

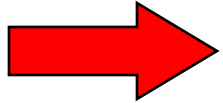
# **Data Management in the Research Laboratory: The Sine Qua Non of 21st-Century Science**

( <http://www.esp.org/rjr/mayo-2005.pdf>)

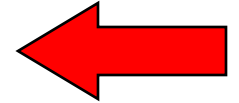
---

Robert J. Robbins  
rrobbins@fhcrc.org  
(206) 667 4778

# **Data Management in the Research Laboratory: The Sine Qua Non of 21st-Century Science**



( <http://www.esp.org/rjr/mayo-2005.pdf>)



---

Robert J. Robbins  
rrobbins@fhcrc.org  
(206) 667 4778

# Abstract

Biomedical researchers are now awash in data. Technological developments, stimulated in part by the successful human genome project, have increased the data-production capabilities of even the smallest laboratory by staggering amounts. High-throughput sequencing facilities can now produce the full genomic sequence of a bacterial pathogen in less than 24 hours. GenBank now adds more sequence data every few hours than it added in the first few years of its existence.

Formal data management is rapidly becoming a requirement in the modern research laboratory. In fact, many laboratories are now finding research-preparation logistics and data management to be the rate-limiting step in their work. Some studies indicate that proper data and logistics management can more than double the scientific output of a small lab. Although commercial laboratory information management systems (LIMs) exist, their cost and complexity make them impractical in the small laboratory. Many laboratories, of necessity, rely upon the "Microsoft LIMs solution" - lots of Excel spreadsheets and the occasional Access database. MS-LIMs is clearly inadequate for meeting the needs ahead.

In this talk we will consider the data-management challenges (and opportunities) faced by the typical research laboratory and some of the options available for meeting those challenges. We will also consider some of the social complexities associated with laboratory data management (whose job is it, anyway?), as well as some of the technical complexities associated with the need for data-management systems to interoperate between laboratories (and even institutions). We will examine some of the large- and small-scale efforts (e.g., caBIG and GeMS) underway to address laboratory data-management issues, and, time permitting, we will offer some predictions about likely future paths in laboratory data management.

# Topics

---

- In-Lab Data Management: Pressing need or bogus issue?
- Things are Different Now
  - Increased (and increasing) data complexity
  - Increased residual data value
  - Increased data volume
- Awash in Data
  - In-Lab Data-Generating Capacity
  - Public Data Explosion

# Topics

---

- 21<sup>st</sup> Century Science: Post-Genome Era
  - New Tools / New Mindset
  - Affects more than just genetics
- Future Vision: Biomedical research is thoroughly data-driven and all researchers have seamless access to vast quantities of reliable data
- Challenge of Lab Data Management
  - Seems too easy to be a real issue
  - Seems too hard to be done well
  - Whose job is it, anyway?

# Topics

---

- Impediments to Biological Data Management
  - Data Source Problems
  - Data Model Problems
  - Philosophical Problems
  - Budget Problems – Reality Check
- The Future
  - Standards
  - caBIG
  - Industry Trends – Information Appliances
  - GeMS

# Introduction

---

Bogus Issue?

# Personal Opinion

---

Data problems are dull and people who work on them are dull.

James Watson, some time in the 1980s.



# Simple Question

---

A biologist says, “Data management is necessary for my research, but not especially important. Personally, I’m just not interested in the details of how it’s done. I have one of my students (or techs) handle it.”

# Simple Question

---

A biologist says, “Data management is necessary for my research, but not especially important. Personally, I’m just not interested in the details of how it’s done. I have one of my students (or techs) handle it.”

Twenty years ago this biologist would have been described as:

# Simple Question

---

A biologist says, “Data management is necessary for my research, but not especially important. Personally, I’m just not interested in the details of how it’s done. I have one of my students (or techs) handle it.”

Twenty years ago this biologist would have been described as:

**TYPICAL**

# Simple Question

---

A biologist says, “Data management is necessary for my research, but not especially important. Personally, I’m just not interested in the details of how it’s done. I have one of my students (or techs) handle it.”

Twenty years from now this biologist will be described as:

# Simple Question

---

A biologist says, “Data management is necessary for my research, but not especially important. Personally, I’m just not interested in the details of how it’s done. I have one of my students (or techs) handle it.”

Twenty years from now this biologist will be described as:

**INCOMPETENT**

# Simple Fact

---

In the post-genomic world, much bio-medical research is impossible without adequate information infrastructure.

# Simple Fact

---

In the post-genomic world, much biomedical research is impossible without adequate information infrastructure.

Quality IT operations, within the institution and within the lab, are now critically important to the mission of biomedical research organizations.

# Introduction

---

Issues



# **Awash in Data**

---

**Massive Local Capacity**

# Small Lab Data-Generation

---

## Typical Projects in Geraghty Lab at FHCRC:

- **Primary data acquisition:** Sequence analysis of the rhesus macaque MHC – how similar/different is the rhesus MHC from human and what are the potential consequences of these differences towards ongoing clinical research?

# Small Lab Data-Generation

---

## Typical Projects in Geraghty Lab at FHCRC:

- **Primary data acquisition:** Sequence analysis of the rhesus macaque MHC – how similar/different is the rhesus MHC from human and what are the potential consequences of these differences towards ongoing clinical research?
- **Variation discovery:** Genomic analysis of the KIR locus in humans – what is the extent of diversity of this locus and can we define a better sequence framework on which to build genetic tests?

# Small Lab Data-Generation

---

## Typical Projects in Geraghty Lab at FHCRC:

- **Primary data acquisition:** Sequence analysis of the rhesus macaque MHC – how similar/different is the rhesus MHC from human and what are the potential consequences of these differences towards ongoing clinical research?
- **Variation discovery:** Genomic analysis of the KIR locus in humans – what is the extent of diversity of this locus and can we define a better sequence framework on which to build genetic tests?
- **Correlating genotype with clinical phenotype:** Host Genomic Polymorphisms and Immune Reconstitution – are genetic factors responsible for immune reconstitution after antiretroviral therapy in AIDS patients?

# Small Lab Data-Generation

## Typical Projects in Geraghty Lab at FHCRC:

- **Primary data acquisition:** Sequence analysis of the rhesus macaque MHC – how similar/different is the rhesus MHC from human and what are the potential consequences of these differences towards ongoing clinical research?
- **How much data will be generated in these studies?**
- **Correlating genotype with clinical phenotype:** Host Genomic Polymorphisms and Immune Reconstitution – are genetic factors responsible for immune reconstitution after antiretroviral therapy in AIDS patients?

# Small Lab Data-Generation

---

## Typical Projects in Geraghty Lab at FHCRC:

- **Primary data acquisition:** Sequence analysis of the rhesus macaque MHC – how similar/different is the rhesus MHC from human and what are the potential consequences of these differences towards ongoing clinical research?

# Small Lab Data-Generation

## Typical Projects in Geraghty Lab at FHCRC:

- **Primary data acquisition:** Sequence analysis of the rhesus macaque MHC – how similar/different is the rhesus MHC from human and what are the potential consequences of these differences towards ongoing clinical research?

### **Data Management Challenge:**

**59 subprojects (each the shotgun sequencing of a 180,000 bp BAC), 150,000 trace files, data-sharing across two collaborating labs, submission of data to public databases.**

# Small Lab Data-Generation

---

## Typical Projects in Geraghty Lab at FHCRC:

- **Primary data acquisition:** Sequence analysis of the rhesus macaque MHC – how similar/different is the rhesus MHC from human and what are the potential consequences of these differences towards ongoing clinical research?
- **Variation discovery:** Genomic analysis of the KIR locus in humans – what is the extent of diversity of this locus and can we define a better sequence framework on which to build genetic tests?



## Data Management Challenge:

50 subprojects (each consisting of 500 sequence traces for each of 5 fosmids from 1 of 50 chromosomes), 125,000 trace files, data analysis, real-time data sharing with multiple collaborators, submission of data to public databases.

- **Variation discovery:** Genomic analysis of the KIR locus in humans – what is the extent of diversity of this locus and can we define a better sequence framework on which to build genetic tests?

# Small Lab Data-Generation

---

## Typical Projects in Geraghty Lab at FHCRC:

- **Primary data acquisition:** Sequence analysis of the rhesus macaque MHC – how similar/different is the rhesus MHC from human and what are the potential consequences of these differences towards ongoing clinical research?
- **Variation discovery:** Genomic analysis of the KIR locus in humans – what is the extent of diversity of this locus and can we define a better sequence framework on which to build genetic tests?
- **Correlating genotype with clinical phenotype:** Host Genomic Polymorphisms and Immune Reconstitution – are genetic factors responsible for immune reconstitution after antiretroviral therapy in AIDS patients?

# Small Lab Data-Generation

## Typical Projects in Geraghty Lab at FHCRC:

- **Primary data acquisition:** Sequence analysis of the rhesus

### **Data Management Challenge:**

**32 separate loci examined; 1,000 individual DNAs;  
64,000 trace files, heterozygous data interpretation,  
data sharing across multiple collaborating labs.**

genetic tests?

- **Correlating genotype with clinical phenotype:** Host Genomic Polymorphisms and Immune Reconstitution – are genetic factors responsible for immune reconstitution after antiretroviral therapy in AIDS patients?

# Small Lab Data-Generation

---

## Data Generation Summary:

- **Rhesus macaque MHC sequencing project:** 59 subprojects, 150,000 trace files, data-sharing across two collaborating labs, submission of data to public databases.
- **Genomic analysis of KIR locus:** 50 subprojects, 125,000 trace files, data analysis, real-time data sharing with multiple collaborators, submission of data to public databases.
- **Host Genomic Polymorphisms and Immune Reconstitution:** 64,000 trace files, heterozygous data interpretation, data sharing across multiple collaborating labs.

# Small Lab Data-Generation

## Data Generation Summary:

- **Rhesus macaque MHC sequencing project:** 59 subprojects, labs,

More than 250,000 trace files generated across more than 100 subprojects, with data to be shared across multiple collaborating laboratories.

This is not a problem to be solved using the MS LIMs solution, implemented by a couple of hard-working students or techs.

# Small Lab Data-Generation

---

## Major Laboratory Challenges:

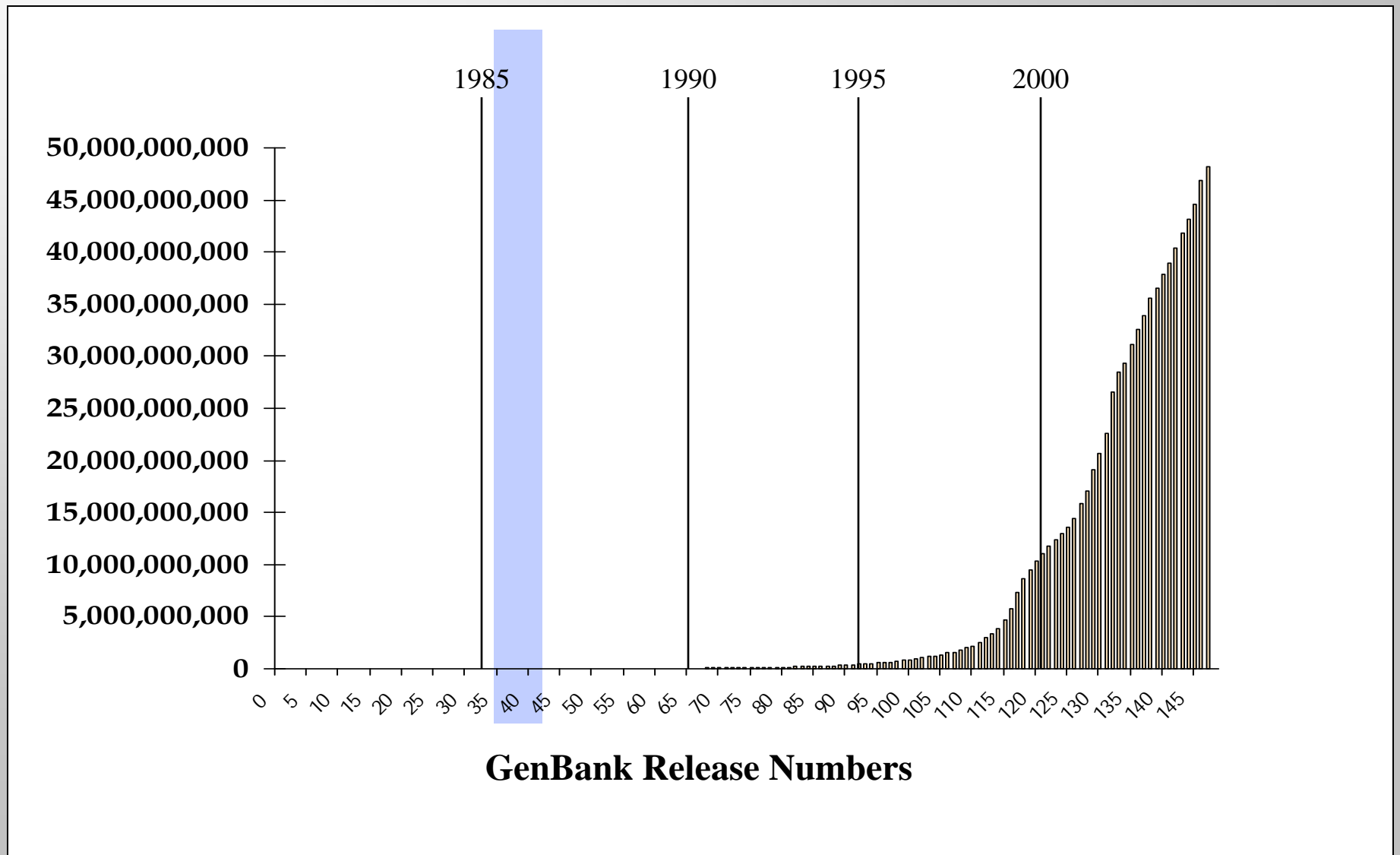
- Tracking laboratory throughput
- Organization of original data and meta data (machine, reagents, quality, etc.)
- Data sharing
- Cost tracking
- Creating a modular and extensible framework for future applications (HTR, Taqman, etc.)

# **Awash in Data**

---

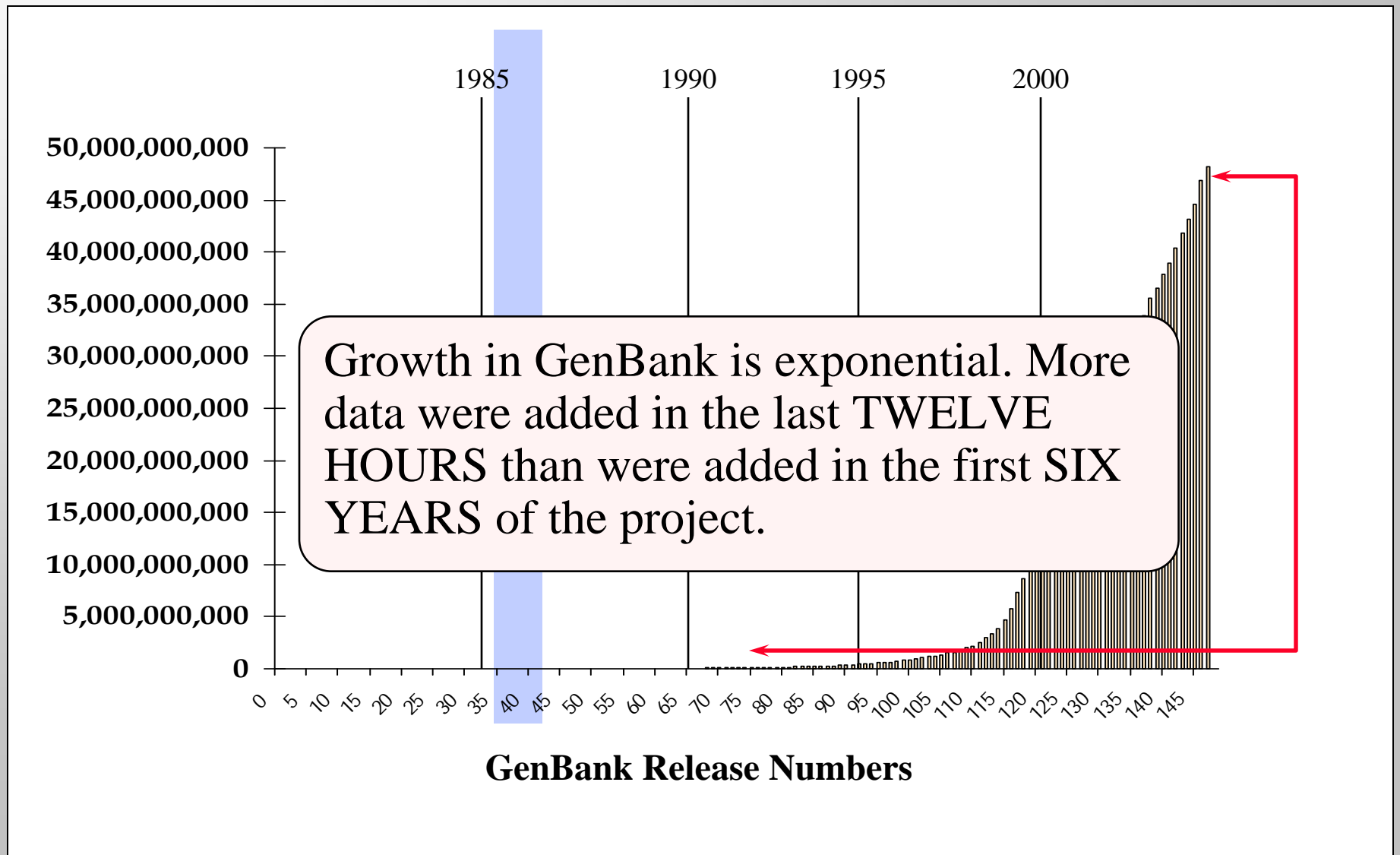
**Public Data Explosion**

# Base Pairs in GenBank

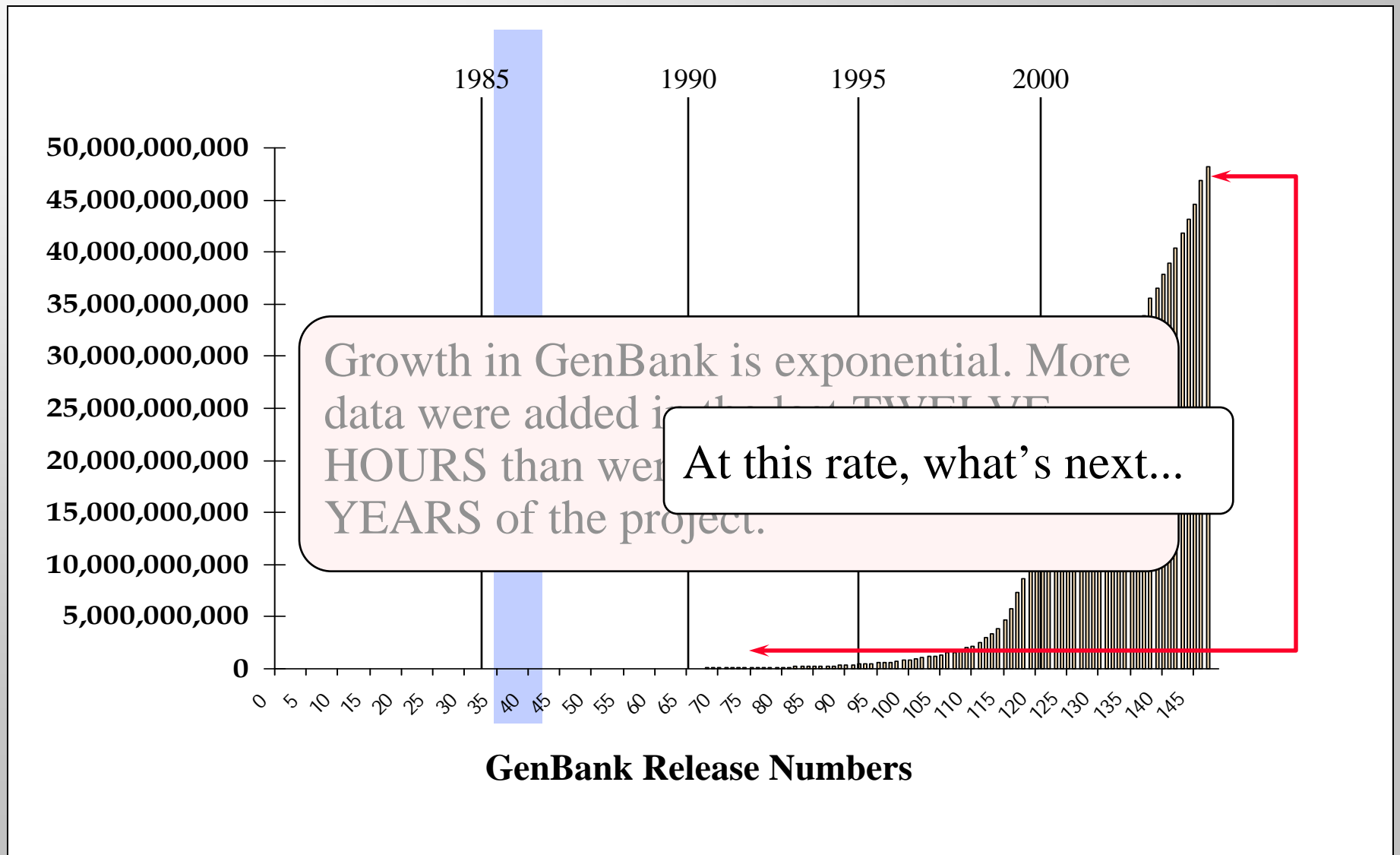




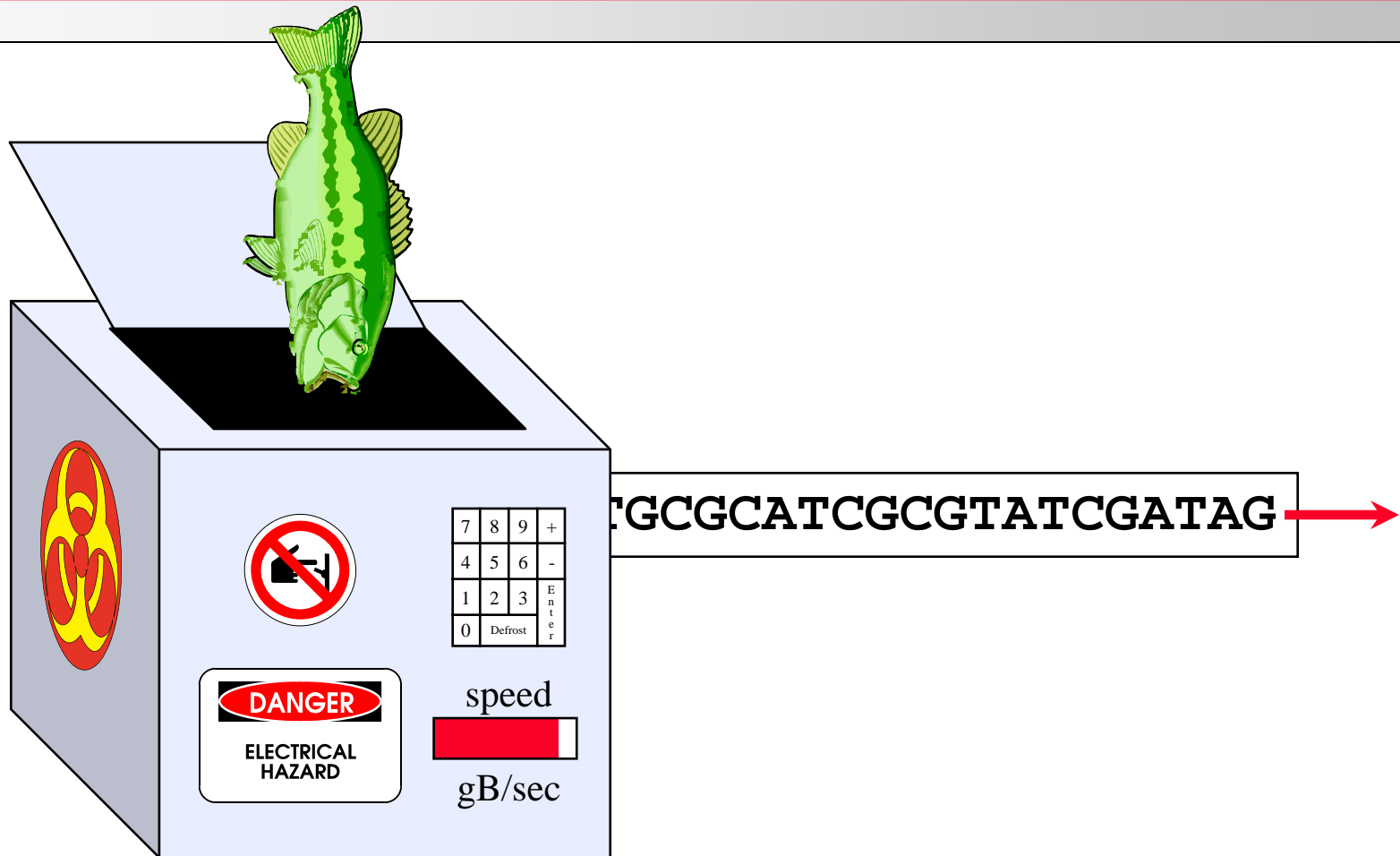
# Base Pairs in GenBank



# Base Pairs in GenBank



# ABI Bass-o-Matic Sequencer



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

# Aside: Joint Genome Institute

---

Like many things in the past ten years, the Bass-o-Matic approach to sequencing has transformed from a joke to reality:



DOE Joint Genome Institute - Microsoft Internet Explorer

File Edit View Favorites Tools Help



JGI brings the expertise of four national laboratories, [Lawrence Berkeley](#), [Lawrence Livermore](#), [Los Alamos](#), and [Oak Ridge](#), and the [Stanford Human Genome Center](#) to bear on the frontiers of genome sequencing and related biology. Our sequencing targets encompass a rapidly expanding range of microbes, animals, and plants. The new [Community Sequencing Program \(CSP\)](#) aims to broaden the range still further. JGI is operated by the [University of California](#) for the U.S. Department of Energy.

ABOUT US  
CSP  
JGI SCIENCE  
JAMBOREES  
NEWS  
EDUCATION  
EMPLOYMENT



**genomes**

[Microbial](#) : [Eukaryotic](#)

[img](#) [Integrated Microbial Genomes system](#)

**latest news**

[JGI Sequences Extinct Cave Bear](#)

[JGI Releases Latest Version of IMG](#)

**sequencing**

This fiscal year : 22.187 billion base pairs sequenced  
[More statistics](#)

DOE Joint Genome Institute - Microsoft Internet Explorer

File Edit View Favorites Tools Help



JGI brings the expertise of four national laboratories, [Lawrence Berkeley](#), [Lawrence Livermore](#), [Los Alamos](#), and [Oak Ridge](#), and the [Stanford Human Genome Center](#) to bear on the frontiers of genome sequencing and related biology. Our sequencing targets encompass a rapidly expanding range of microbes, animals, and plants. The new [Community Sequencing Program \(CSP\)](#) aims to broaden the range still further. JGI is operated by the [University of California](#) for the U.S. Department of Energy.

ABOUT US  
CSP  
JGI SCIENCE  
JAMBOREES  
NEWS  
EDUCATION  
EMPLOYMENT



**genomes**

[Microbial](#) : [Eukaryotic](#)

 [Integrated Microbial Genomes system](#)

**latest news**

[JGI Sequences Extinct Cave Bear](#)

[JGI Releases Latest Version of IMG](#)

**sequencing**

This fiscal year : 22.187 billion base pairs sequenced  
[More statistics](#)



DOE Joint Genome Institute - Microsoft Internet Explorer

File Edit View Favorites Tools Help



**JGI**  
DOE JOINT GENOME INSTITUTE  
US DEPARTMENT OF ENERGY  
OFFICE OF SCIENCE

JGI brings the expertise of four national laboratories, [Lawrence Berkeley](#), [Lawrence Livermore](#), [Los Alamos](#), and [Oak Ridge](#), and the [Stanford Human Genome Center](#) to bear on the frontiers of genome sequencing and related biology. Our sequencing targets encompass a rapidly expanding range of microbes, animals, and plants. The new [Community Sequencing Program \(CSP\)](#) aims to broaden the range still further. JGI is operated by the [University of California](#) for the U.S. Department of Energy.

ABOUT US  
CSP  
JGI SCIENCE  
JAMBOREES  
NEWS  
EDUCATION  
EMPLOYMENT



genomes

[Microbial](#) : [Eukaryotic](#)

[img](#) [Integrated Microbial Genomes system](#)

latest news

[JGI Sequences Extinct Cave Bear](#)

[JGI Releases Latest Version of IMG](#)

sequencing

This fiscal year : 22.187 billion base pairs sequenced  
[More statistics](#)

And this is only part way through the third quarter of the fiscal year...

# Aside: Joint Genome Institute

http://www.jgi.doe.gov - JGI - Statistics - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Overall Sequencing Progress, Updated Quarterly

Quarter	Q20* Bases (Billions)			Operating Hours**		
	Goal	Actual Total	Actual % Goal	Goal	Actual Total	Actual % Goal
Q1	7	7.248	104%	2100	2,208	105%
Q2	7	8.000	114%	2100	2,160	103%
Q3	7			2100		
Q4	7			2100		
FY2005	28			8400		

\*Q20 indicates good confidence in the assignment of a base.

\*\*Number of hours a week that sequencing machines are producing data.



# Aside: Joint Genome Institute

http://www.jgi.doe.gov - JGI - Statistics - Microsoft Internet Explorer

File Edit View Favorites Tools Help

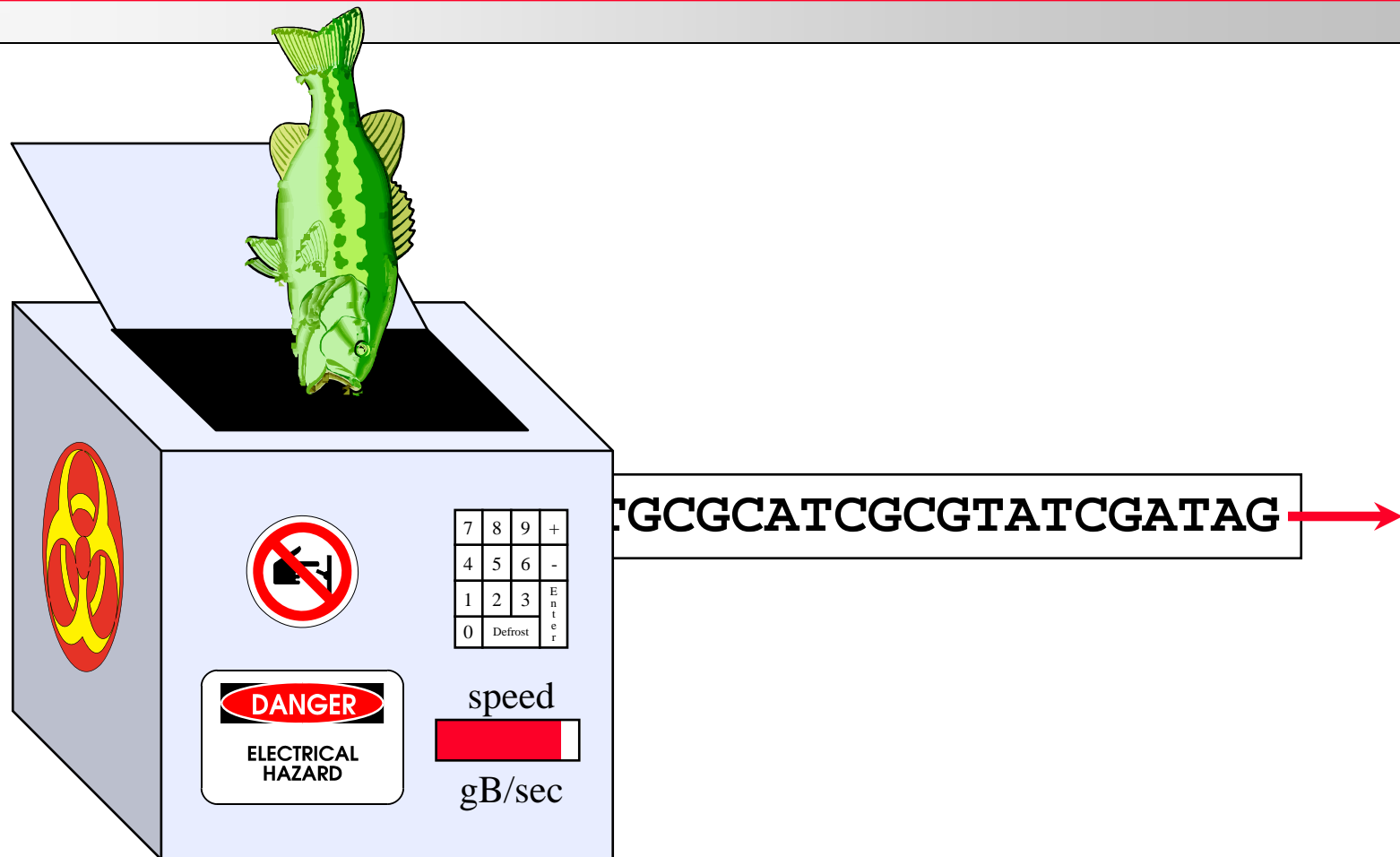
Overall Sequencing Progress, Updated Quarterly

Quarter	Q20* Bases (Billions)			Operating Hours**		
	Goal	Actual Total	Actual % Goal	Goal	Actual Total	Actual % Goal
Q1	7	7.248	104%	2100	2,208	105%
Q2	7	8.000	114%	2100	2,160	103%
Q3	7			2100		
Q4	7			2100		
				8400		

8,000,000,000 bases in 2160 hours =  
3,703,703 bases per hour

assignment of a base.  
ng machines are producing  
data.

# DOE/JGI Bass-o-Matic Sequencer



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

DOE Joint Genome Institute - Microsoft Internet Explorer

File Edit View Favorites Tools Help



**JGI**  
DOE JOINT GENOME INSTITUTE  
US DEPARTMENT OF ENERGY  
OFFICE OF SCIENCE

JGI brings the expertise of four national laboratories, [Lawrence Berkeley](#), [Lawrence Livermore](#), [Los Alamos](#), and [Oak Ridge](#), and the [Stanford Human Genome Center](#) to bear on the frontiers of genome sequencing and related biology. Our sequencing targets encompass a rapidly expanding range of microbes, animals, and plants. The new [Community Sequencing Program \(CSP\)](#) aims to broaden the range still further. JGI is operated by the [University of California](#) for the U.S. Department of Energy.

**ABOUT US**

**CSP**

**JGI SCIENCE**

**JAMBOREES**

**NEWS**

**EDUCATION**

**EMPLOYMENT**



**[ genomes ]**

[Microbial](#) : [Eukaryotic](#)

[img](#) [Integrated Microbial Genomes system](#)

**[ latest news ]**

[JGI Sequences Extinct Cave Bear](#)

[JGI Releases Latest Version of IMG](#)


**[ sequencing ]**

This fiscal year : 22.187 billion base pairs sequenced  
[More statistics](#)

JGI - Community Sequencing Program - Microsoft Internet Explorer




File Edit View Favorites Tools Help

# CSP



## The Community Sequencing Program

- [Overview](#)  
What the Community Sequencing Program is and how it works.
- [How to Propose a Project](#)  
Types of projects accepted, information to include in a proposal, and how to submit it.
- [Review Process and User Agreement](#)  
The review process, scoring criteria, technical reviews, and User Agreements.
- [Sequencing and Project Management](#)  
How your project will be managed and who is responsible for what.
- [Results and Publications](#)  
Information about results, our data release policy, and publications
- [Forms](#)  
Proposal templates, User Agreements, and documentation for DNA preparation and shipping.
- [FAQ](#)  
Answers to commonly asked questions about the CSP.
- [People and Contacts](#)  
Whom to contact for more information, and information about advisory groups.



