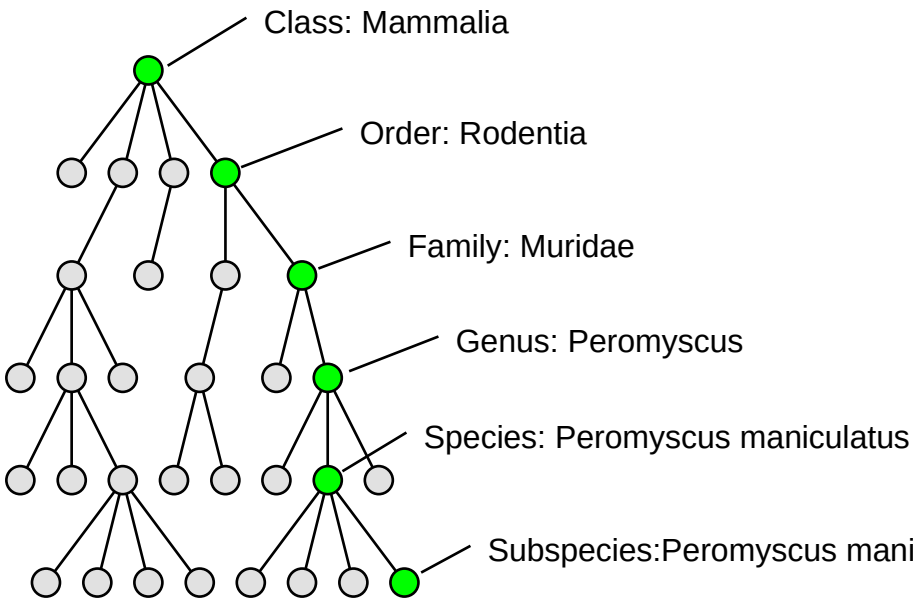
Peromyscus maniculatus bairdii

Data object
(DNA sequences?)

Suppose we permit querying at any level, but require classification of objects at leaf level.

Classification Challenges

Classification Hierarchy



Data Objects to be Classified

Classified as:

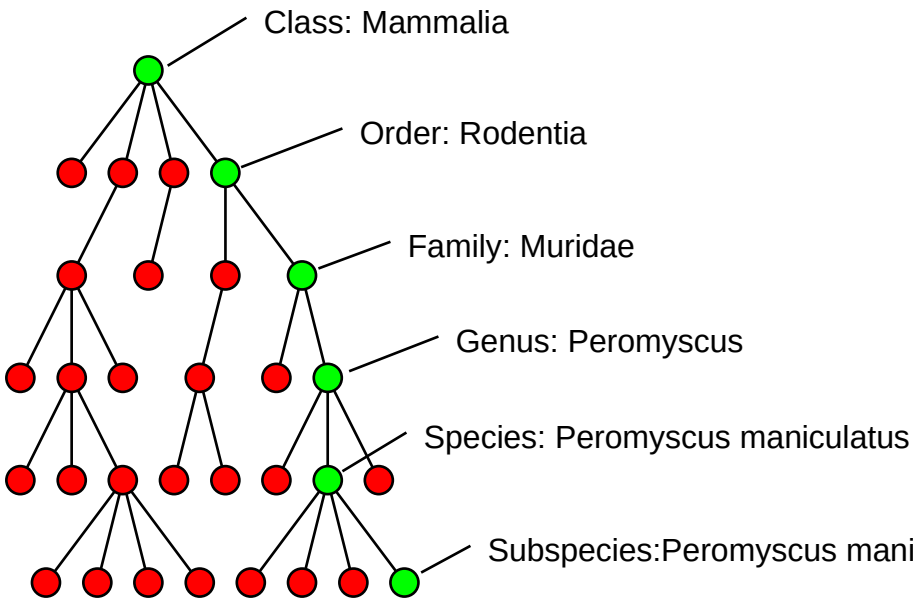
Peromyscus maniculatus bairdii

Data object
(DNA sequences?)

Suppose we permit querying at any level, but require classification of objects at leaf level. Then all questions referring to nodes on the path from the classification point to the top return **TRUE**,

Classification Challenges

Classification Hierarchy



Data Objects to be Classified

Classified as:

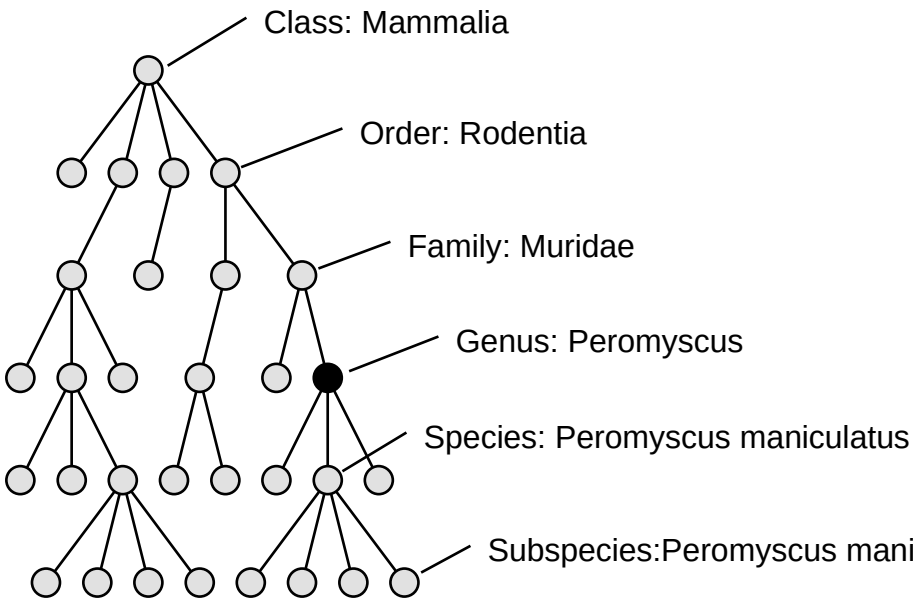
Peromyscus maniculatus bairdii

Data object
(DNA sequences?)

Suppose we permit querying at any level, but require classification of objects at leaf level. Then all questions referring to nodes on the path from the classification point to the top return **TRUE**, all others **FALSE**.

Classification Challenges

Classification Hierarchy



Data Objects to be Classified

Classified as:

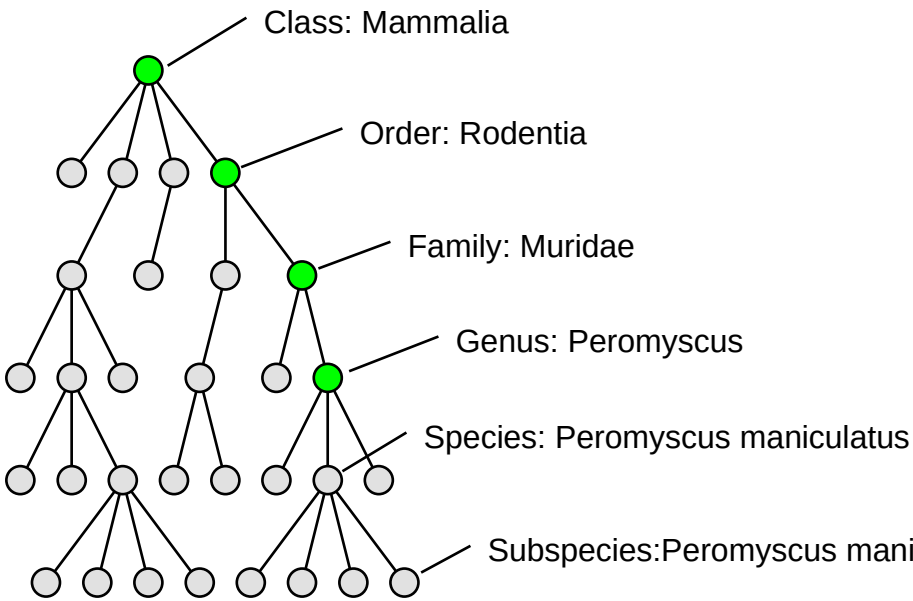
Peromyscus

Data object
(DNA sequences?)

Now, suppose the we permit querying at any level, and also that we allow classification of objects at any level.

Classification Challenges

Classification Hierarchy



Data Objects to be Classified

Classified as:

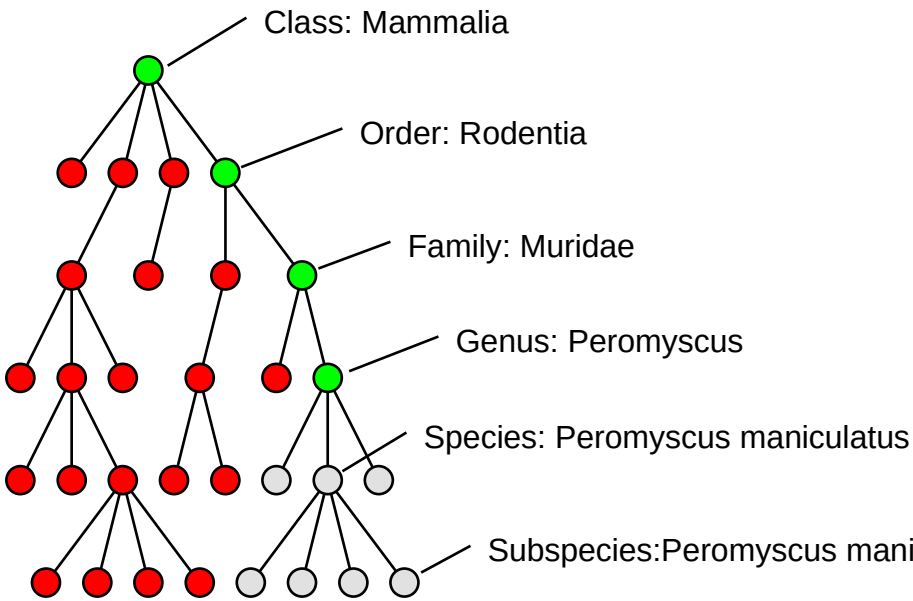
Peromyscus

Data object
(DNA sequences?)

Now, suppose the we permit querying at any level, and also that we allow classification of objects at any level. Then all questions referring to nodes on the path from the classification point to the top return **TRUE**,

Classification Challenges

Classification Hierarchy



Data Objects to be Classified

Classified as:

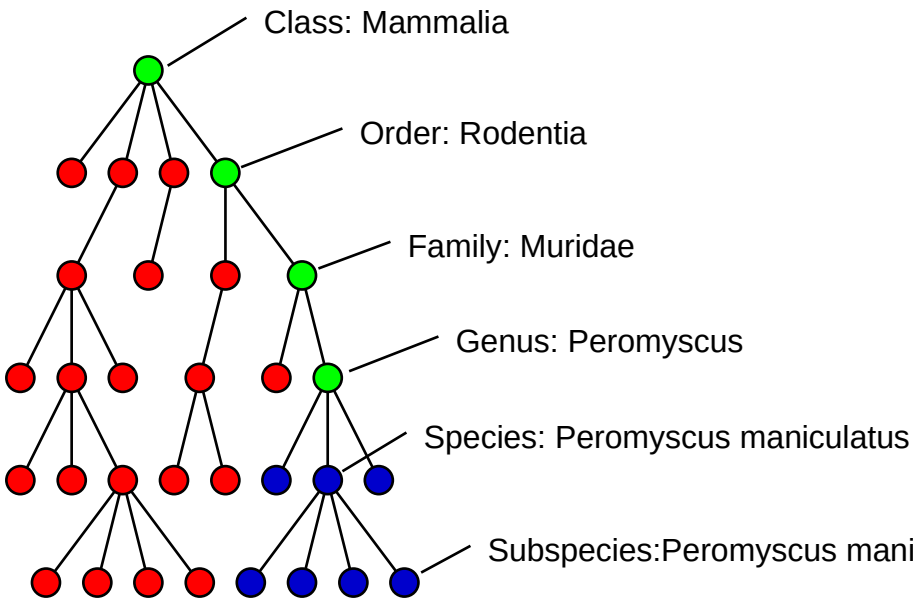
Peromyscus

Data object
(DNA sequences?)

Now, suppose the we permit querying at any level, and also that we allow classification of objects at any level. Then all questions referring to nodes on the path from the classification point to the top return **TRUE**, all questions referring to nodes lateral to this path return **FALSE**,

Classification Challenges

Classification Hierarchy



Data Objects to be Classified

Classified as:

Peromyscus

Data object
(DNA sequences?)

Now, suppose the we permit querying at any level, and also that we allow classification of objects at any level. Then all questions referring to nodes on the path from the classification point to the top return **TRUE**, all questions referring to nodes lateral to this path return **FALSE**, and all questions referring to nodes below the classification point return **MAYBE**.

Classification Challenges

Classification Hierarchy

Data Objects to be Classified

Tri-state logic required:

If hierarchical classification schemes are used, then tri-state logic may be required.

Now, suppose the we permit querying at any level, and also that we allow classification of objects at any level. Then all questions referring to nodes on the path from the classification point to the top return **TRUE**, all questions referring to nodes lateral to this path return **FALSE**, and all questions referring to nodes below the classification point return **MAYBE**.

Database III

Data Integration

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 biomedical data.

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

Data Integration Impediments

Technical: Integrating distributed, heterogeneous databases is not easy.

Sociological: Local incentives encourage competition, not cooperation.

Conceptual: Semantic mismatches exist among databases.

Schema Change

Schema-change Issues

Problems occur at many levels:

- Bio-database schemas evolve at a high rate (cf. failure of IGD as cited by Stein).
- We need systematic support for inter-database referential integrity.
- We need support for intra-database referential integrity following lumping or splitting actions.
- More issues…

Schema-change Issues

Problems occur at many levels:

Schema Evolution:

Schemas of scientific databases evolve at a high rate. Without tools to support referential integrity in the face of these changes, long-term data integration is impossible.

- More issues…

Data Source Problems

Topics

- Data-source problems

  Biology is a small-instrument, multi-source science.

  Integrating multi-source data is hard.

