Sharing Digital Biological Data:

Historical Successes and Continuing Challenges

(http://www.esp.org/rjr/MGED-2006.pdf)

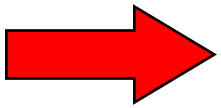
Robert J. Robbins

rrobbins@fhcrc.org

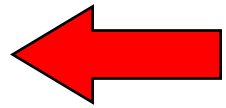
(206) 667 4778

Sharing Digital Biological Data:

Historical Successes and Continuing Challenges



(http://www.esp.org/rjr/MGED-2006.pdf)



Robert J. Robbins

rrobbins@fhcrc.org

(206) 667 4778

Abstract

Forty years ago, bioinformatics and digital data sharing (DDS) were unknown. Twenty years ago, bioinformatics and DDS were in a crisis of scalability. Today, that problem has (largely) been solved.

Now, bioinformatics and DDS are ubiquitous, and some are beginning to envision a future where digital biological data are (a) fully sharable, and (b) embedded in a semantic web and jointly accessible in a meaningful way.

Despite the successes of the past and the promises of the future, many unresolved problems still impede the useful sharing of biological data. Some of these unsolved problems were first recognized many years ago. Others are just becoming apparent. Some of the problems are technical, others sociological. A few even trace their roots to problems in metaphysics.

As we rush into the future, we must take care not to forget the lessons of the past. In this talk, we will consider several of the outstanding challenges still facing digital data sharing.

Topics: Past Successes

- Past Success
  - Physical Interoperability
  - Data Sharing is now Easy
- The Model of NCBI/GenBank
  - Tremendous success
  - Special case (of sequence data)
  - Special case (of NCBI)
  - Doesn’t generalize

Topics: Future Challenges

- Database Technology

Topics: Future Challenges

- Database Technology
- Metaphysics: The Concept of Identity

Topics: Future Challenges

- Database Technology
- Metaphysics: The Concept of Identity
- Science Itself

Topics: Future Challenges

- Database Technology
- Metaphysics: The Concept of Identity
- Science Itself
- Inappropriate Standards

Topics: Future Challenges

- Database Technology
- Metaphysics: The Concept of Identity
- Science Itself
- Inappropriate Standards
- IT Industry Trends

Topics: Future Challenges

- Database Technology
- Metaphysics: The Concept of Identity
- Science Itself
- Inappropriate Standards
- IT Industry Trends
- Inevitability of Change