Page last updated 5/12/2005 · [Disclaimer](#) · [Comments/Questions](#)  
©2005 The Regents of the University of California

Done Internet

JGI - Community Sequencing Program - Microsoft Internet Explorer

File Edit View Favorites Tools Help

HOME ABOUT US CSP SEQUENCING JGI SCIENCE JAMBOREES NEWS EDUCATION EMPLOYMENT

csp

Overview

How to Propose a Project

Review Process and User Agreement

Sequencing and Project Management

Results and Publications

Forms

FAQ

People and Contacts




## Overview

### What is the Community Sequencing Program?

The Community Sequencing Program (CSP) was created to provide the scientific community at large with access to high-throughput sequencing at the Department of Energy's Joint Genome Institute (JGI). Sequencing projects will be chosen based on scientific merit, judged through independent peer review. Criteria for participation in this program, the review process, and interactions between JGI and participants are outlined on this web site. Through this program, the Department of Energy aims to assist and further sequence-based scientific research from a broad range of disciplines.

The CSP consists of two programs:

- a small-genome program for shotgun sequencing of genomes smaller than 250 Mb and other sequencing projects with a total request of less than 1 Gb.
- a large-genome program for shotgun sequencing of genomes larger than 250 Mb. Proposals to the large-genome program must address relevance to the DOE missions of environmental remediation, carbon sequestration, and alternative energy production.

Page last updated 5/12/2005 · [Disclaimer](#) · [Comments/Questions](#)  
©2005 The Regents of the University of California

Done Internet

JGI - Community Sequencing Program - Microsoft Internet Explorer

File Edit View Favorites Tools Help

HOME ABOUT US **CSP** SEQUENCING JGI SCIENCE JAMBOREES NEWS EDUCATION EMPLOYMENT

## csp Overview

Overview  
How to Propose a Project  
Review Process and User Agreement  
Sequencing and Project Management  
Results and Publications  
Forms  
FAQ  
People and Contacts

### What is the Community Sequencing Program?

The Community Sequencing Program (CSP) was created to provide the scientific community at large with access to high-throughput sequencing at the Department of Energy's Joint Genome Institute (JGI). Sequencing projects will be chosen based on scientific merit, judged through independent peer review. Criteria for participation in this program, the review process, and interactions between JGI and participants are outlined on this web site. Through this program, the Department of Energy aims to assist and further sequence-based scientific research from a broad range of disciplines.

The CSP consists of two programs:

- a small-genome program for shotgun sequencing of genomes smaller than 250 Mb and other sequencing projects with a total request of less than 1 Gb.
- a large-genome program for shotgun sequencing of genomes larger than 250 Mb. Proposals to the large-genome program must address relevance to the DOE missions of environmental remediation, carbon sequestration, and alternative energy production.

Page last updated 5/12/2005 · [Disclaimer](#) · [Comments/Questions](#)  
©2005 The Regents of the University of California

Done Internet

# Aside: Joint Genome Institute

---

## CSP Project Description Limits:

- Limit 5 pages for total shotgun sequencing of less than 400 Mb (e.g., 8x coverage of genomes < 50 Mb, microbial communities or directed sequencing projects).
- Limit 10 pages for sequencing requests between 400 Mb and 2 Gb (e.g., 8x coverage of genomes between 50 and 250 Mb).
- Limit 15 pages for sequencing requests greater than 2 Gb (e.g., 8x coverage of genomes larger than 250 Mb).

# What's Really Next

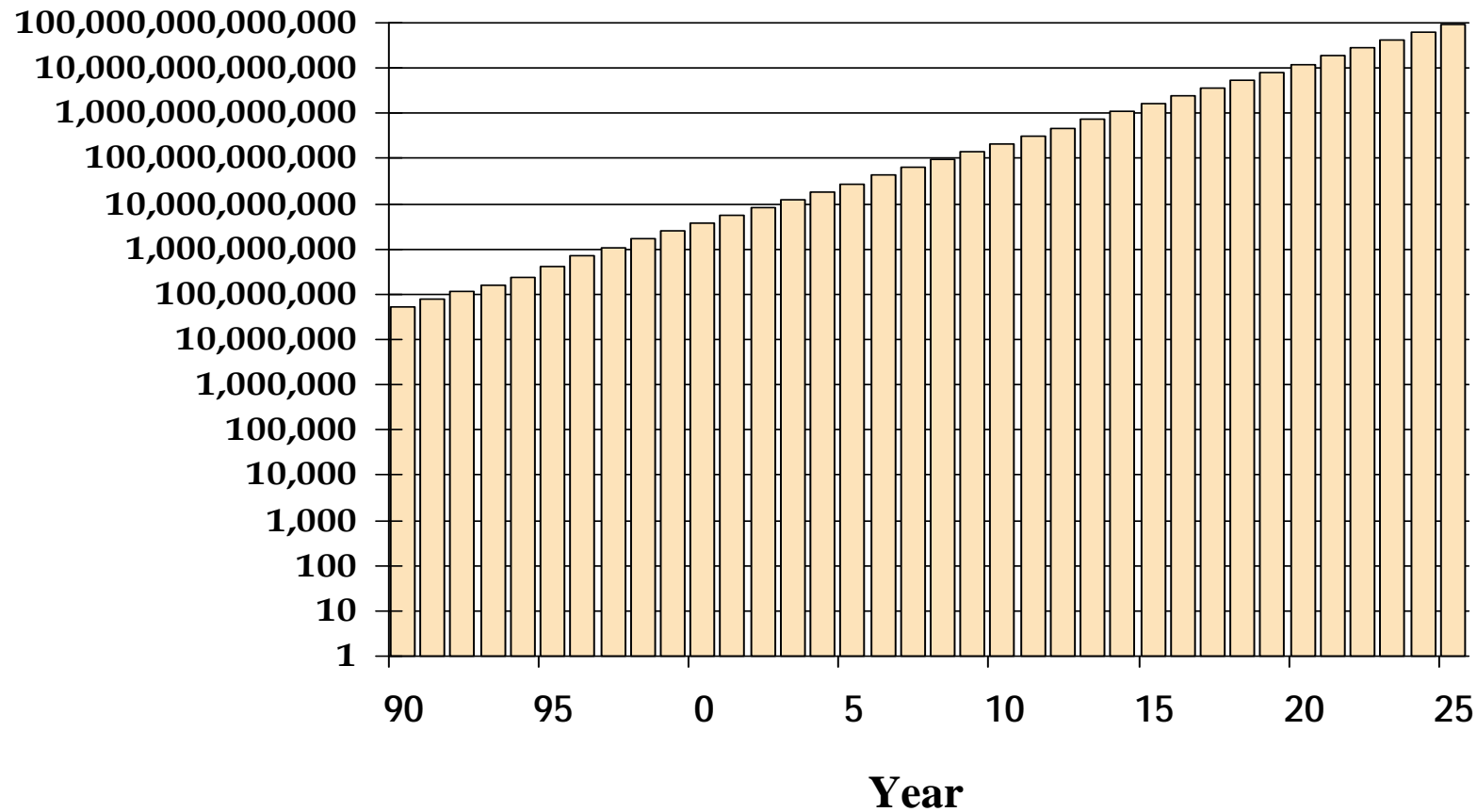
---

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

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

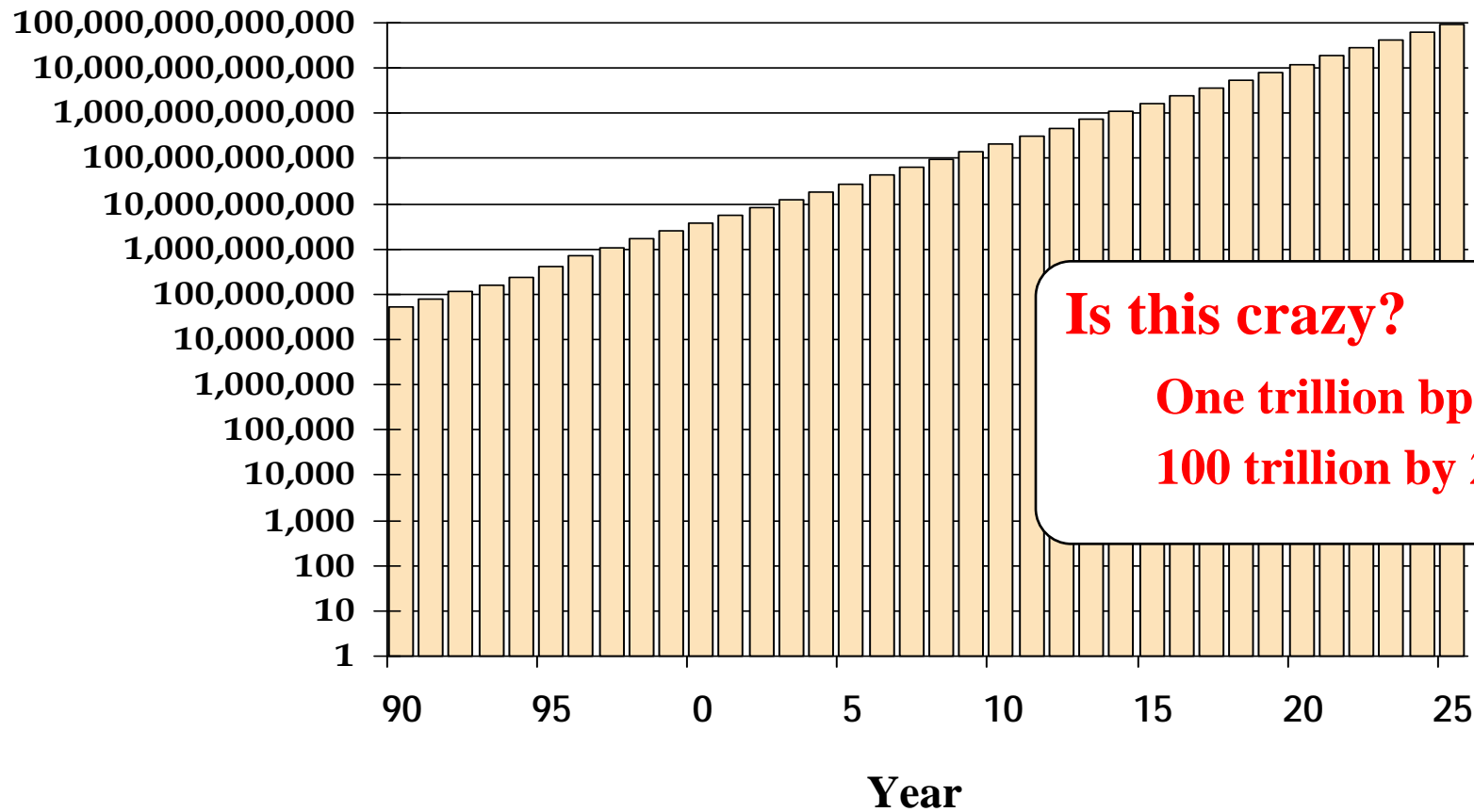


# Projected Base Pairs in GenBank



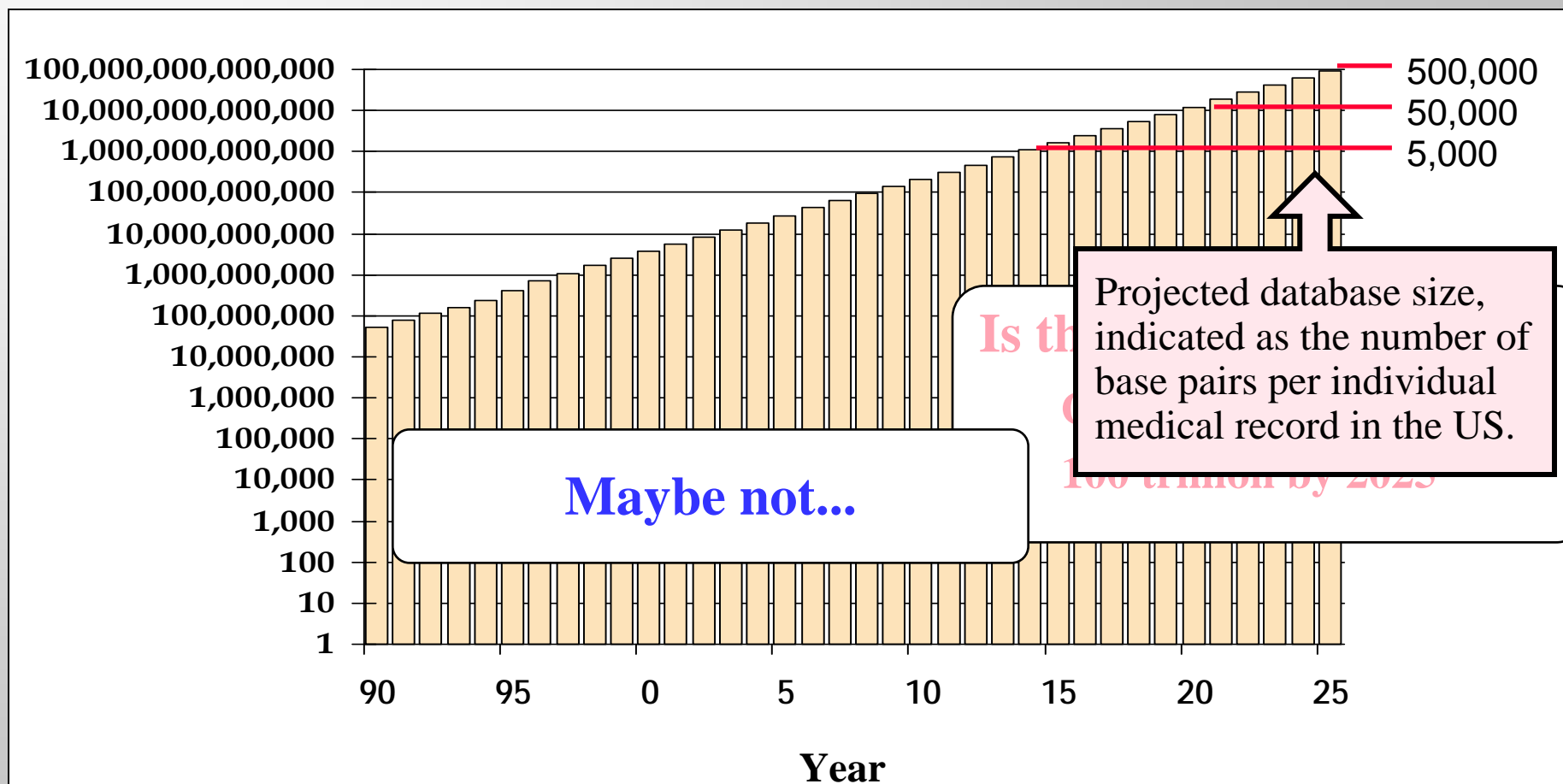
Assumed annual growth rate: 50%  
(*less than current rate*)

# Projected Base Pairs in GenBank



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

# Projected Base Pairs in GenBank



*21<sup>st</sup> Century Biology*  
*Post Genome Era*

# 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

# Post-Genome Era

---

## Suggested definition:

- functional genomics = biology

# 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.

# Paradigm Shift in Biology

---

## Case of Microbiology

A serious suggestion has been made that the DOE/JGI should consider sequencing **ALL KNOWN** and **CULTURABLE** bacteria.

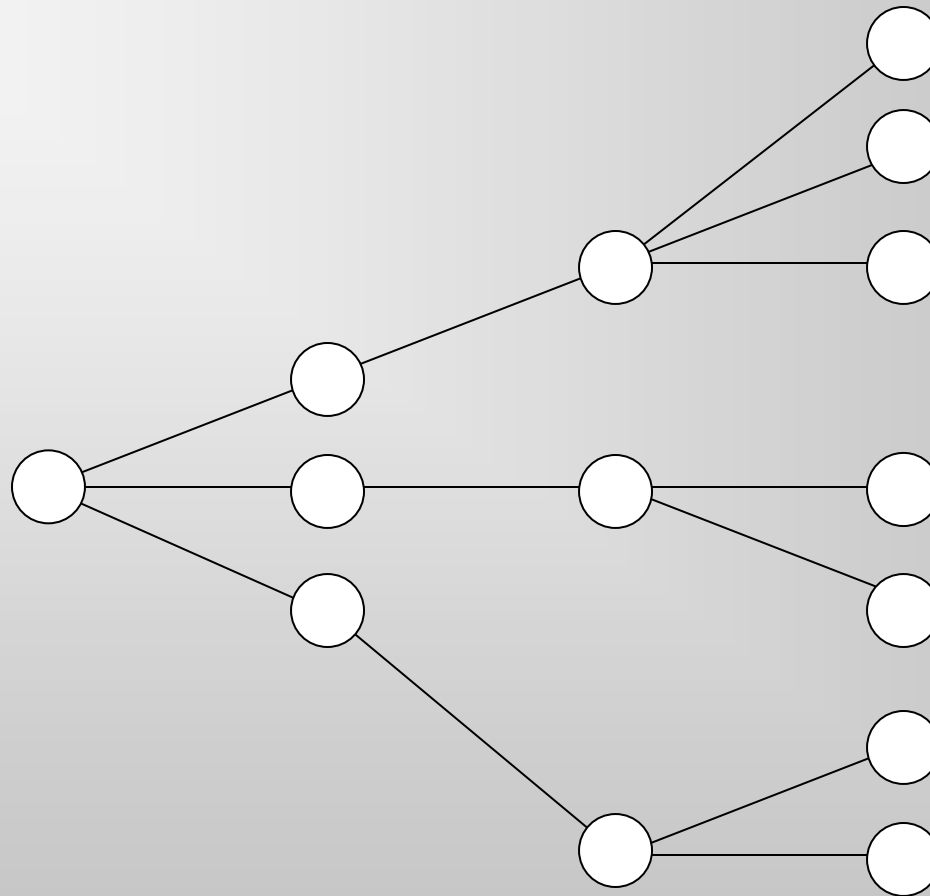
If a full, annotated sequence were available for all known bacteria, the practice of microbiology would match Gilbert's prediction.

# *Data Source Problems*

# Single-Instrument Science

instrument

researchers



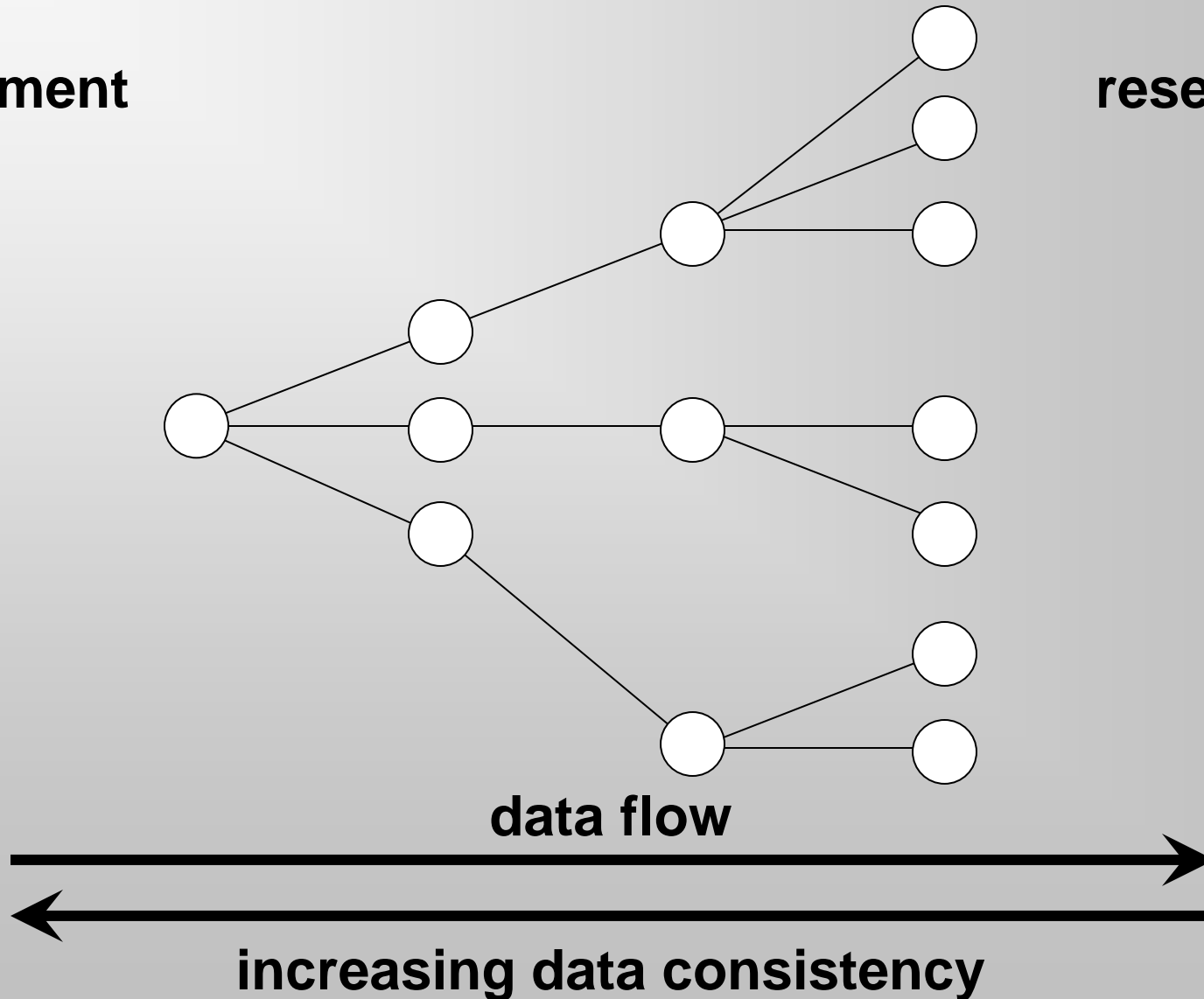
data flow



# Single-Instrument Science

instrument

researchers



# Single-Instrument Science

instrument

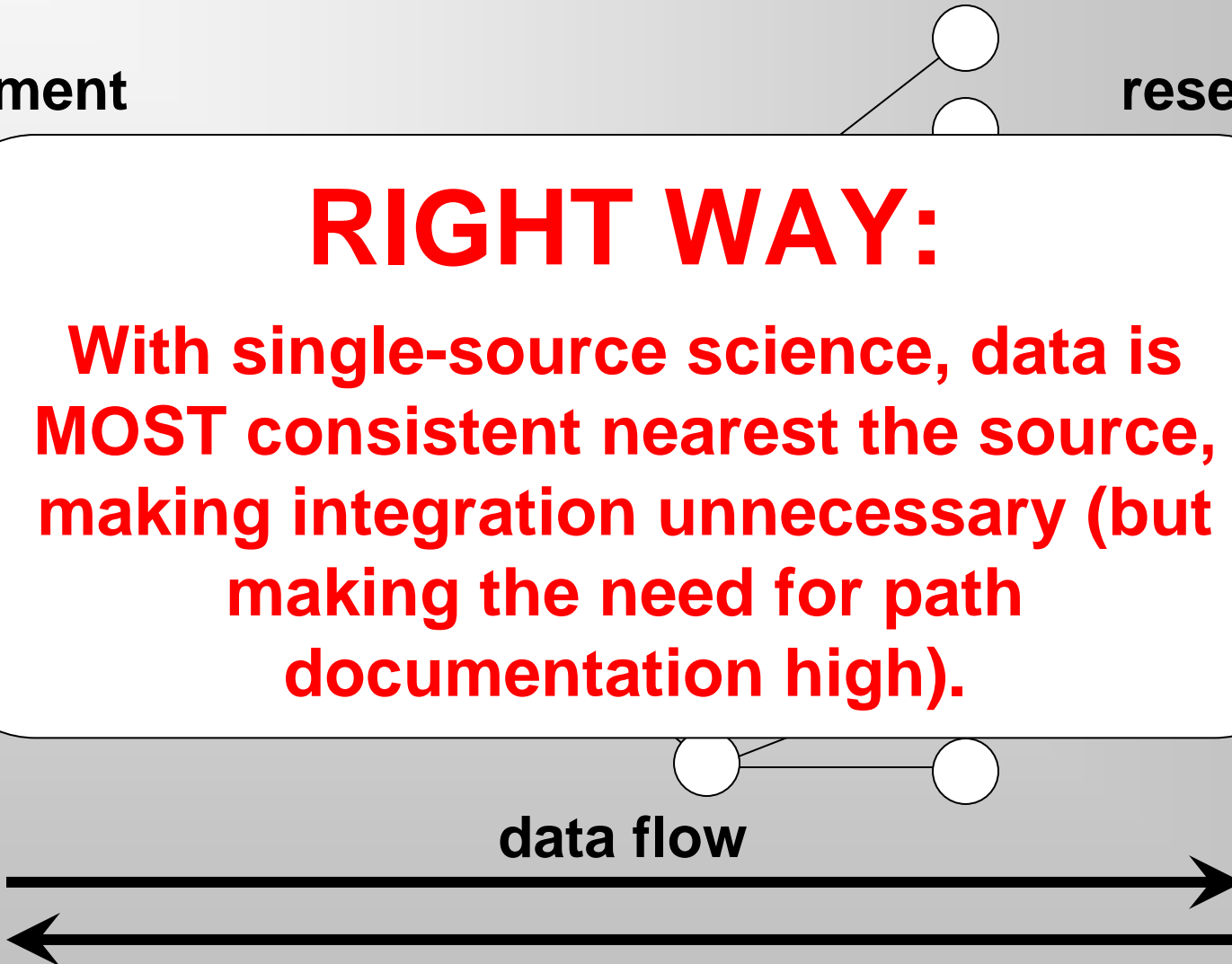
researchers

## RIGHT WAY:

With single-source science, data is **MOST** consistent nearest the source, making integration unnecessary (but making the need for path documentation high).

data flow

increasing data consistency

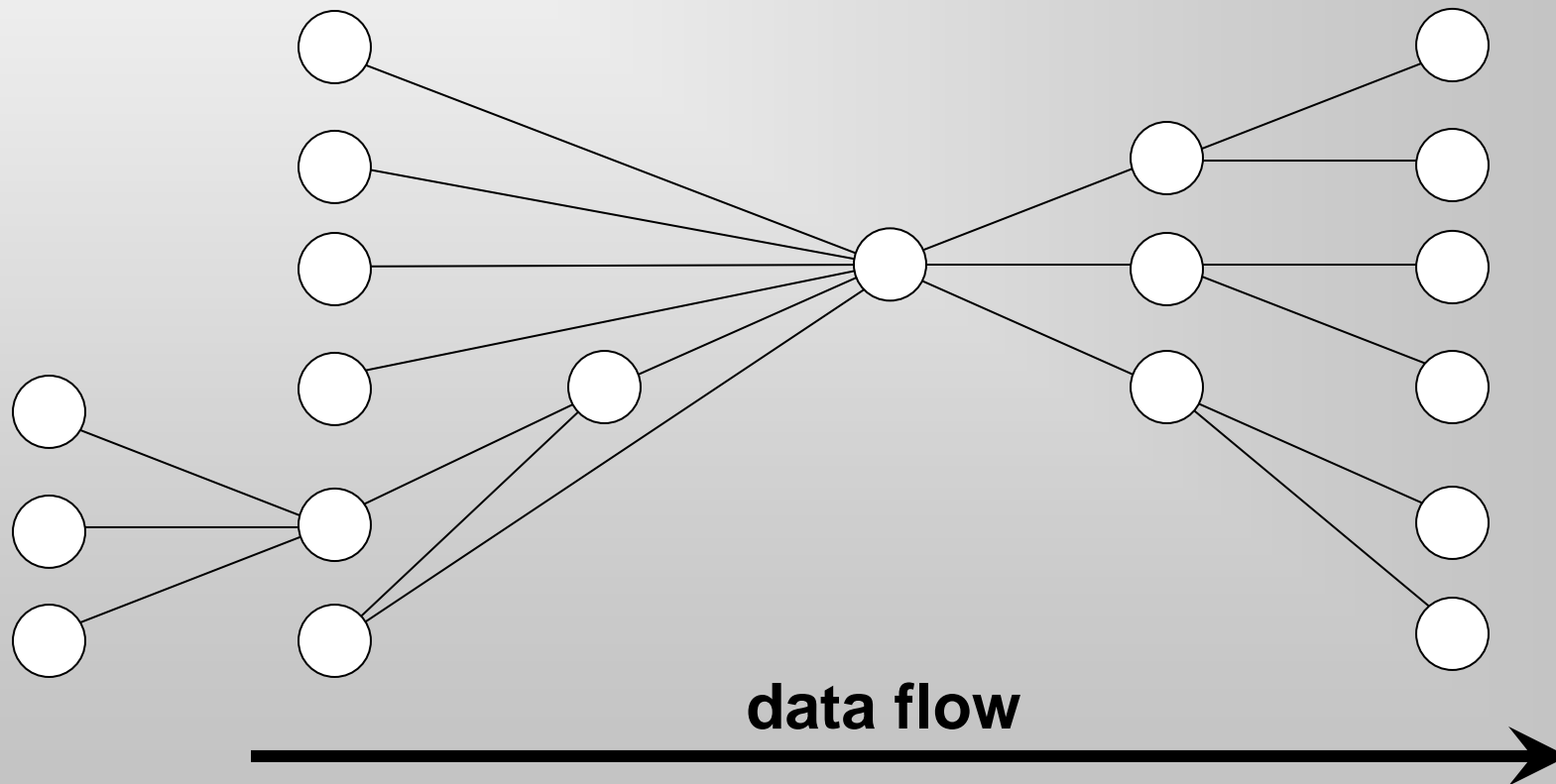


# Multi-Instrument Science

researchers

data resource(s)

researchers



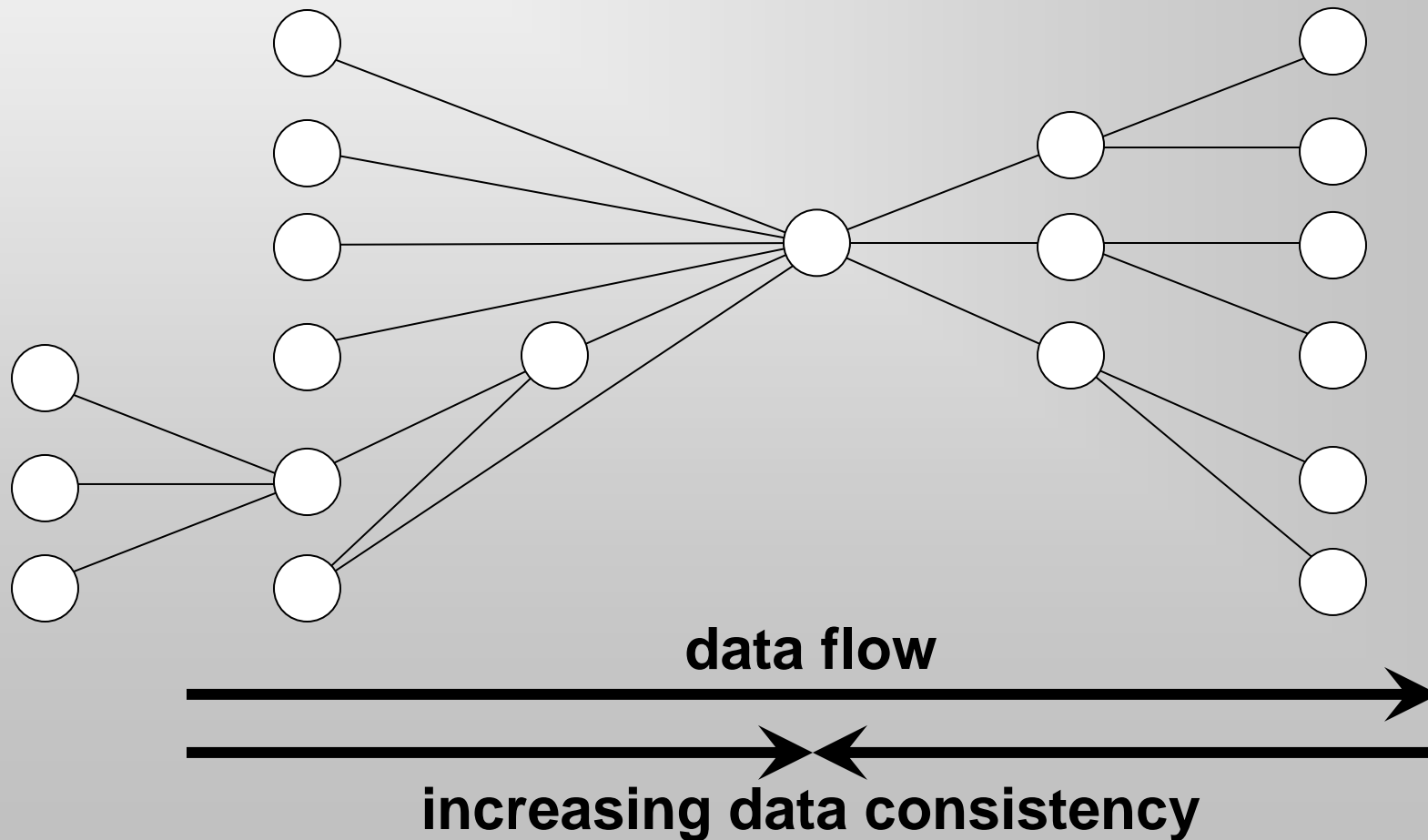


# Multi-Instrument Science

researchers

data resource(s)

researchers



# Multi-Instrument Science

researchers

data resource(s)

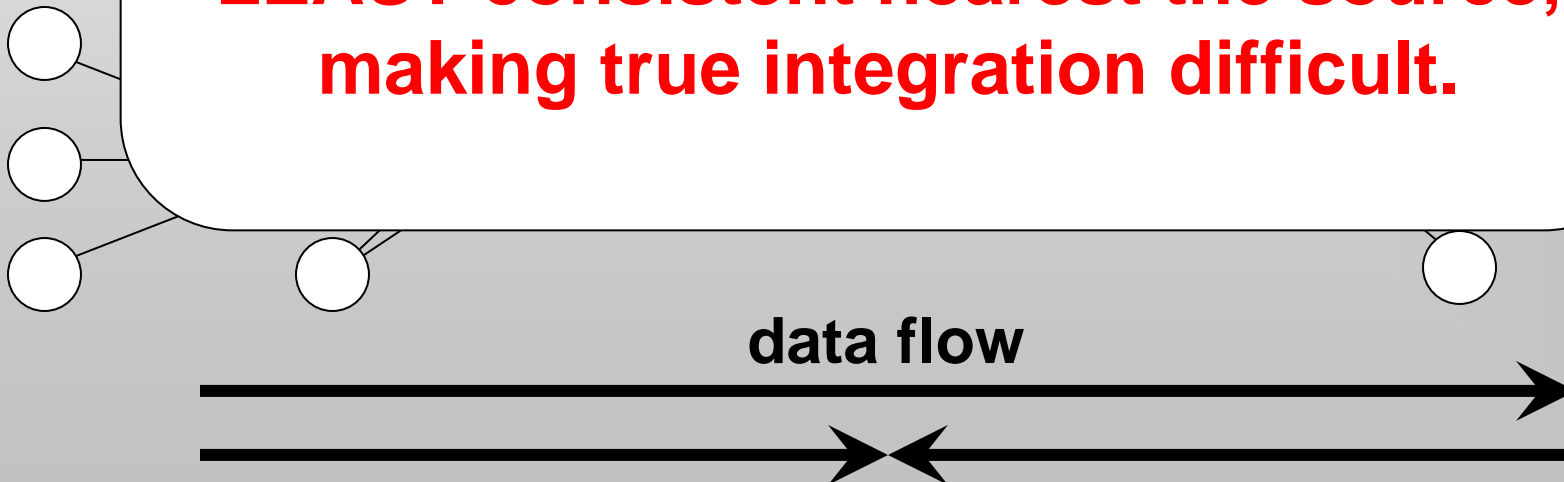
researchers

**STOP – WRONG WAY:**

With multi-source science, data is  
**LEAST** consistent nearest the source,  
making true integration difficult.