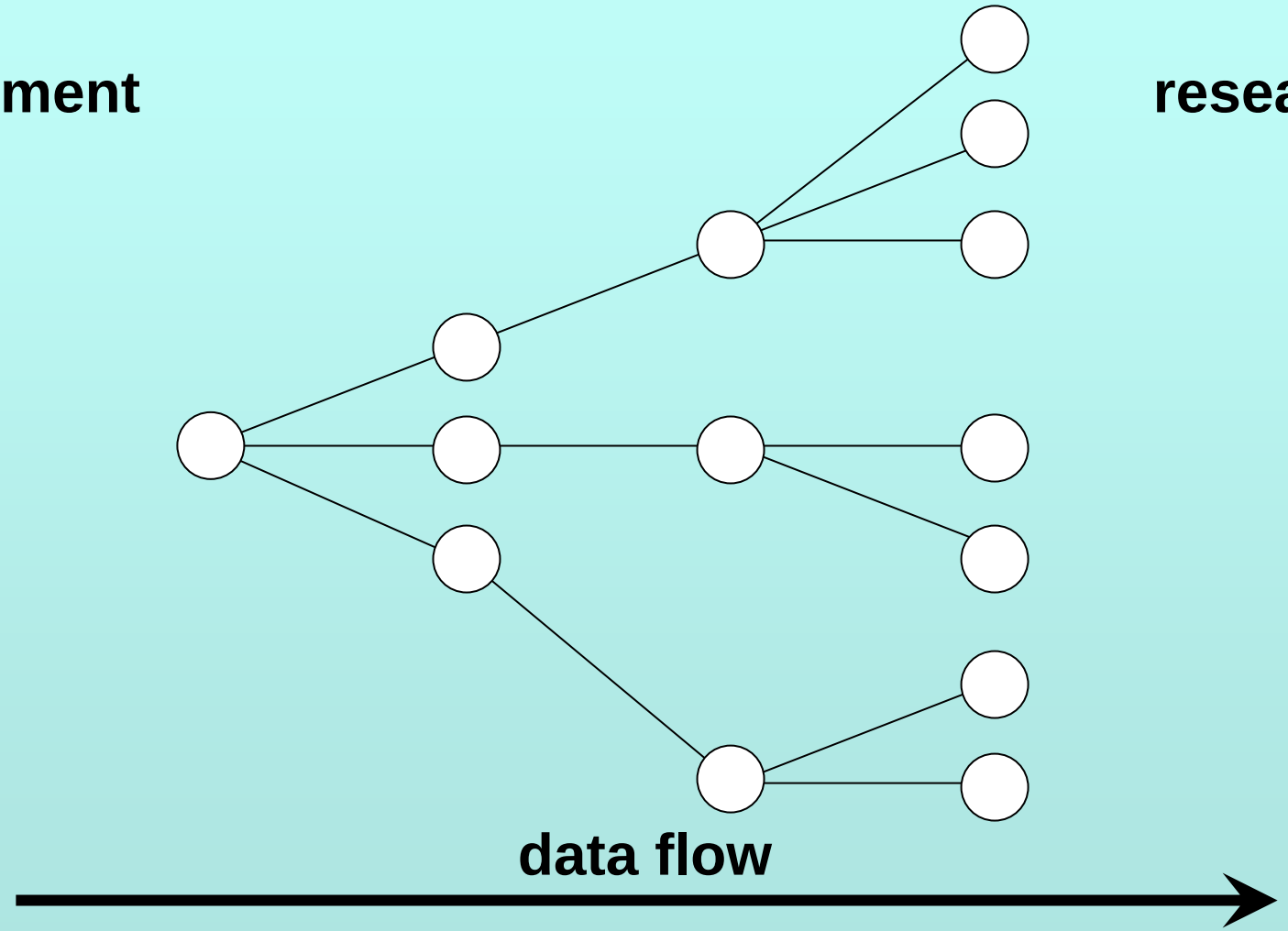
  Consistency flows in the wrong direction.

  GenBank is a false model.