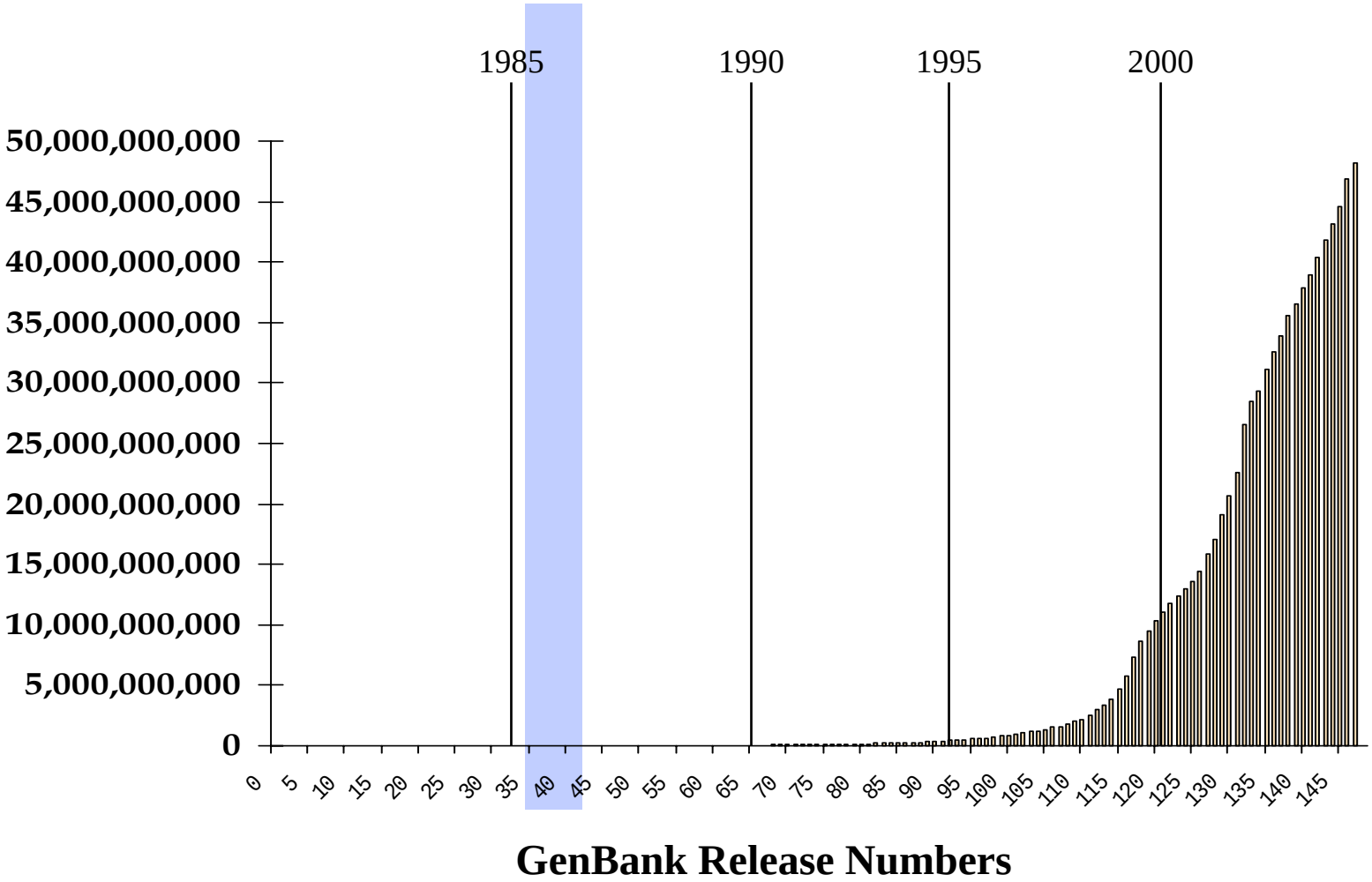
Topics: Future Challenges

- Database Technology
- Metaphysics: The Concept of Identity
- Science Itself
- Inappropriate Standards
- IT Industry Trends
- Inevitability of Change
- Social Scalability