data flow

increasing data consistency

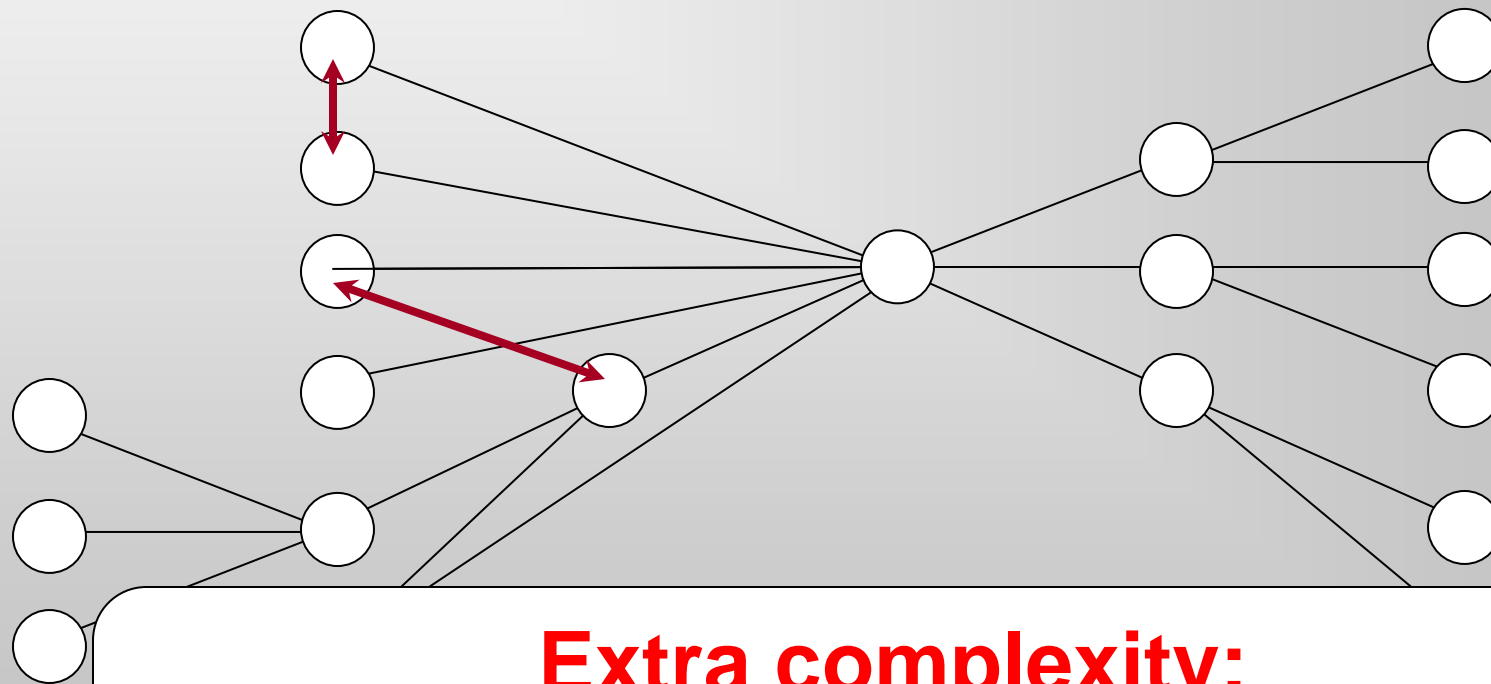


# Multi-Instrument Science

researchers

data resource(s)

researchers



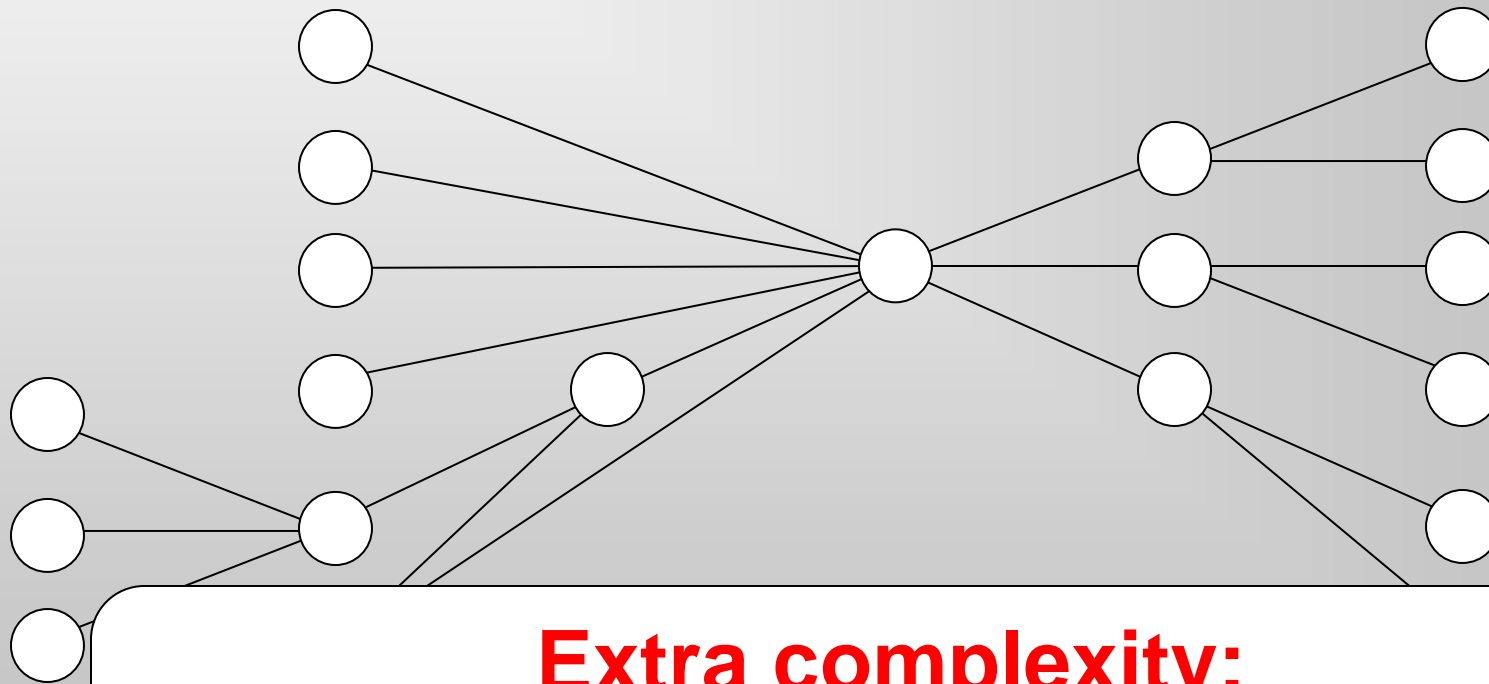
**Extra complexity:**  
**Undocumented, uncoordinated local data  
exchange**

# Multi-Instrument Science

researchers

data resource(s)

researchers



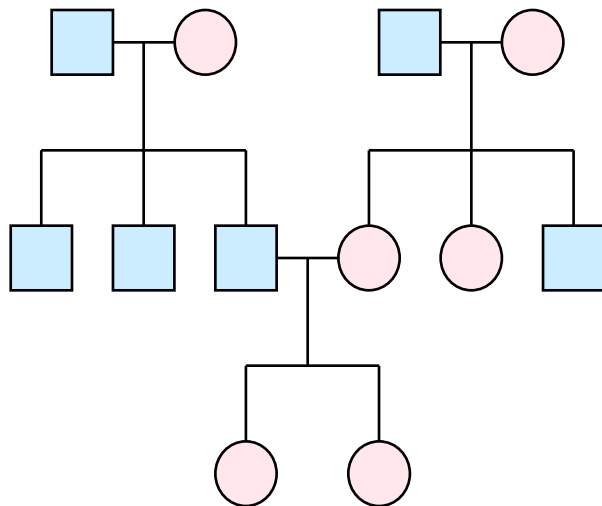
**Extra complexity:**

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

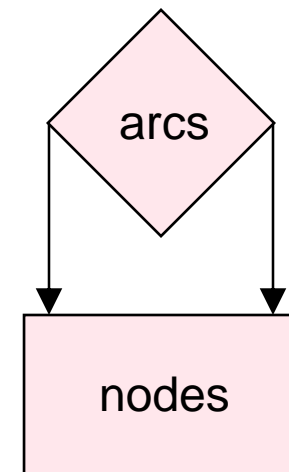
# *Data Model Problems*

# Graph Challenges

Pedigree

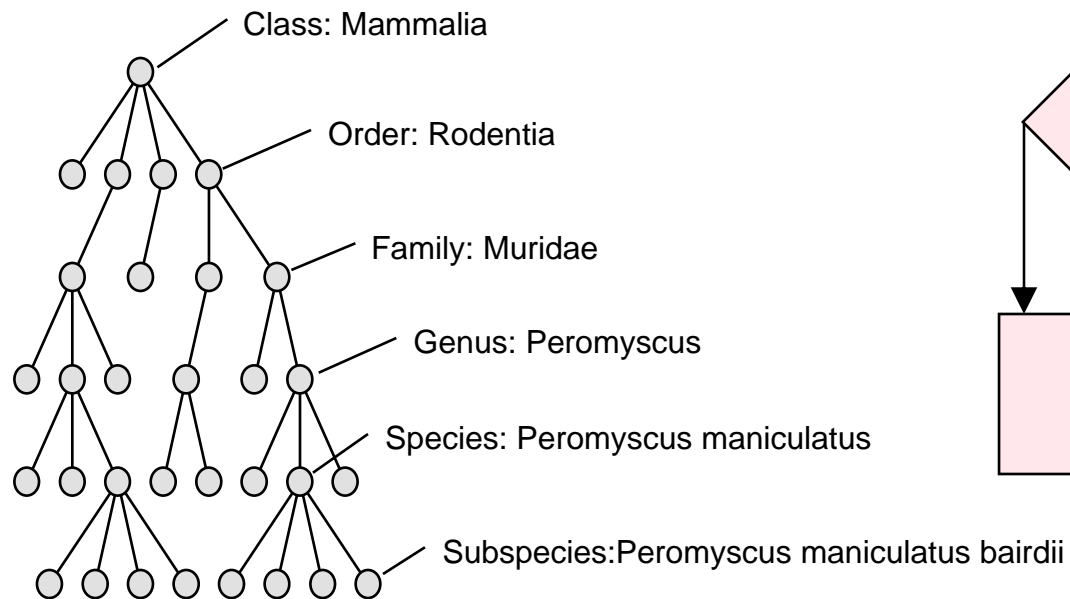


Relational Representation

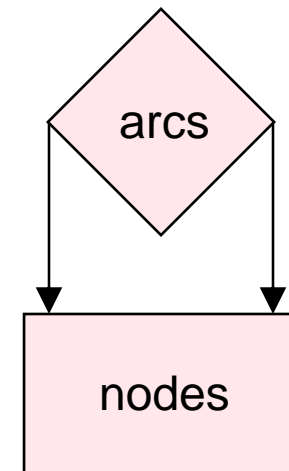


# Graph Challenges

## Classification Hierarchy

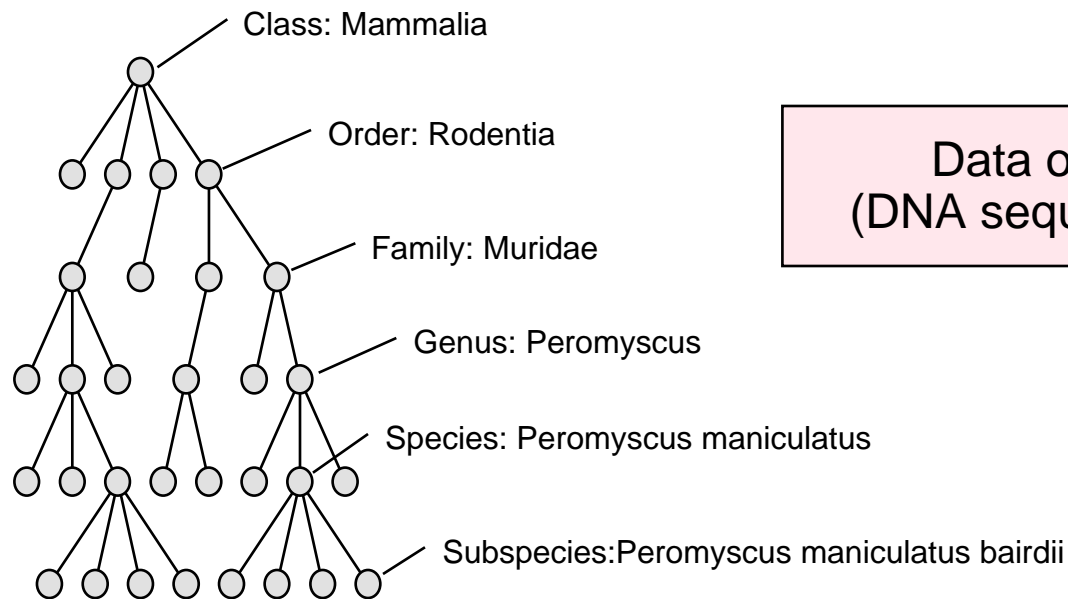


## Relational Representation



# Classification Challenges

## Classification Hierarchy



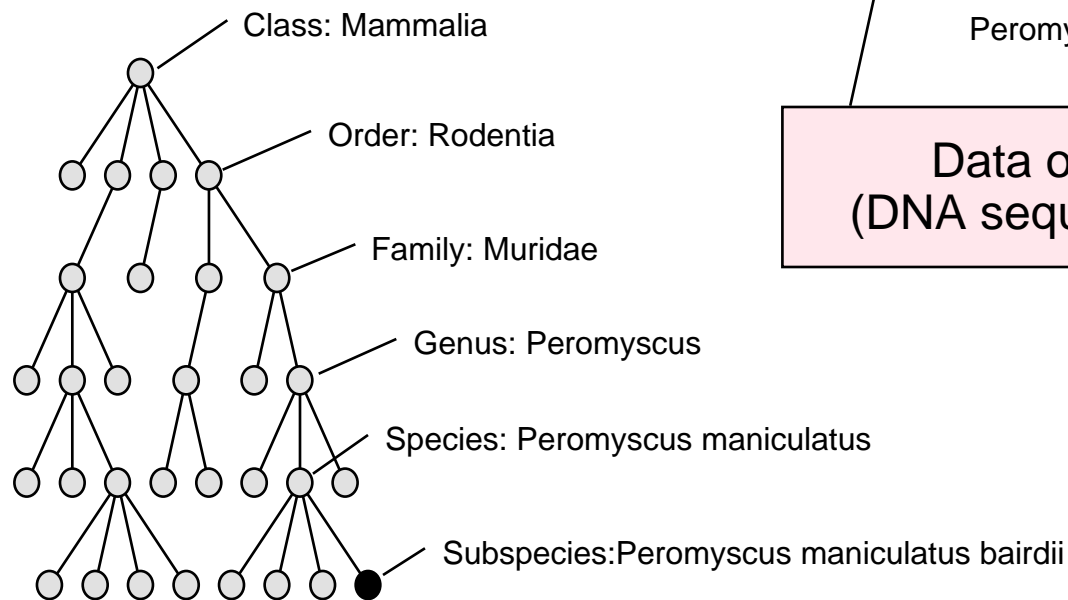
## Data Objects to be Classified

Data object  
(DNA sequences?)

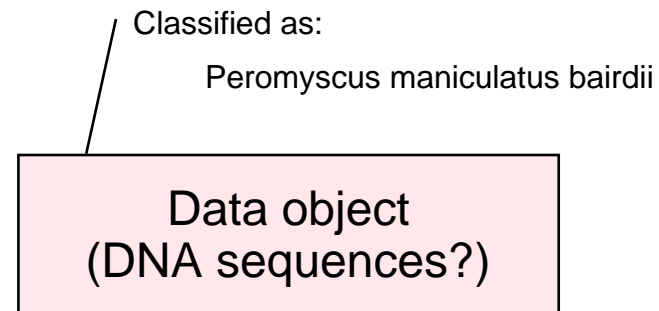


# Classification Challenges

## Classification Hierarchy



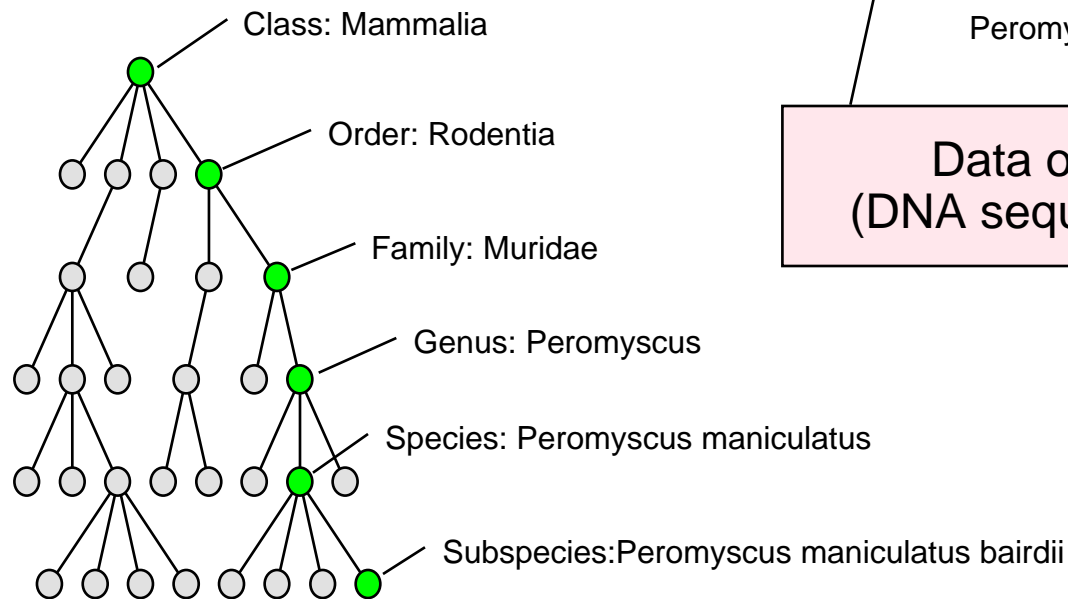
## Data Objects to be Classified



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:

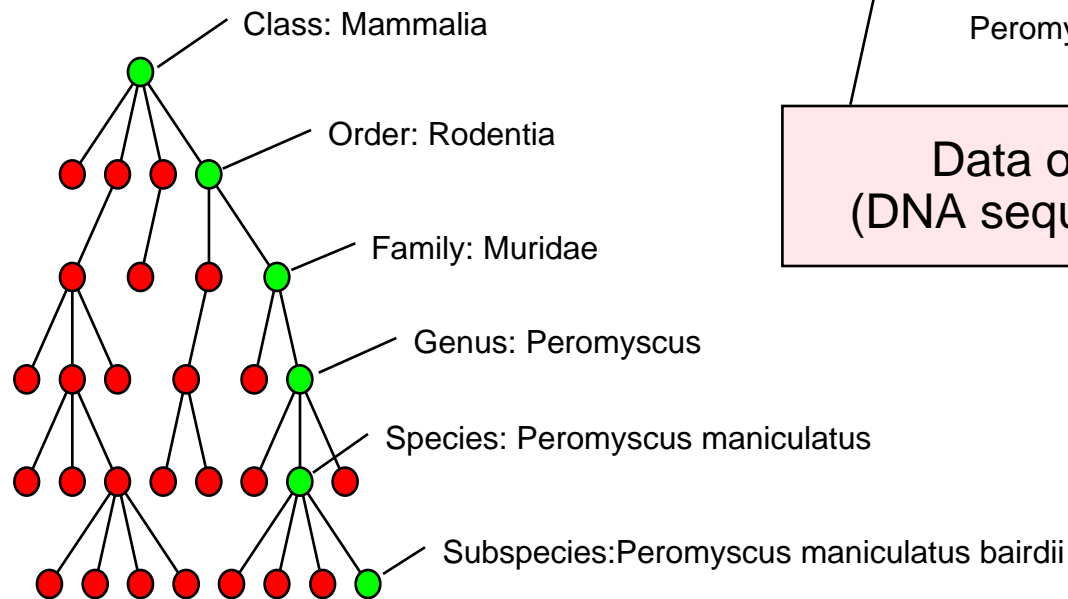
*Peromyscus maniculatus bairdii*

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:

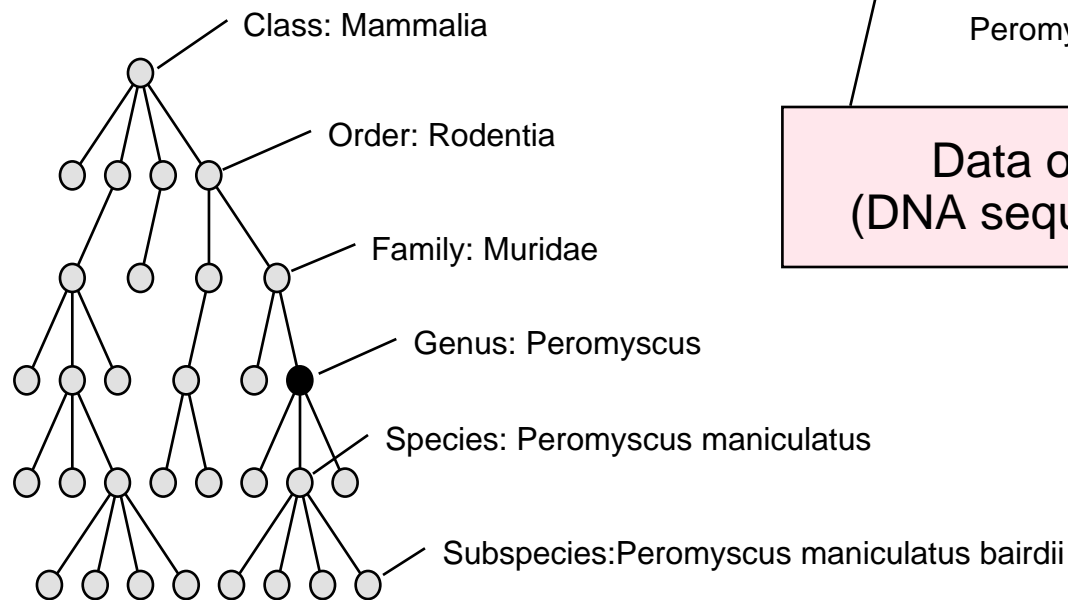
*Peromyscus maniculatus bairdii*

Data object  
(DNA sequences?)

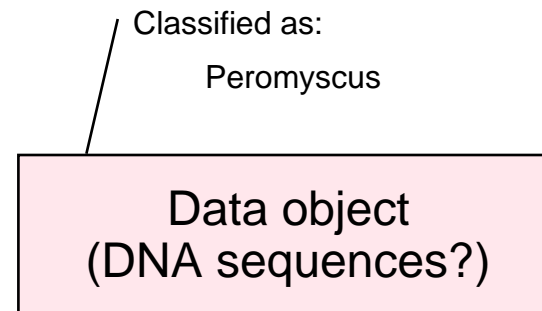
Suppose we permit querying at any level, but require classification of objects at leaf level. Then all questions referring to nodes on the path from the classification point to the top return **TRUE**, all others **FALSE**.

# Classification Challenges

## Classification Hierarchy



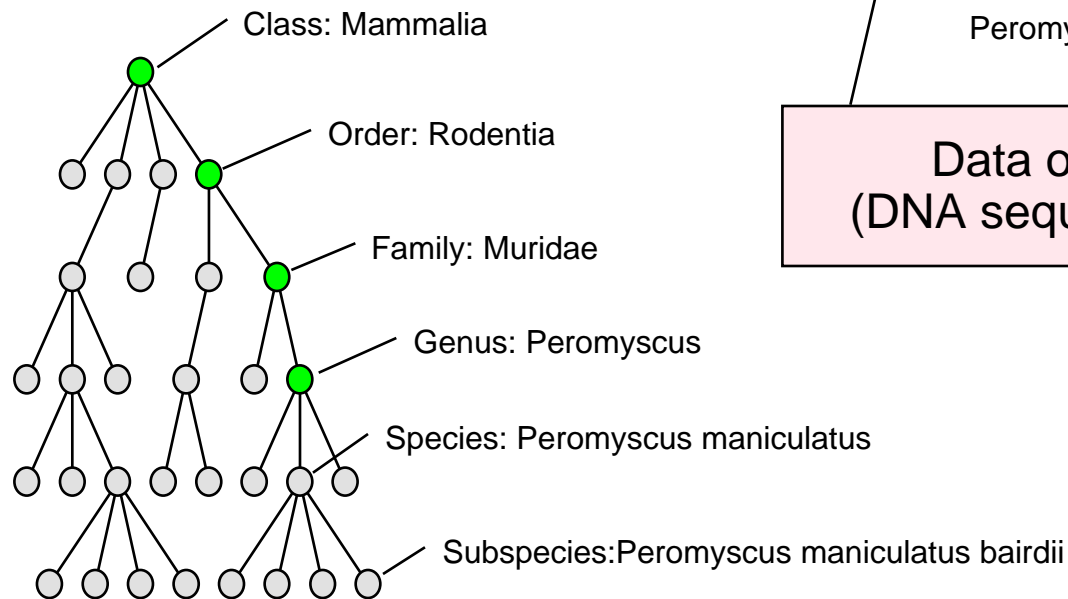
## Data Objects to be Classified



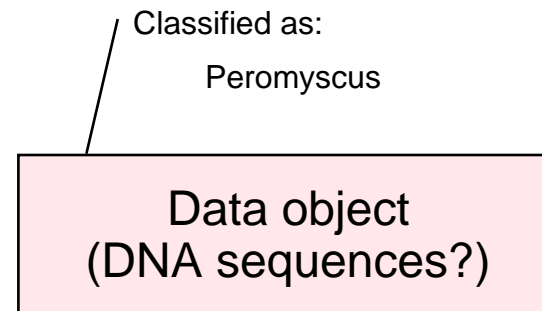
Now, suppose the we permit querying at any level, and also that we allow classification of objects at any level.

# Classification Challenges

## Classification Hierarchy



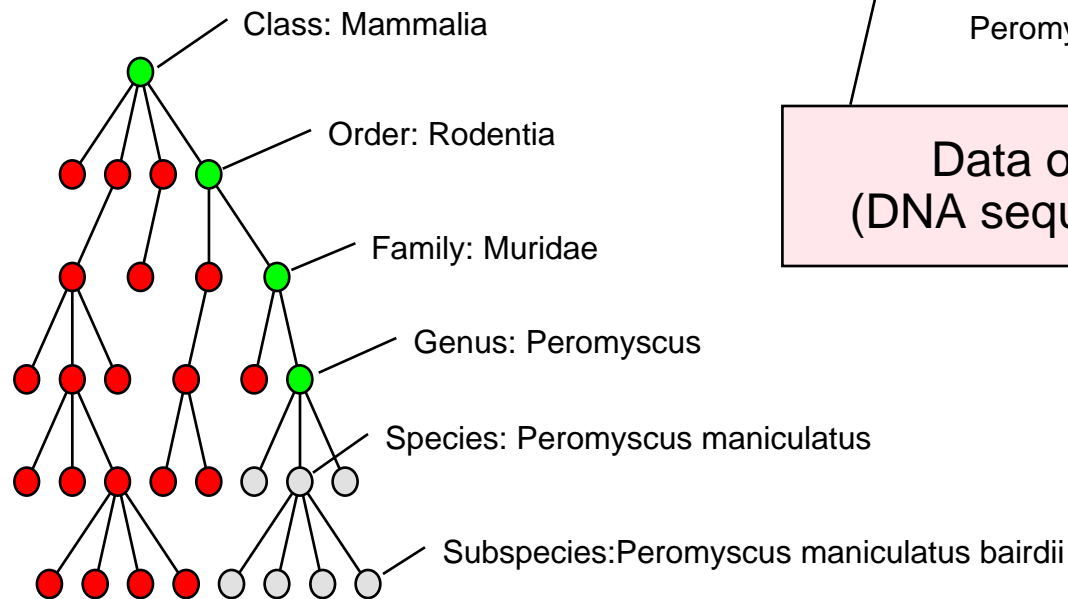
## Data Objects to be Classified



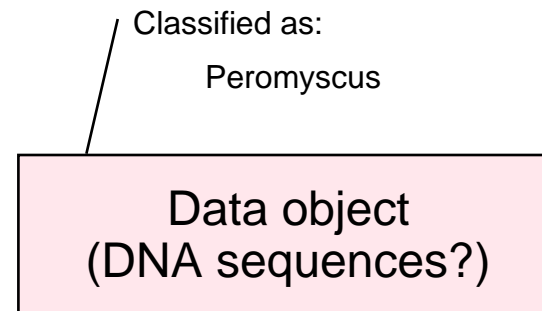
Now, suppose we permit querying at any level, and also that we allow classification of objects at any level. Then all questions referring to nodes on the path from the classification point to the top return **TRUE**,

# Classification Challenges

## Classification Hierarchy



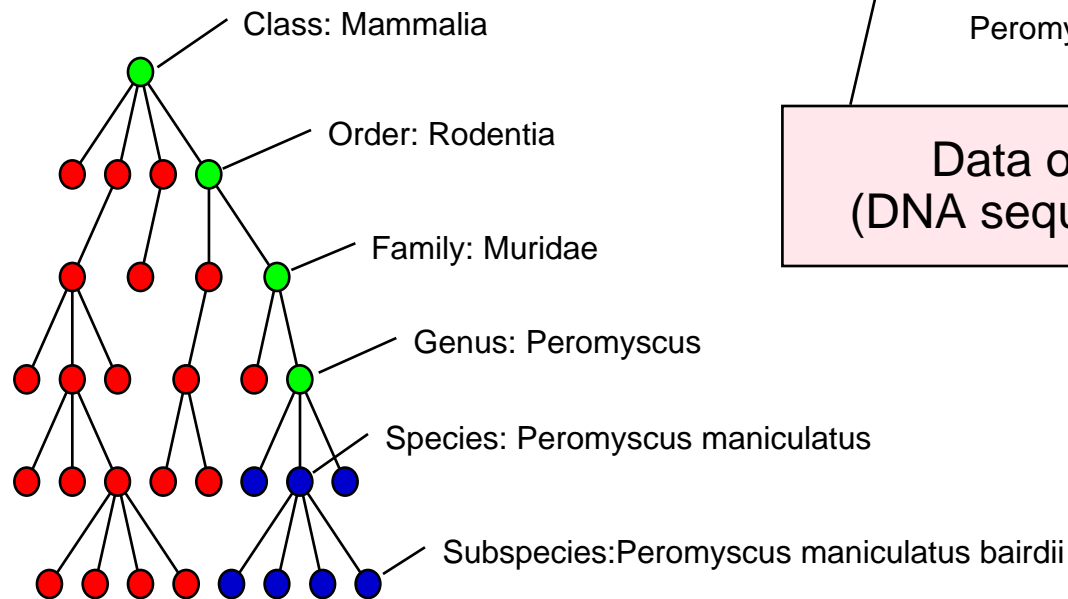
## Data Objects to be Classified



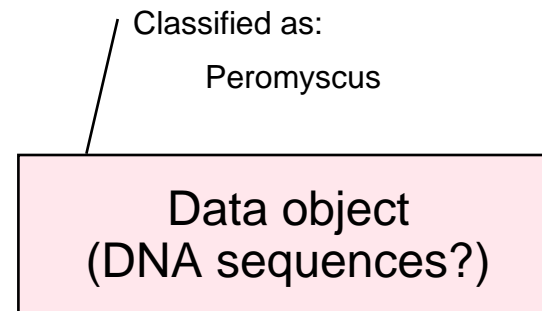
Now, suppose we permit querying at any level, and also that we allow classification of objects at any level. Then all questions referring to nodes on the path from the classification point to the top return **TRUE**, all questions referring to nodes lateral to this path return **FALSE**,

# Classification Challenges

## Classification Hierarchy



## Data Objects to be Classified



Now, suppose 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**.

*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 is critical when one is developing a successful identifier scheme for any data resource likely to involve data sharing for unanticipated uses.**

**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 different ways:
  - INDIVIDUAL SPECIFICATION: denoting an individual object without identifying either its class membership or its individuality - e.g., “this thing”
  - CLASS IDENTIFICATION: specifying that 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 different 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.**



# *Budget Problems: Reality Check*

# Information-Intensive Business

---

One can barely begin to read a current journal without finding a reference to the fact that biomedical research has become an information-intensive field.

# Information-Intensive Business

---

One can barely begin to read a current journal without finding a reference to the fact that biomedical research has become an information-intensive field.

Maybe we should look to information-intensive fields for operational ideas...

# Reality Check I

---

## 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

# Reality Check I

---

## 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.

# Reality Check I

---

## **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

# Reality Check I

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 %

# Reality Check II

---

One biotech company, Celera, spent more money on IT in its first year of business than all of NCI has spent on IT in the last five years.



# Reality Check III

---

## Resource Availability

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

# Reality Check III

---

## 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 biomedical information infrastructure is tiny.

# Reality Check III

---

## **Appropriate overall funding level:**

- approx. 5-15% of total public-sector biomedical research funding

# Reality Check III

---

## Appropriate overall funding level:

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

# Reality Check III

---

## Appropriate overall funding level:

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

### Seem high?

What percent of enterprise operating budgets goes to IT in those industries where IT makes a strategic difference?

# Reality Check III

---

Appropriate overall funding level:

**Warning:**

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

to IT in those industries where IT makes a strategic difference?