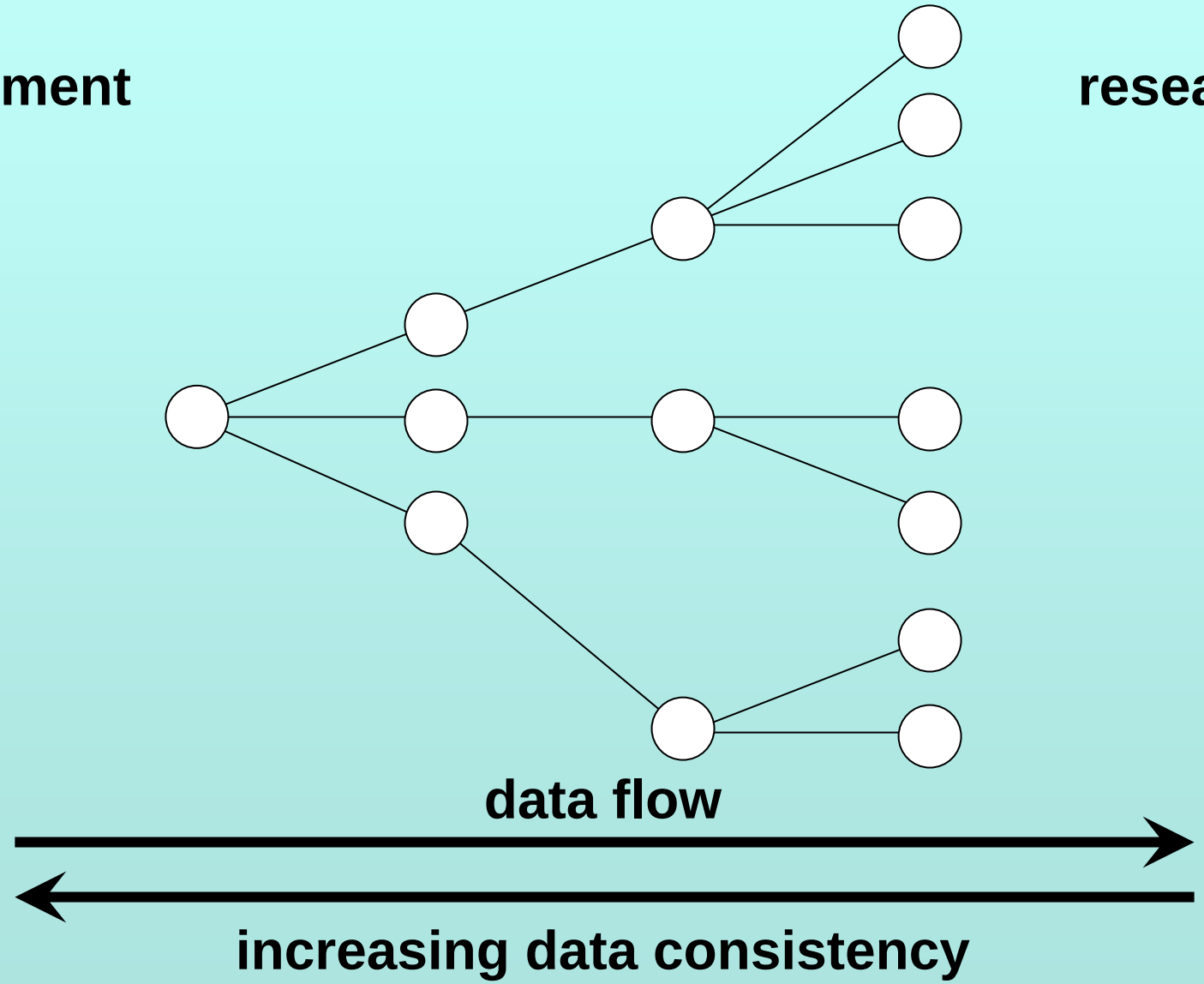
Source I

Basics

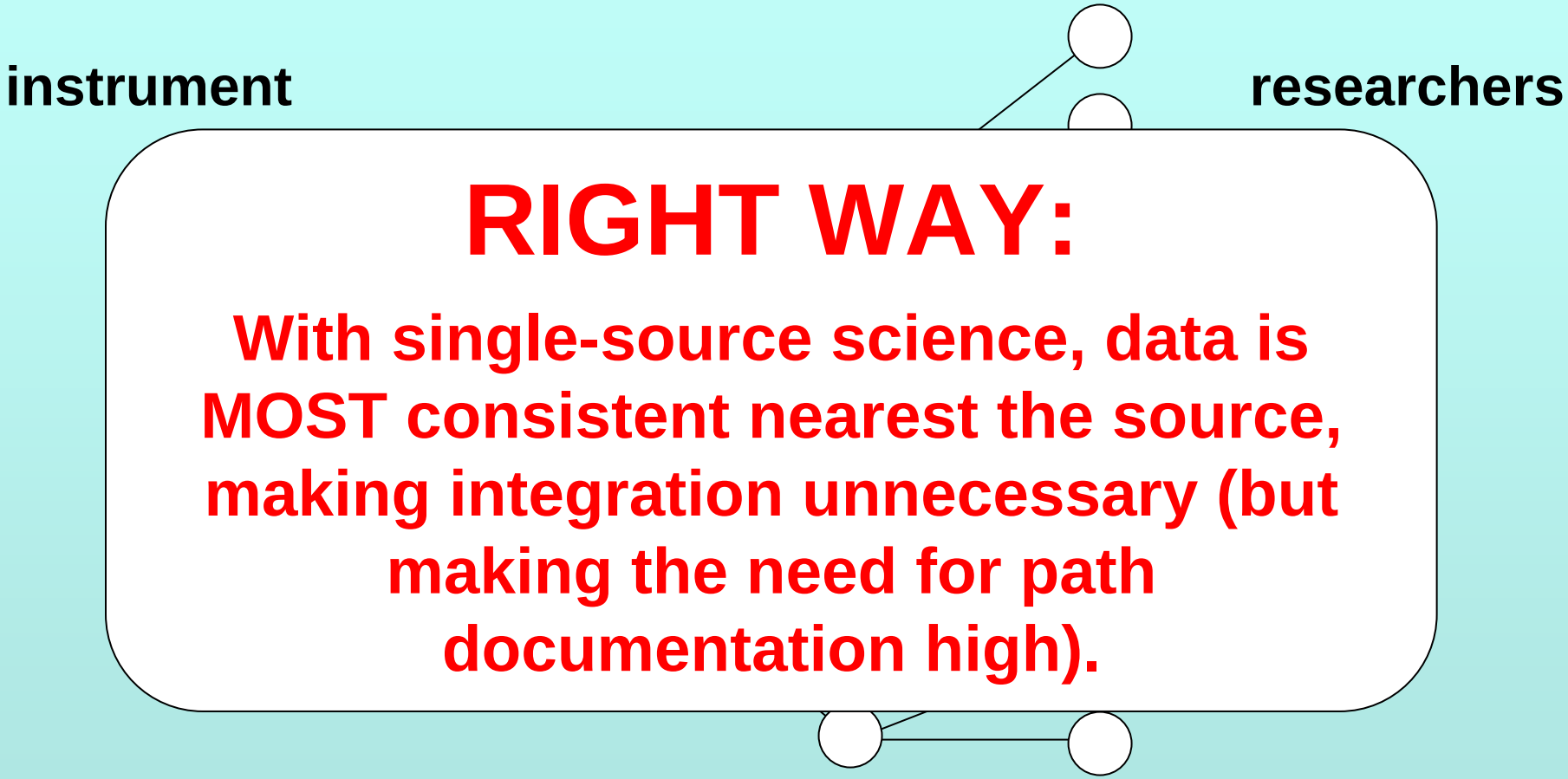
Single-instrument Science



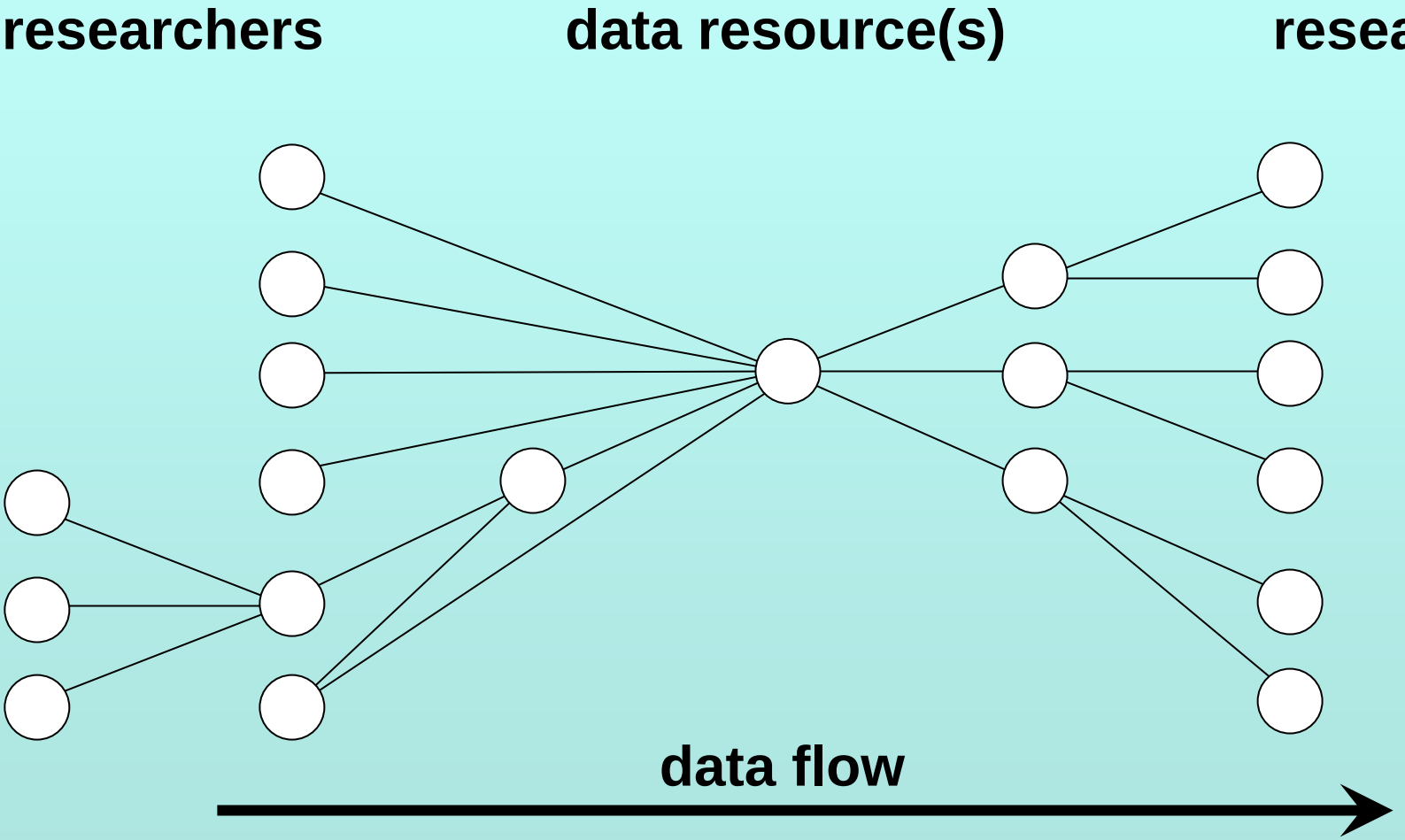
Single-instrument Science



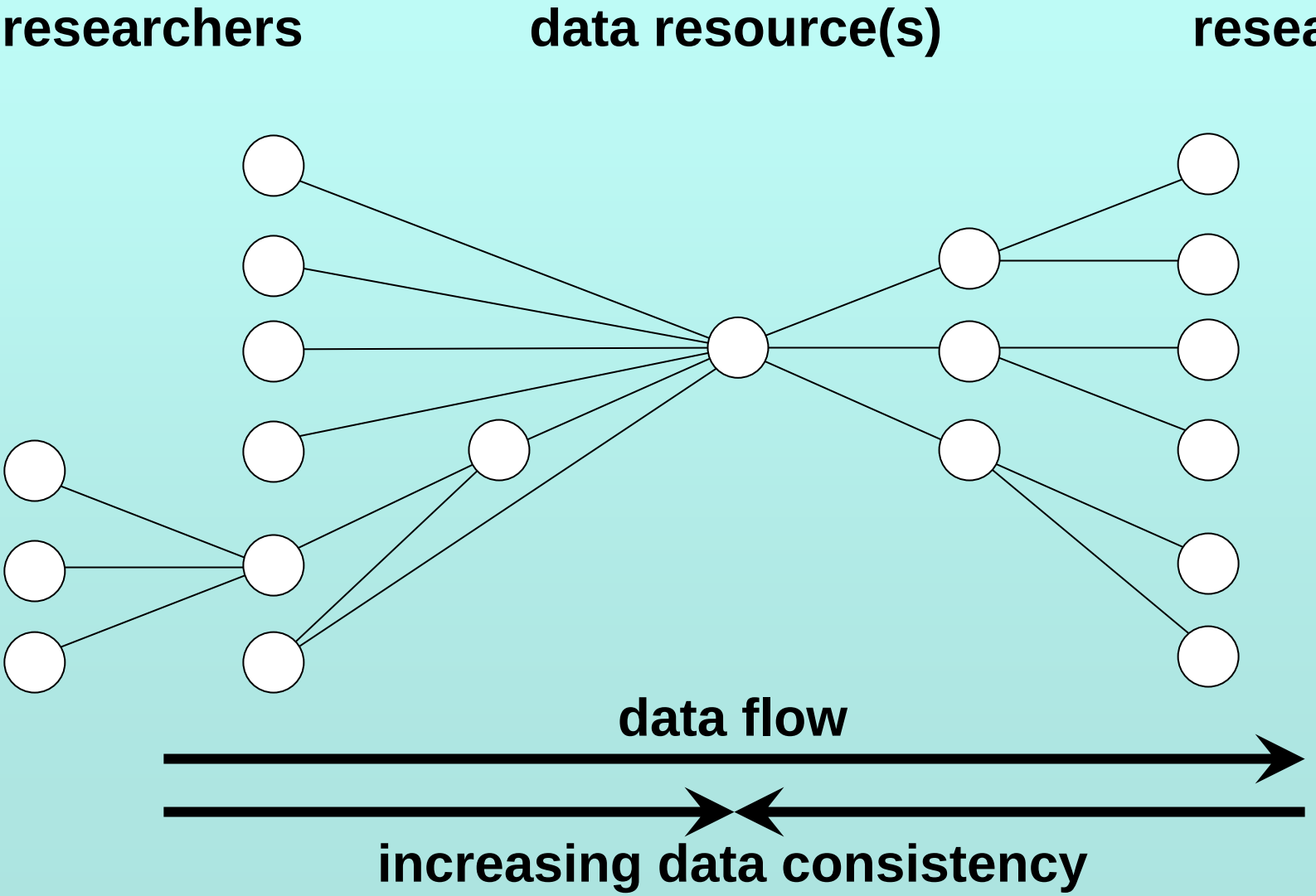
Single-instrument Science



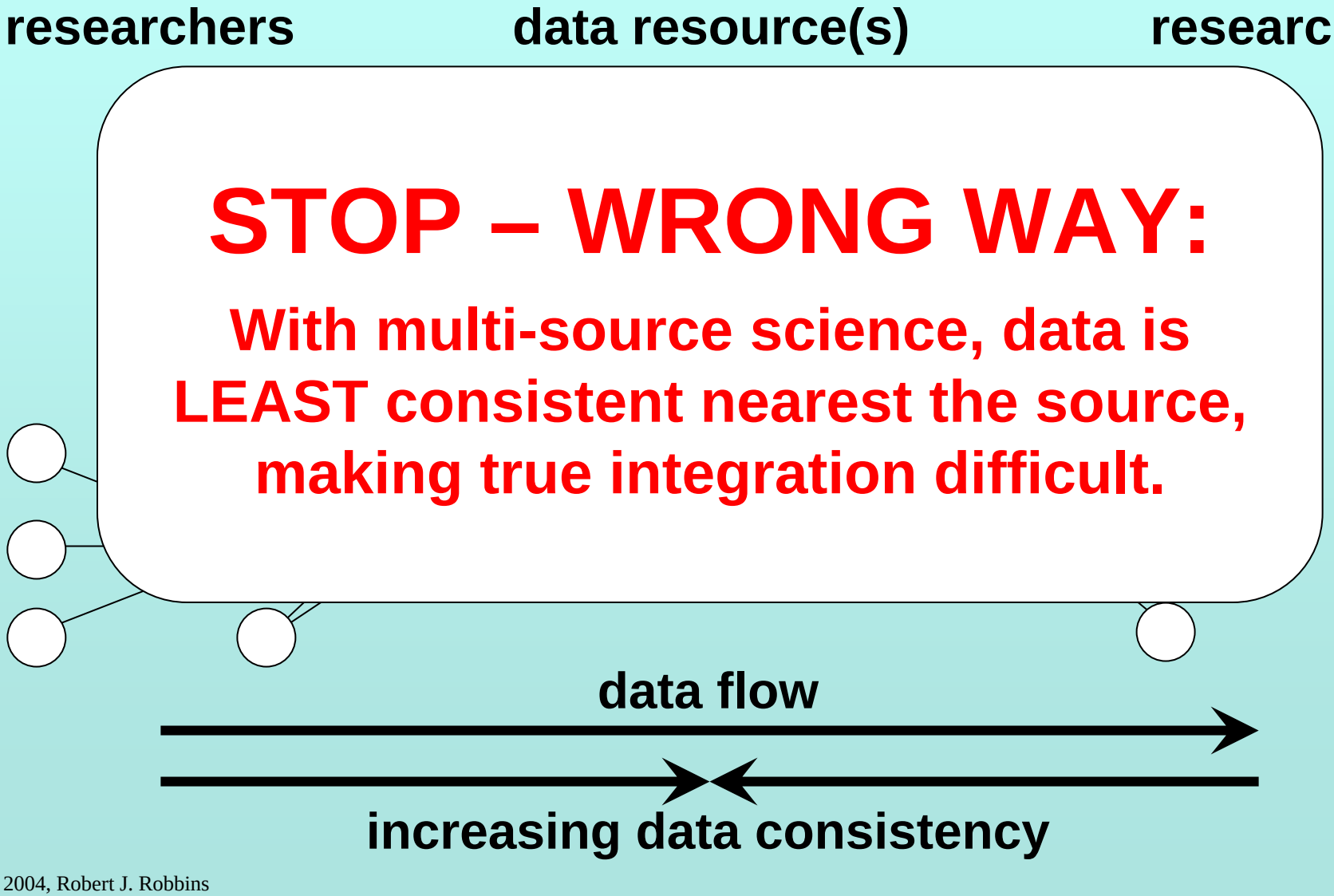
Multi-instrument Science



Multi-instrument Science

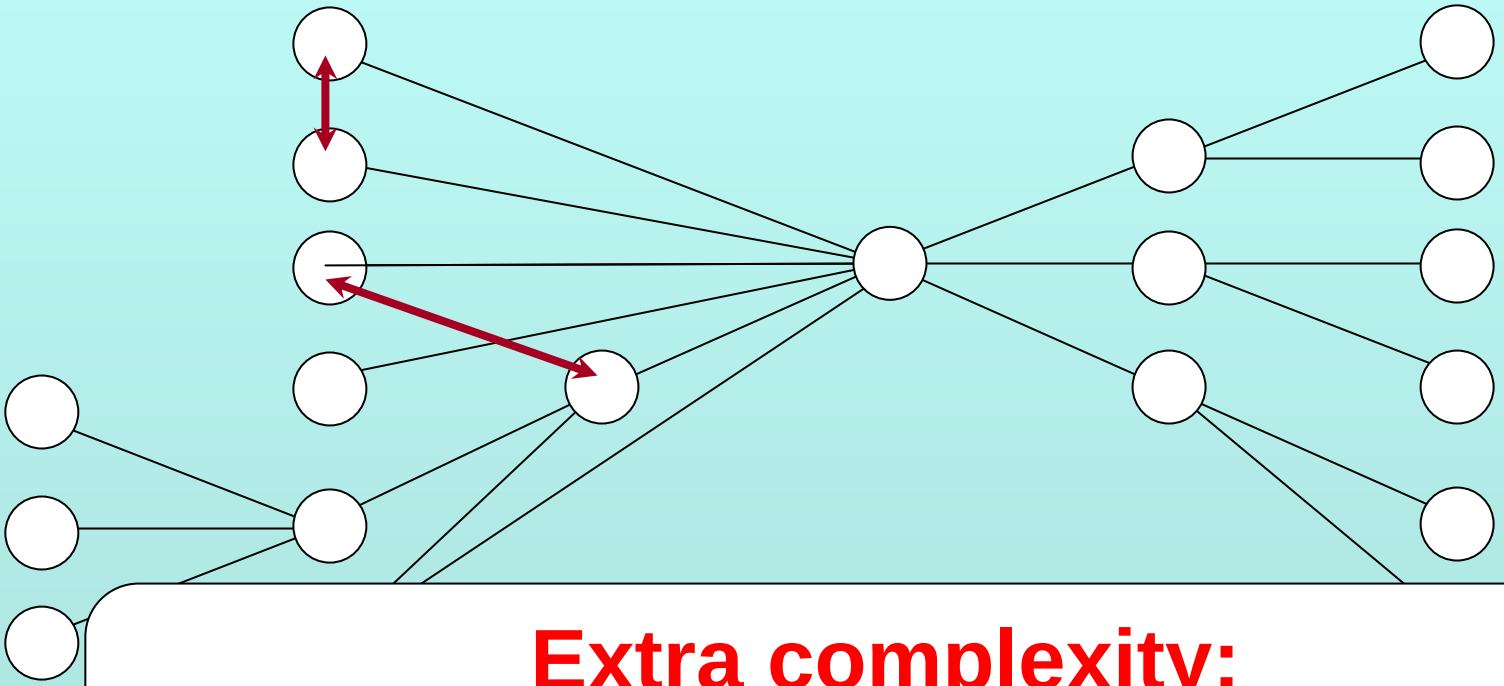


Multi-instrument Science



Multi-instrument Science

researchers **data resource(s)** **researchers**

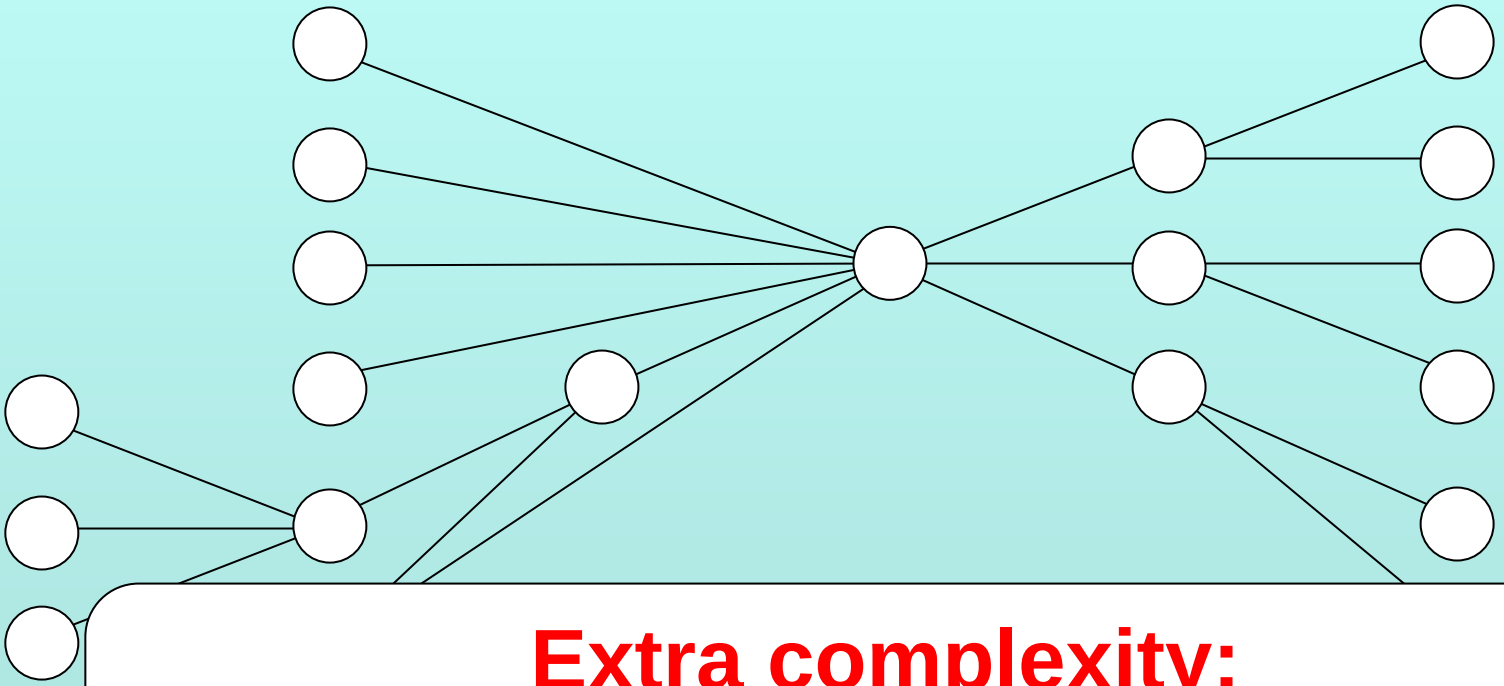


Extra complexity:

Undocumented, uncoordinated local data exchange

Multi-instrument Science

researchers **data resource(s)** **researchers**

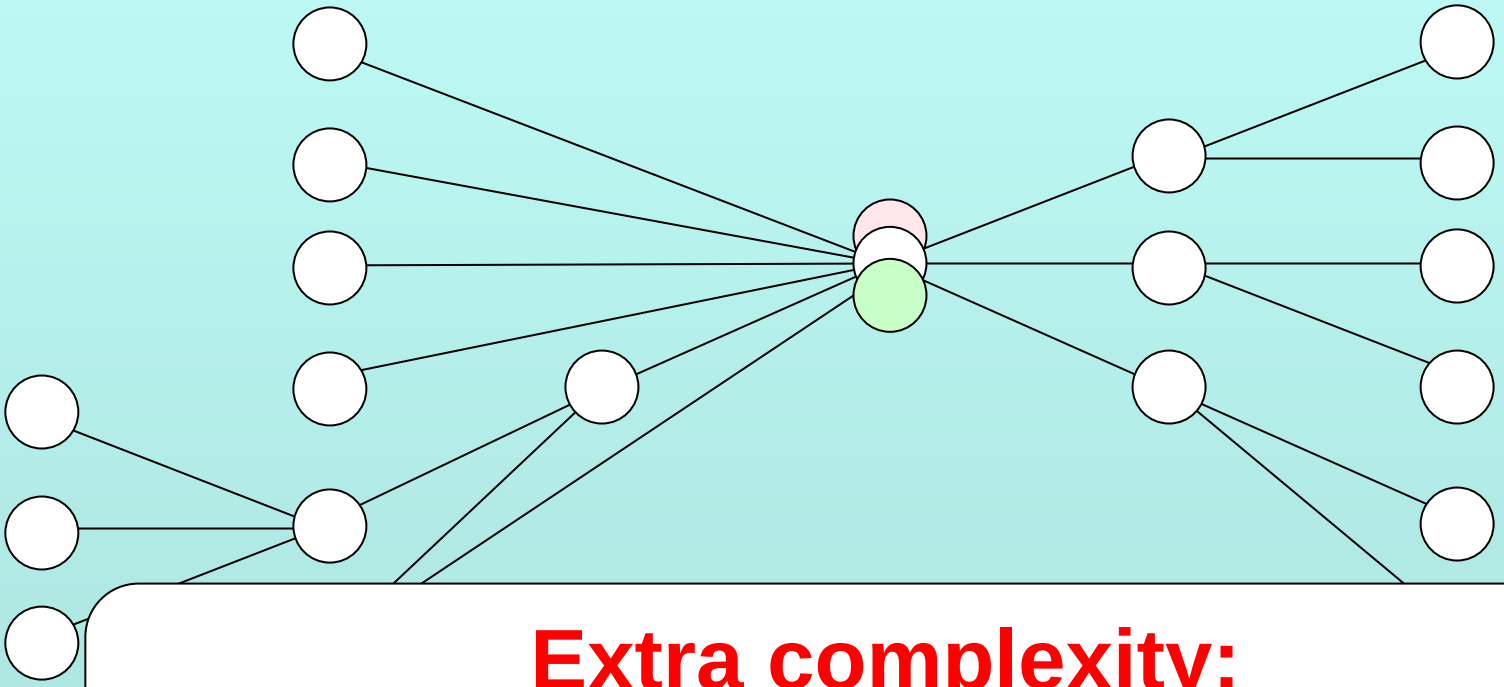


Extra complexity:

Data collected locally to meet local needs are not globally consistent - or even equivalent.