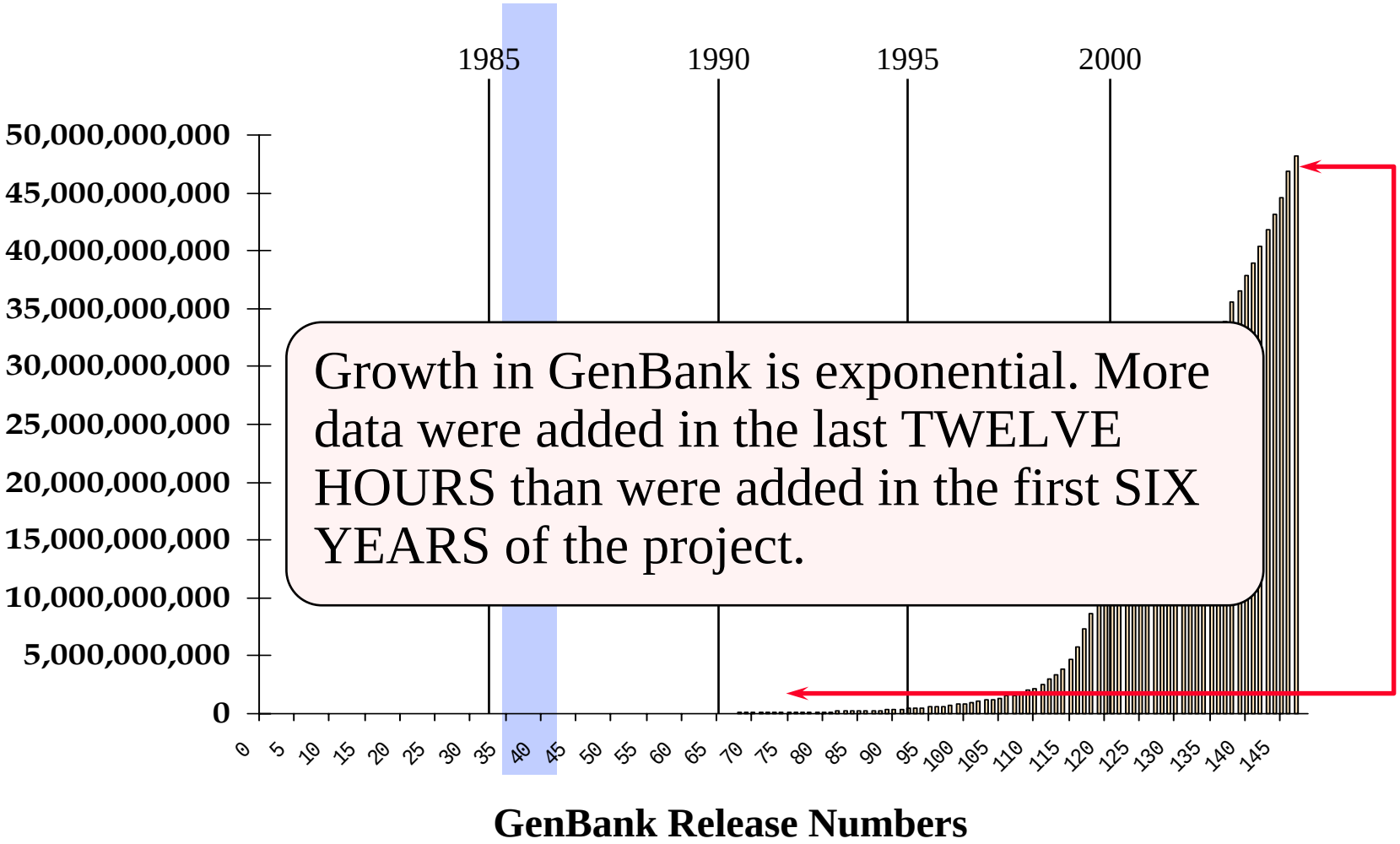
Past Success

GenBank Model

Base Pairs in GenBank



Base Pairs in GenBank



Growth in GenBank is exponential. More data were added in the last TWELVE HOURS than were added in the first SIX YEARS of the project.

Future Challenge

GenBank Model

The GenBank Model

- In some ways, GenBank provides a good model for other bioinformatics efforts…

The GenBank Model

- In some ways, GenBank provides a good model for other bioinformatics efforts…

Track record of success

Single source for critical data

Integrated query tools

Integration with other relevant data sets

Well defined notion of what it is doing

…

The GenBank Model

- In many other ways, GenBank provides a very bad model for other efforts…

The GenBank Model

- In many other ways, GenBank provides a very bad model for other efforts…

Single, almost trivial data type

Monolithic, data-warehouse mechanism

Supports only observations, not “facts”

Highly constrained update mechanism

Huge (and growing) budget

…

GenBank as a False Model

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

GenBank as a False Model

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

NOUNS: Design a database of dogs.

GenBank as a False Model

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

NOUNS: Design a database of dogs.

VERBS: Design a database of dogs jumping.

GenBank as a False Model

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

NOUNS: Design a database of genes.

GenBank as a False Model

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

NOUNS: Design a database of genes.

VERBS: Design a database of gene expression.

The PROBLEM

The Problem

- Sharing raw digital data is now easy.
- Integrating independent digital data into information is hard.
- Cohering independent digital information in knowledge is very hard.

Challenges Due To Limits of Database Technology

Caution from the Past

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.

  http://www.esp.org/foundations/bioinformatics/holdings/CS-90-21.pdf

- CS-90-22

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

  http://www.esp.org/foundations/bioinformatics/holdings/CS-90-22.pdf

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.
  - Schema flexibility is essential.
  - Better formal data models are required, with support for more complex logic.