---

# The Future

---

# *Standards*



# Standards

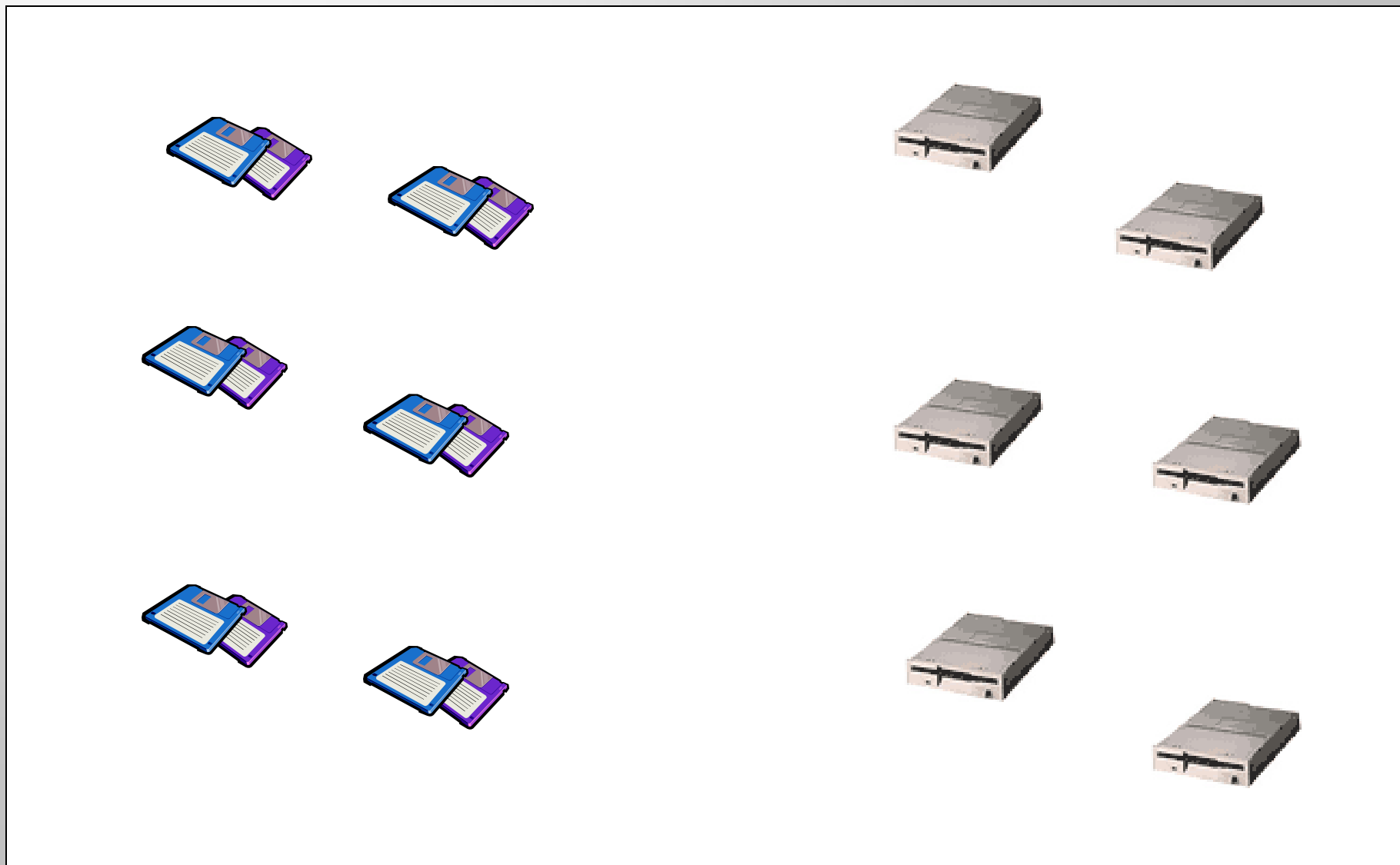
---

## **Standards are Useful, but:**

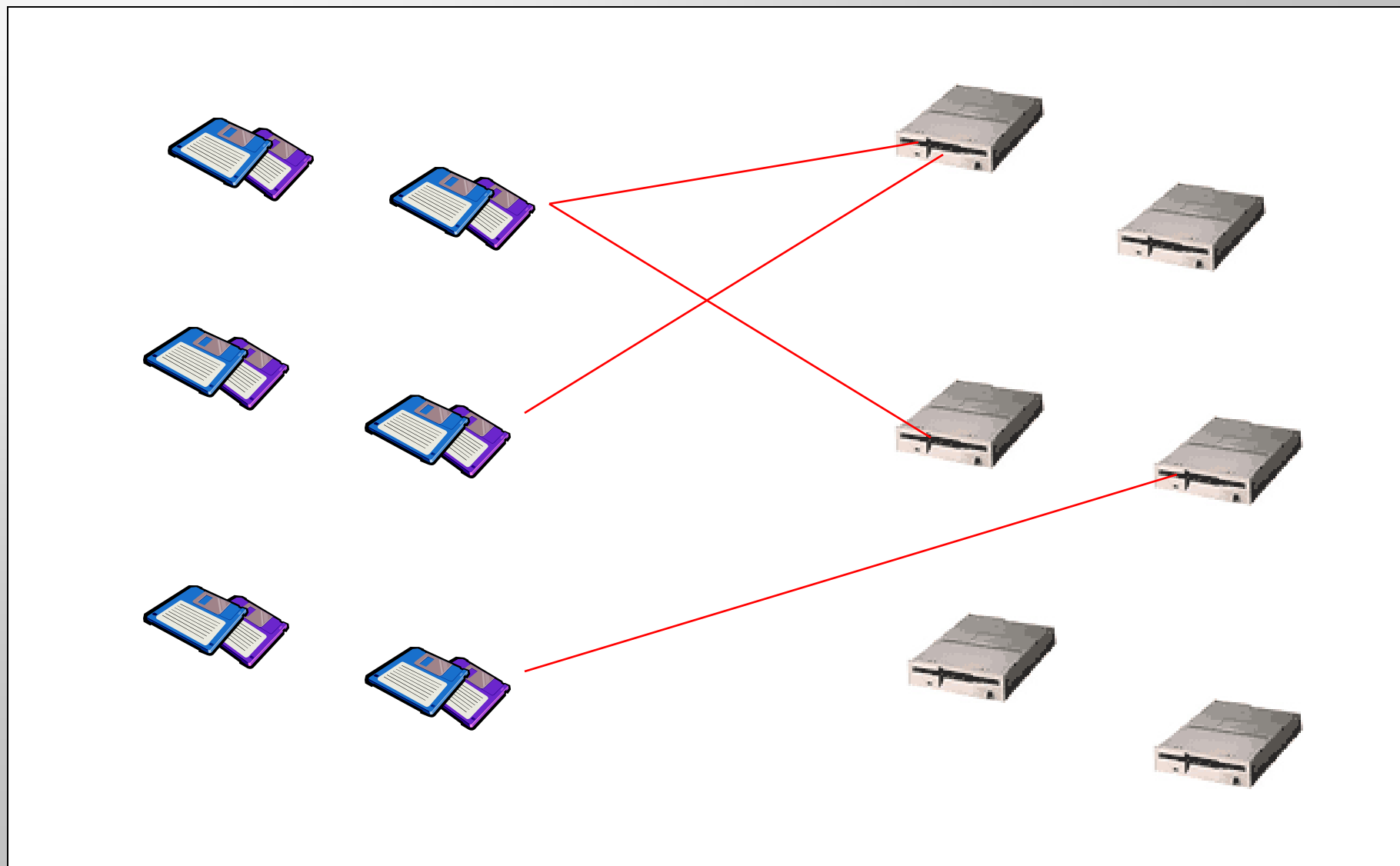
- It is important to avoid premature standards.
- Constraining standards should be avoided.
- Enabling standards should be embraced.
- The utility of having many standards to choose from is not a joke.