Multi-instrument Science

researchers **data resource(s)** **researchers**



Extra complexity:

Multiple centralized resources may exist, meaning there is no authoritative source.

Source II

Scope

Data-source Scope Issues

Problems occur at many levels:

- Integrating sequence data into GenBank
- Connecting GenBank with other genomic resources
- Connecting genomic data with other biological data
- Connecting all biological data with medical data
- Connecting all biomedical data with…

Source III

Solution: GenBank

GenBank as a False Model

- Classic Kuhnian paradigm science
- Simple, unambiguous data type (string)
- Symbiotic relationship with publishers
- Sequences are nouns, not verbs

Source IV

Real Solutions

Data-source Solutions

Institutional Solutions:

- Getting from RO1 science to international standards is too big a step
- We need solutions at the research institution level.
- Biomedical research organizations need to provide coherent support for biomedical IT, just as they do for biomedical bench research.
- Integrating institutional solutions is feasible; integrating individual lab solutions is not.

Philosophical Problems IDENTITY

Object Identity and Bioinformatics

- In any semantic web for the life sciences, no matter what technology is used, several needs must be met:
  - IDENTITY MANAGEMENT: It must be possible to identify unambiguously biological objects (more precisely to identify digital objects and associate them unambiguously with real-world biological objects).
  - IDENTITY ADJUDICATION: It must be possible to determine whether two different digital objects describe the same or different real world objects
  - REFERENTIAL INTEGRITY: It must be possible to make unambiguous, semantically well-defined assertions linking an object in one information resource to one or more objects in other information resources.

Object Identity and Bioinformatics

- In any semantic web for the life sciences, no matter what technology is used, several needs must be met:
  - RETAIL VS WHOLESALE CUSTOMERS: The semantic web must support the retail needs for coherence and the wholesale need for variation and disagreement (cf elephant and blind men story)
  - TRI_STATE LOGIC: Systems involving the classification of biological objects need tri-state logic to handle queries.
  - NO CURATION: In all but the best-funded public databases, there are no funded resources available for information curation.
  - CONSISTENCY IS IMPOSSIBLE: science consists of assertions and observations, not facts; assertions and observations can differ without being untrue.

Object Identity and Bioinformatics

- In any semantic web for the life sciences, no matter what technology is used, several needs must be met:
  - FINAL ONTOLOGY REQUIRES PERFECT KNOWLEDGE: In a context-free global environment, the data model must meet the requirements of all possible users (or fail for some users).
  - REALITY IS NOT NEGOTIABLE: The requirements for scientific information systems are determined by discovery, not negotiation.
  - SOCIOLOGICAL IMPEDIMENTS: Technological solutions must also meet sociological requirements; an information system that could manage useful information is a failure if many are unwilling to participate.
  - EXPECTATIONS MUST BE MANAGED: never forget,

    success = deliverables / expectations

Object Identity and Bioinformatics

- Concept of identity still subject to metaphysical distinctions:
  - NUMERICAL IDENTITY: one thing being the one and only such thing in the universe - e.g., there should be one and only human being associated with a patient ID
  - QUALITATIVE IDENTITY: two things being identical (sufficiently similar) in enough properties to be perfectly interchangeable (for some purpose) – e.g., there are many books associated with an ISBN identifier

Object Identity and Bioinformatics

- Properties are subject to identity-related distinctions:
  - ACCIDENTAL PROPERTIES: properties of an object that are contingent – that is, properties that are free to change without affecting the identity of the object
  - ESSENTIAL PROPERTIES: non-contingent properties – that is, properties which DEFINE the identity of the object and thus which cannot change without affecting the identity of the object (for some purpose)

Object Identity and Bioinformatics

- Properties are subject to identity-related distinctions:

Recognizing the distinction between essential and accidental properties will be critical in developing a successful identifier scheme for caBIG.

Especially challenging will be the fact that whether a particular property is essential or not is often context dependent.

Object Identity and Bioinformatics

- Properties are subject to identity-related distinctions:
  - INTRINSIC PROPERTIES: properties of an object that are properties of the thing itself
  - EXTRINSIC PROPERTIES: properties of the object that are properties of the object's relationship to other objects external to itself

Object Identity and Bioinformatics

- Properties are subject to identity-related distinctions:
  - INTRINSIC PROPERTIES: properties of an object that are properties of the thing itself
  - EXTRINSIC PROPERTIES: properties of the object that are properties of the object's relationship to other objects external to itself

Identifying tandemly duplicated genes is a perfect example of the need to distinguish between extrinsic and intrinsic properties.

Object Identity and Bioinformatics

- "Identification" is a process that reduces ambiguity. Ambiguity reducing identification can occur in a number of differ ways:
  - INDIVIDUAL SPECIFICATION: denoting an individual object without identifying either its class membership or its individuality - e.g., "this thing"
  - CLASS IDENTIFICATION: specifying than an object is a member of a class of objects that are sufficiently similar that the objects may be considered interchangeable (for some purpose) – e.g., "this book is Darwin's *Origin of Species"*
  - INDIVIDUAL IDENTIFICATION: specifying that an object is in fact a PARTICULAR genuinely unique object in the universe – e.g., this book is Darwin's own personally annotated copy of *Origin of Species*"

Object Identity and Bioinformatics

- "Identification" is a process that reduces ambiguity. Ambiguity reducing identification can occur in a number of differ ways:

Note that as we move along this continuum our notion of "essential properties" changes.

This shows again that the concept of identity can be context dependent.

PARTICULAR genuinely unique object in the universe – e.g., this book is Darwin's own personally annotated copy of *Origin of Species*"

Object Identity and Bioinformatics

- Digital identifiers (IDs) perform different kinds of identification:
  - REAL-WORLD IDENTIFIER: identifier serves as a digital token representing a real-world (i.e., non-digital) object (e.g., patient ID); this kind of identifier is often used to associated a digital object (bag of properties) with a real-world object
  - DIGITAL IDENTIFIER: identifier serves as a digital token representing a (published?) digital object (e.g., LSID or URL)

Object Identity and Bioinformatics

- Digital identifiers (IDs) perform different kinds of identification:
  - REAL-WORLD IDENTIFIER: identifier serves as a digital token representing a real-world (i.e., non-digital) object (e.g., patient ID); this kind of identifier is often used to associated a digital object (bag of properties) with a real-world object
  - DIGITAL IDENTIFIER: identifier serves as a digital token representing a (published?) digital object (e.g., LSID or URL)

This distinction can be hard to make:
What does an IP address identify?

Object Identity and Bioinformatics

- Digital identifiers (IDs) can truly identify particular objects or they can merely specify singular objects, with no guarantee of what that singular object is:
  - IDENTIFICATION: the same LSID should always return exactly the same (bit for bit) digital object
  - SPECIFICATION: the same URL is not guaranteed to return the same thing twice

Object Identity and Bioinformatics

Note that these two situations really just represent the opposite ends of a continuum:

At one end EVERY property is essential – at the other end NO property is essential.

At both ends, the relationship of identifier to object is clear. In between, this clarity does not exist and contention can and will exist between identifiers and properties (e.g., the same human being could accidentally be assigned two patient IDs, but we could infer identity from the essential properties).

Object Identity and Bioinformatics

- Different methods exist for answering the question whether or not two objects are the same:
  - DEMONSTRATED IDENTITY: the identifiers are the same and the essential properties are the same
  - INFERRED IDENTITY: the identifiers are different but the essential properties are the same
  - INFERRED NON-IDENTITY: the identifiers are the same, but the essential properties are different
  - ASSERTED IDENTITY: the identifiers are the same, but the state of the essential properties are unknown

Object Identity and Bioinformatics

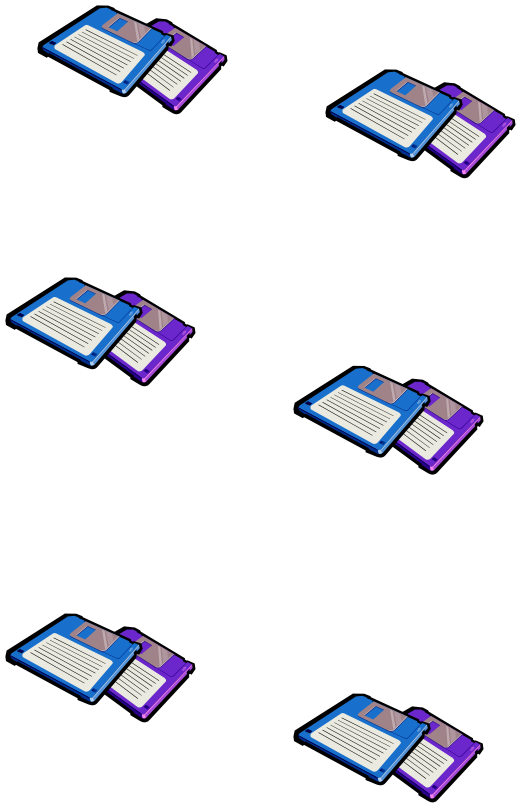
- Different methods exist for answering the question whether or not two objects are the same:
  - DEMONSTRATED IDENTITY: the identifiers are the same and the

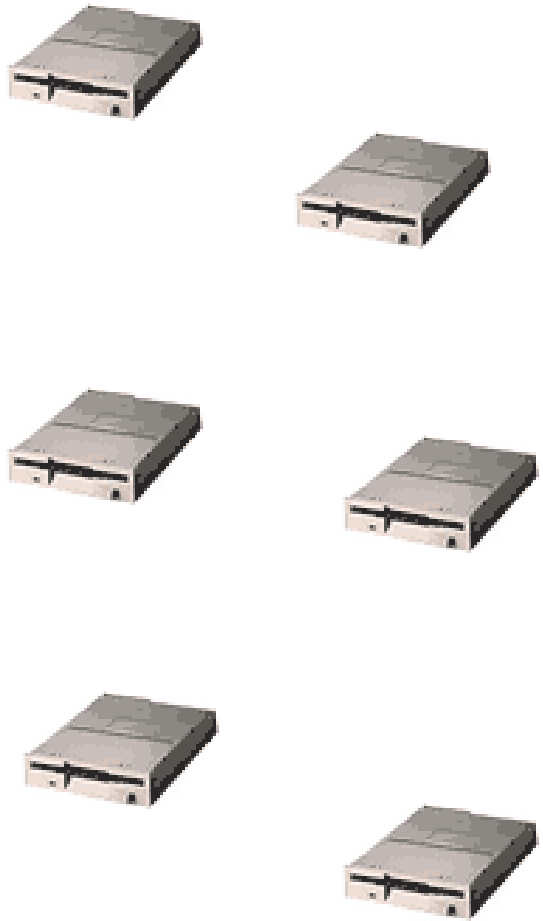
With checksums, LSIDs are an instance of DEMONSTRATED identity.

Without checksums, LSIDs are an instance of ASSERTED identity.