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

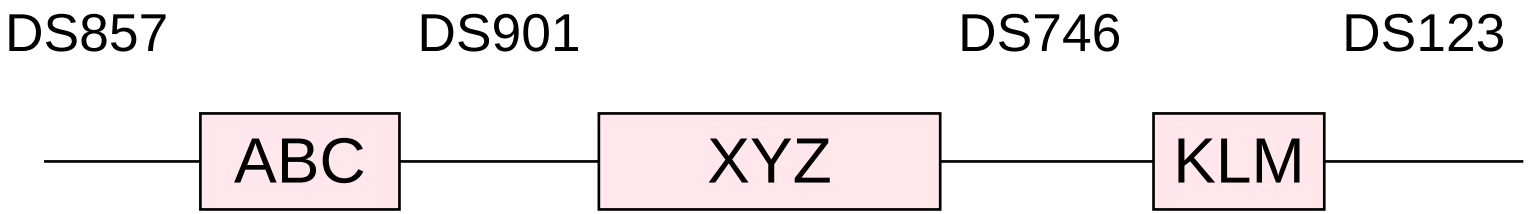
Relational Databases

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

Relational Databases

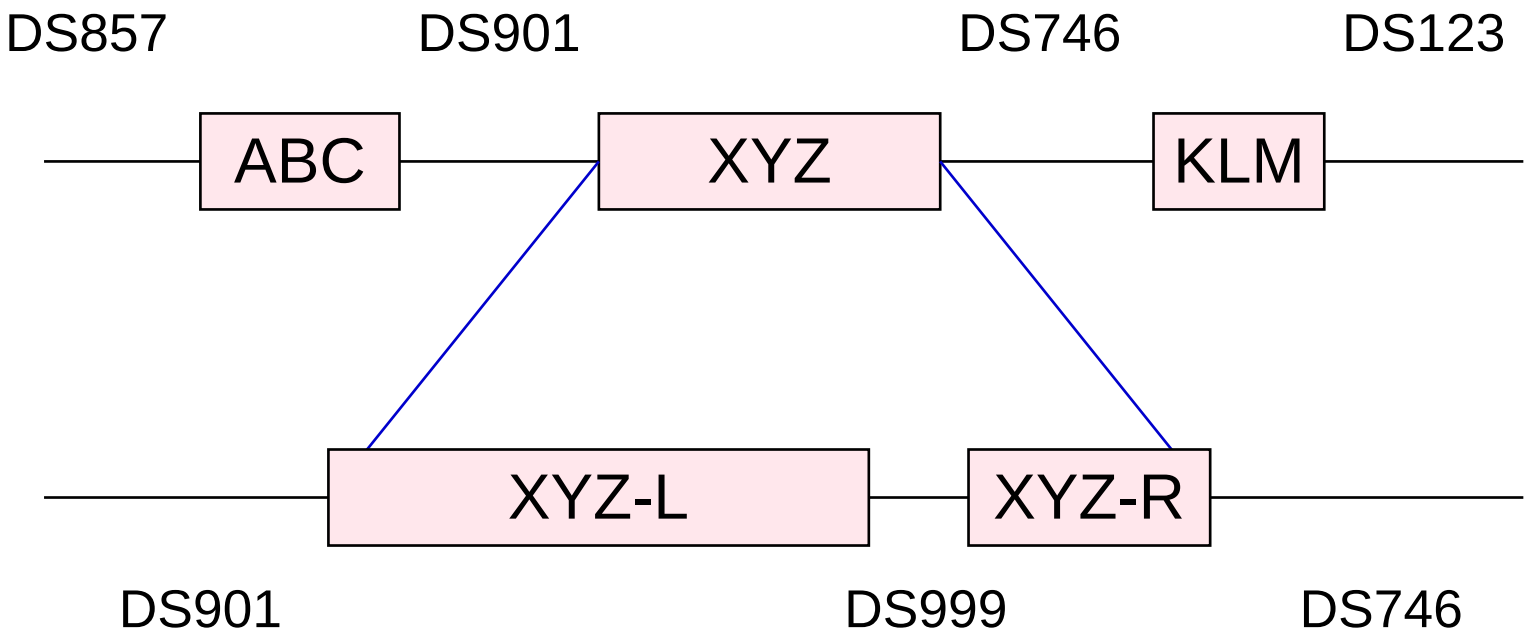
Real Objects:	Hypothetical Objects:
• CONCRETE • STABLE (or known instability) • IMMUTABLE (more or less)	• INSUBSTANTIAL • UNSTABLE • HIGHLY MUTABLE (lumping and splitting)