# Bad Data-exchange Standard

---

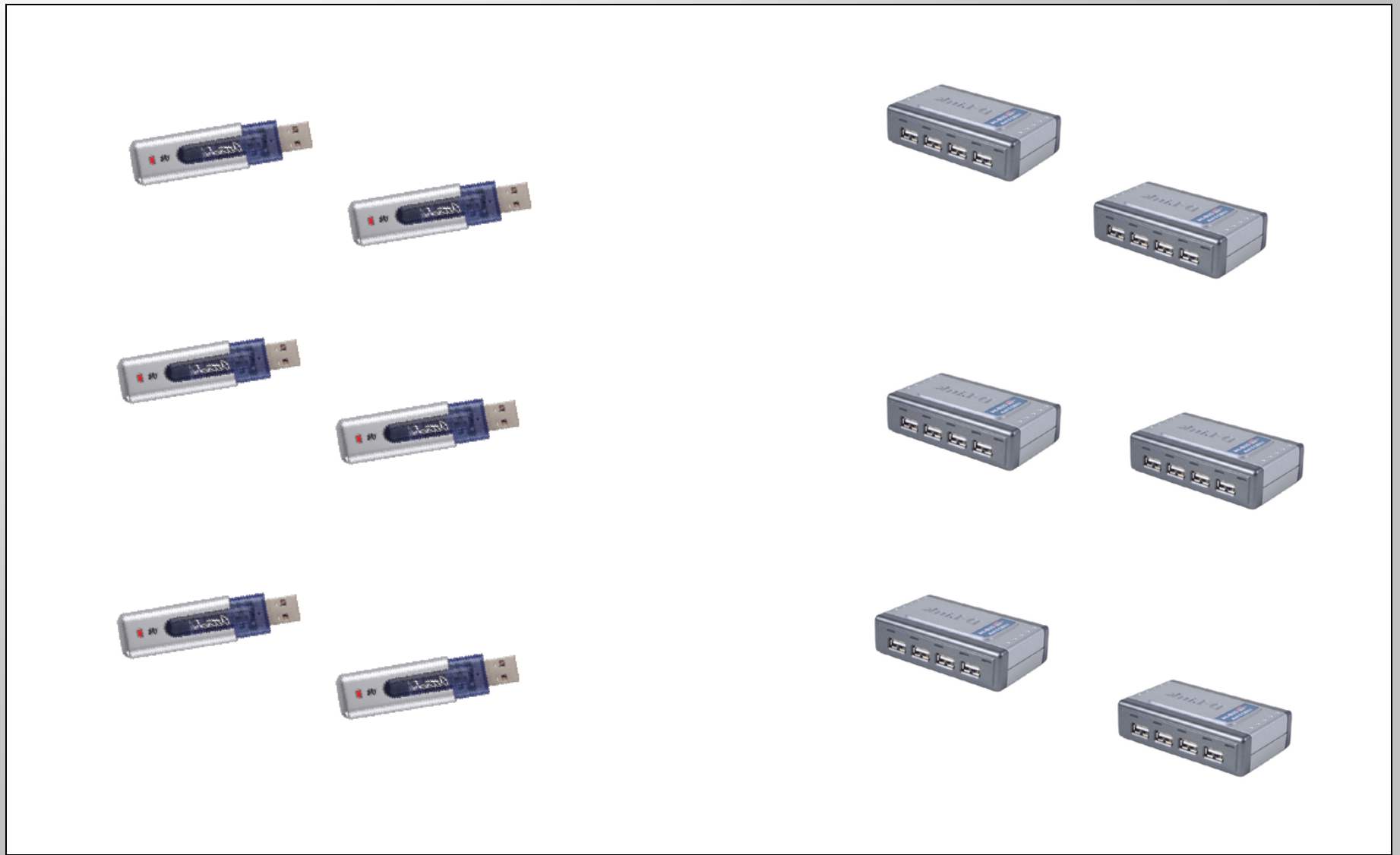


# Bad Data-exchange Standard

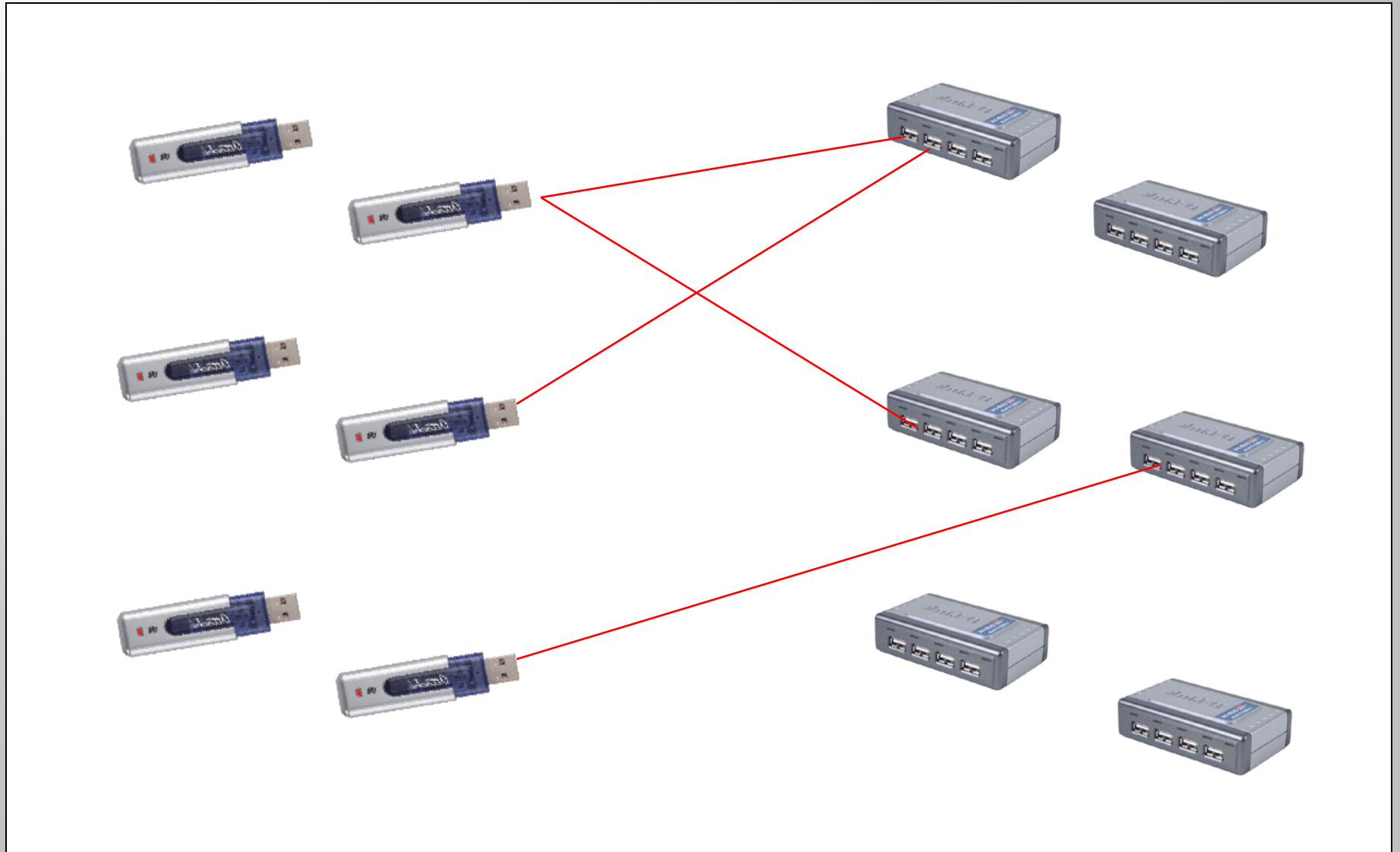


# Good Data-exchange Standard

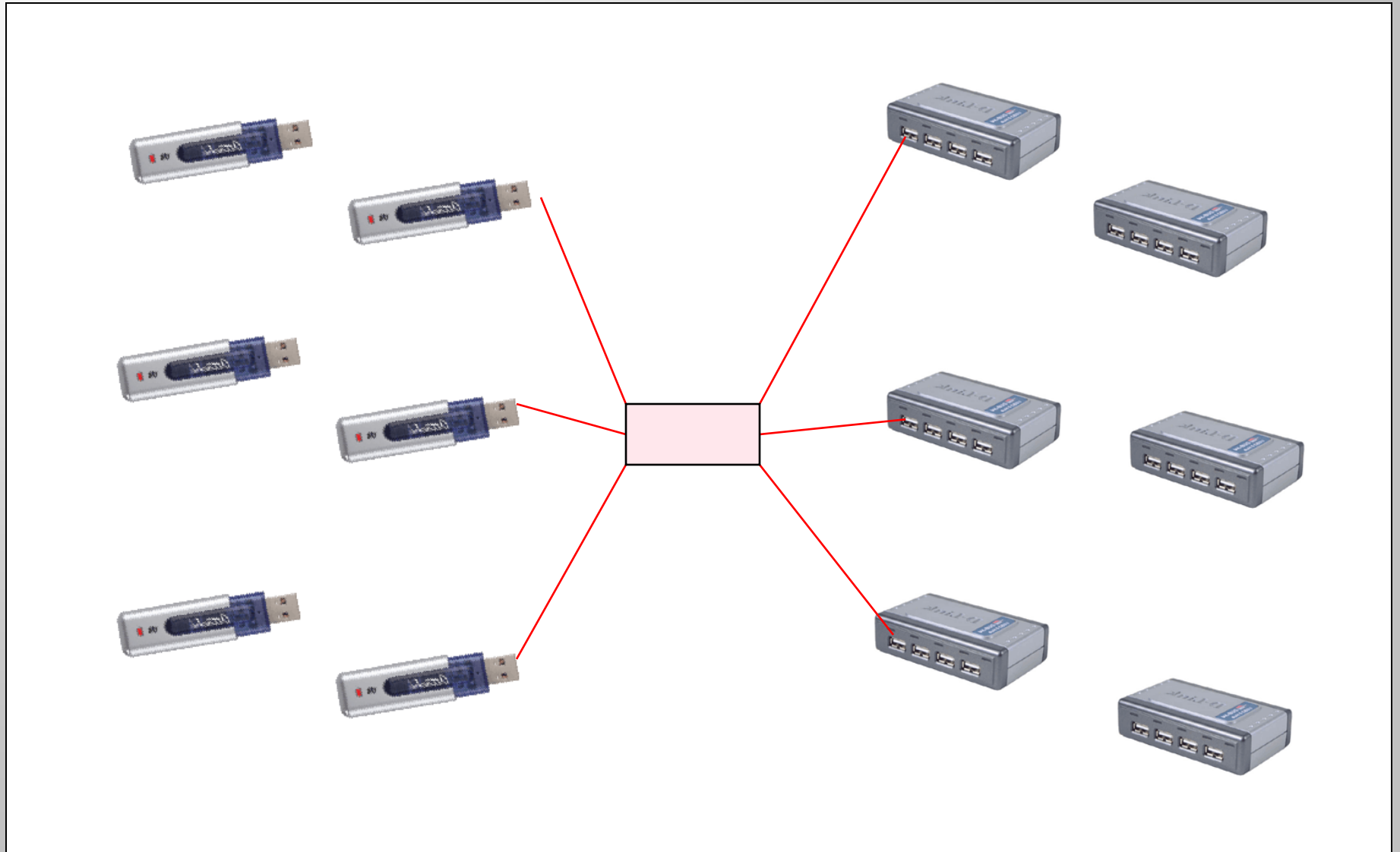
---



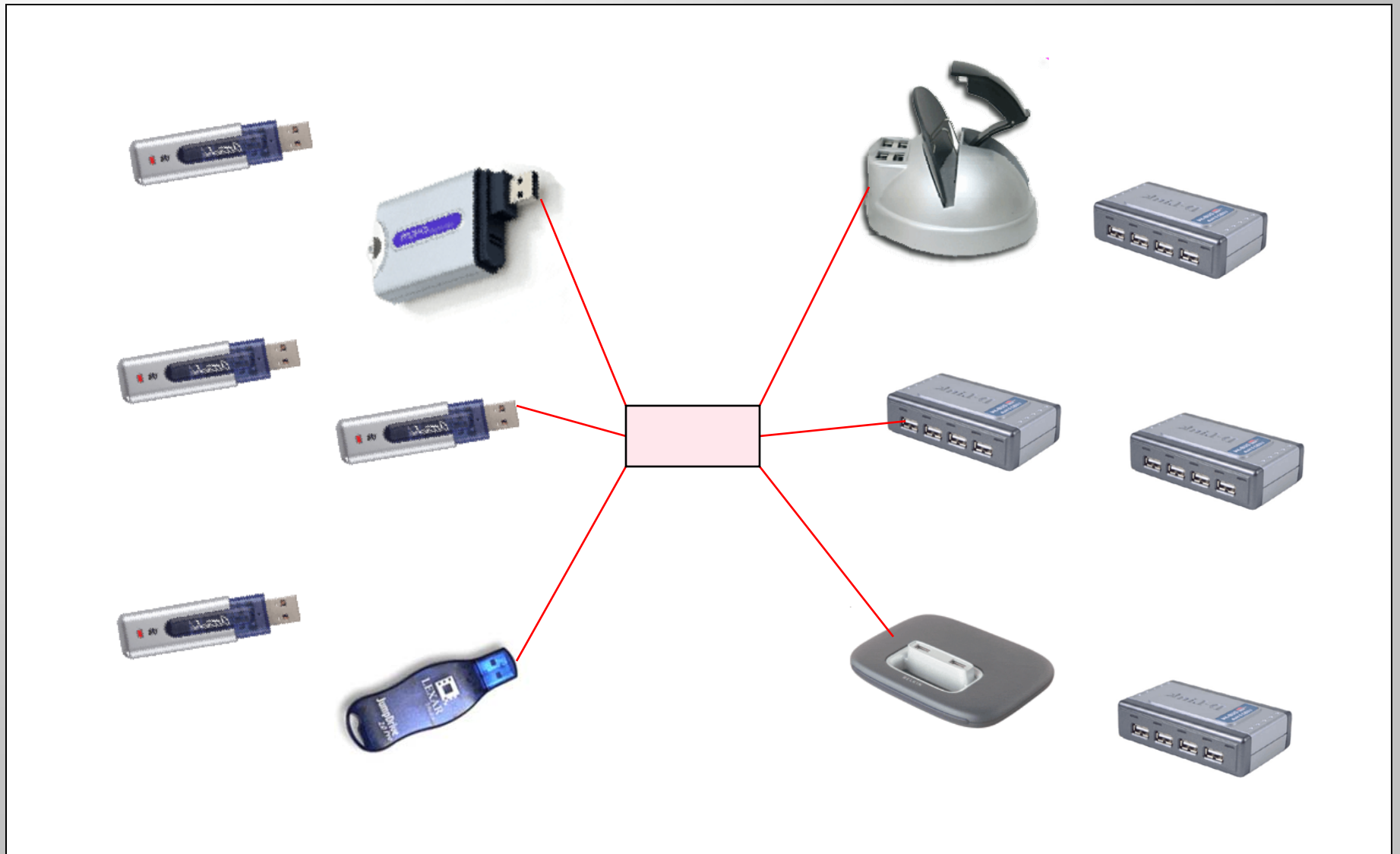
# Good Data-exchange Standard



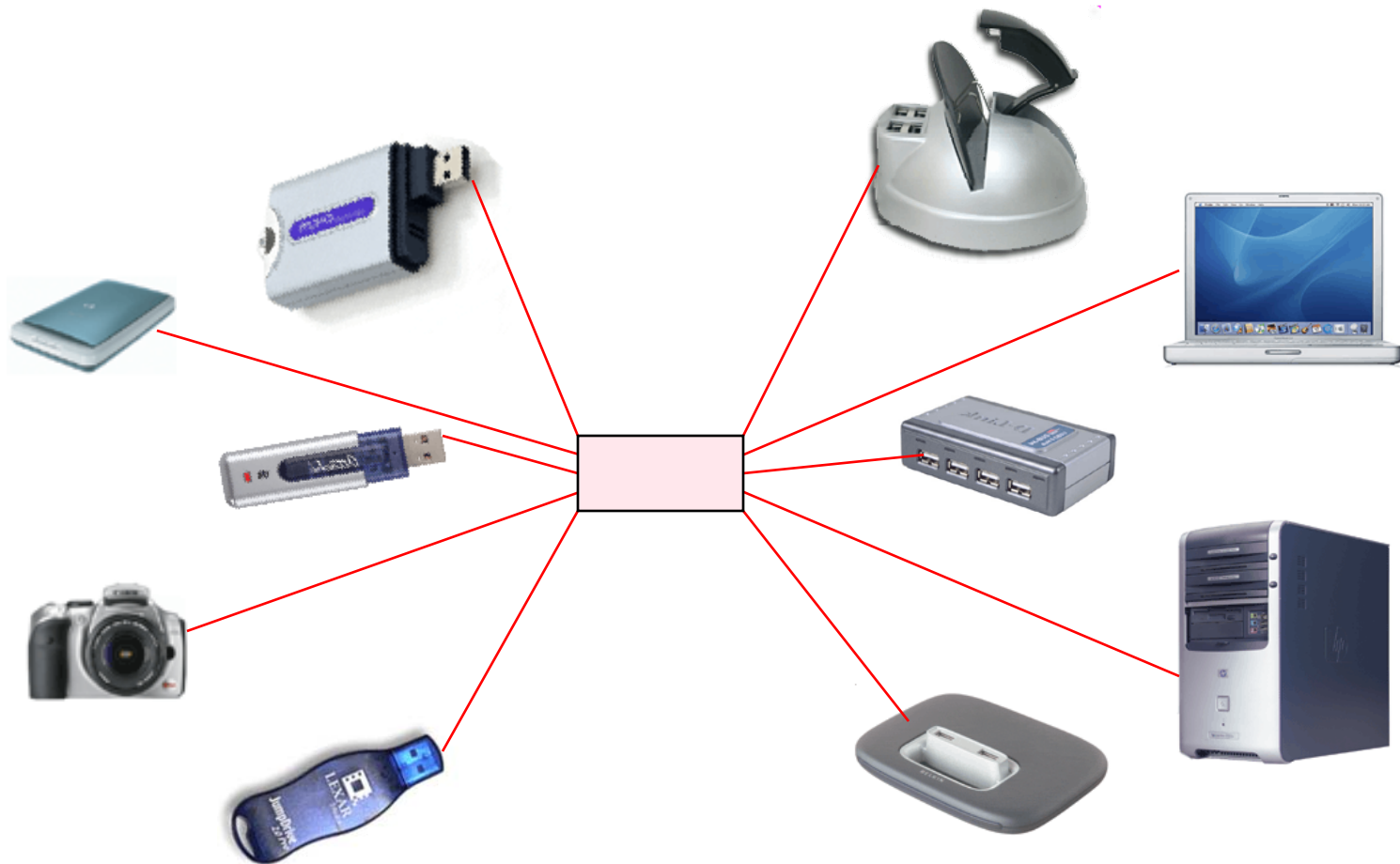
# Good Data-exchange Standard



# Good Data-exchange Standard



# Good Data-exchange Standard



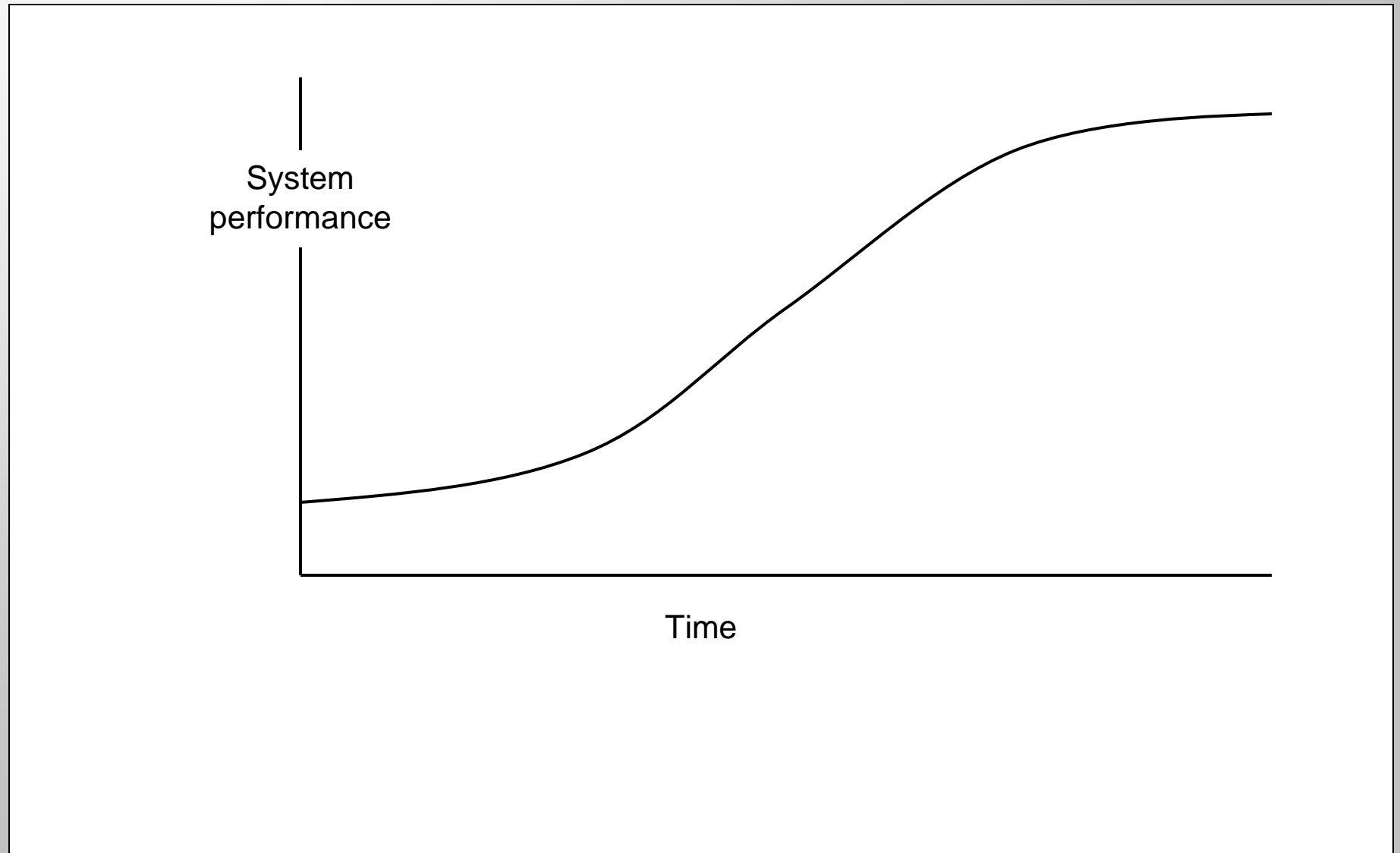


---

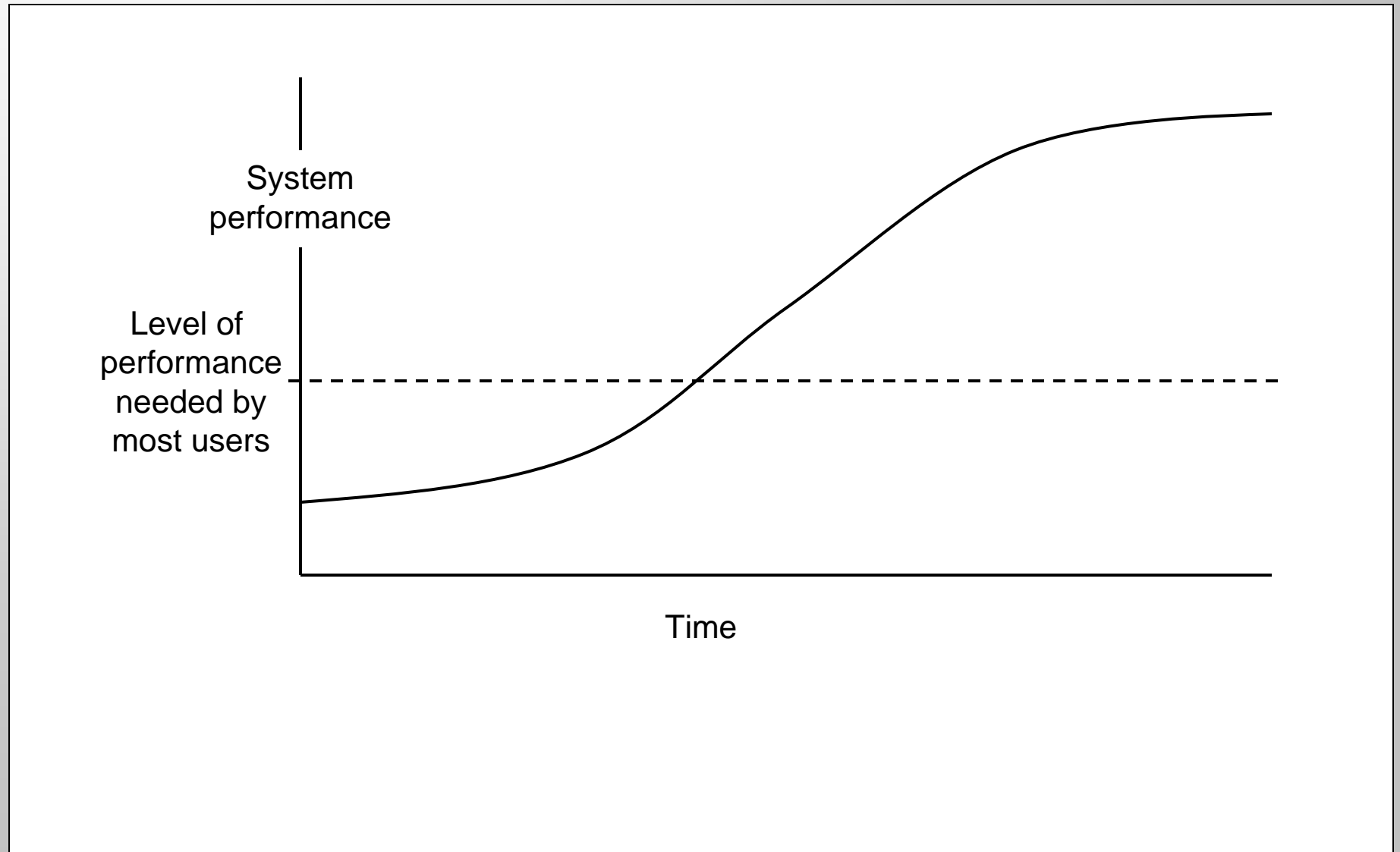
# *Industry Trends*

# Industry Trends

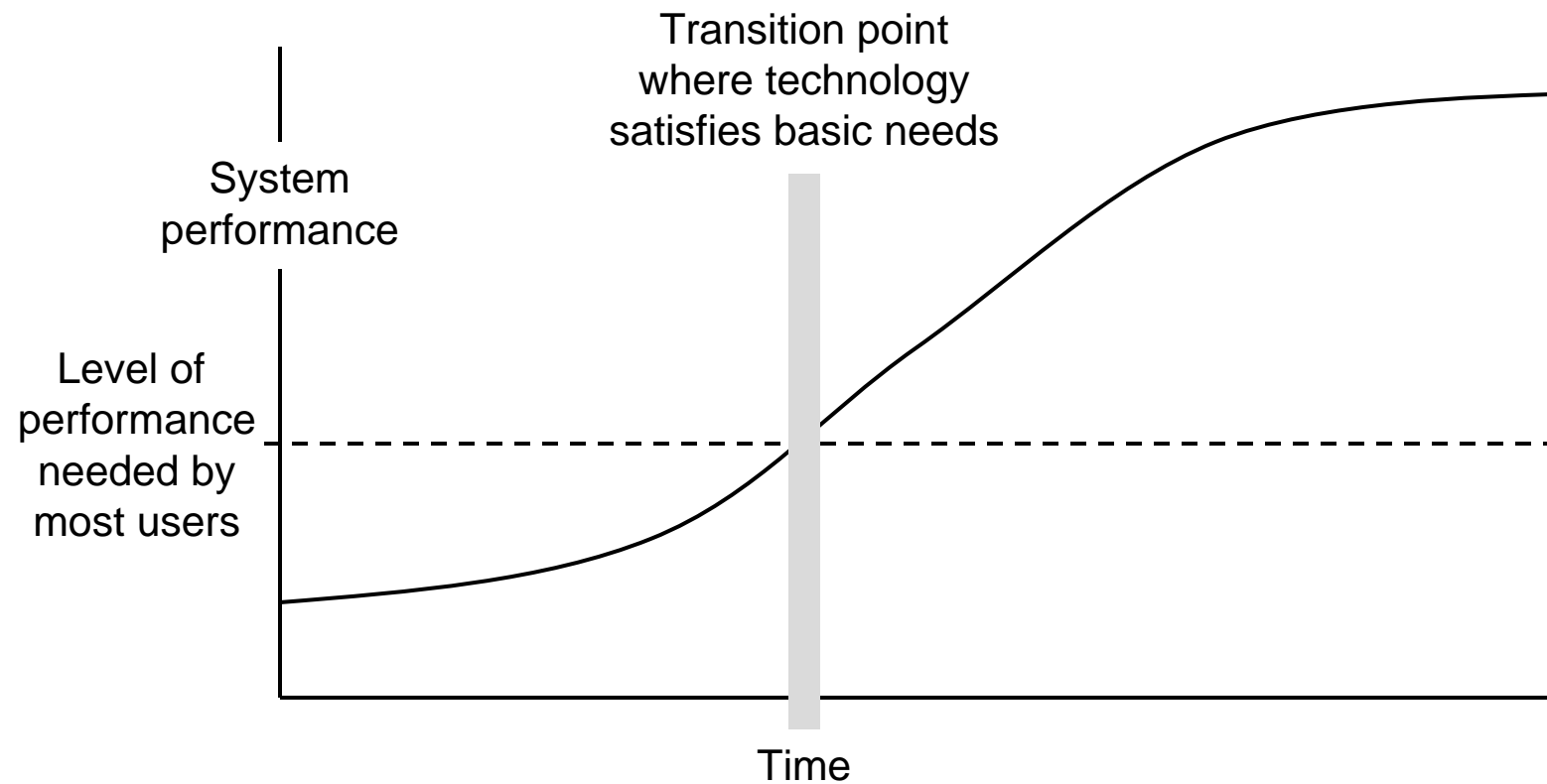
---



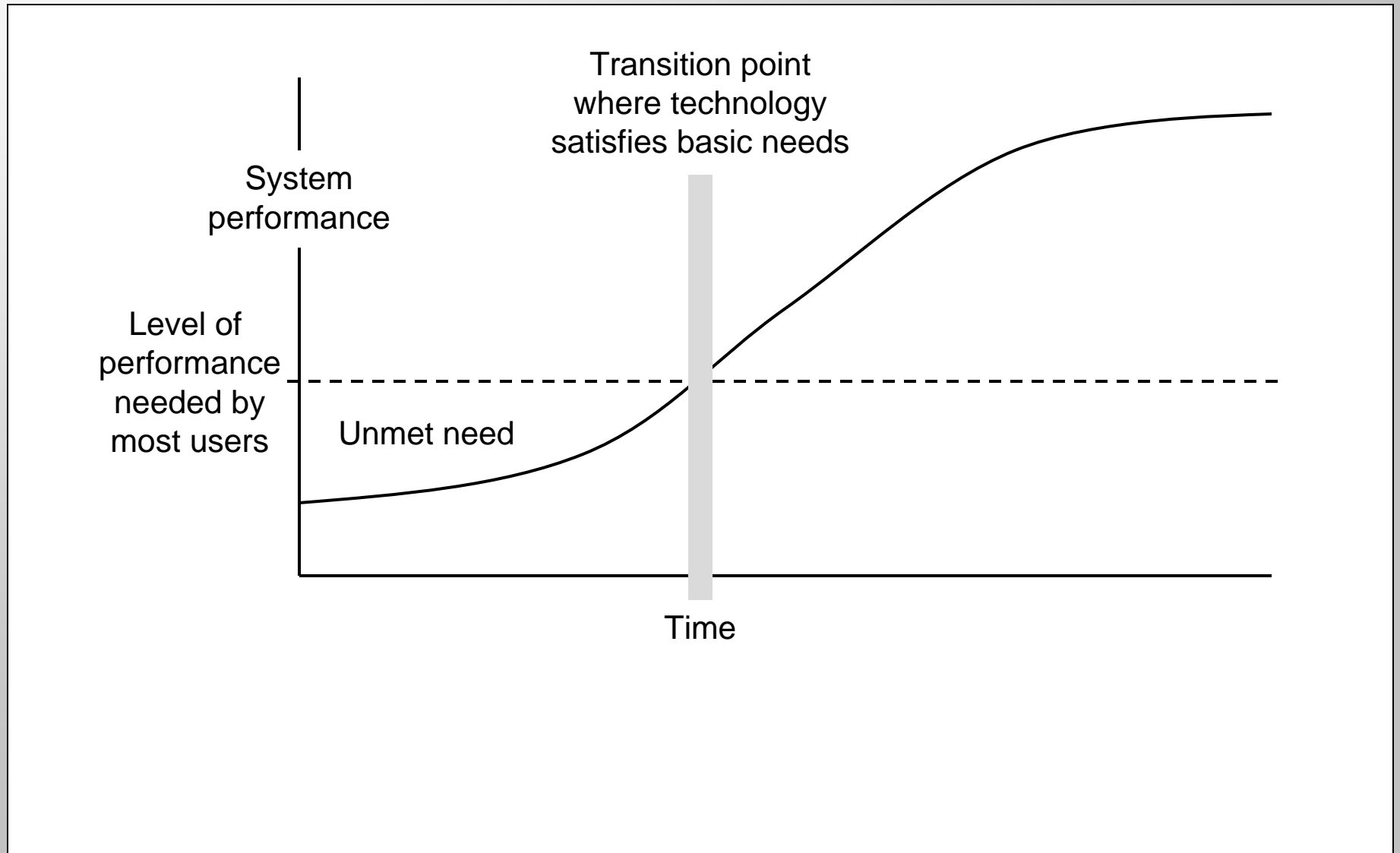
# Industry Trends



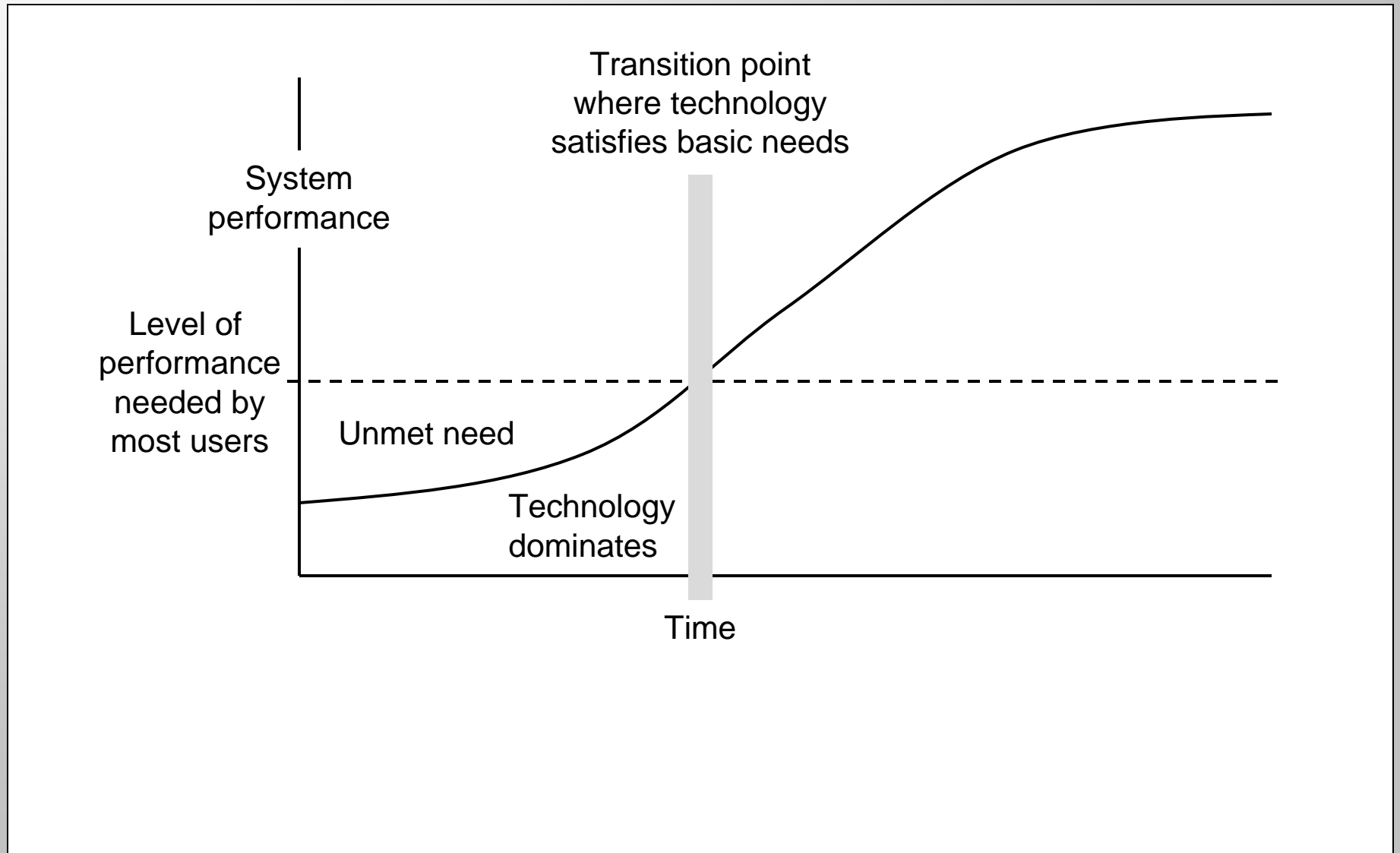
# Industry Trends



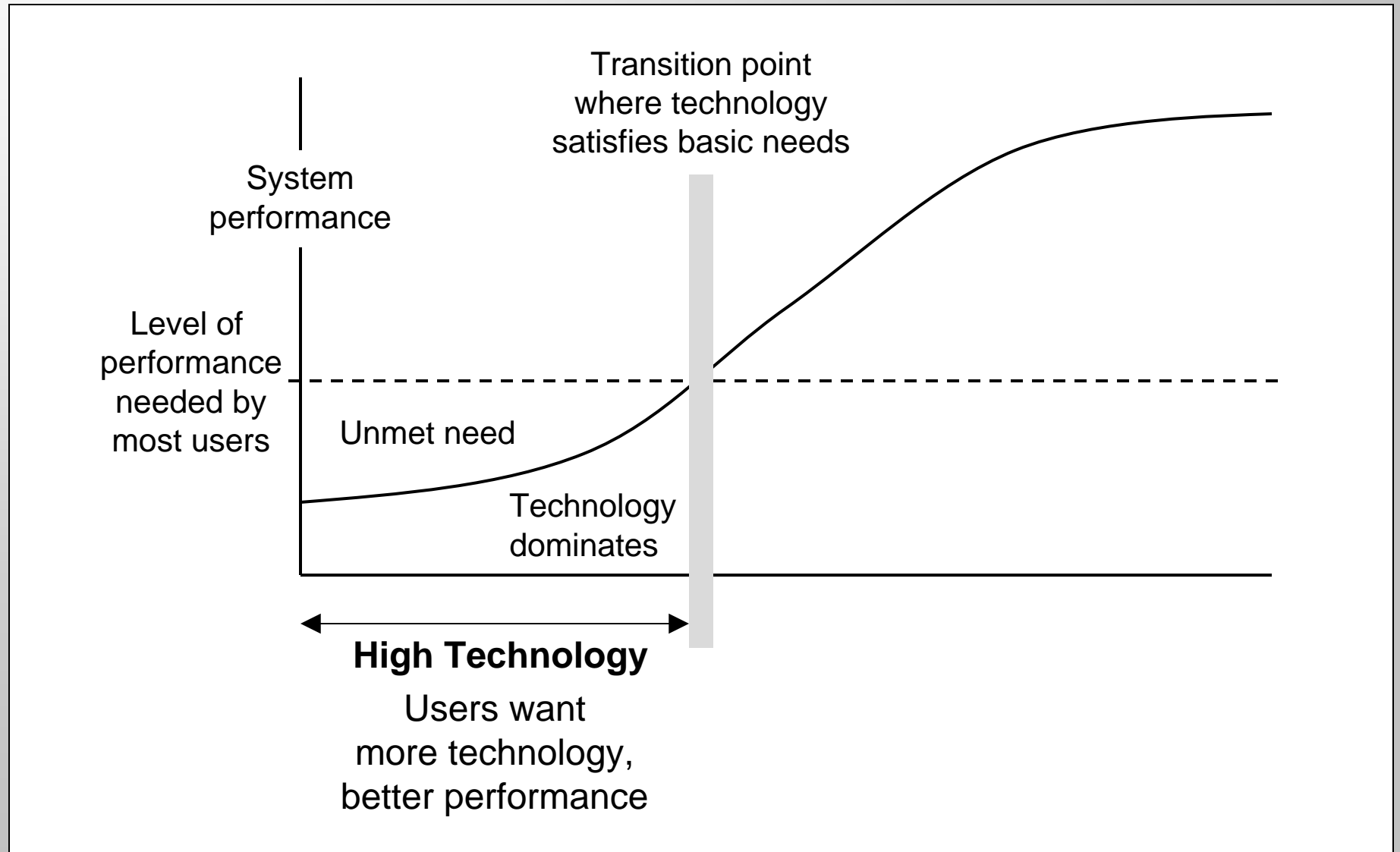
# Industry Trends



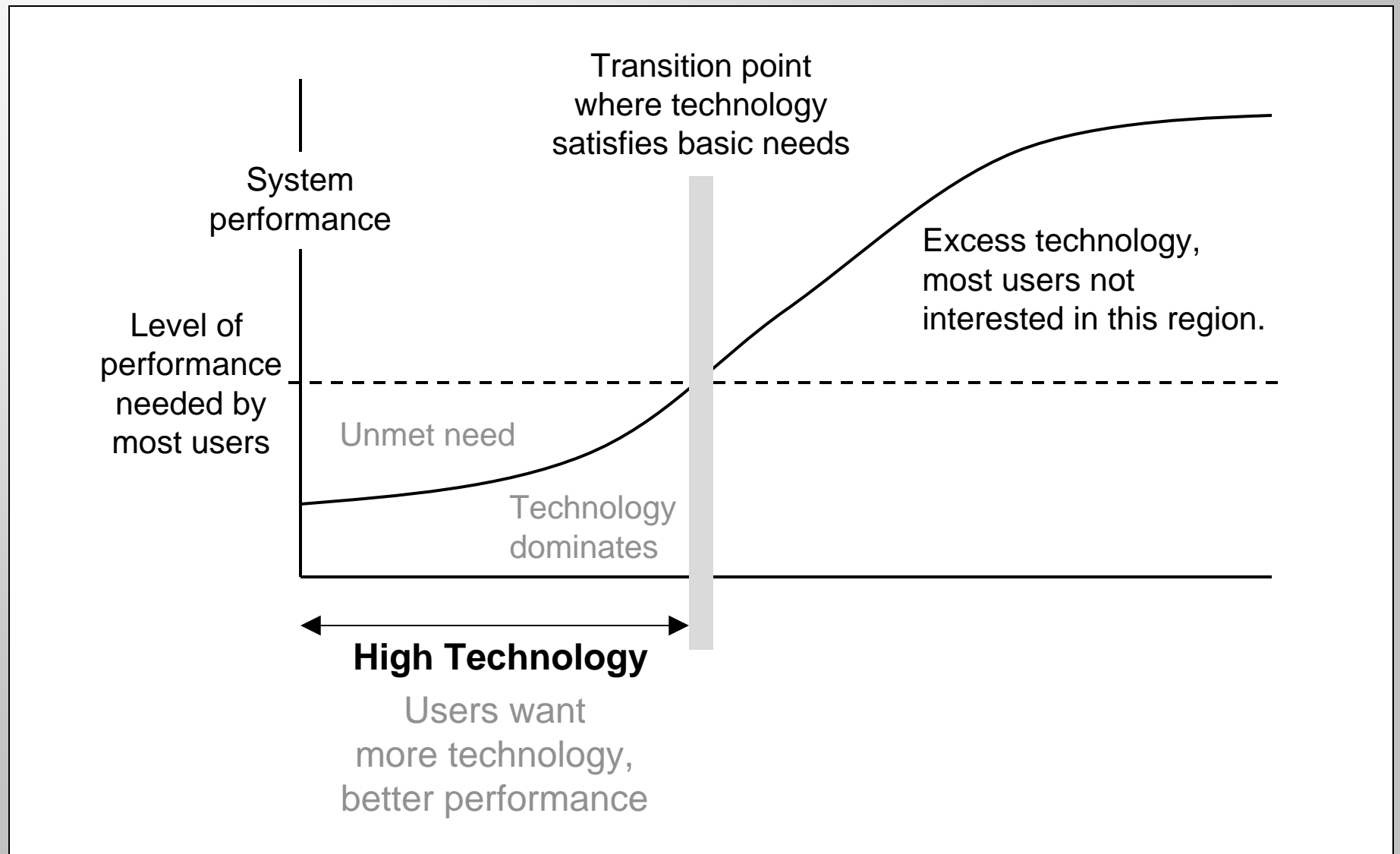
# Industry Trends



# Industry Trends

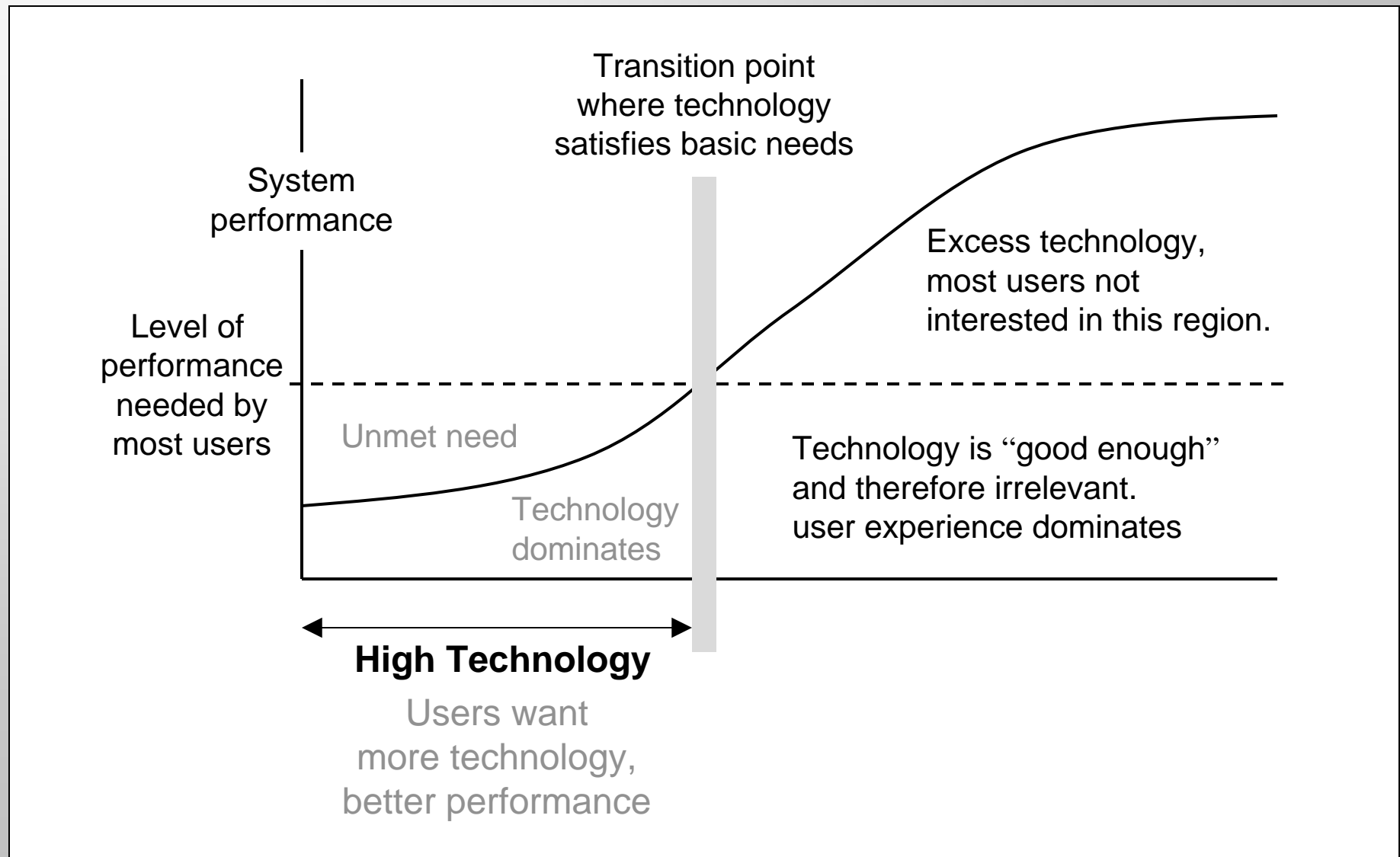


# Industry Trends

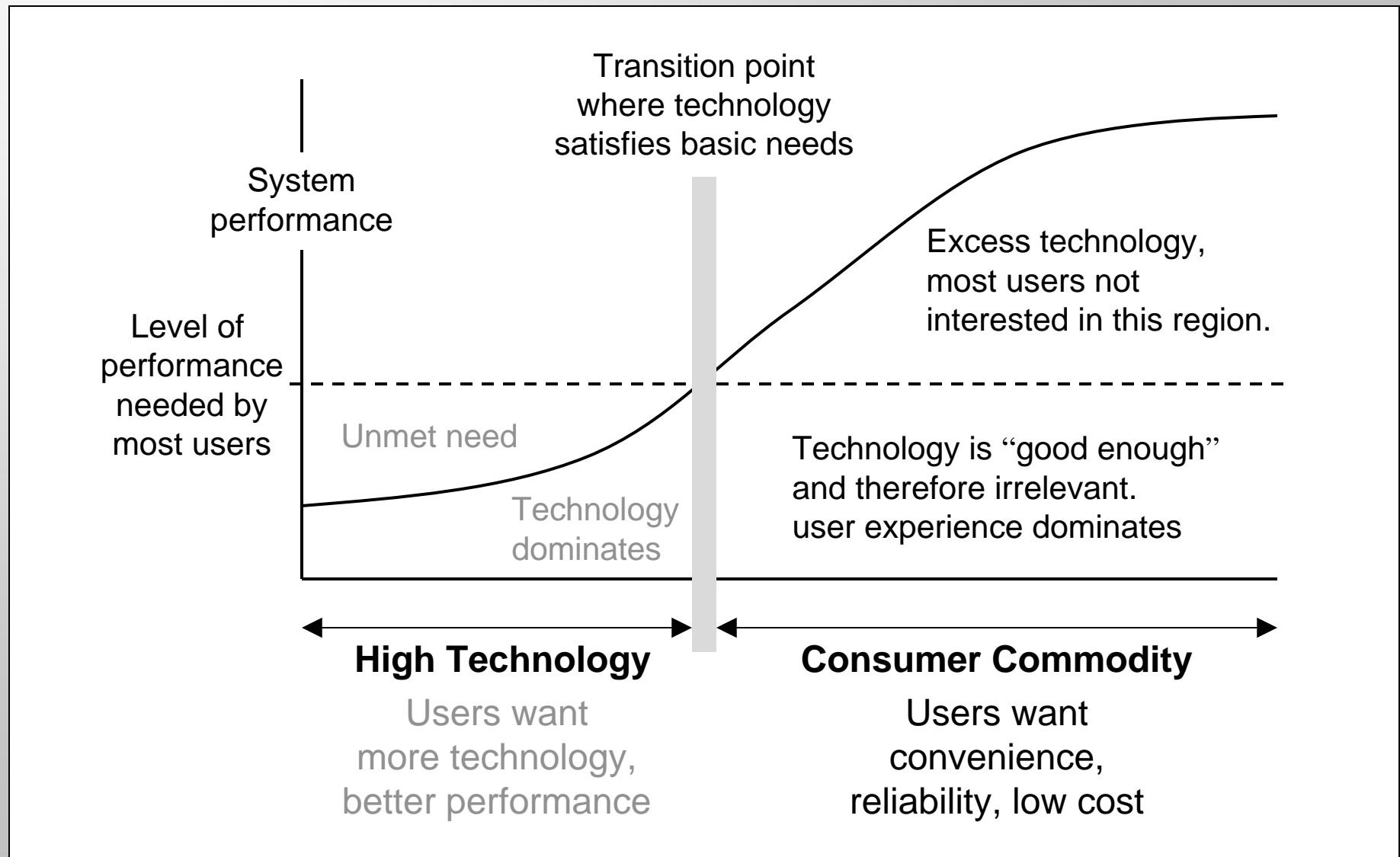




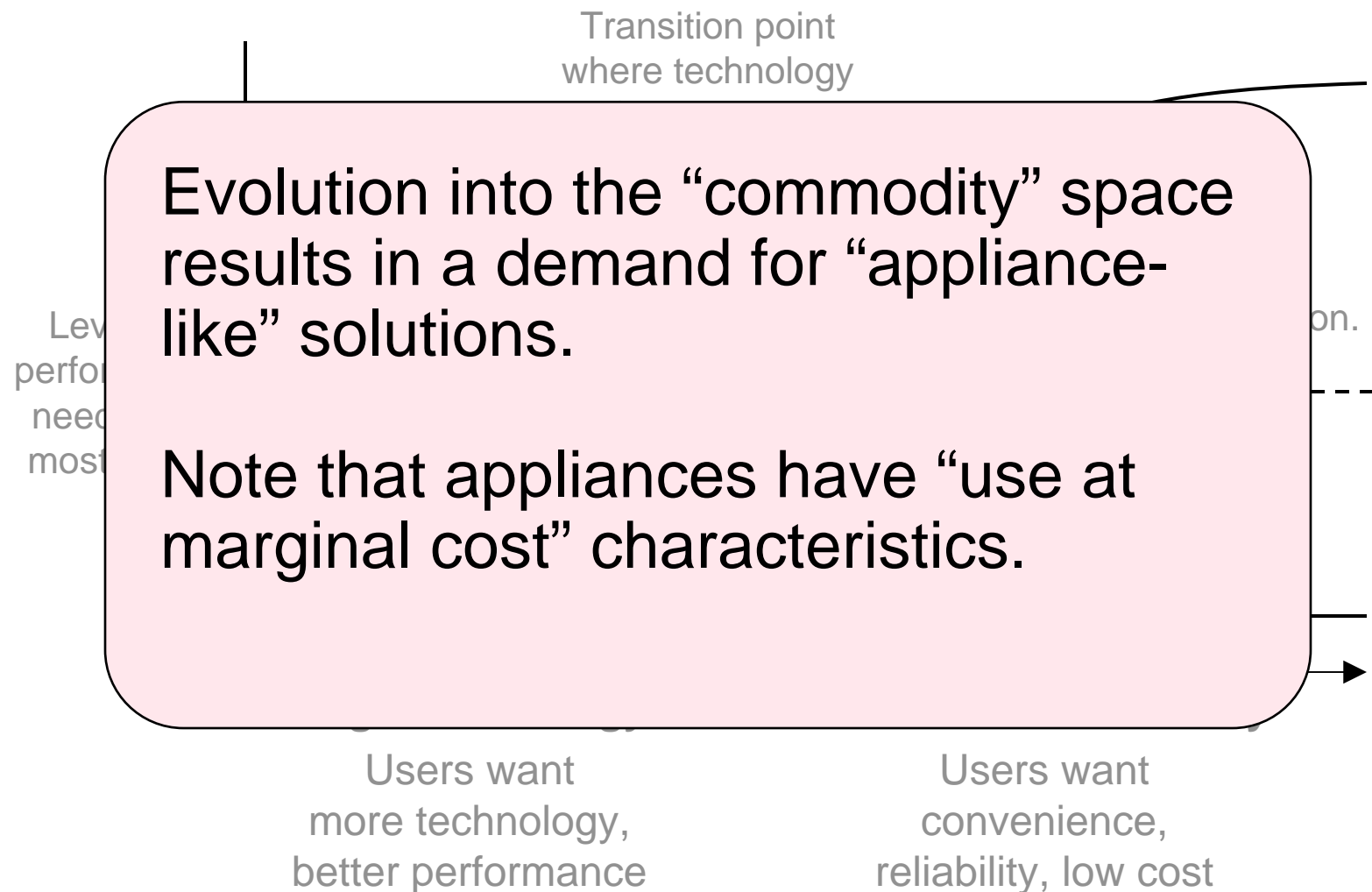
# Industry Trends



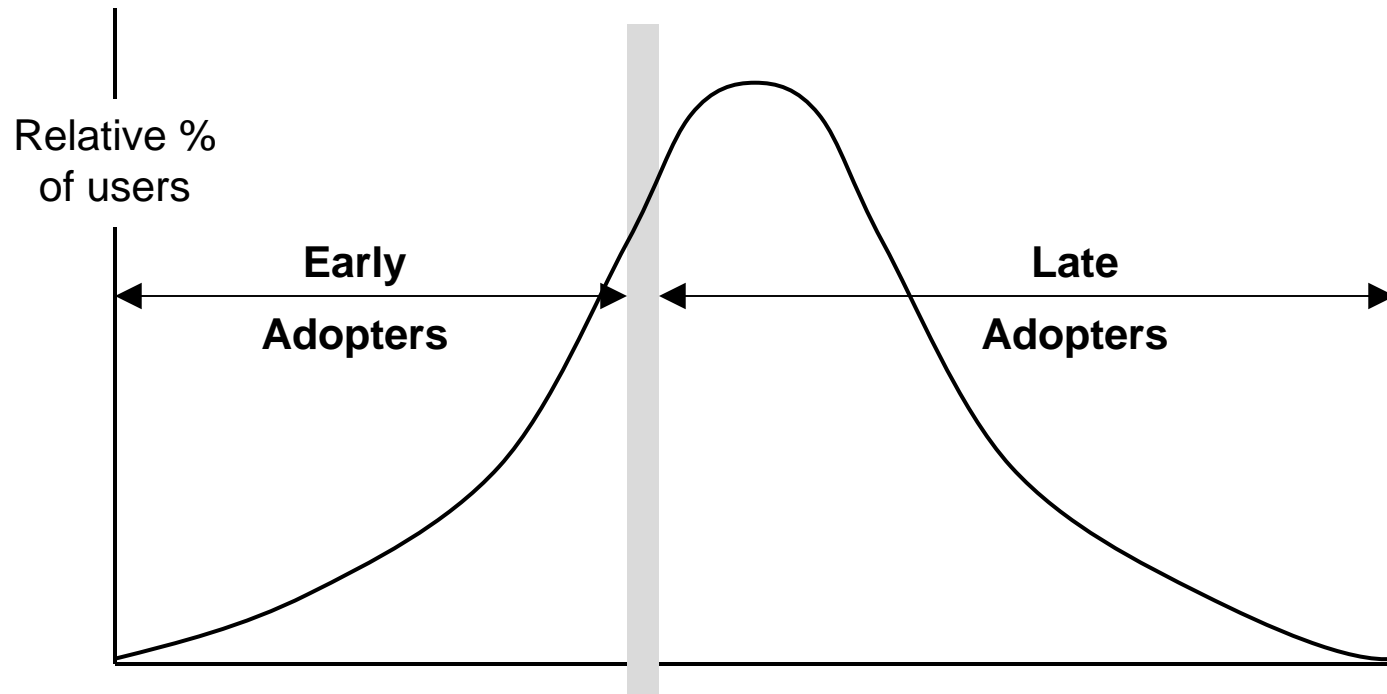
# Industry Trends



# Industry Trends



# 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.

---

*GeMS*

## **Small-Lab Sequencing:**

- An estimate, based on data supplied by ABI, suggests that there are approximately 5,000 small laboratory efforts in the US that are equipped with one or more medium to high throughput sequencers.
- These labs need effective data management systems to deal with the complexities of operating the instrument efficiently and to manage the data produced by the instrument.

## **Possible Data-Management Solutions:**

- Commercial LIMS systems: expensive
- “Roll-your own” LIMS: difficult to achieve, not extensible, prone to failure when developer leaves
- MS-LIMs: not up to the task
- ????

## **Commercially Available Solutions:**

- Applied Biosystem's SQL-GT and Sequence Collector: \$250,000 for the whole set-up.
- Scierra Laboratory workflow system: \$145,000 plus 18% per year maintenance, plus unknown customization fees.
- Geospiza finch server: Costs for various packages from \$60,000 plus \$24,000 per year to over \$100,000 plus 28% of cost/year.



## Quotes from survey respondents:

- “After we obtain the raw sequence data, it is sent on to our users.”
- “Traces are data-based haphazardly by individuals.”
- “As far as I know, there are no low-cost commercial sequence managers available.”
- “We have also 'rolled our own' software here.”
- “Unfortunately, there isn't much out there to the best of my knowledge.”
- “...download into Microsoft Access”
- “There is an in-lab, home constructed, FileMaker Pro database of text files.”

# GeMS

---

To address the data management needs of the small sequencing laboratory, the Geraghty lab at FHCRC has been developing a Genetics Management Software suite (GeMS) - an information-appliance approach to managing sequencing data.

## Goals of GeMS:

- Implement an information-appliance approach to provide targeted support for high-throughput laboratory devices

## Information Appliances:

- are designed to support a specific activity, such as music, photography, or writing.
- combine powerful software applications with the ease of use of household appliances.
- are controlled by simple, intuitive user interfaces that require minimal training to use.
- can be used “out of the box”, without requiring complex configuration or set-up activities.
- connect to digital networks for the purpose of gathering or distributing information.
- manage data in standard formats and can share information easily with other similar systems.

Donald Norman. 1998. *The Invisible Computer: Why Good Products Can Fail, the Personal Computer is So Complex, and Information Appliances are the Solution*. MIT Press. Cambridge, Mass

# GeMS

Small sequencing and genotyping laboratories need IT solutions to help them deal with their sequencing and genotype data. These labs need data management systems that:

- are designed to support a specific activity, such as music, photography, or writing.
- combine powerful software applications with the ease of use of household appliances.
- are controlled by simple, intuitive user interfaces that require minimal training to use.
- can be used “out of the box”, without requiring complex configuration or set-up activities.
- connect to digital networks for the purpose of gathering or distributing information.
- manage data in standard formats and can share information easily with other similar systems.

# GeMS

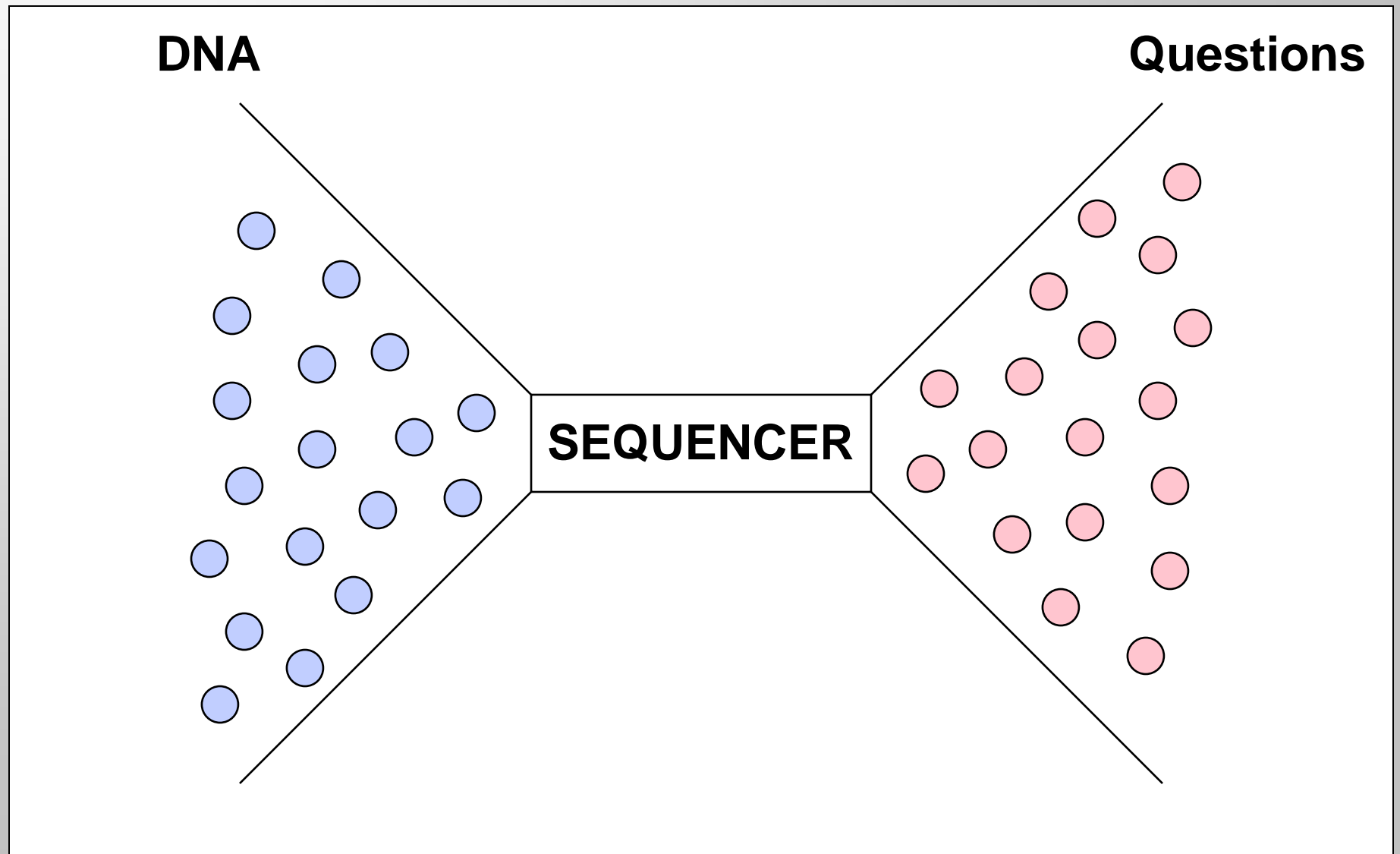
---

## **GeMS Services:**

- Tracking laboratory throughput
- Organizing original data and meta data (machine, reagents, quality, etc.)
- Tracking costs
- Sharing data

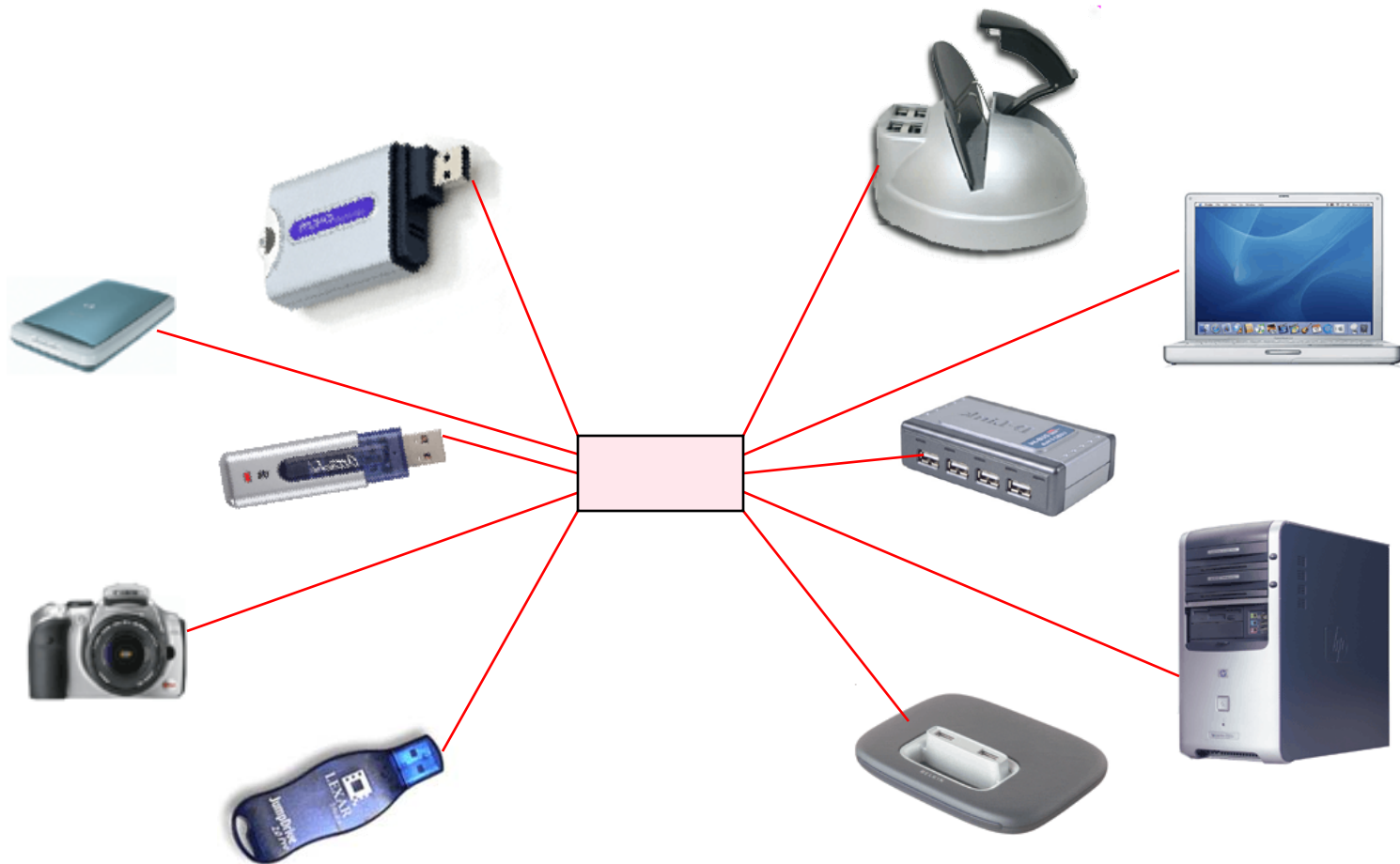
## GeMS Operational Foci:

- Emphasizes commonalities of sequencer-based research.
- Provides a modular and extensible framework for future applications (HTR, Taqman, etc.)
- Views the specifics of the research (organisms, DNA source, scientific questions) as a *detail* to be managed as parameters within a common framework





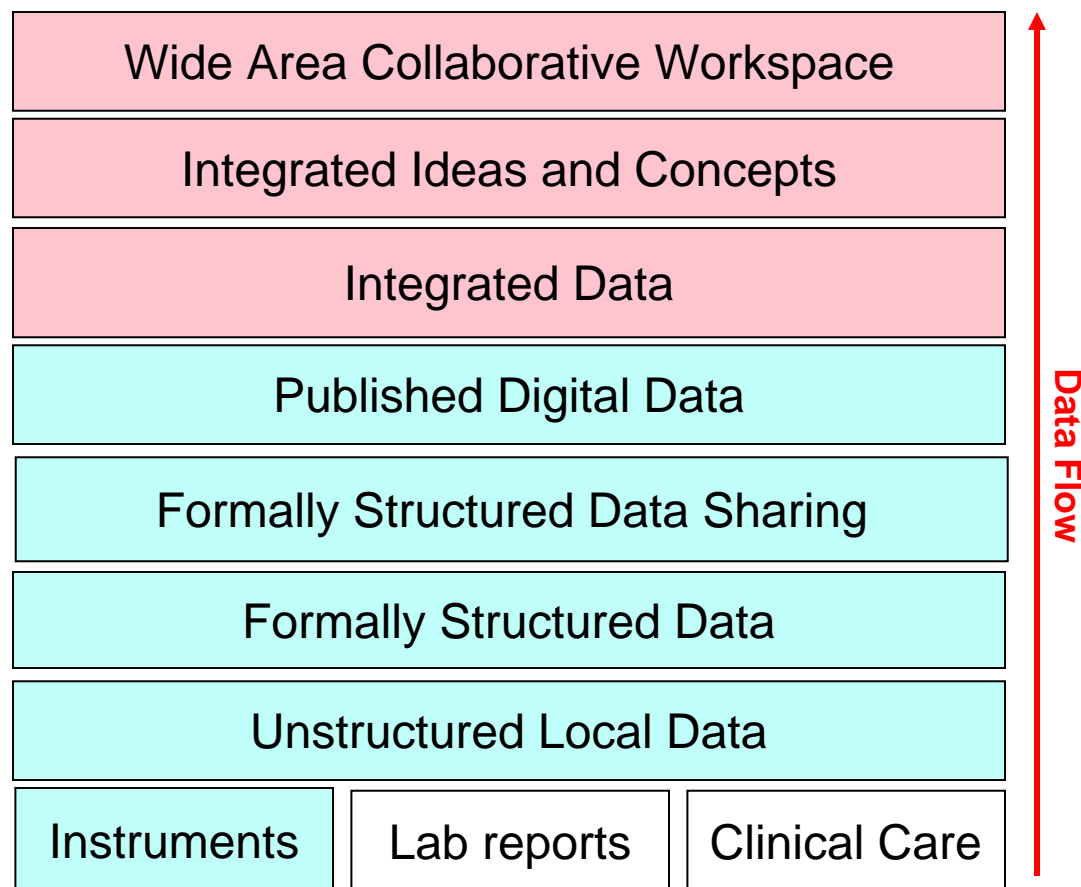
# Good Data-exchange Standard



# GeMS

GeMS uses a layered architecture to match the various processes as data move from initial collection through increasing layers of refinement.