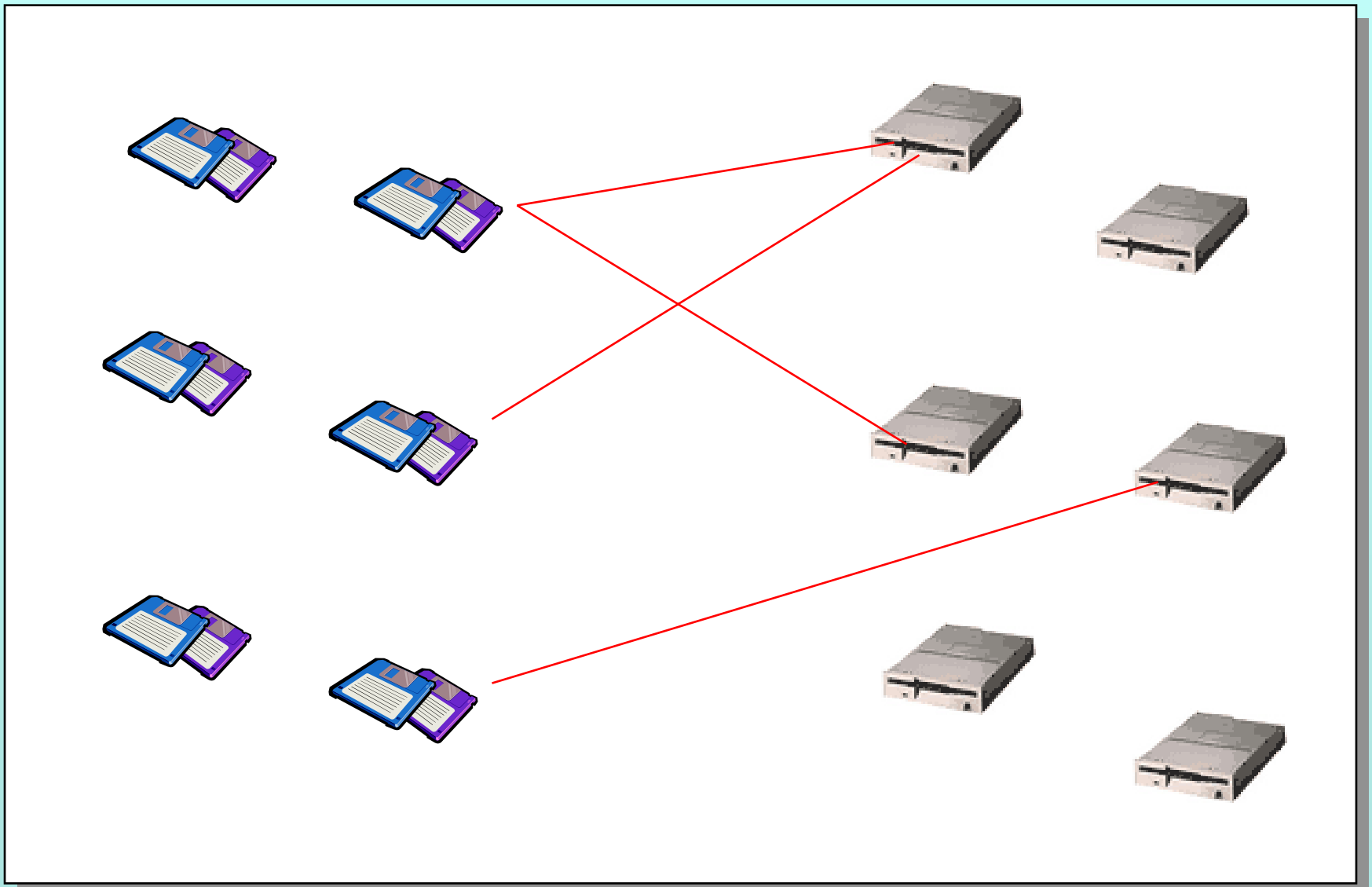
Standards

Bad Data-exchange Standard

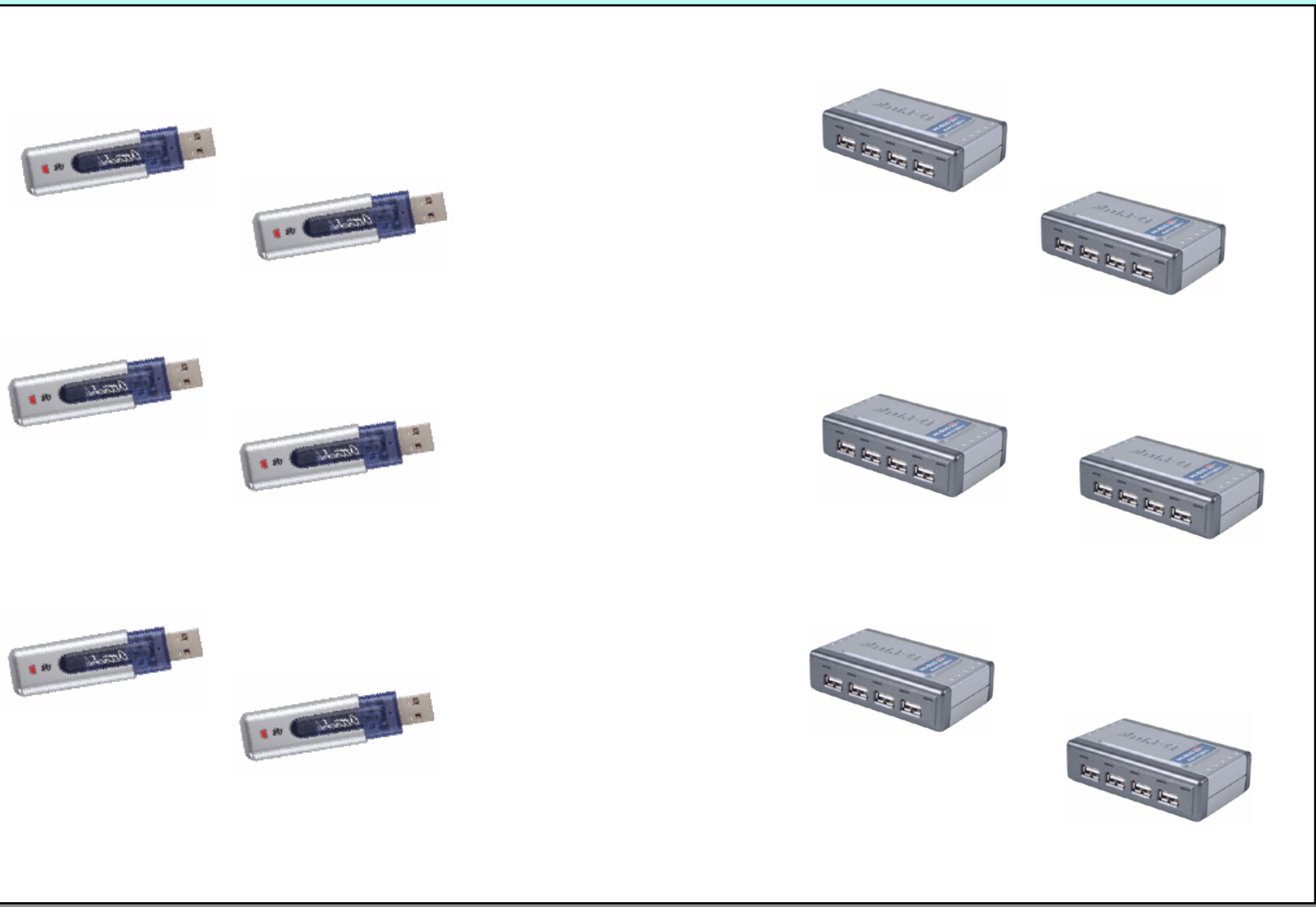




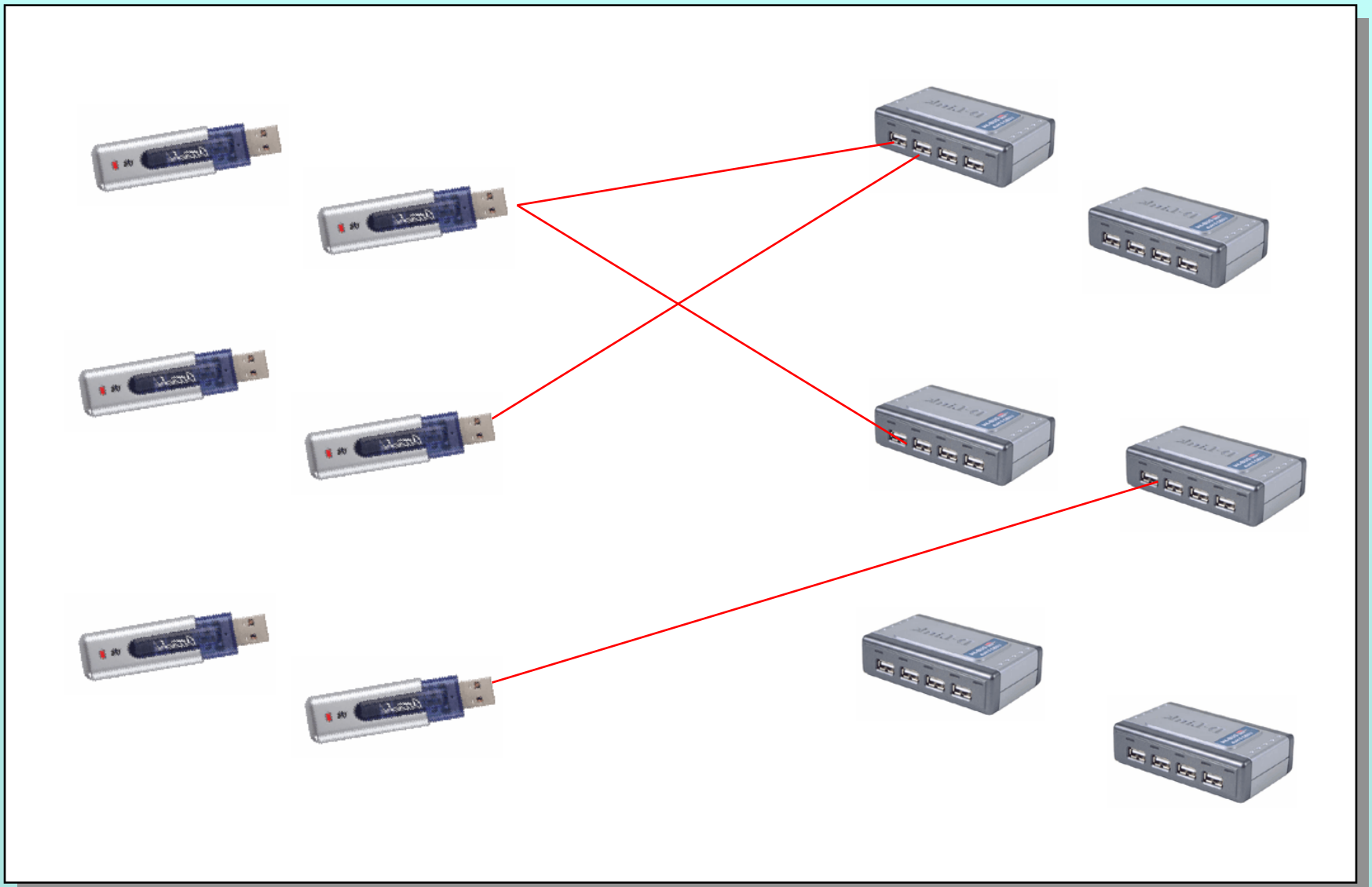
Bad Data-exchange Standard



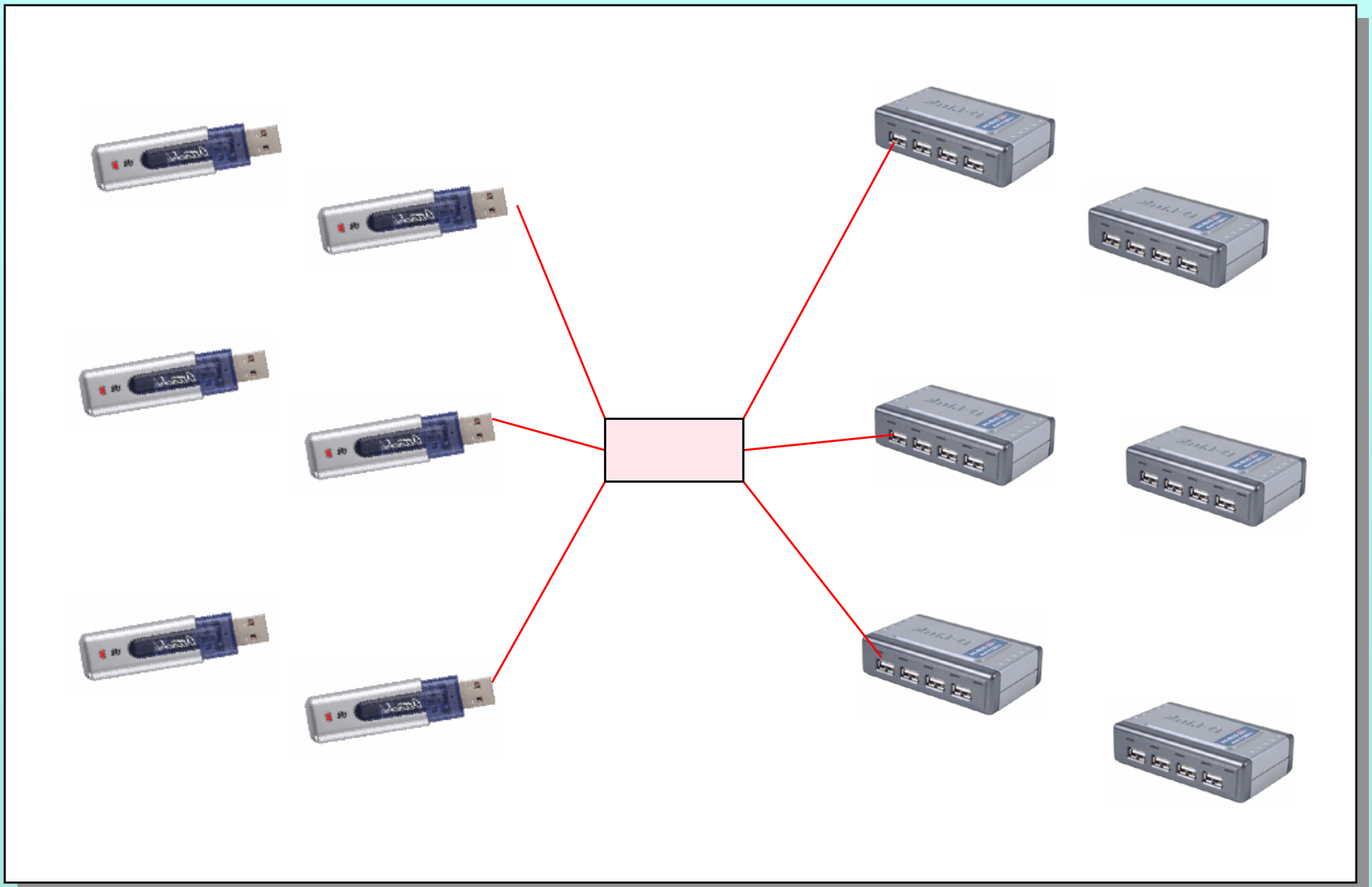
Good Data-exchange Standard



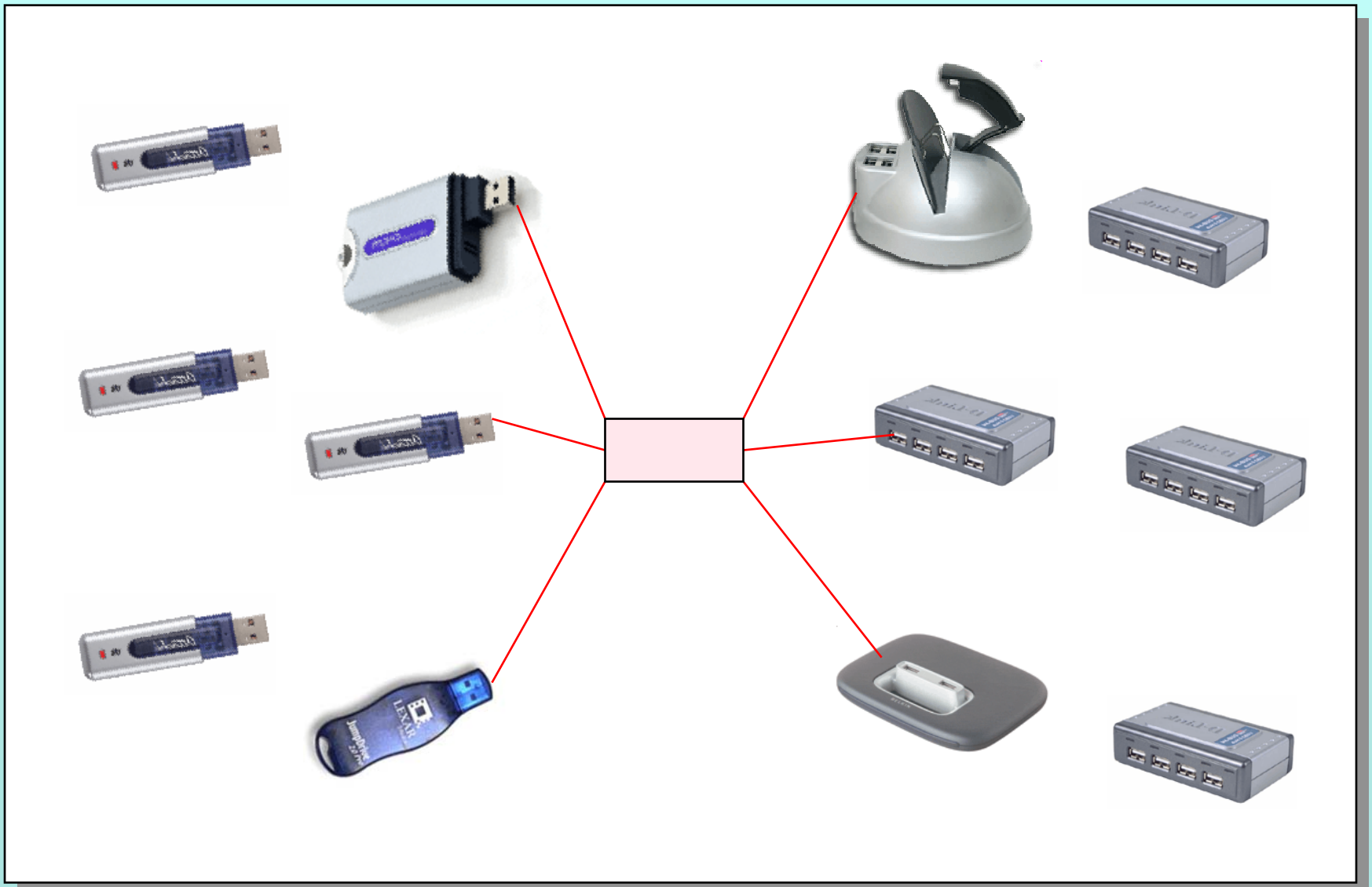
Good Data-exchange Standard



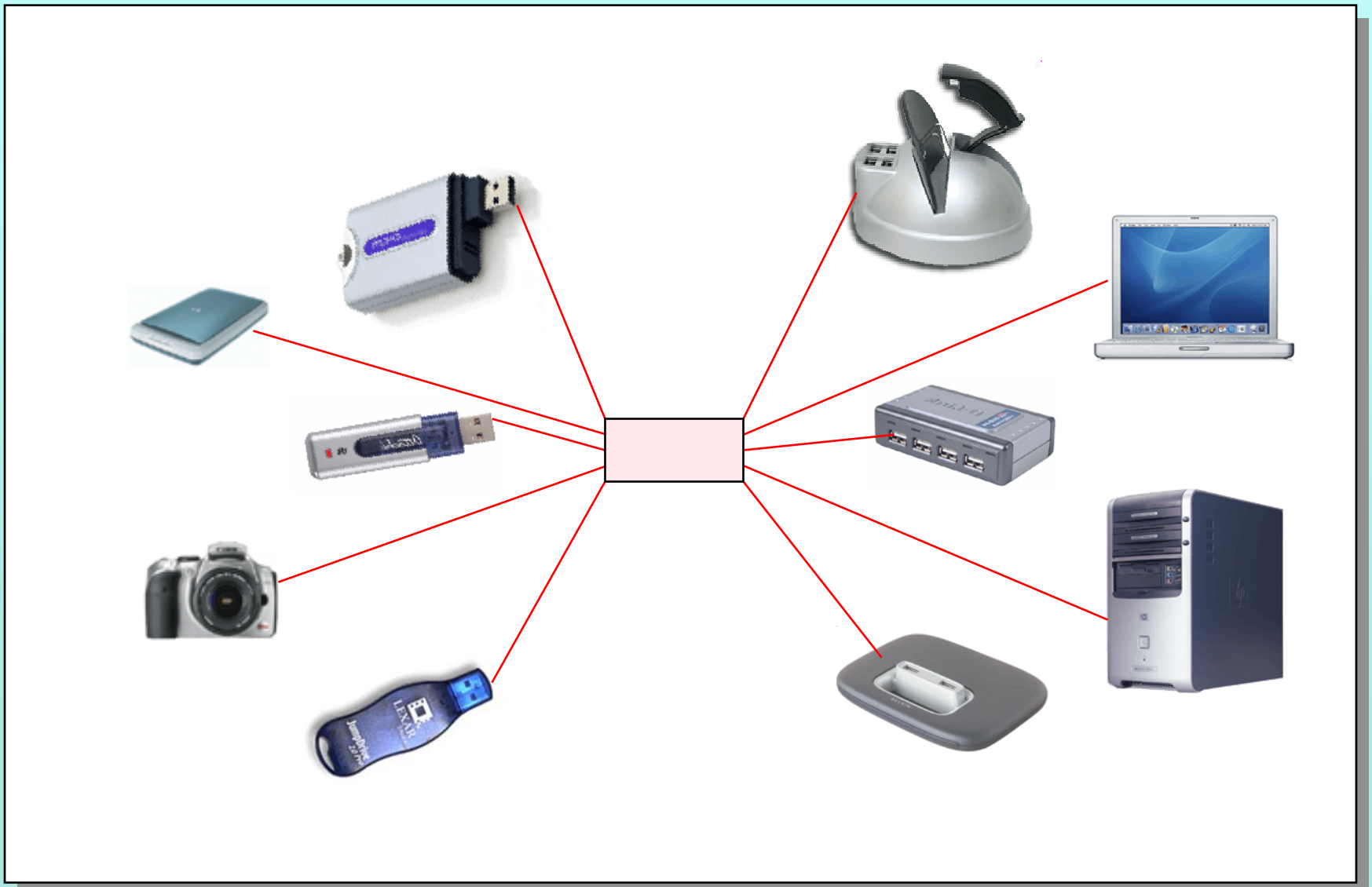
Good Data-exchange Standard