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

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

Schema Evolution:

Schemas of scientific databases evolve at a high rate. And, data objects within scientific databases lump or split or even change class. Without tools to support referential integrity in the face of these changes, long-term data integration is impossible.

Database III

Data Models & Complex Logic

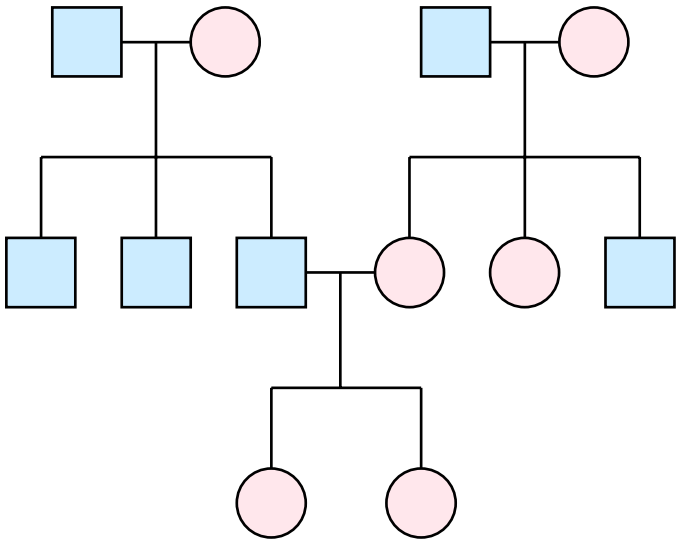
Data-model Challenges

Many bio-data problems involve:

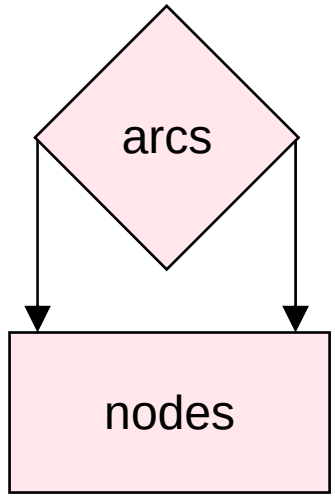
- Graphs: pedigrees, taxonomies, partial orderings, etc…
- Repeat time series observations, with inconsistent results
- Provisional conclusions
- Universal linking tables