- Wide area data integration is seen as stack of activities
- Local data mgt layers focus on bringing full power of high throughput DNA sequencing instruments into hands of small (R01-funded) laboratory

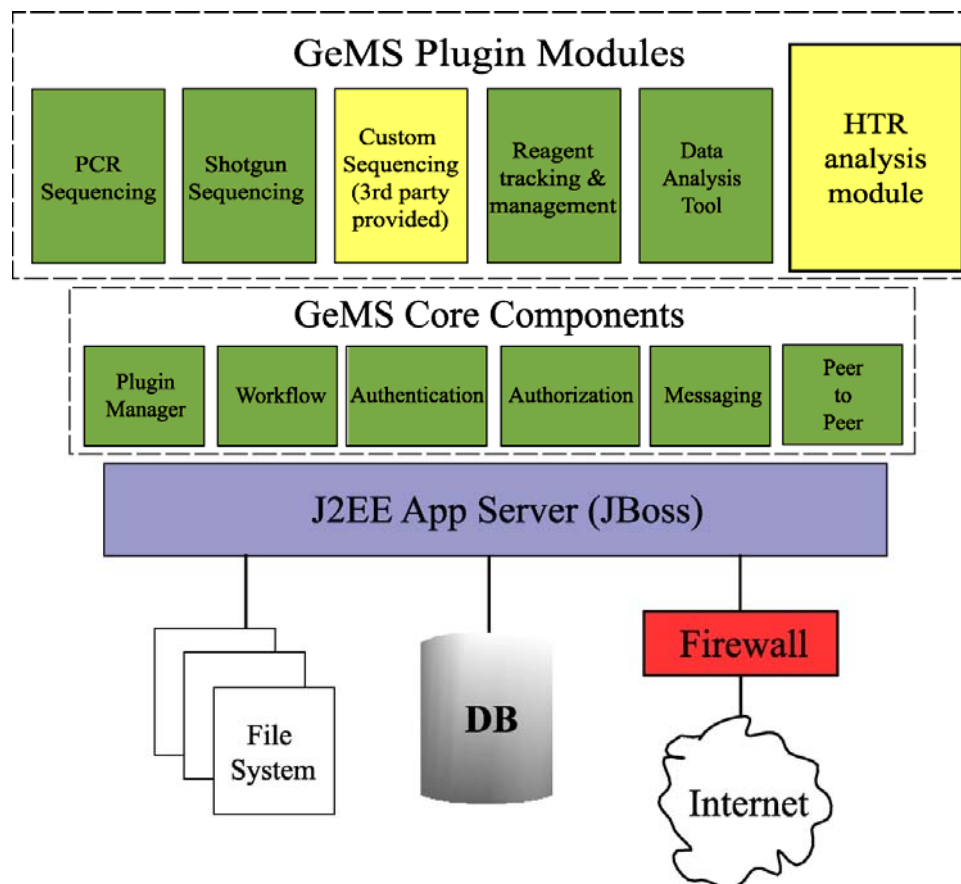


# GeMS

GeMS/IA uses a modular N-tier approach, making it easier to implement and giving it useful flexibility.

- The data store is accessed through a service API.
- Core services are made available using a J2EE framework. These services are used by the plugins to carry out their functions.
- Plugins represent the functional components that use the core services.

At the base, we have developed a Linux-based “turn-key server” to provide an easy to administer foundation. The GeMS-IA core consists of a PostgreSQL database, a J2EE/JBoss application server.



## **GeMS-IA Technical Implementation: Open Source Components**

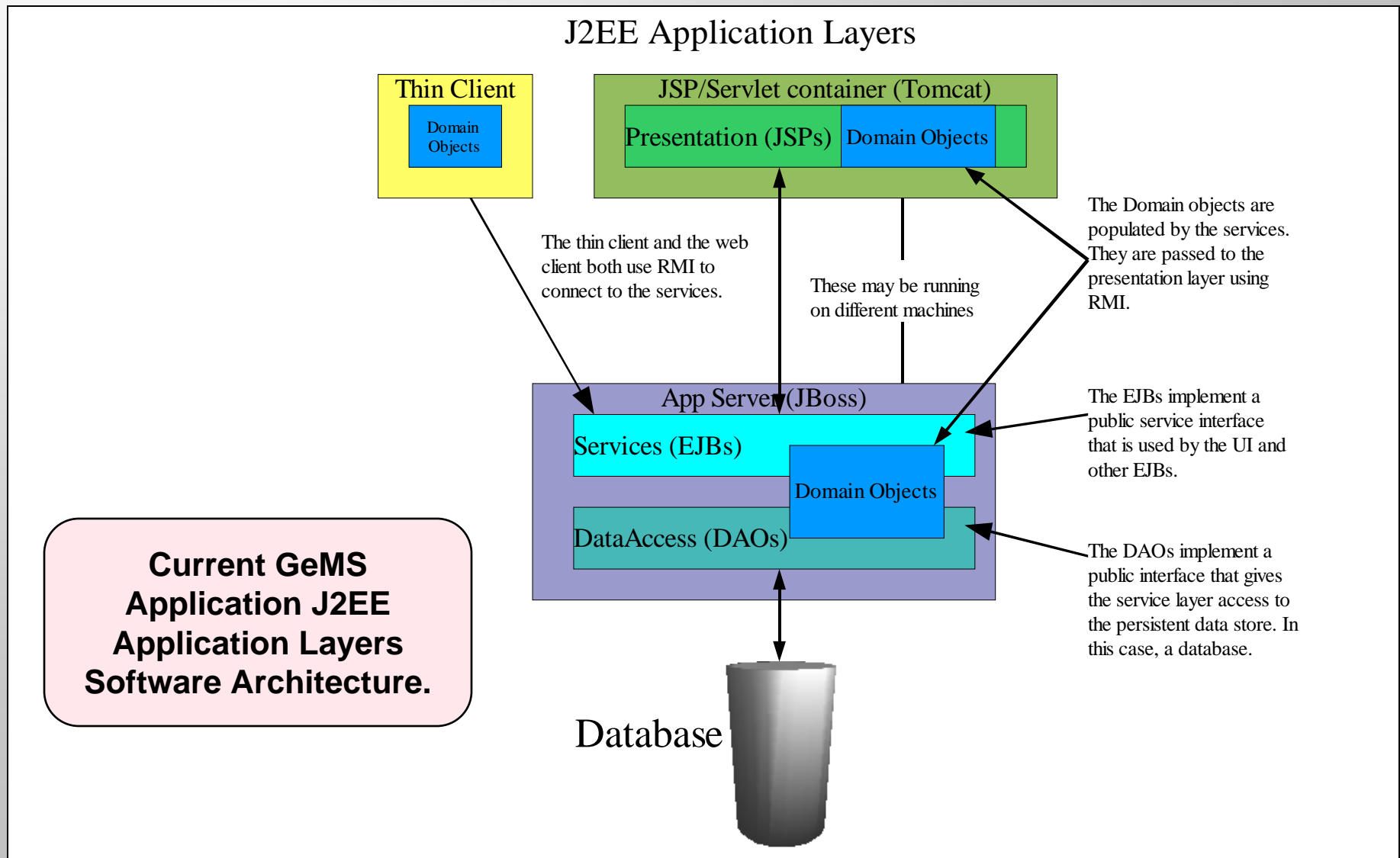
To maximize its cost-effectiveness and extensibility, GeMS/IA has been designed and implemented using open source systems and tools. Specifically,

Operating system:	Linux
System Admin Support:	WebMin
Database:	Postgres
Web server:	Tomcat
J2EE Server:	JBoss
Client Development:	Java

Currently GeMS-IA has 850 classes, and about 140,000 lines of code.

The database has 98 tables.

# GeMS



# GeMS

## GeMS core components

### *Authentication*

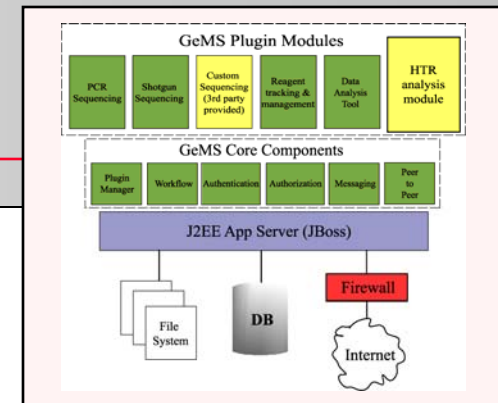
Authentication will be implemented using the J2EE Pluggable Authentication Module (PAM) mechanism.

### *Authorization*

The security requirements of this project require much more flexibility than the standard user/group security model. The requirements specify that access control apply to individual data elements

### *GeMS-IA Messaging*

The messaging component will allow users of the GeMS system to communicate easily and effectively. Users will be able to send and receive messages via email, secure file transfer, adding a message or URL to a web page, and by instant messaging. Recipients may be specified as an individual user or group of users.



# GeMS

## GeMS core components

### *GeMS-IA Work Flow*

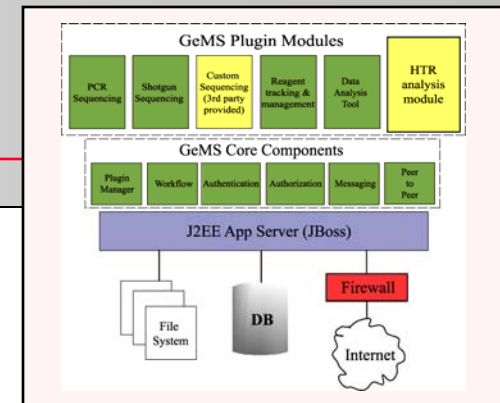
The work flow component will allow users of the GeMS system to collect a series of different tasks into a “work flow.” This will free up the user to perform other work since they will not have to monitor the system as each individual task is completed.

### *Plug-in management*

Support for different protocols and analysis tools will be provided in pluggable modules. These modules are basically J2EE EARs (Enterprise Archives) that build upon the services provided by the platform.

### *GeMS-IA Peer to peer*

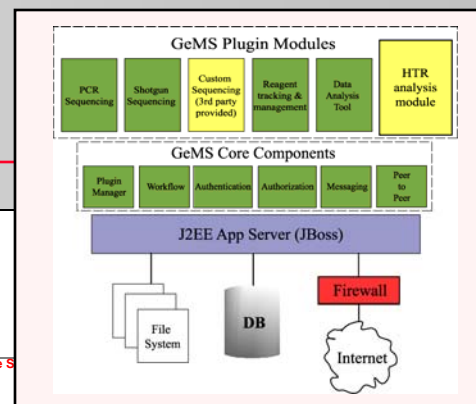
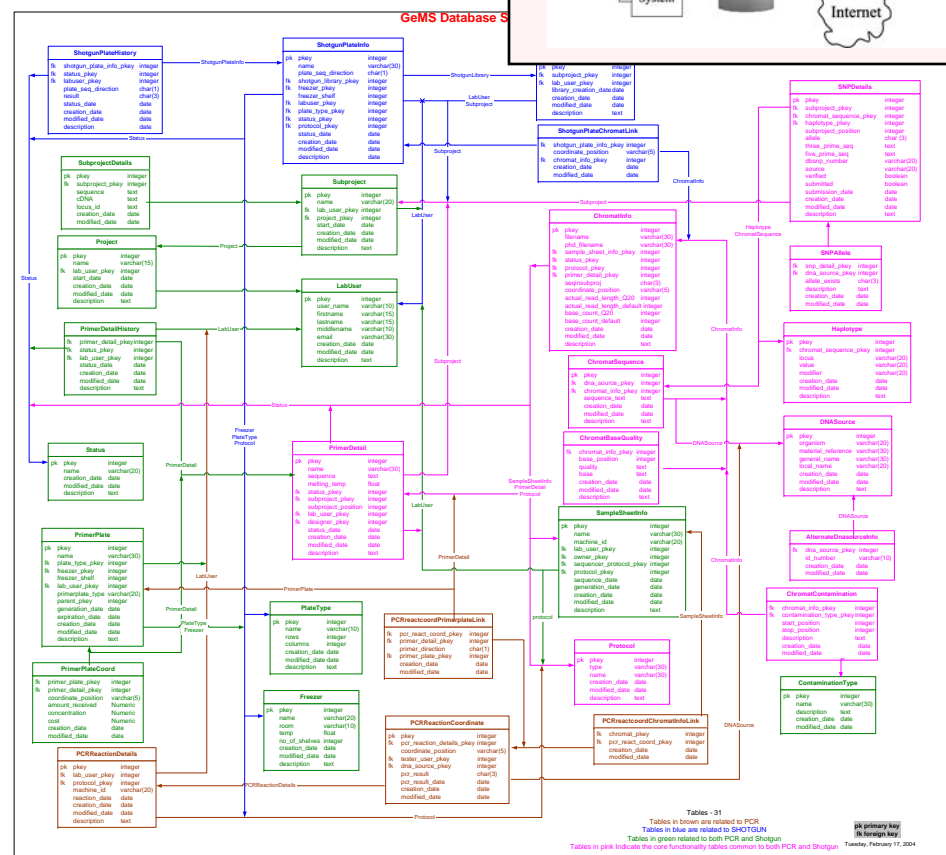
The peer to peer component will allow users of the GeMS-IA system to invoke services on other GeMS-IA instances that are set up as collaborators.



# GeMS Data Schema

**The GeMS schema currently relates all key variables in automated high throughput DNA sequencing to the output files for data analysis, sharing and comparison including**

- **DNA Source information**
- **SNP Identification**
- **Primers**
- **amplicons**
- **Haplotypes**
- **Sequencers**
- **Technicians**
- **PCR Thermocyclers**





# GeMS

## GeMS Plug-in modules

1) *PCR sequencing and 2) Shotgun sequencing.*

In the current GeMS, the PCR and Shotgun modules are packaged together in the GemsSequencing Module.

*Chromatogram quality reporting*

*Sequence assembly reporting*

3) *GeMS sequence analysis tools*

The programming is now broken down into four modular functions, with three directly used for primer design for PCR sequencing.

*Assemble/View Chromats:*

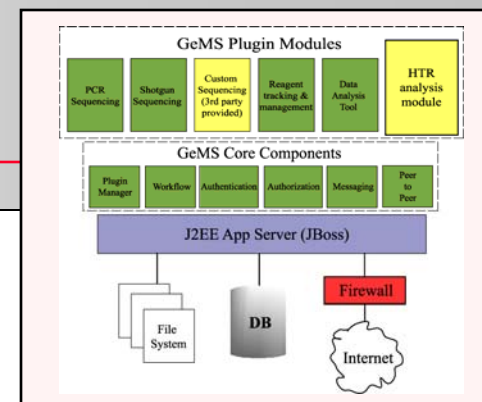
*RepeatMasker utility:*

*Primer3 utility:*

*Blast Primer utility:*

4) *Cost accounting and reagent tracking.*

The primary function of cost tracking is to determine the cost of running a particular protocol.



MyGems - Systems Analyst - Mike McCormick

https://asclepius:8443/GemsWebApp/myGems.do

Google

MyGems - Systems Analys...

**GeMS** [Logout](#)

[Analysis Tools](#) [Cost Tracking](#) [PCR Experiment Design](#) [Reports](#) [Search](#) [Setup](#) [User Admin](#)

### PCR Experiment Design

Project	ACTG
Plate Type*	96_WELL
Plate Orientation( <a href="#">help</a> )*	Top to Bottom, Left to Right
Plate Design( <a href="#">help</a> )*	Cell Line
PCR Procedure*	PCRREACTION
PCR Protocol*	PCRREACTION
Sequencing Procedure*	SEQUENCING_REACTION
Sequencing Protocol*	SEQUENCING_REACTION
Plate Name Prefix*	ACTG
Record Materials	<input type="checkbox"/>
Record Machines	<input checked="" type="checkbox"/>

[Select Amplicons](#) [Select Cell Lines](#) [Design Experiments](#)

MyGems - Systems Analyst - Mike McCormick

https://asclepius:8443/GemsWebApp/myGems.do

Google

bookmarks ▾

MyGems - Systems Analys...

GeMS

[Logout](#)

[Analysis Tools](#)

[Cost Tracking](#)

[PCR Experiment Design](#)

[Reports](#)

[Search](#)

[Setup](#)

[User Admin](#)

Generate Report for PCR Sequencing

Sequencing Date

MM-DD-YYYY  
MM-DD-YYYY\*MM-DD-YYYY for ranges  
% and \_ wildcards not supported

Amplicon Name

Cell Line Name

Sequencing Primer

Project Name

Primer Designer

MIGRATIONX  
FARIBA BARAHMAND  
CHRIS BLANKLEY  
Eileen Ball

PCR Operator

MIGRATIONX  
FARIBA BARAHMAND  
CHRIS BLANKLEY  
Eileen Ball

Sequencing Rxn Operator

MIGRATIONX  
FARIBA BARAHMAND  
CHRIS BLANKLEY  
Eileen Ball

PCR Machine

OLD\_Thermalcycler  
Thermalcycler\_01  
Thermalcycler\_02  
Thermalcycler\_03

Sequencing Machine

ABI3730\_Artemis  
ABI3730\_Hephaestus  
OLD\_ABI3700

Good chromat min length

100

Display

10 results

per page

Submit

Cancel

Save Query

Load Query

Choose File

no file selected

The query supports standard SQL query wildcards: \* is NOT a wild card).

**Plate Level:** [Return to Query Screen](#)

Plate	#of Chromats	#of Good Chromats	Read Length		Operators		Sequencers	PCR Machines
			Q20	Q40	PCR	Sequencing		
<a href="#">990314</a>	96	96	741	735		qvu	ABI3730_Artemis	
<a href="#">990694</a>	64	10	156	148		qvu	ABI3730_Artemis	
<a href="#">990772</a>	64	10	157	149		qvu	ABI3730_Artemis	
<a href="#">992009</a>	80	39	141	136		scnelson	ABI3730_Artemis	
<a href="#">992642</a>	96	92	665	620		mmccormi	ABI3730_Artemis	
<a href="#">994772</a>	54	49	742	711		rdaza	ABI3730_Artemis	
<a href="#">994882</a>	54	51	742	708		rdaza	ABI3730_Artemis	
<a href="#">999404</a>	64	52	514	494		rdaza	ABI3730_Artemis	
<a href="#">999514</a>	64	52	496	463		rdaza	ABI3730_Artemis	
<a href="#">999633</a>	68	56	503	484		rdaza	ABI3730_Artemis	

11 - 20 Of 25

[firstPage](#)

[prevPage](#)

[nextPage](#)

[lastPage](#)

# GeMS: Productivity Gains

Parameter	Improvement
Homologies Mapping	time reduced four fold (estimated 20 hours/year).
Primer Quality	eliminated design errors (start/end pos.) from 5% of all primers to none. reduced strand errors from 1% of all primers to none (combined estimated 100 hours/year including laboratory time saved).
Primer Ordering	automation saved one hour per plate (40 hours/year).
Sample Sheet Creation	automation saved 5 minutes per plate (200 hours/year).
PCR/Seq plate map creation	shows user which cell lines, primer(s), go in each well. Reduces user errors and save time setting up experiments (estimated 400 hours/year including laboratory time saved).
Chromatogram Quality Reports	saved 30 minutes per quality output summary (estimated 100 hours/year) eliminated naming errors – saved variable time depending on number and complexity of naming errors (estimated total 200 hours/year including laboratory time saved).
Data Organization	Able to easily group together chromats based on a list of criteria. (e.g. group all chromats from one cell line, or all chromats from one amplicon, etc.) Saved variable time and reagent cost checking quality criteria depending on the size of the project. (estimated \$10,000 reagent costs and 100 hours/year laboratory time saved).
Managing assemblies	GeMS saves chromats in a centralized place and can dynamically create assemblies in any combination desired. This avoids data duplications and saves both space and analysis time (estimated 200 hours/year).
Cost Tracking	Improved from general estimation to detailed tracking that related work effort and reagent cost to a specific protocol being run over time (saved 20% reagent costs or \$30,000 annualized).

# GeMS: Productivity Gains

Parameter	Improvement
Homologies Mapping	time reduced four fold (estimated 20 hours/year).
Primer Quality	eliminated design errors (start/end pos.) from 5% of all primers to none. reduced strand errors from 1% of all primers to none (combined estimated 100 hours/year including laboratory time saved).
Primer	
Sequencing	
PCR	
Chromatograms	
Data Organization	Able to easily group together chromatograms based on a list of criteria. (e.g. group all chromatograms from one cell line, or all chromatograms from one amplicon, etc.) Saved variable time and reagent cost checking quality criteria depending on the size of the project. (estimated \$10,000 reagent costs and 100 hours/year laboratory time saved).
Managing assemblies	GeMS saves chromatograms in a centralized place and can dynamically create assemblies in any combination desired. This avoids data duplications and saves both space and analysis time (estimated 200 hours/year).
Cost Tracking	Improved from general estimation to detailed tracking that related work effort and reagent cost to a specific protocol being run over time (saved 20% reagent costs or \$30,000 annualized).

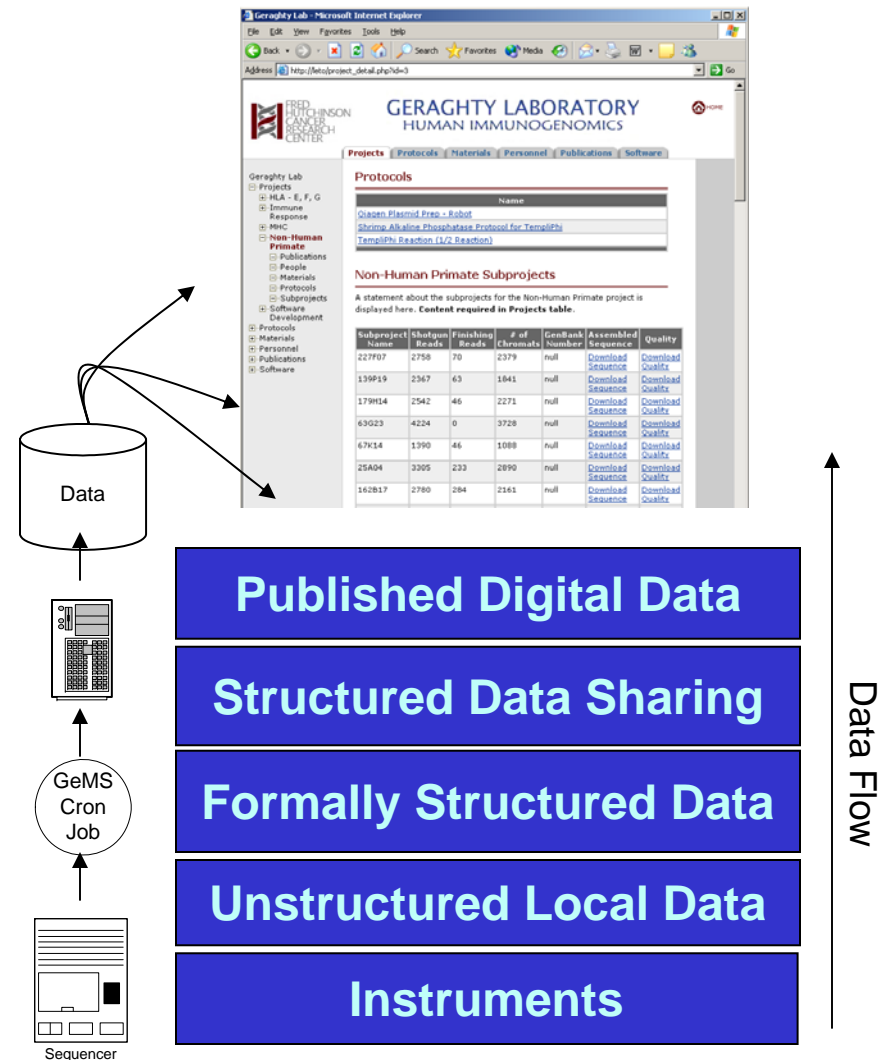
**Total Savings:**

**1360 hrs + \$40,000 reagent costs/yr**

# GeMS

## From Data Generation to Data Publication:

- Nightly Data pick up by system
- Unstructured and unrelated data sent to GeMS server for processing
- Data related to associated parameters
- Subset of data made available to the Geraghty website





# GeMS Extensibility

---

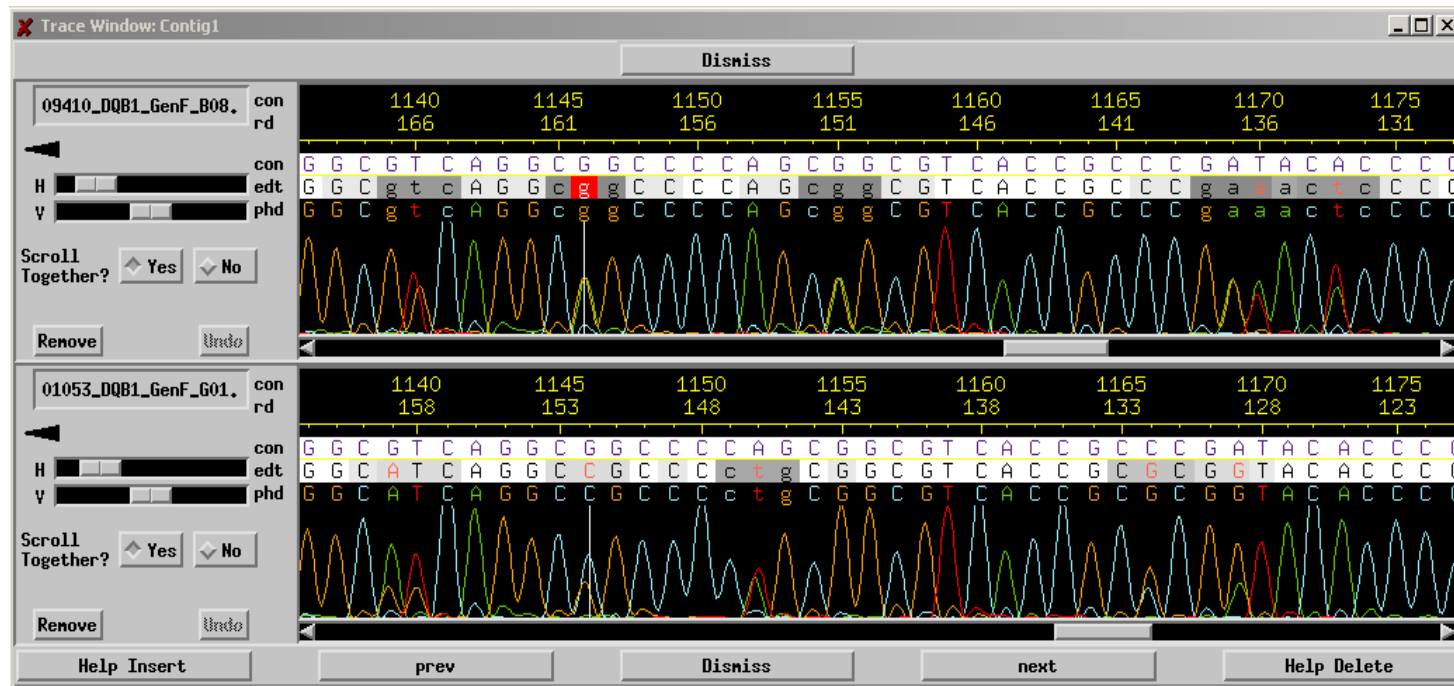
## **Heterozygous Trace Resolution software (HTR):**

- Interprets heterozygous DNA sequence data directly from the chromatogram without manual interpretation.
- Written in Java
- Current implementation does not have user interface.
- Undergoing upgrades to improve accuracy and to deliver data quality metrics.



# GeMS Extensibility

## Two cell lines - multiple polymorphic positions



# GeMS Extensibility

---

## More Future Plans for GeMS/IA:

- To build a new module for an additional genetics data generating instrument (Taqman, Sequenome(?), FACSAN).
- To create and maintain the ability to connect distributed installations supporting the two distinct types of genetics instruments (sequencers and Taqman).

NOTE: Many of the problems associated with data sharing between labs simply disappear if the labs employ common informatics systems and common data models.

- To create and maintain an adaptation of the existing EDRN Research Network Exchange (ERNE using OODT) that will assist Import/export functions for distributed GeMS installations with other widely available databases containing genetics data.