Good Data-exchange Standard

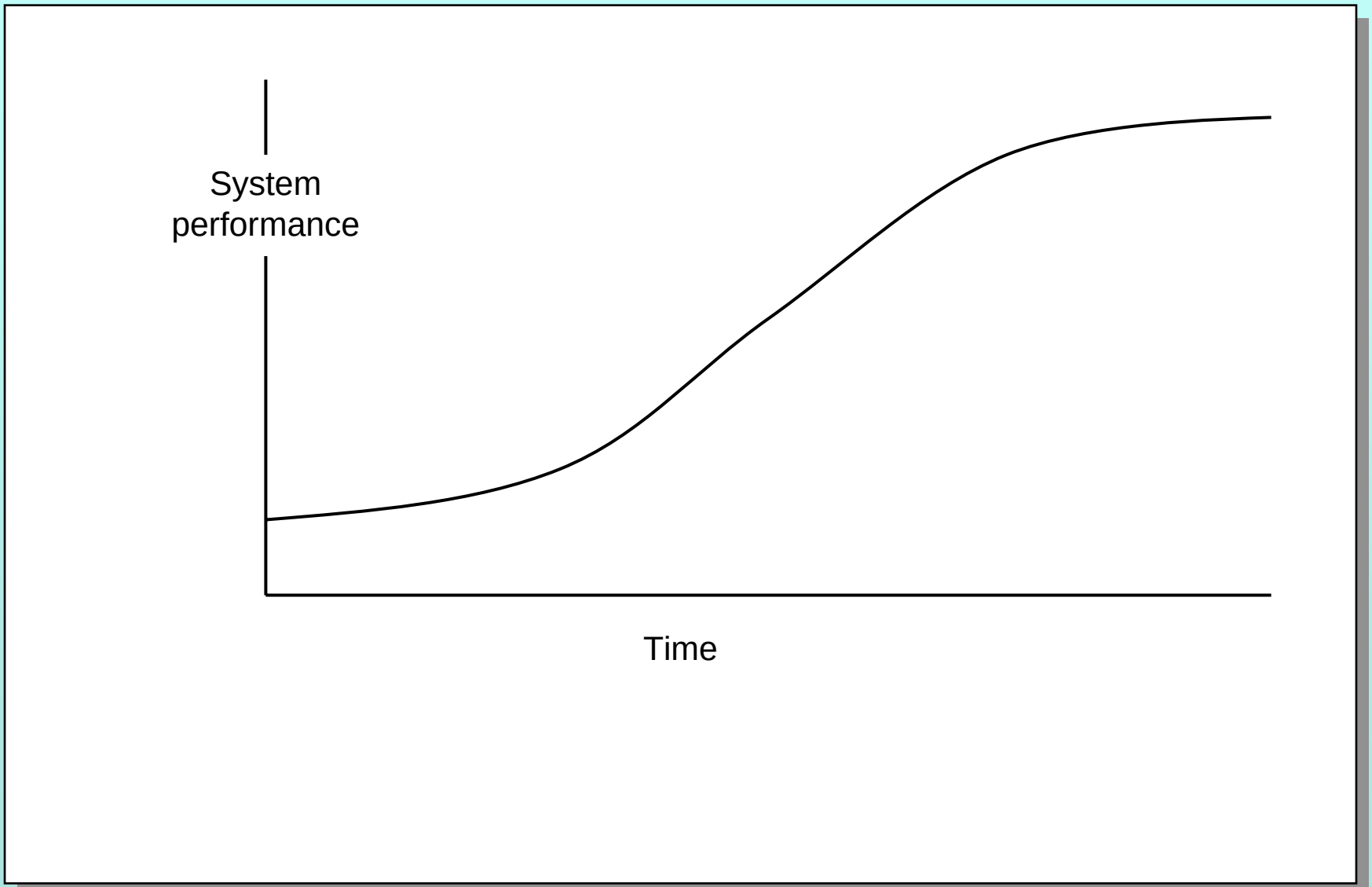


Good Data-exchange Standard

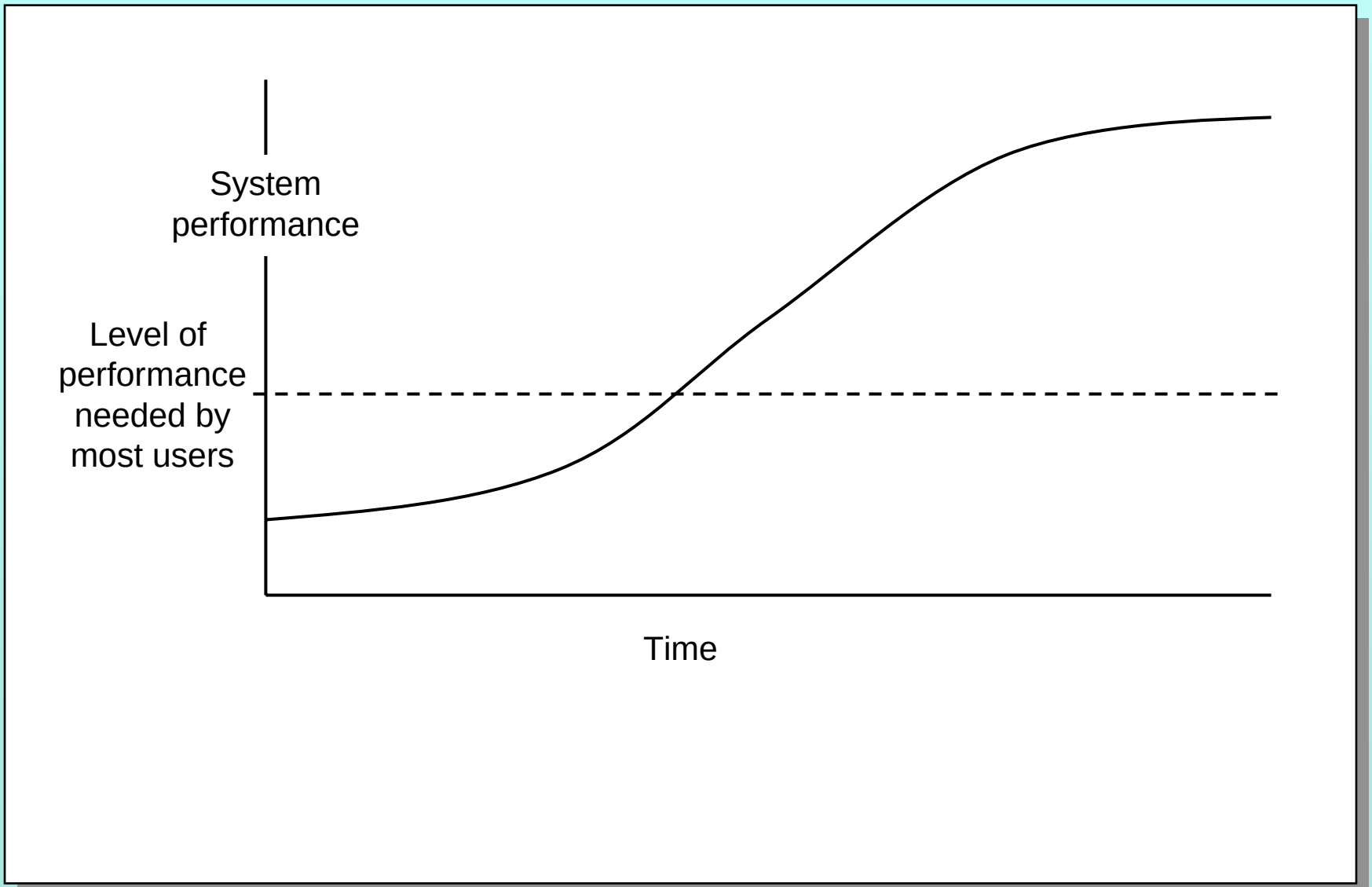


Industry Trends

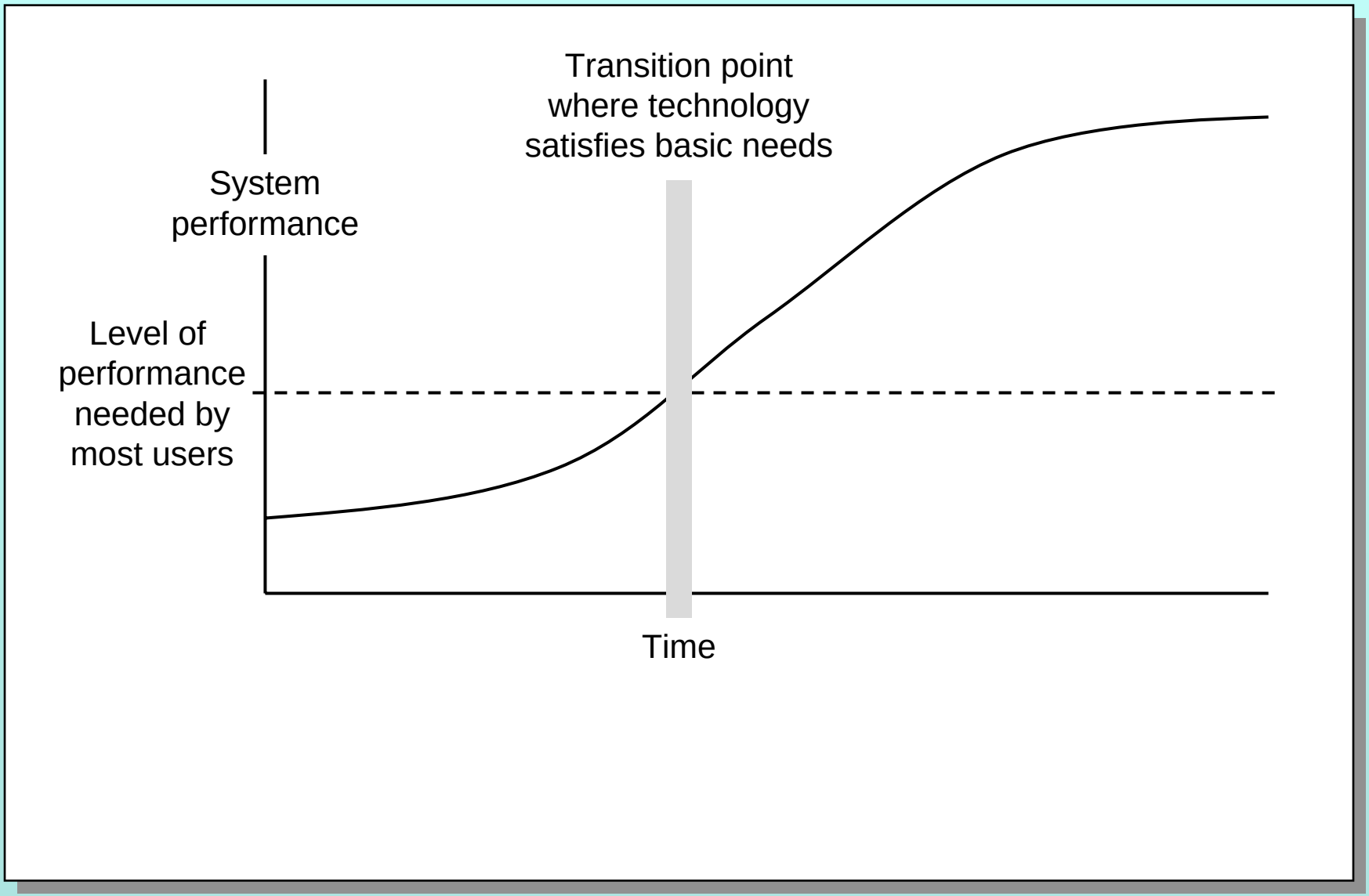
Industry Trends



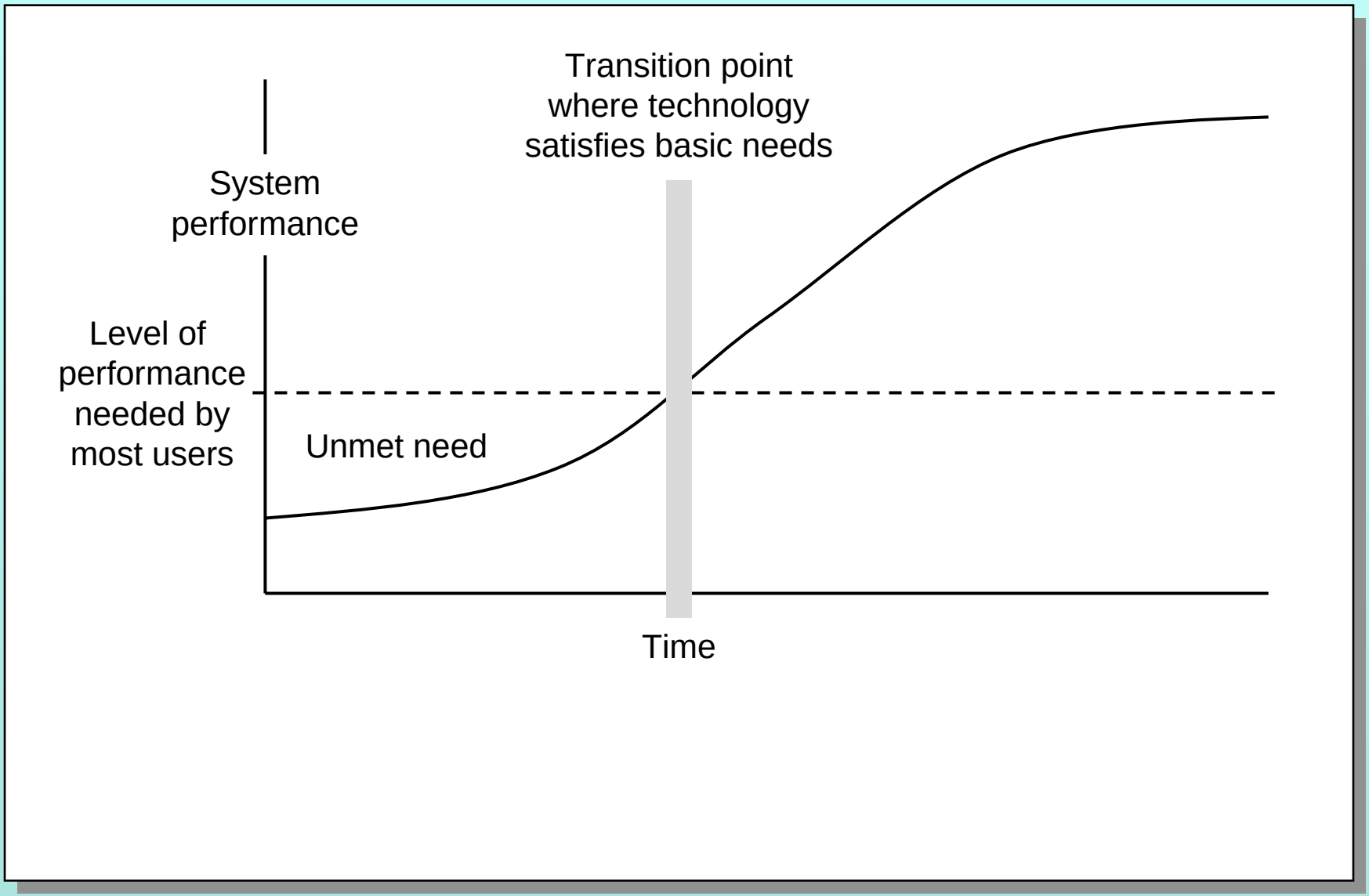
Industry Trends



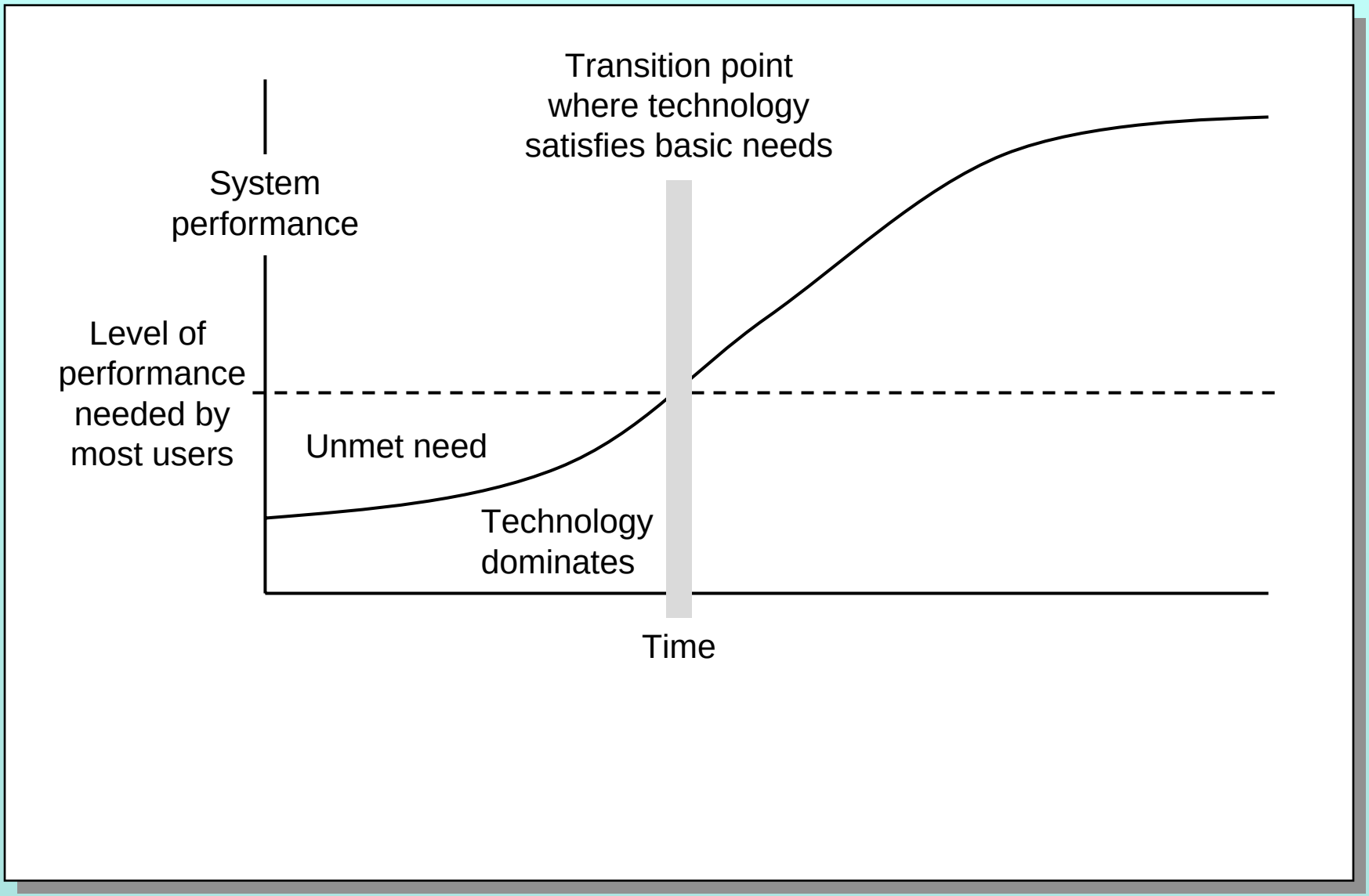
Industry Trends



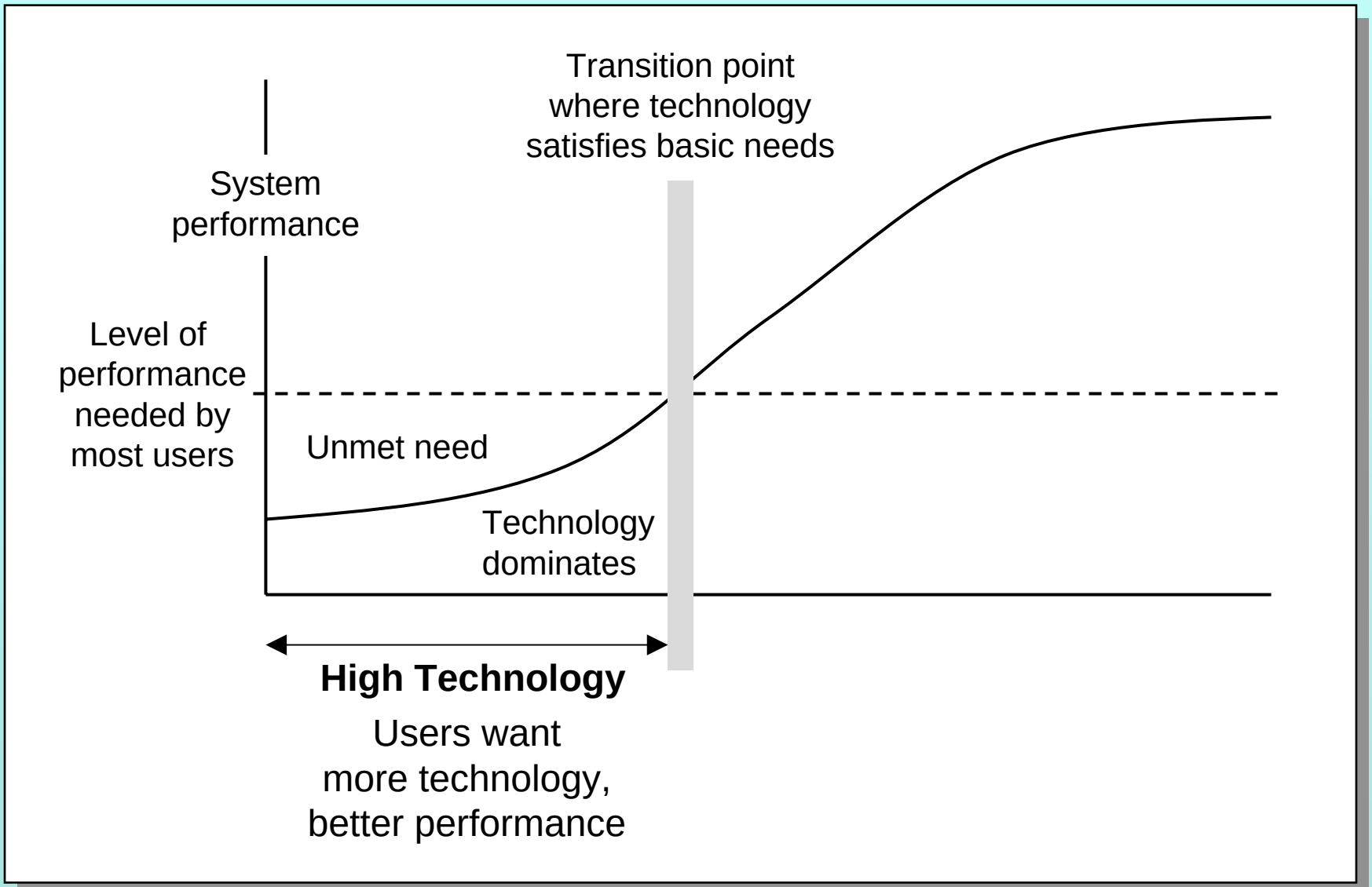