# GeMS People

---

## ***Immune response genes***

Quyen Vu  
Skylar Nelson

Dan Geraghty, Ph.D.  
PI / Lab Director  
geraghty@fhcrc.org  
206 667 4668

## ***GeMS software development***

Lee Davis\*  
Mike McCormick\*  
Simon Fortelny\*  
Ruihan Wang\*

Mark Thornquist, Ph.D.  
(Public Health Sciences Division,  
FHCRC), EDRN related initiative

## ***HTR software package***

Ruihan Wang\*  
Wade Smith\*

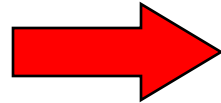
Thomas Geraghty\*, (COO,  
Immunogenomics Inc.) Off-site  
testing, support, and requirements  
gathering

# GeMS People

---

## ***Immune response genes***

Quyen Vu  
Skylar Nelson



Dan Geraghty, Ph.D.  
PI / Lab Director  
geraghty@fhcrc.org  
206 667 4668

## ***GeMS software development***

Lee Davis\*  
Mike McCormick\*  
Simon Fortelny\*  
Ruihan Wang\*

Mark Thornquist, Ph.D.  
(Public Health Sciences Division,  
FHCRC), EDRN related initiative

## ***HTR software package***

Ruihan Wang\*  
Wade Smith\*

Thomas Geraghty\*, (COO,  
Immunogenomics Inc.) Off-site  
testing, support, and requirements  
gathering

# **The Solution**

# Location of Solution Components

---

**LABORATORY:** QA/QC; basic data management and analysis

**INSTITUTION:** Shared resources; basic storage & management; statistics and analysis support; digital publishing support;

**RES. COMMUNITY:** Information appliances; public data collections; analytical support

**FUNDING AGENCY:** Core grant support; caBIG; BISTI

**GLOBAL:** Identity management; authentication, authorization, auditing

**END**

# **Understanding Research**



# **Understanding Research**

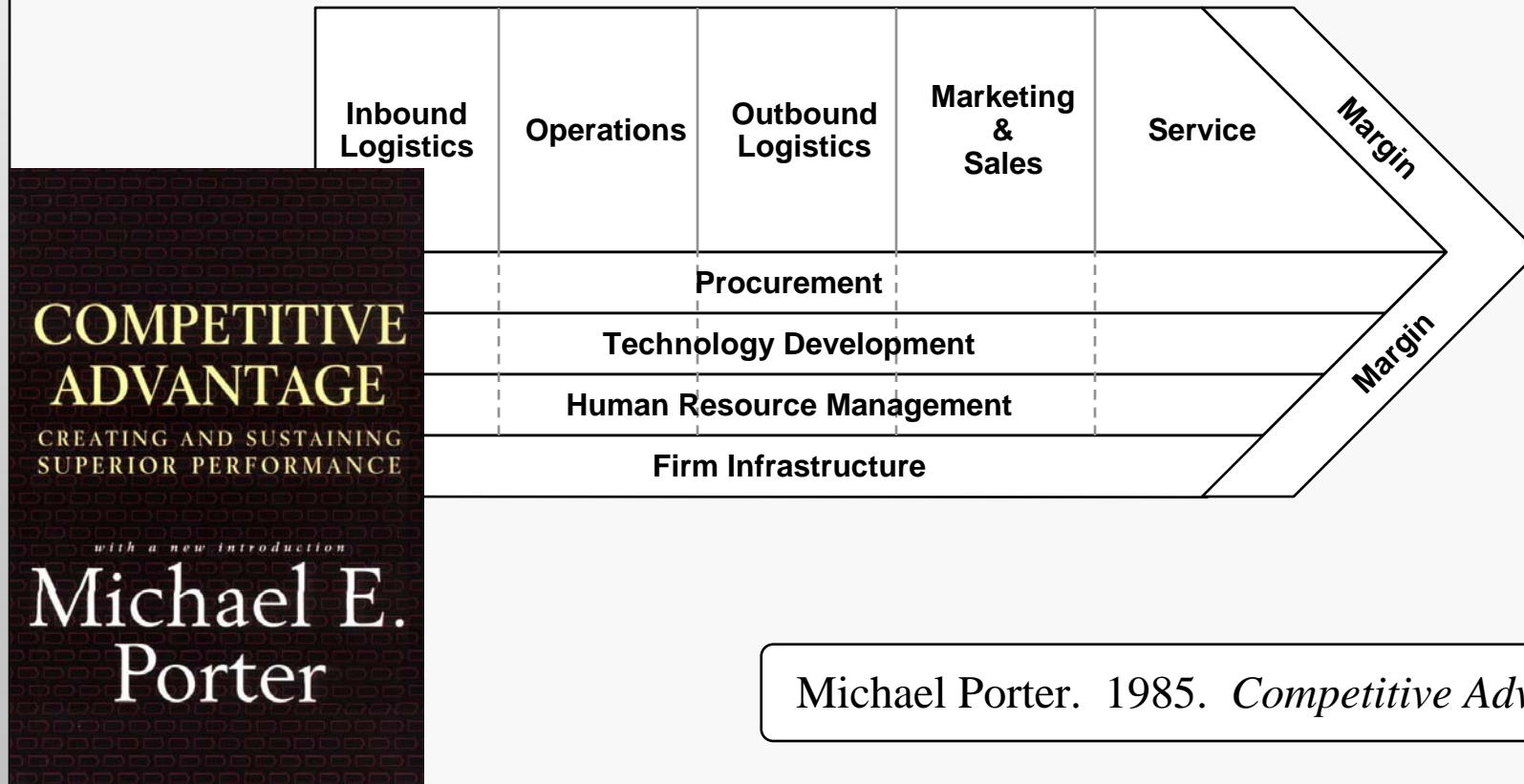
---

**Business Model**

# **Michael Porter Value-Chain Analysis**

# The Value Chain: *Commerce*

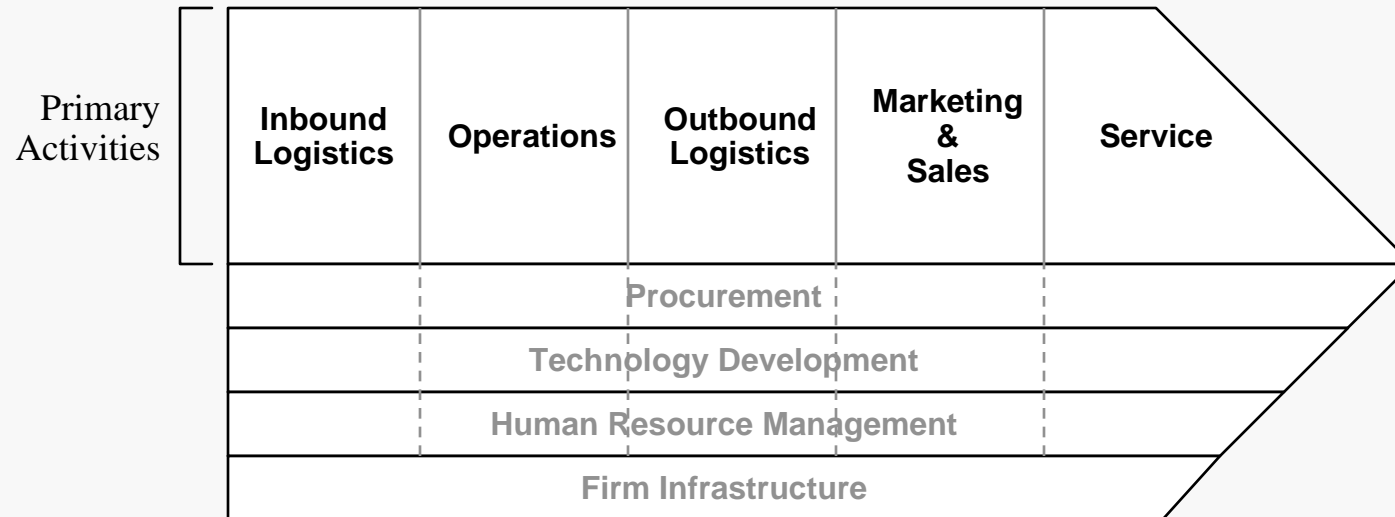
Michael Porter's works on competitive advantage contain a compelling analysis of the various components of operational activities in a competitive enterprise.



Michael Porter. 1985. *Competitive Advantage*.

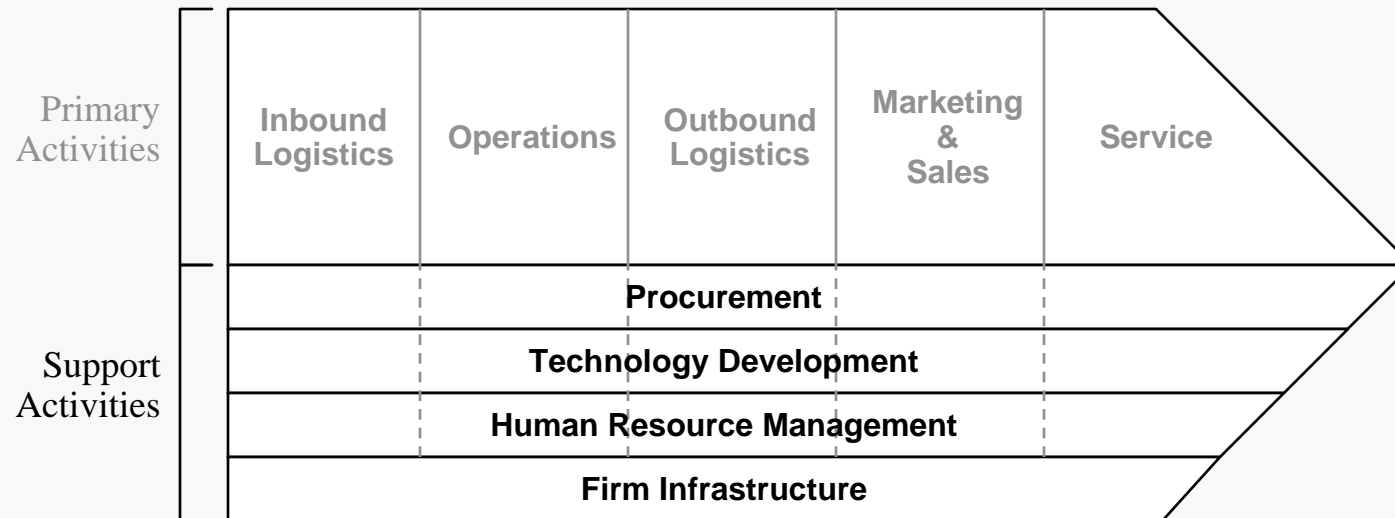
# The Value Chain: *Commerce*

According to Porter, the value-adding **primary activities** of the enterprise define the enterprise. Primary activities must be managed to deliver maximum strategic competitive advantage.



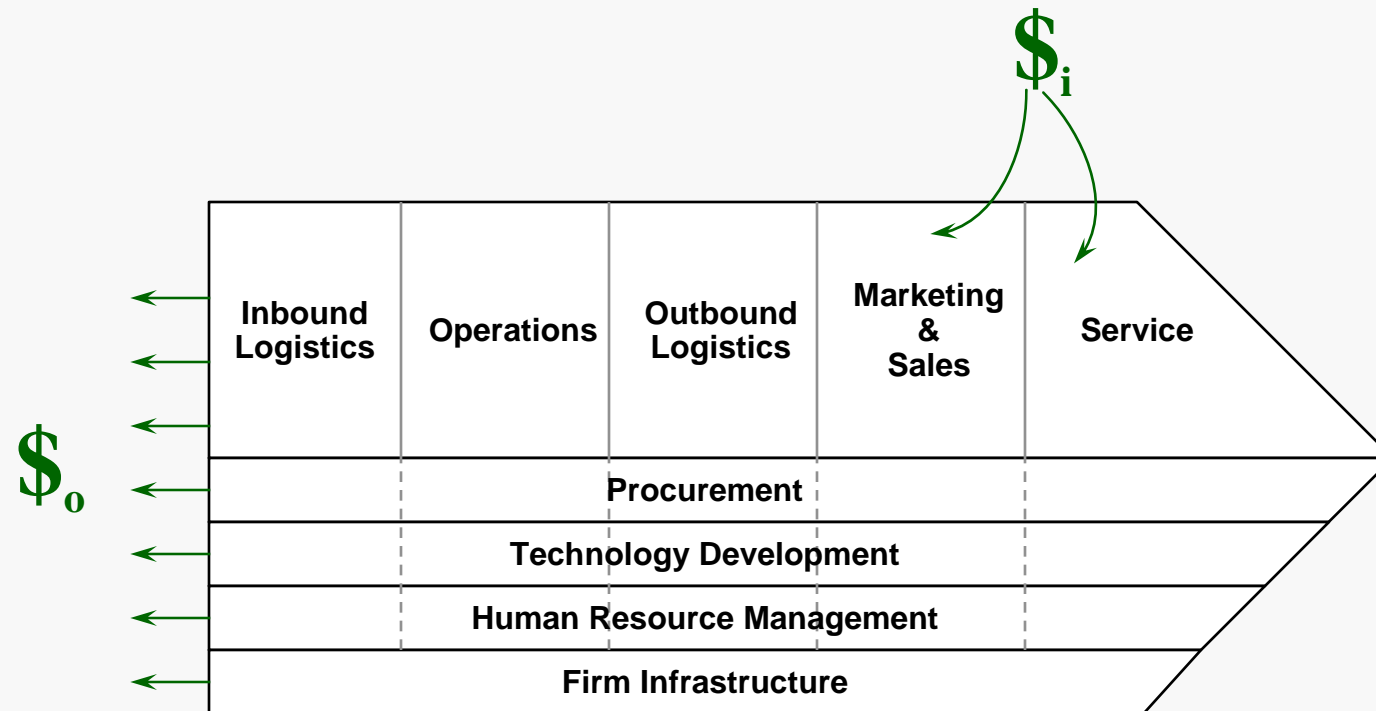
# The Value Chain: *Commerce*

According to Porter, the value-adding **primary activities** of the enterprise define the enterprise. Primary activities must be managed to deliver maximum strategic competitive advantage.



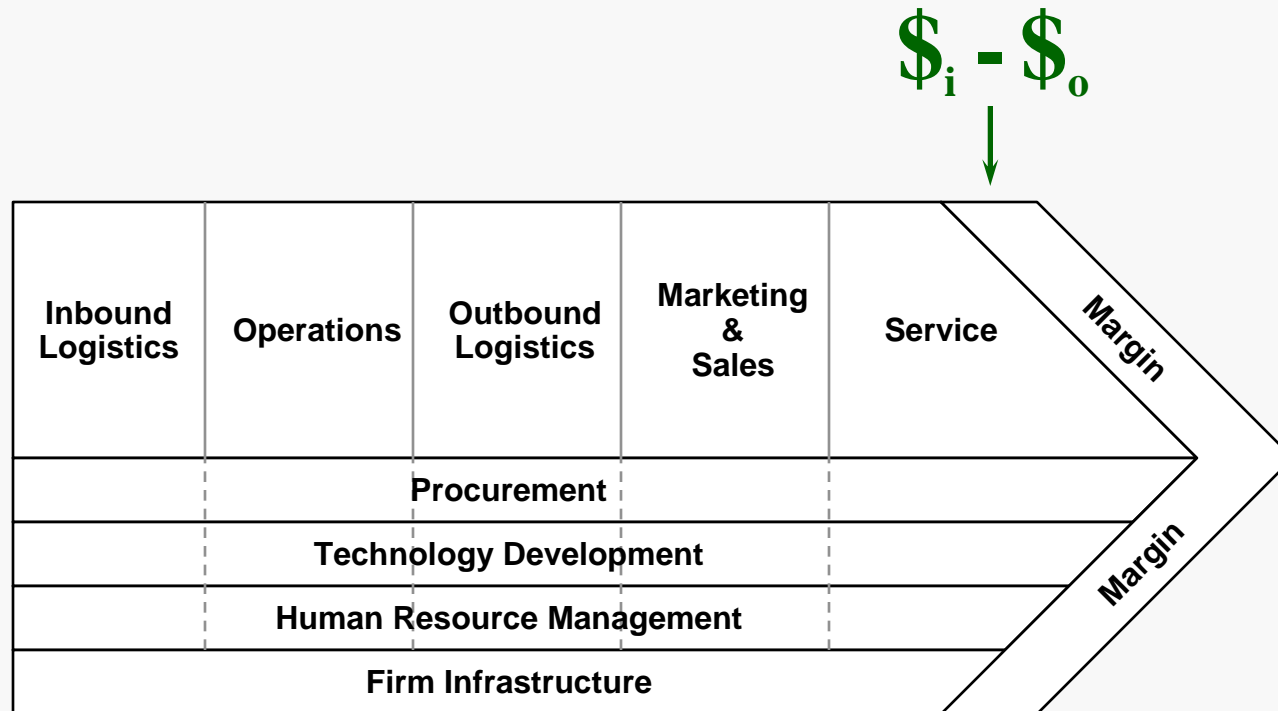
Conversely, **support activities** are necessary but not sufficient for the success of the enterprise. Support activities must be managed for maximum cost-effectiveness.

# The Value Chain: *Commerce*



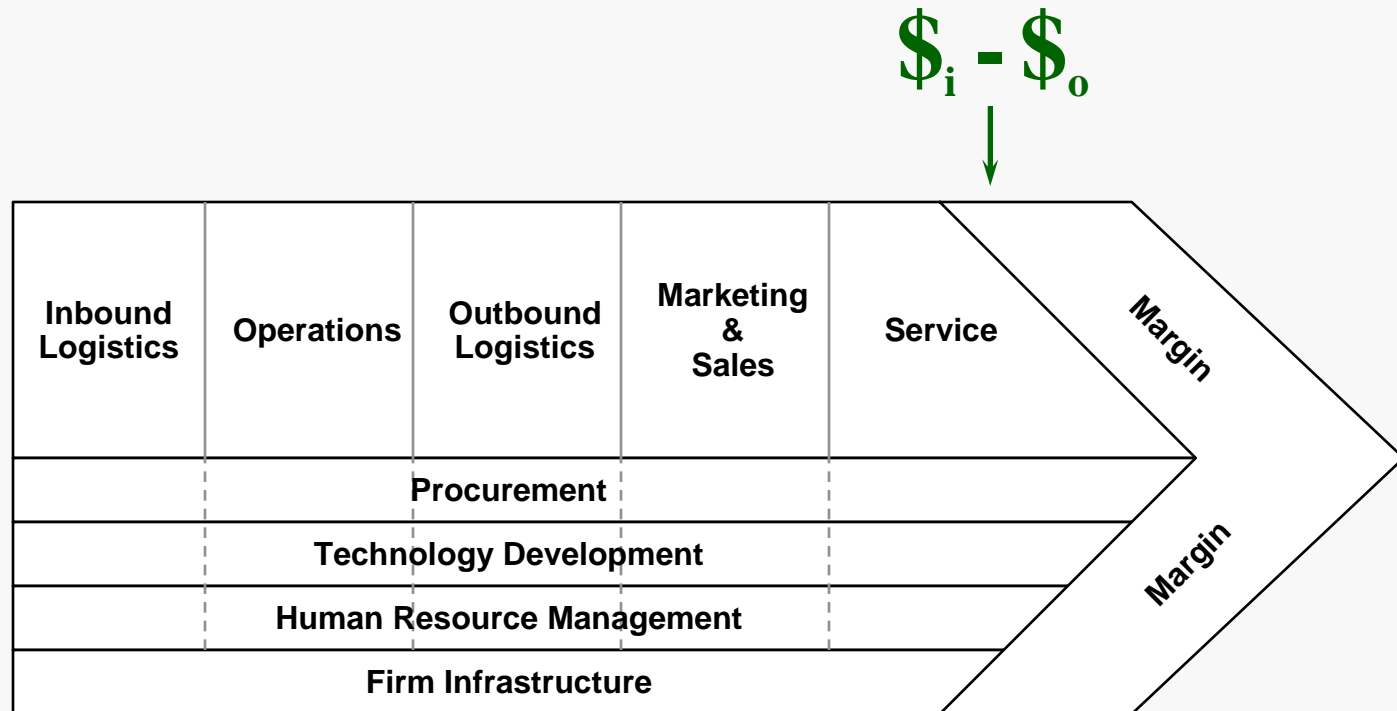
Cash outflow  $\$0$  occurs during the value-adding and support processes. Cash inflow  $\$i$  occurs when the value-added products are sold to customers.

# The Value Chain: *Commerce*



Simplistically speaking, the difference between cash inflow and outflow ( $\$i - \$o$ ) provides the margin of profit.

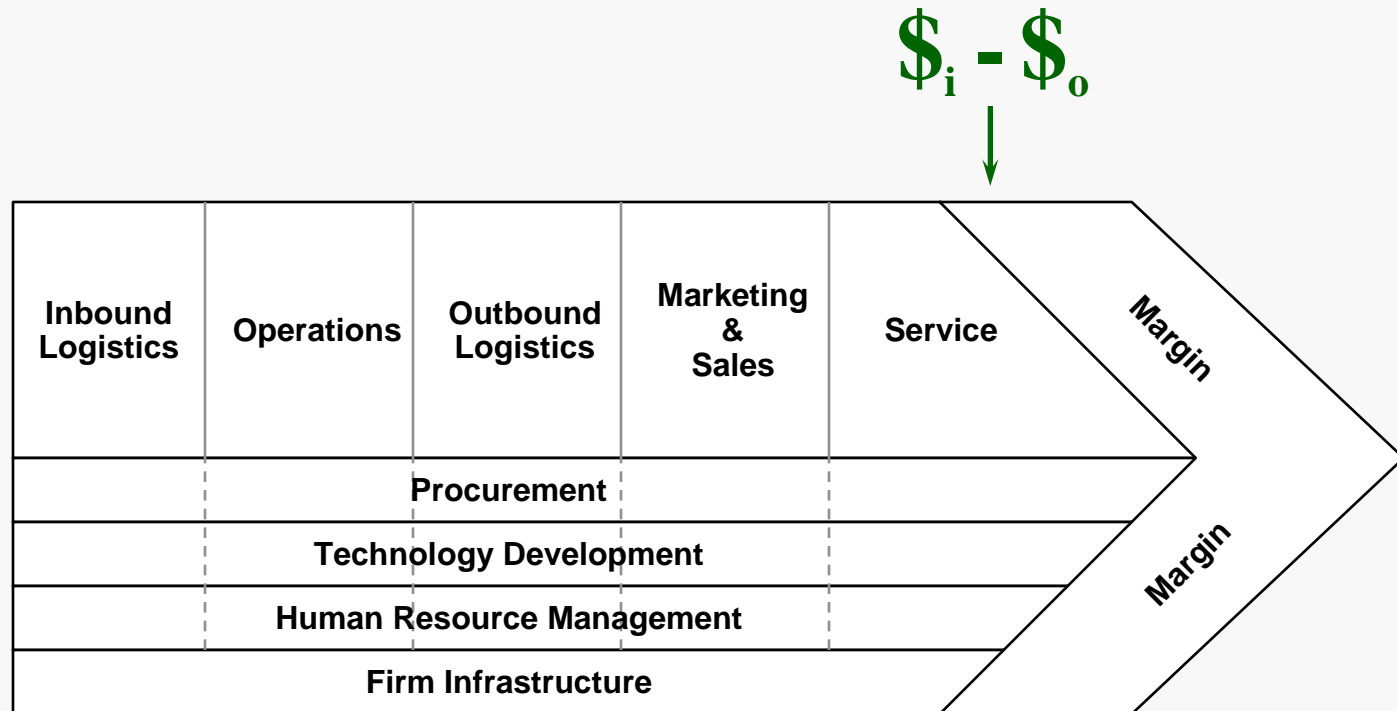
# The Value Chain: *Commerce*



Increased expenses (strategic investment) can lead to increased profits, if the expenses generate more value than they cost.



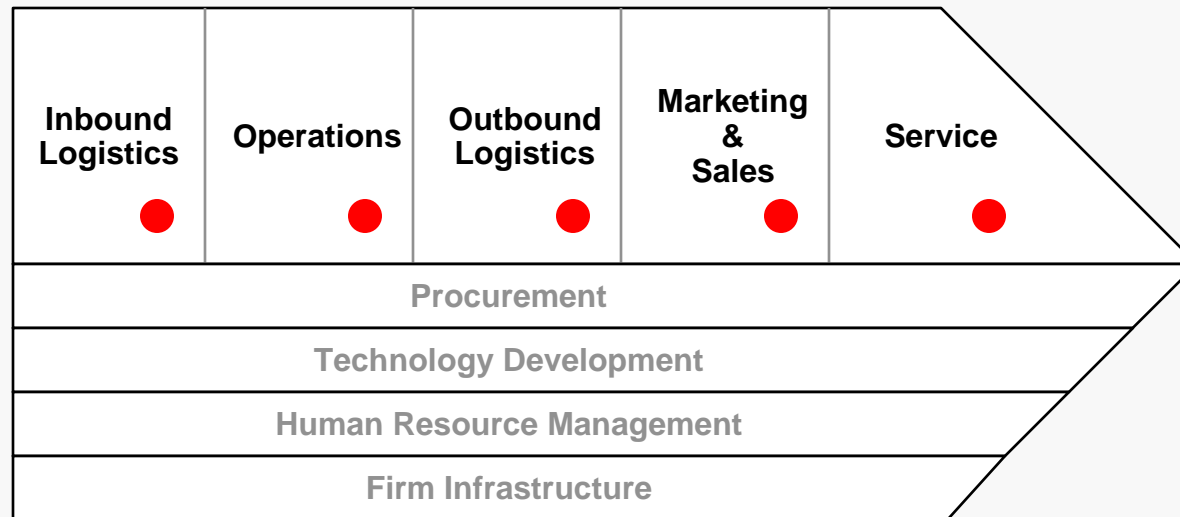
# The Value Chain: *Commerce*



Note: Because  $\$o$  usually occurs before  $\$i$ , we can judge the appropriateness of cost-incurring activities to the extent that we can measure the effect of a particular  $\$o$  upon overall  $\$i$ .

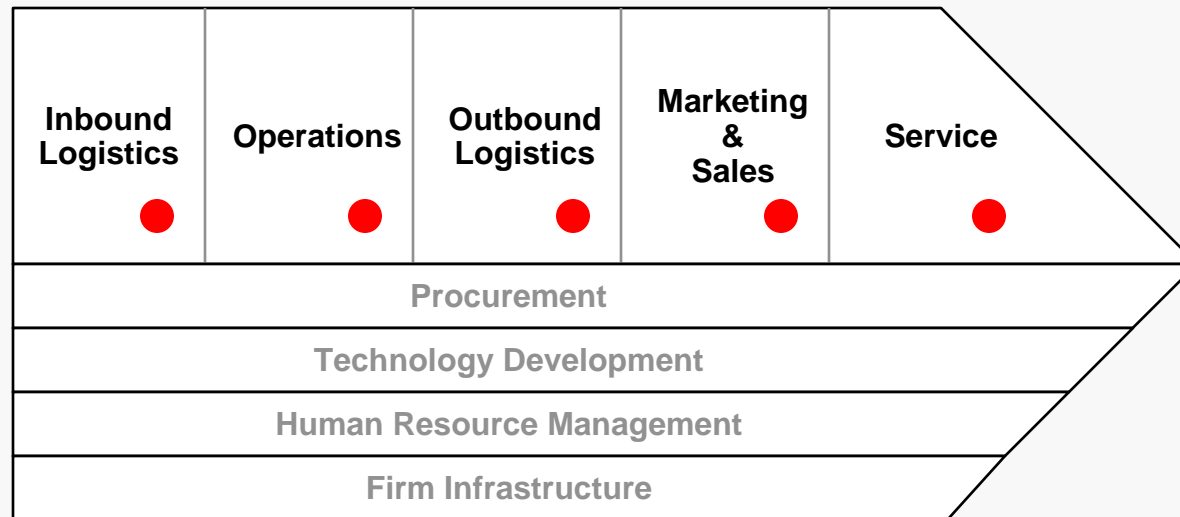
# The Value Chain: *Commerce*

As highly plastic tools, computers can play useful roles in both the accomplishment and the management of tasks. Thus, computers have potential roles in all phases of the value chain.



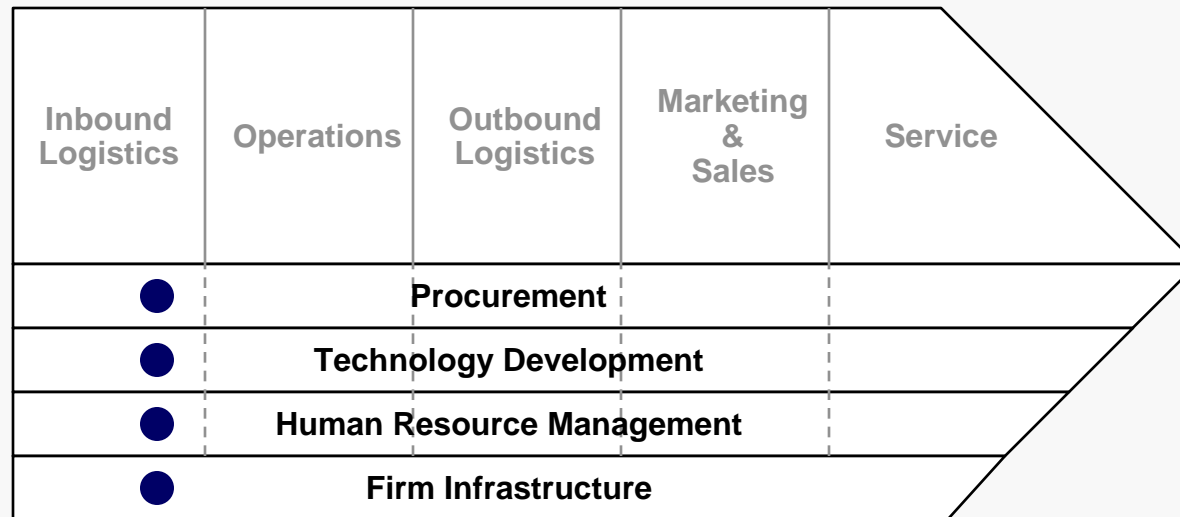
# The Value Chain: *Commerce*

As highly plastic tools, computers can play useful roles in both the accomplishment and the management of tasks. Thus, computers have potential roles in all phases of the value chain.



Many of the most successful companies of the last fifteen years have achieved that success through the skilled deployment of IT to great competitive advantage.

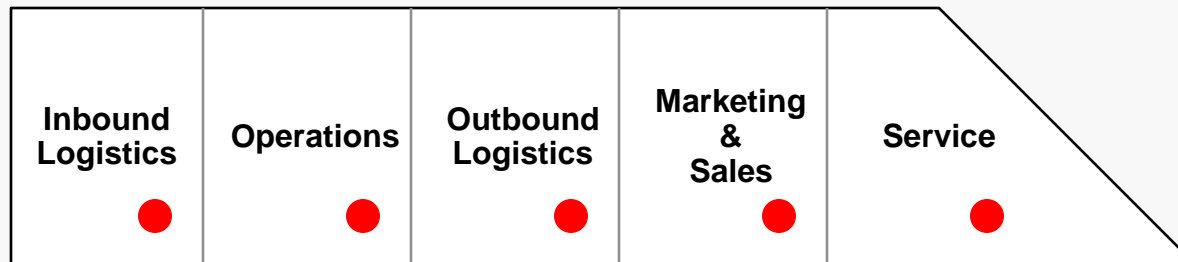
# The Value Chain: *Commerce*



Computers can also play useful roles in many support activities. Here, IT delivers infrastructure strength and may contribute to competitive advantage through cost containment.

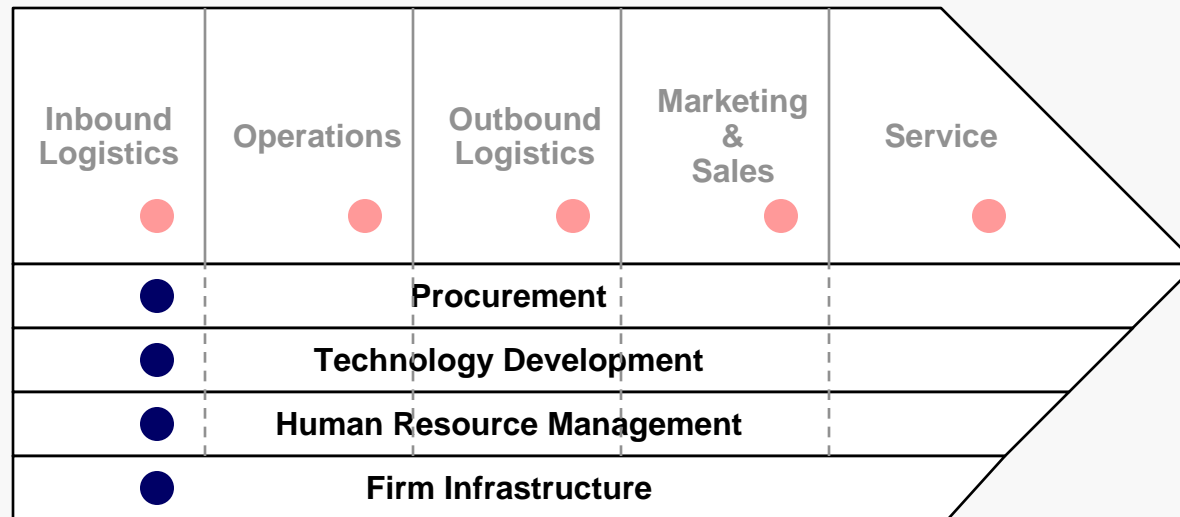
# The Value Chain: *Commerce*

In the value-adding chain, IT is a strategic asset and must be managed accordingly. Investment is made to maximize strategic competitive **effectiveness**.



# The Value Chain: *Commerce*

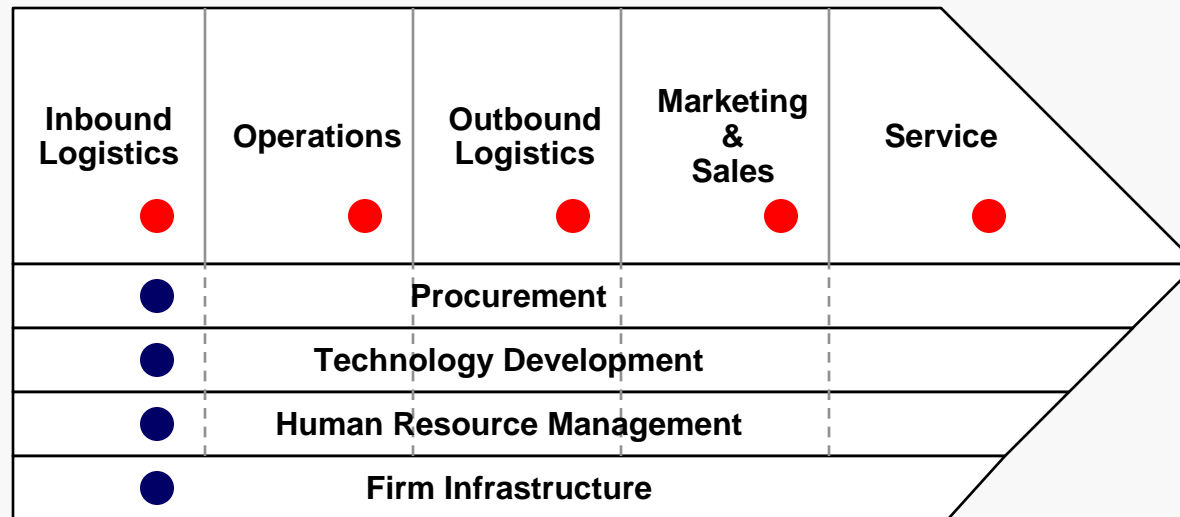
In the value-adding chain, IT is a strategic asset and must be managed accordingly. Investment is made to maximize strategic competitive **effectiveness**.



In support activities, IT is a cost-center component and must be managed accordingly. Costs must be contained and the entire operation tuned to achieve maximum operational **efficiency**.

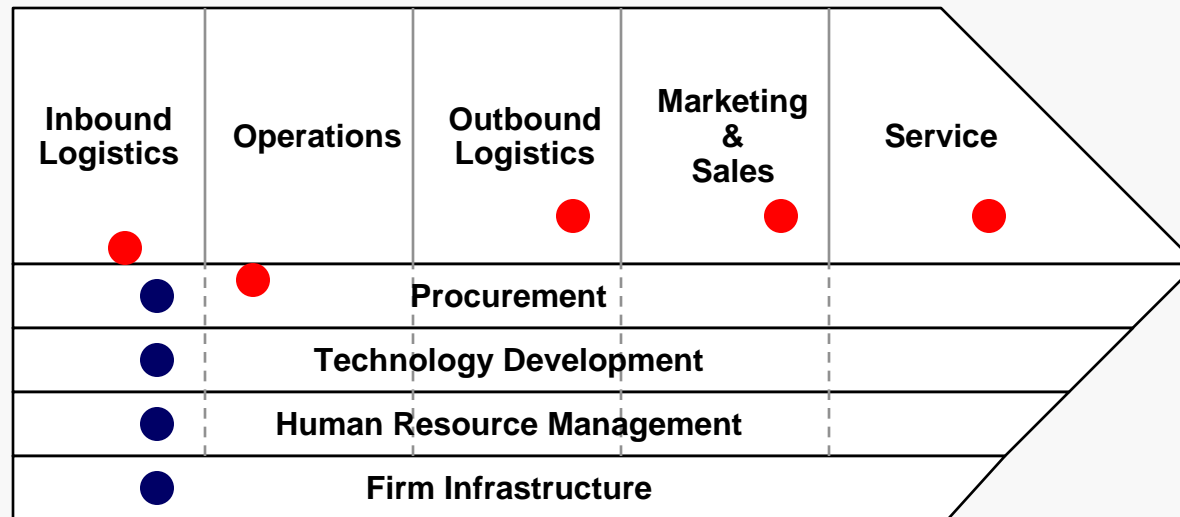
# The Value Chain: *Commerce*

The rapid rate of technological change adds another complexity.



# The Value Chain: *Commerce*

The rapid rate of technological change adds another complexity.

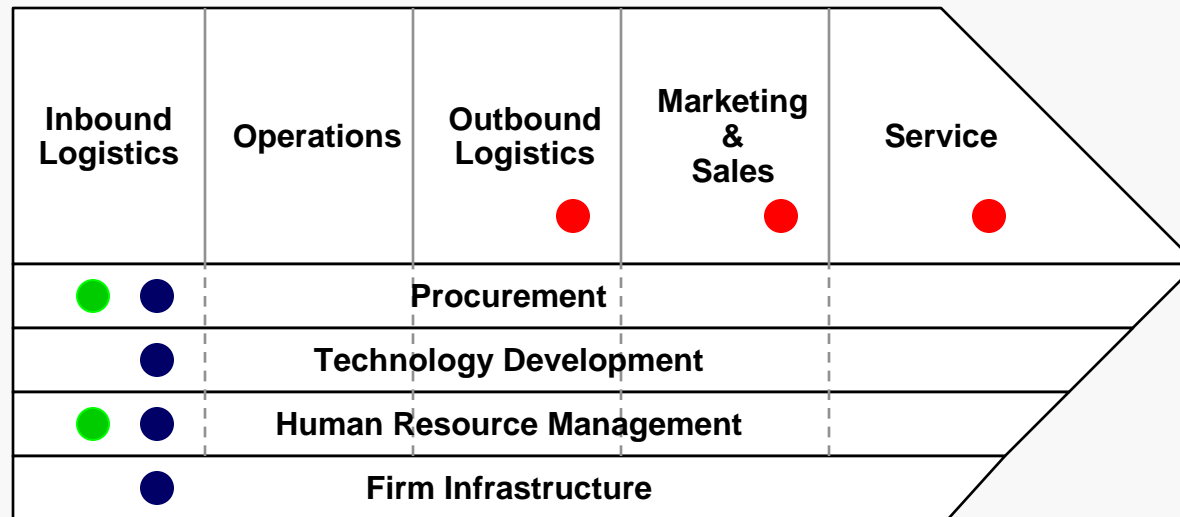


Value-adding activities can become support activities overnight.



# The Value Chain: *Commerce*

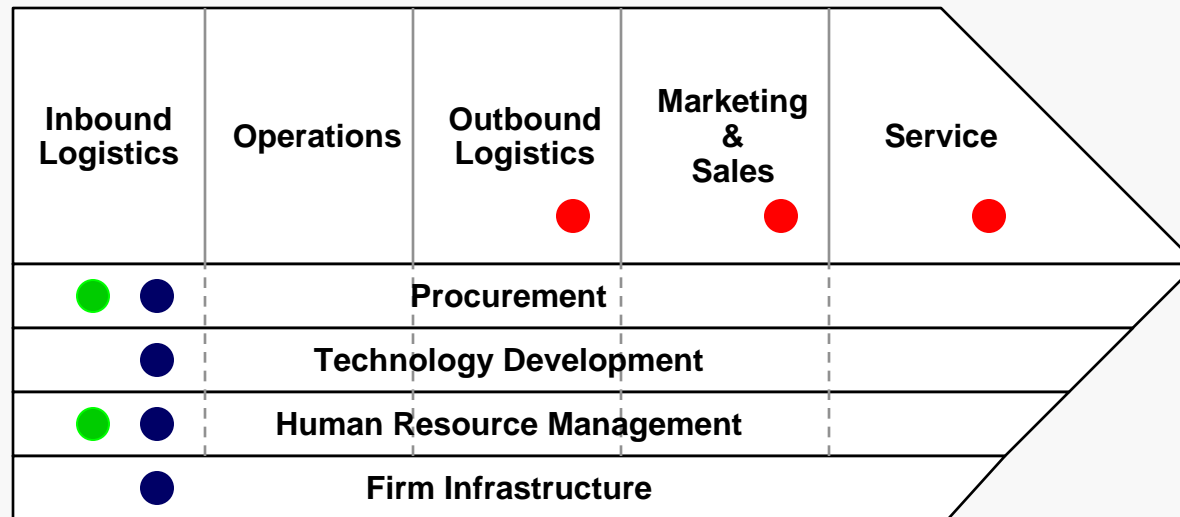
The rapid rate of technological change adds another complexity.



Value-adding activities can become support activities overnight.

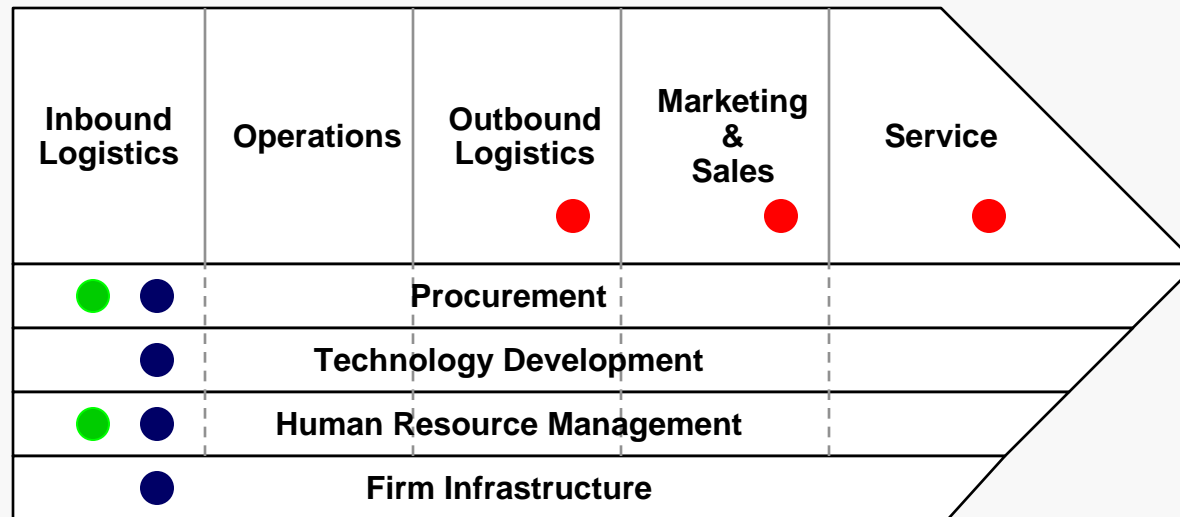
# The Value Chain: *Commerce*

Although this change complicates IT operational management in any organization, the problem is exacerbated in a grant-funded research organization.



# The Value Chain: *Commerce*

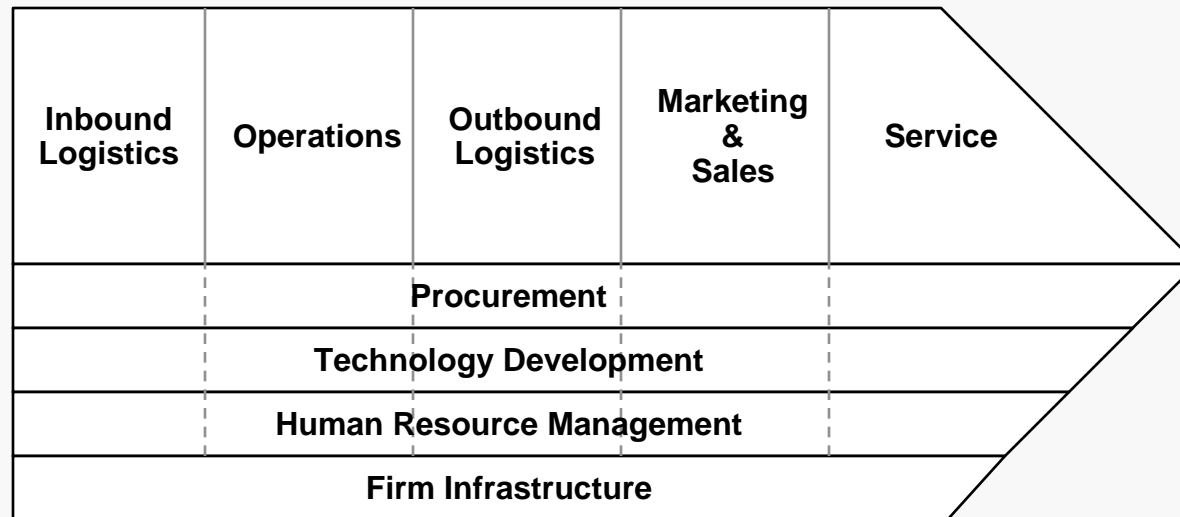
Although this change complicates IT operational management in any organization, the problem is exacerbated in a grant-funded research organization.



In a grant-funded environment, the primary value-adding activities are funded with **direct** dollars, whereas the support activities are funded with **indirect** dollars.

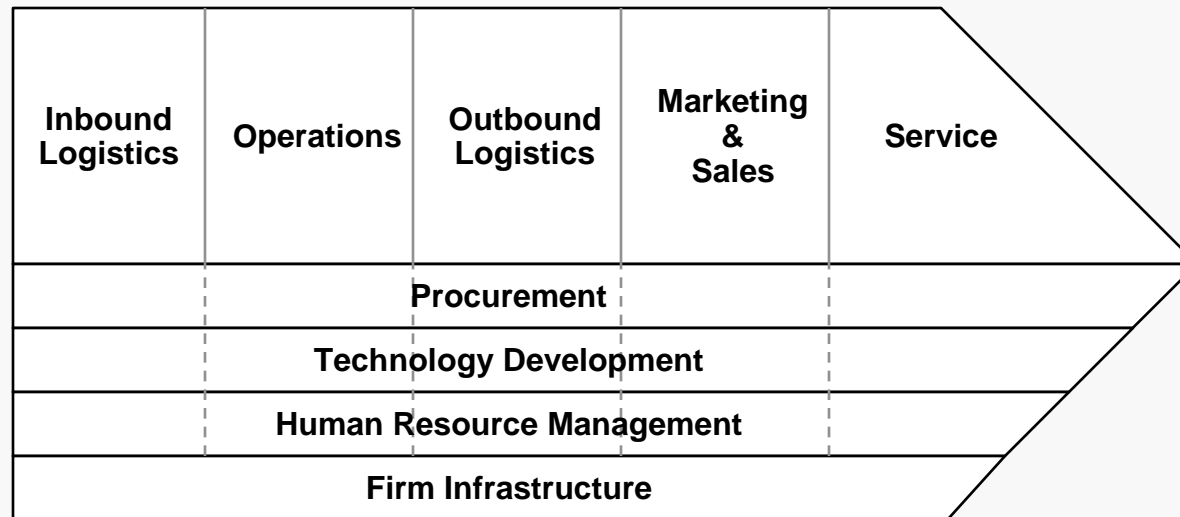
# The Value Chain: *Commerce*

Despite increased recognition of its importance, investment in IT to support public-sector, grant-funded research is currently falling behind the private sector. Why?



# The Value Chain: *Commerce*

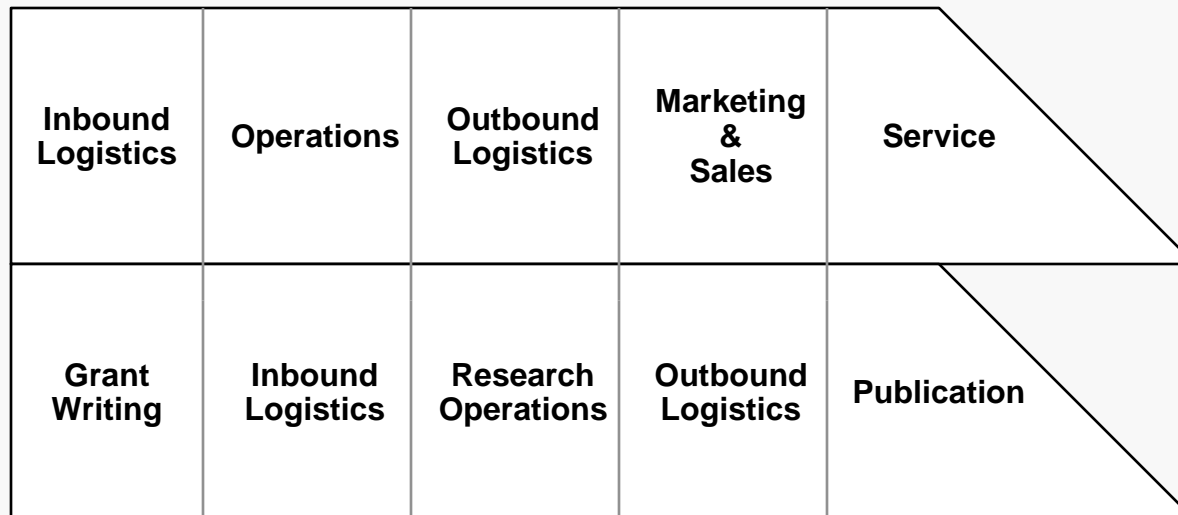
Despite increased recognition of its importance, investment in IT to support public-sector, grant-funded research is currently falling behind the private sector. Why?



Other factors complicate the daily management of and the long-term planning for IT operations in a biomedical research organization.

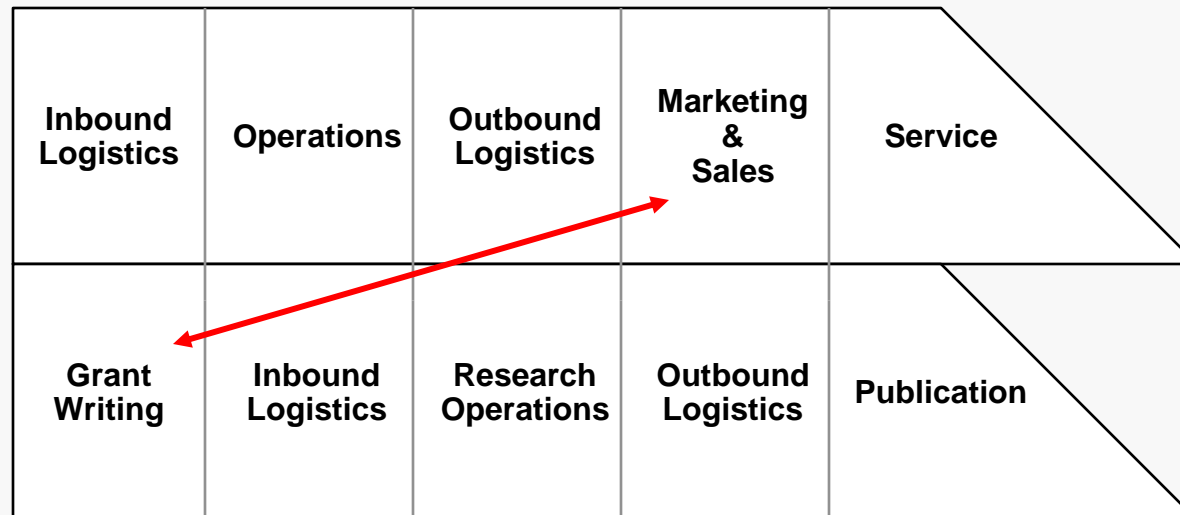
# The Value Chain: *Commerce*

First, even at a generic level, the value-adding activities for research are different from those of commerce.



# The Value Chain: *Commerce*

First, even at a generic level, the value-adding activities for research are different from those of commerce.

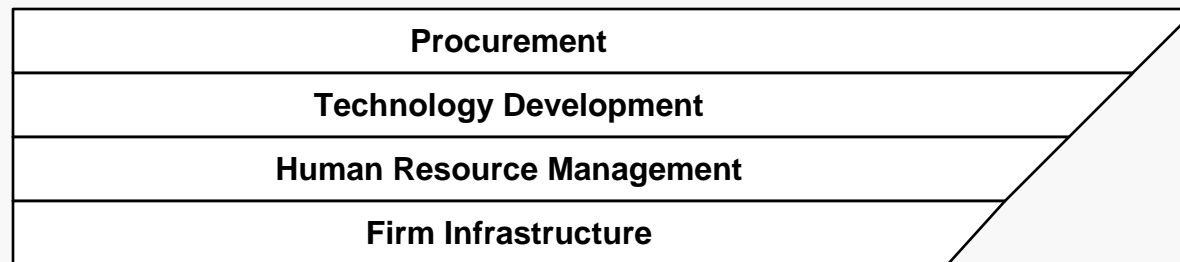


Not only are the categories somewhat different, but there is a significant reversal in time sequence of some components.

# The Value Chain: *Commerce*

---

Although some differences exist in the support activities, these are not as significant as those in the primary activities.

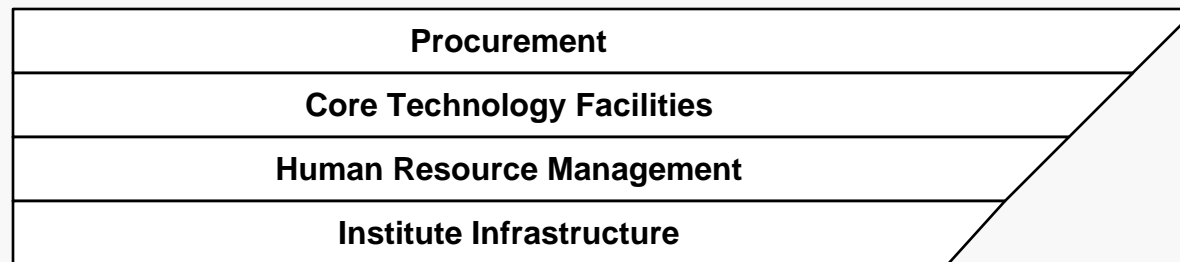




# The Value Chain: *Commerce*

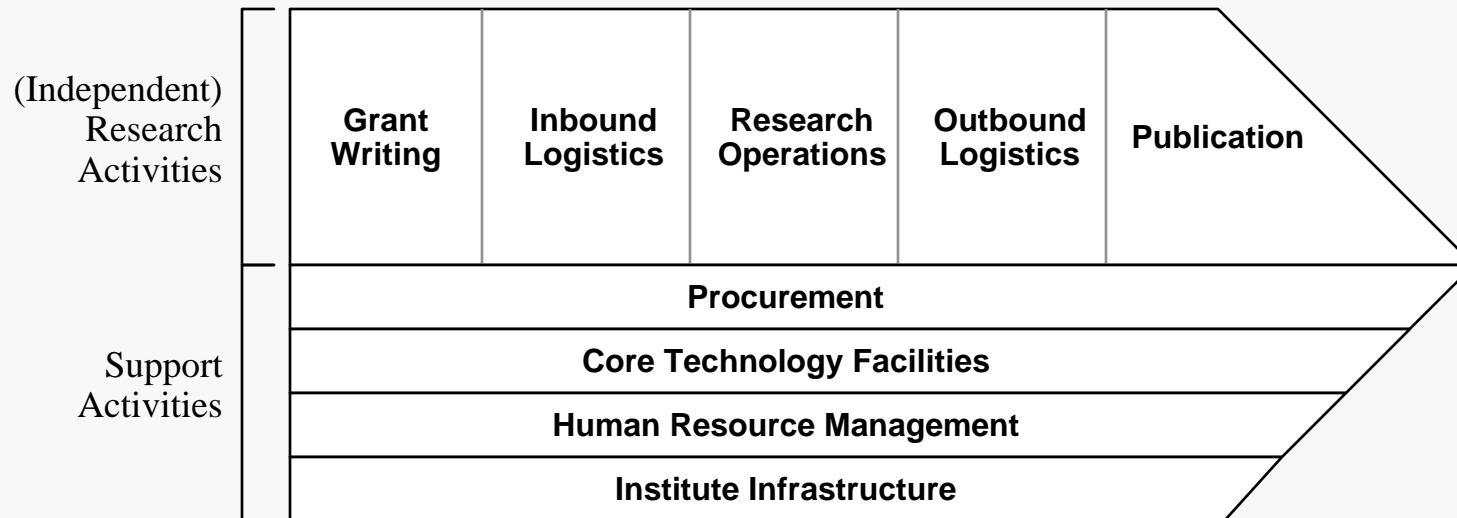
---

Although some differences exist in the support activities, these are not as significant as those in the primary activities.



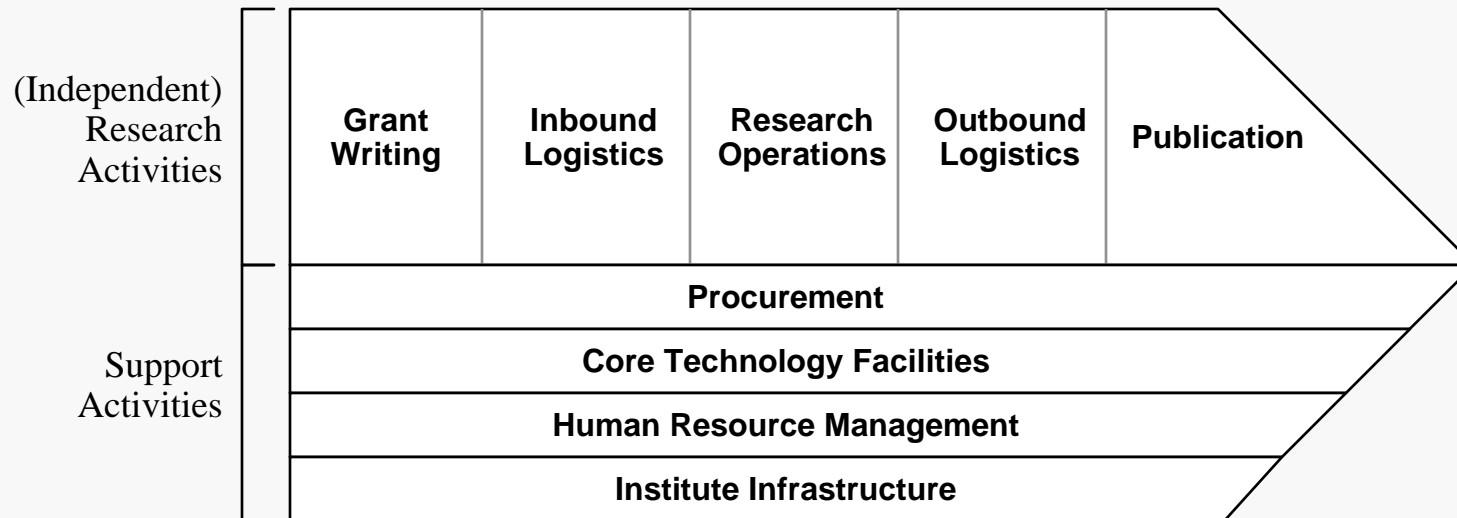
# The Value Chain: *Research*

Combining these adjustments we get the following “Porter diagram” for research.



# The Value Chain: *Research*

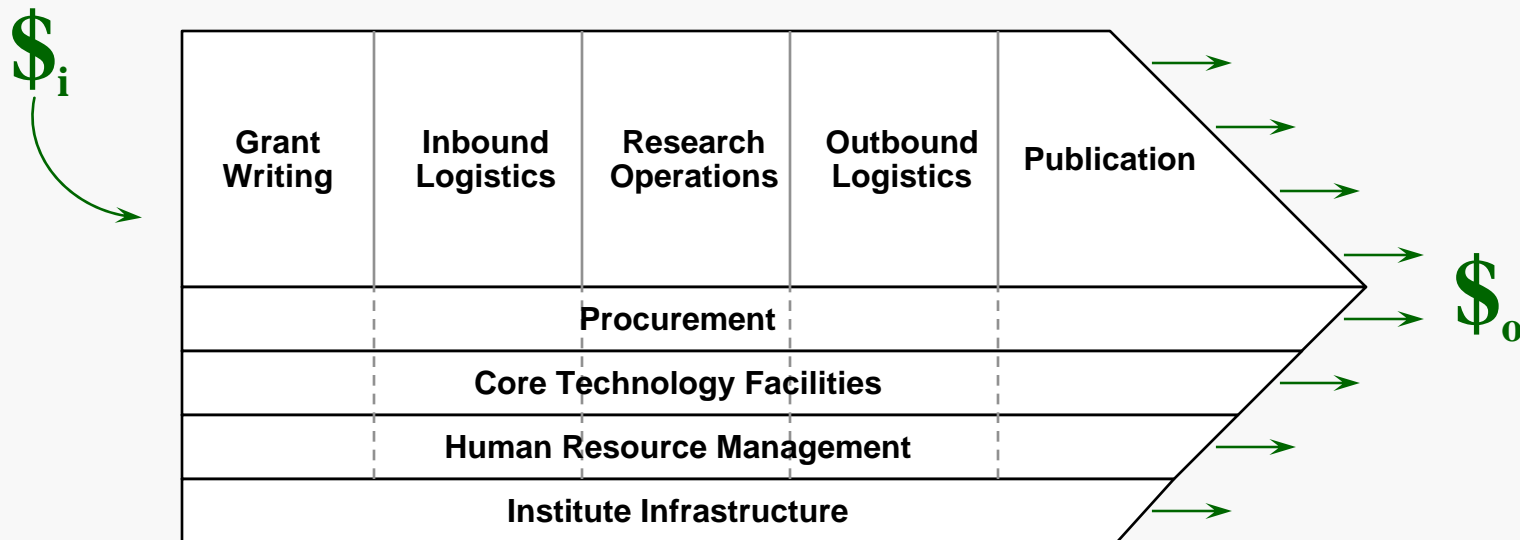
Combining these adjustments we get the following “Porter diagram” for research.



Now we can consider some other complicating factors...

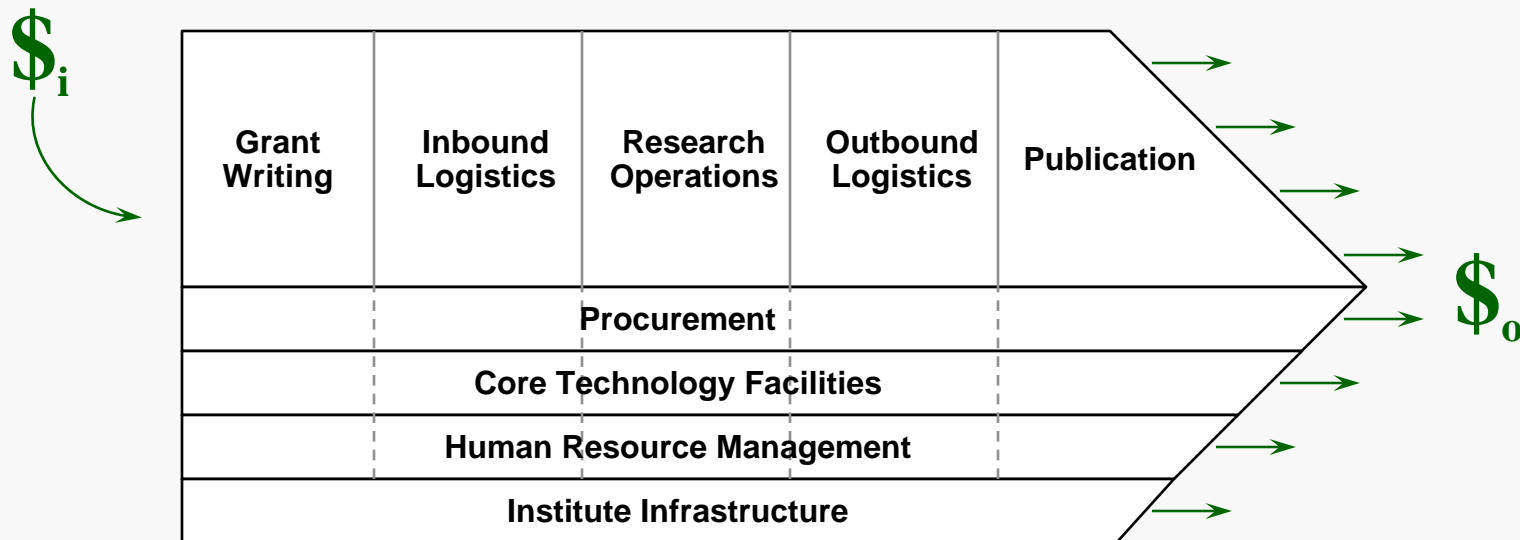
# The Value Chain: *Research*

Cash flow is backwards, in that “income” precedes expenses. Furthermore, “income” is really just authorization to request reimbursement for appropriate expenses.



# The Value Chain: *Research*

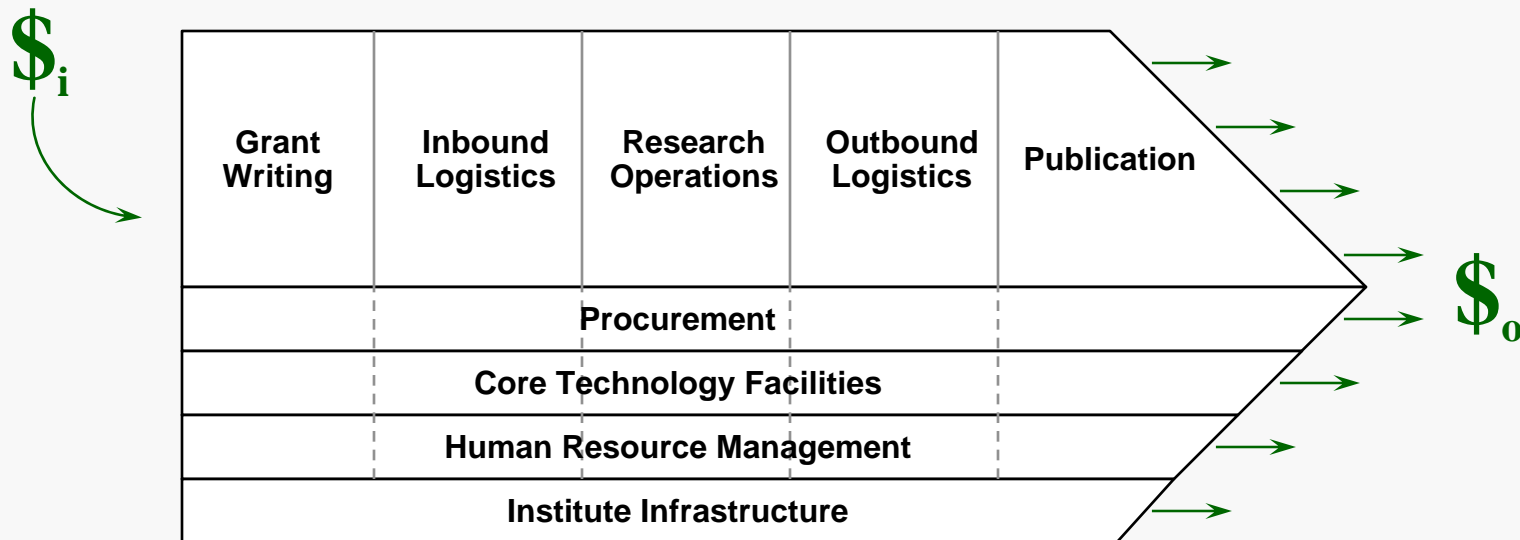
Cash flow is backwards, in that “income” precedes expenses. Furthermore, “income” is really just authorization to request reimbursement for appropriate expenses.



Because  $\$_I$  is capped as a reimbursement for an approved subset of  $\$_o$ ,  $\$_I$  must always be less than or at best equal to  $\$_o$ . This means there can never be a real profit margin.

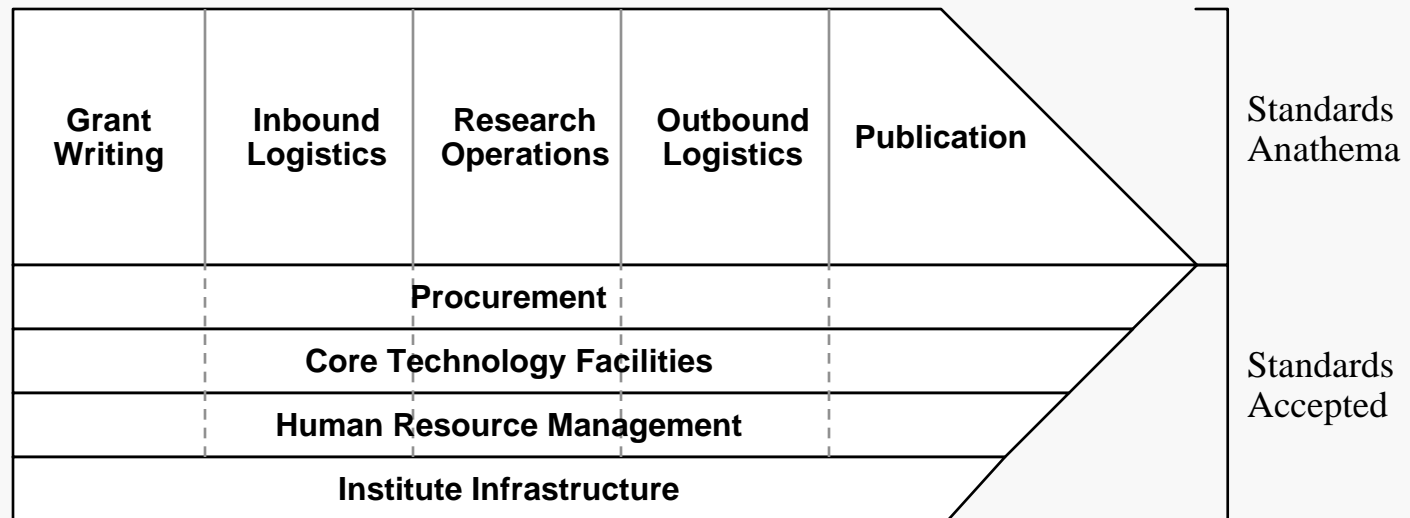
# The Value Chain: *Research*

Without a profit margin, true strategic investment in IT is difficult, if not impossible.



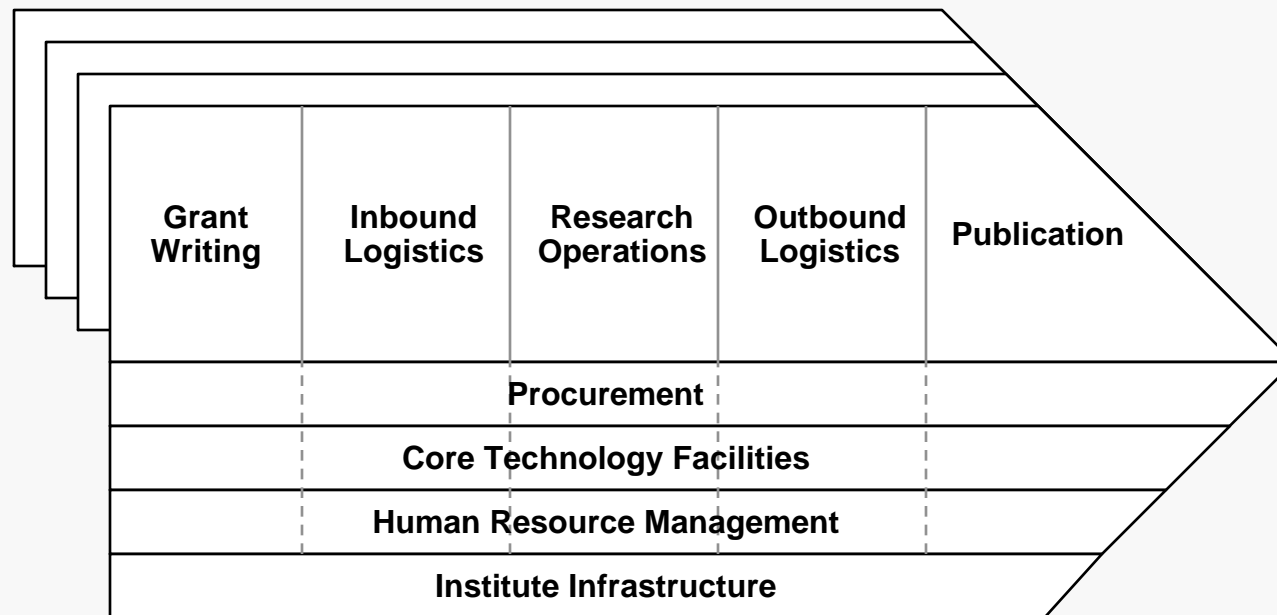
Thus, compared with private-sector enterprises, IT investment in grant-funded research organizations is often trivial and ineffective.

# The Value Chain: *Research*



The sociology of public-funded research activities resists efficiencies in the value-adding chain. Much of this resistance is **legitimate**.

# The Value Chain: *Research*



In a grant-funded research organization, there are multiple value-adding chains, one for each independently funded research activity.



# The Value Chain: *Research*

---

**Aligning IT operations with hundreds of independent research activities (each with its own dynamic goals, budgets, staff, and timelines) is not easy. Indeed, efforts to achieve *specific* alignment with all of these activities must fail.**

value-adding chains, one for each independently funded research activity.

# The Value Chain: *Research*

---

**The trick is UNDERSTANDING the process and values of research.**

**With understanding, and acceptance, real alignment can be achieved.**

value-adding chains, one for each independently funded research activity.

# Understanding Research

---

## Business Model

- The Porter value-chain analysis shows that the funding model, and the value-adding process of grant-funded research is fundamentally different from that of businesses that sell goods or services to consumers.
- Measuring ROI is metaphorical (at best)
- No common measurement for success – i.e., no bottom line

# **Understanding Research**

---

**Operational Practices**

# Understanding Research

---

## Operational Practices

- Independent
- Portable
- Third-party pay; Third-party rewards
- Deals with the unknown, cannot be standards driven
- Intensely opportunistic
- Pan-enterprise collaboration

# **Understanding Research**

---

**Cultural Norms**

# Understanding Research

---

## Cultural Norms

- Ultimate goal: extraction of new knowledge from nature
- Values-based life style
- Strong differences among fields (and researchers)
- One-off solutions are common

# Understanding Research

---

## Cultural Norms

- Skepticism is a given
- Evidence is expected
- Logic is required
- Criticism is a primary form of discourse
- Understanding is the goal: NT triumphant