Industry Trends



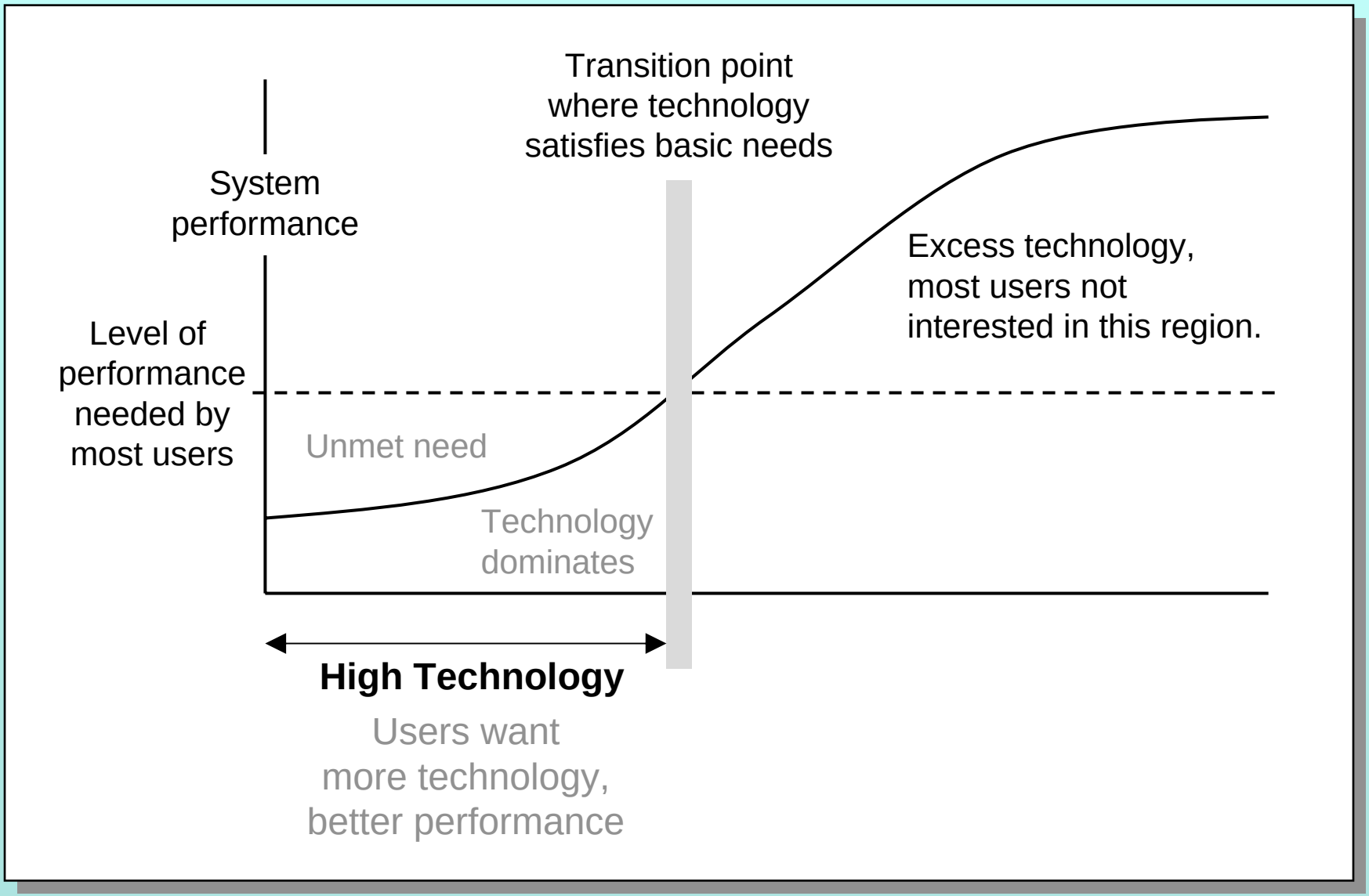
Industry Trends



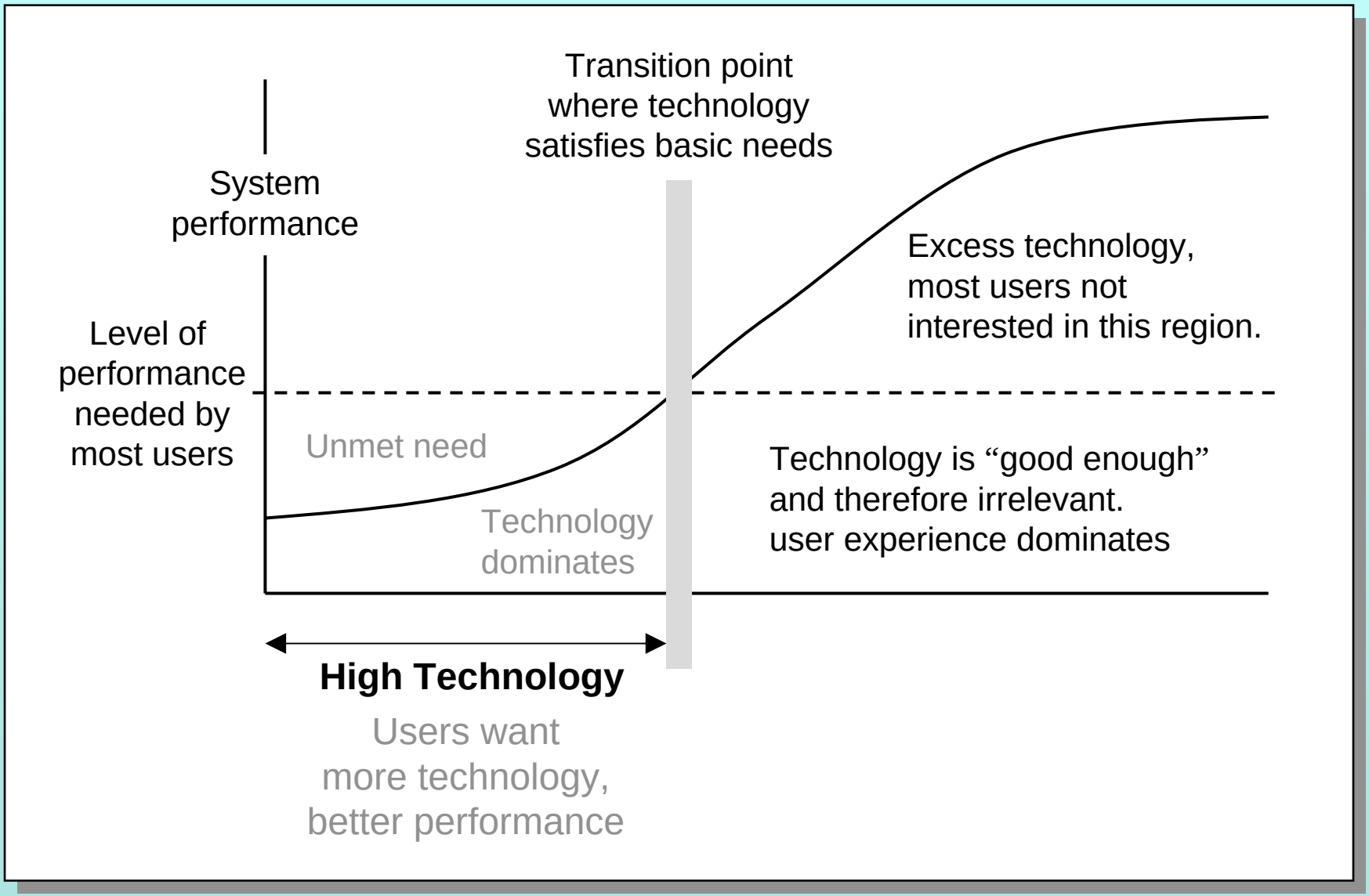
Industry Trends



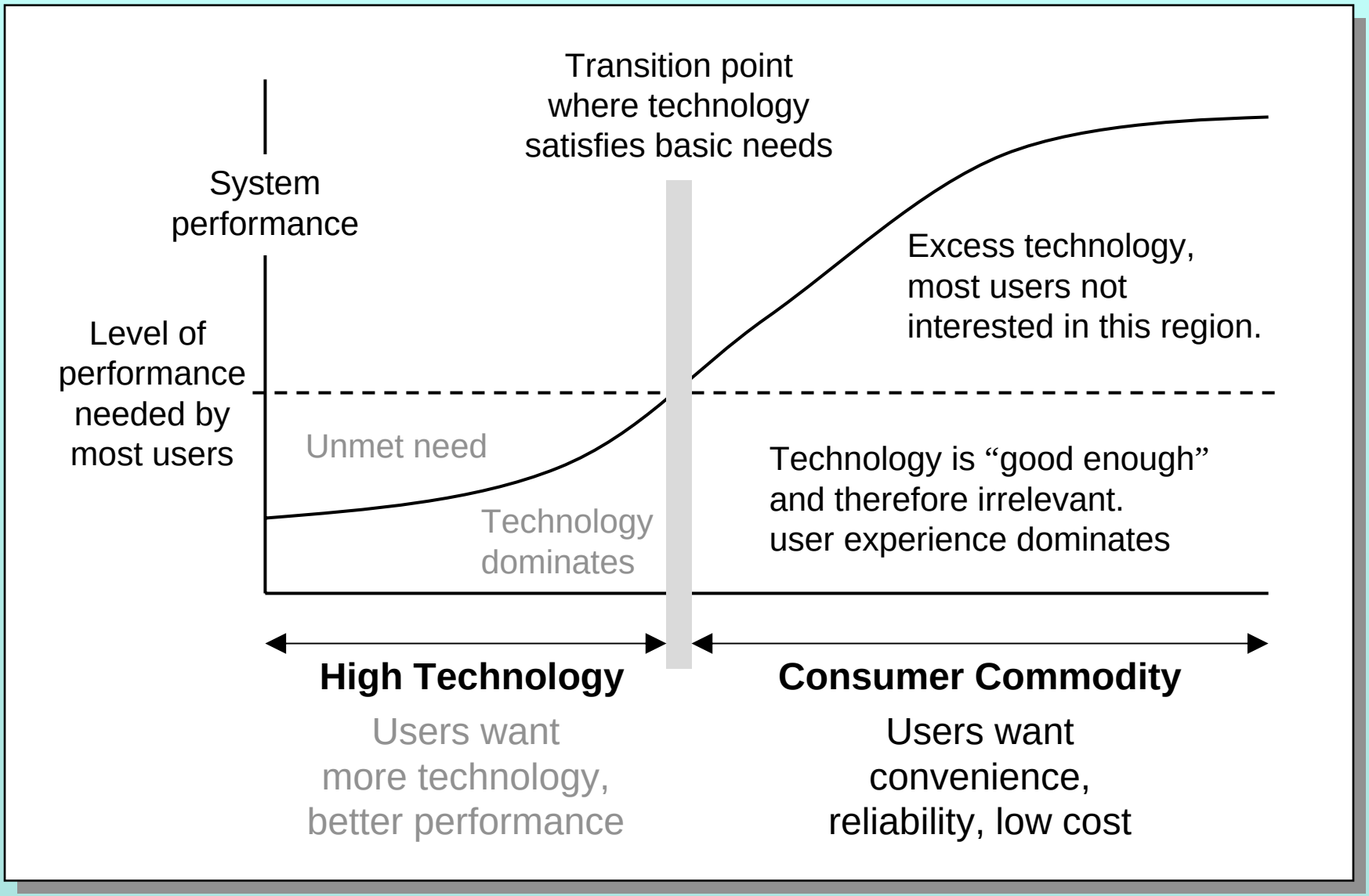
Industry Trends



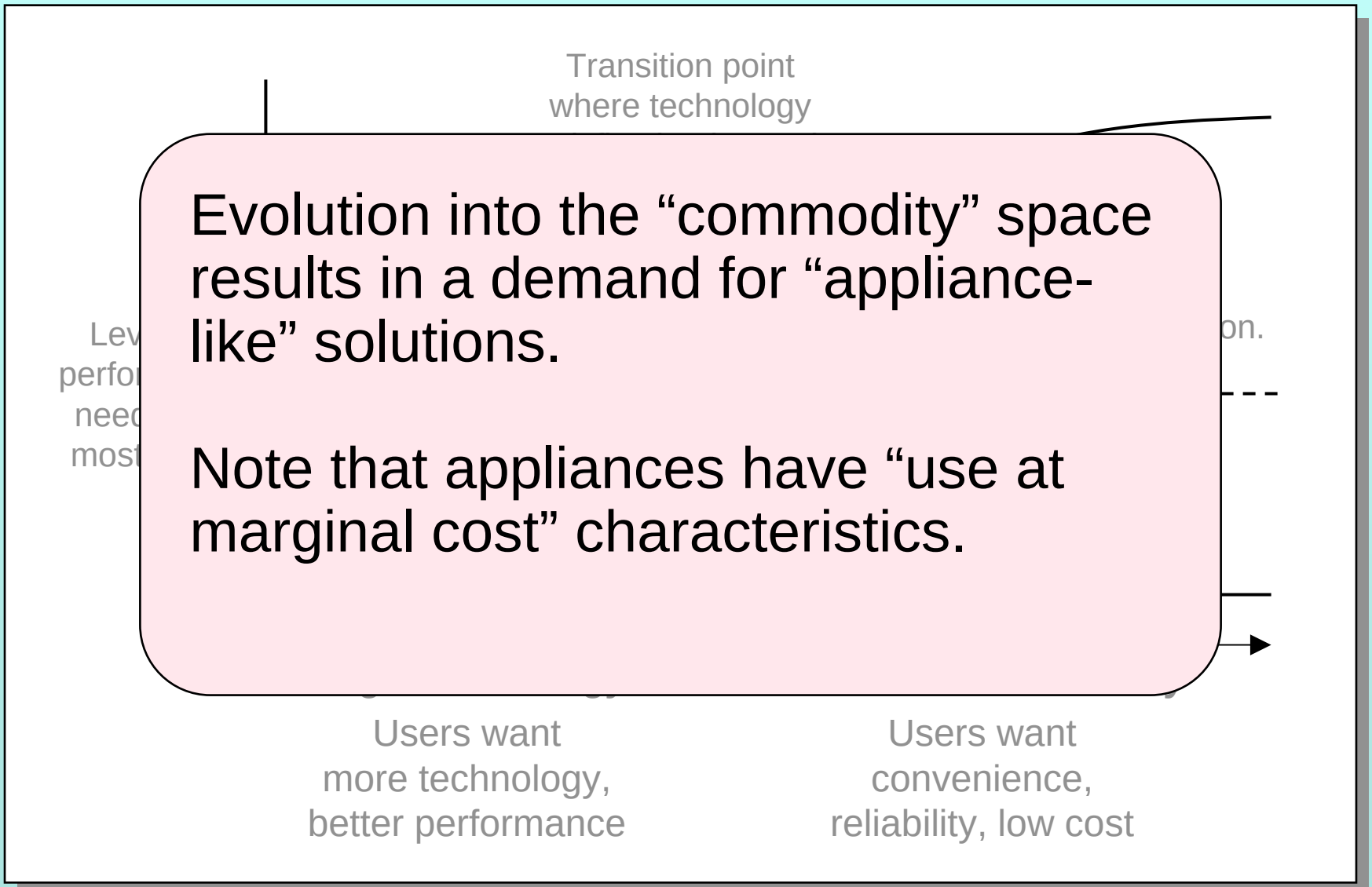
Industry Trends



Industry Trends



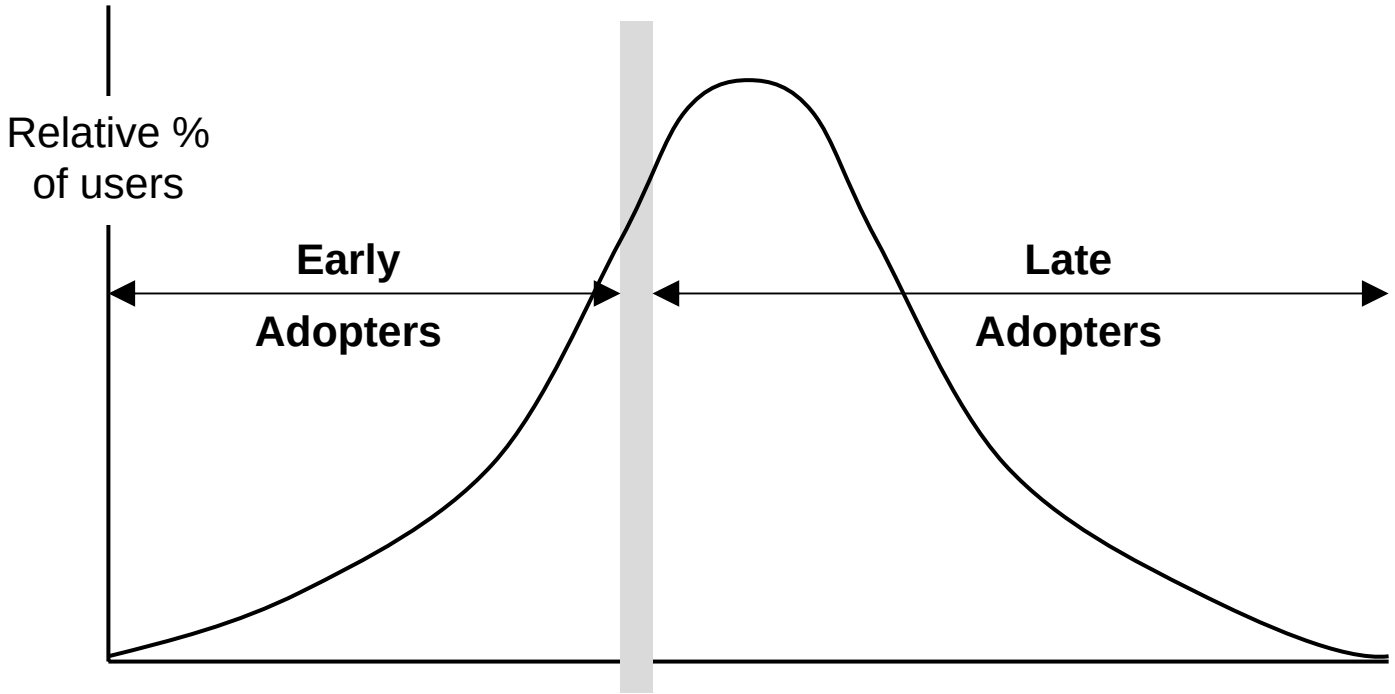
Industry Trends



Evolution into the “commodity” space results in a demand for “appliance-like” solutions.

Note that appliances have “use at marginal cost” characteristics.

Industry Trends



Early adopters drive the technical capabilities of the system, forcing the bar of acceptable performance upward. However, at some point the bar stabilizes and late adopters come to dominate the market for (and hence the design of) technology products.

http://www.esp.org/rjr/bcc-2004.pdf