Graph Challenges

Pedigree

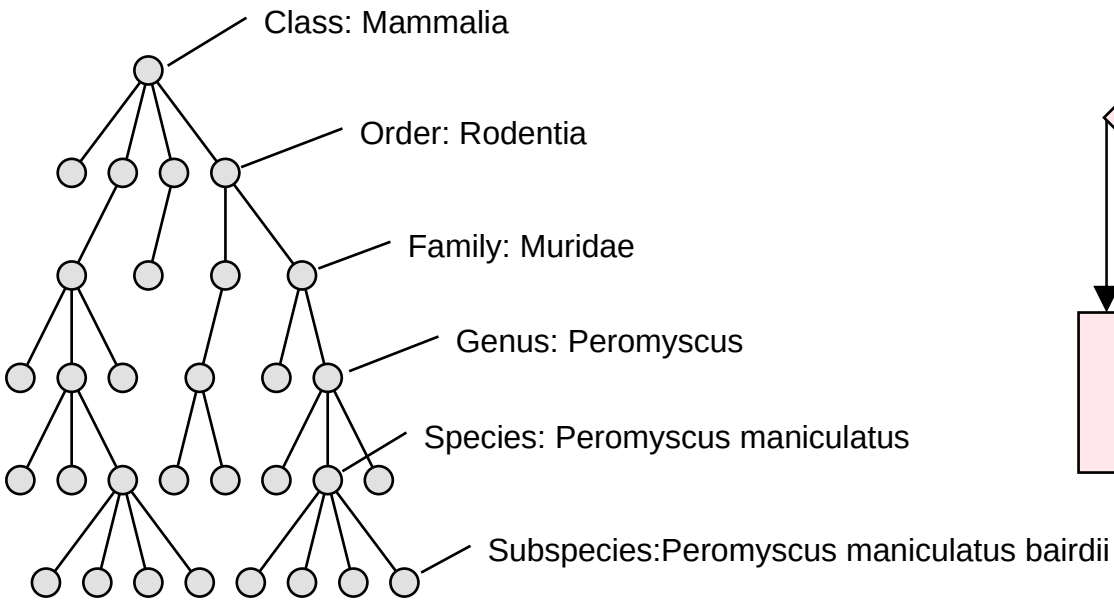


Relational Representation

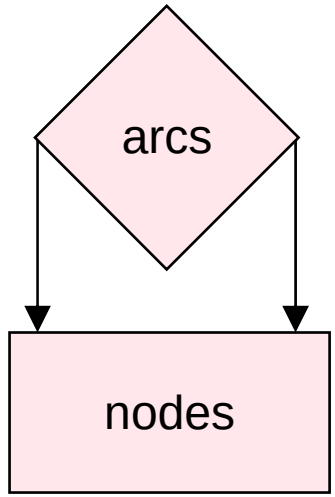


Graph Challenges

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)

Graph Challenges

Classification Hierarchy

Relational Representation

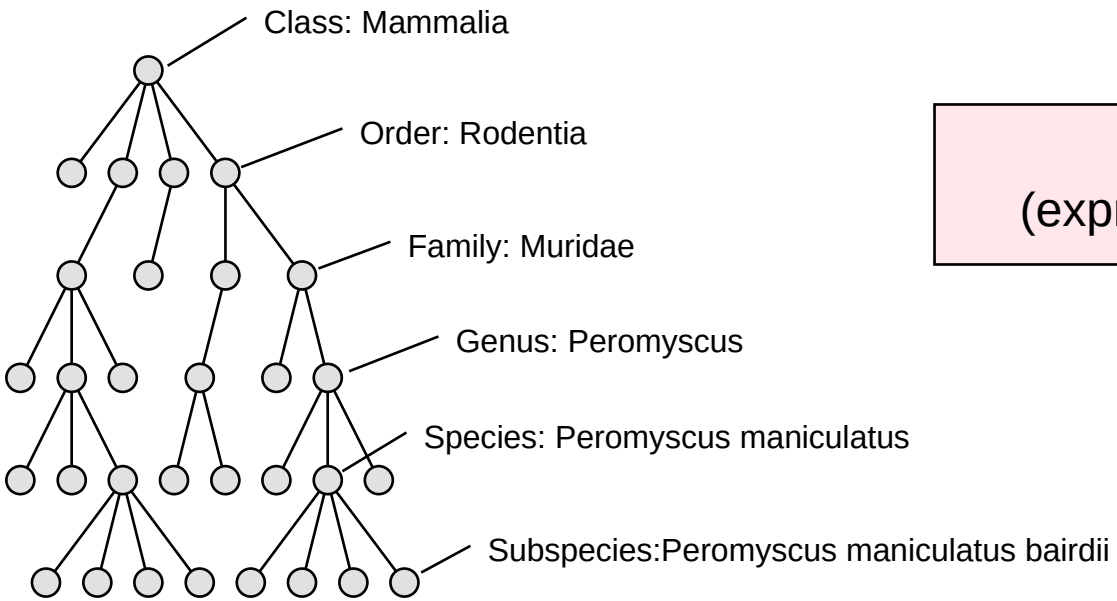
Graph solutions needed:

For efficient updating, graphs are best stored as transitive reductions.

For efficient querying, graphs are best stored as transitive closures.

Classification Challenges

Classification Hierarchy

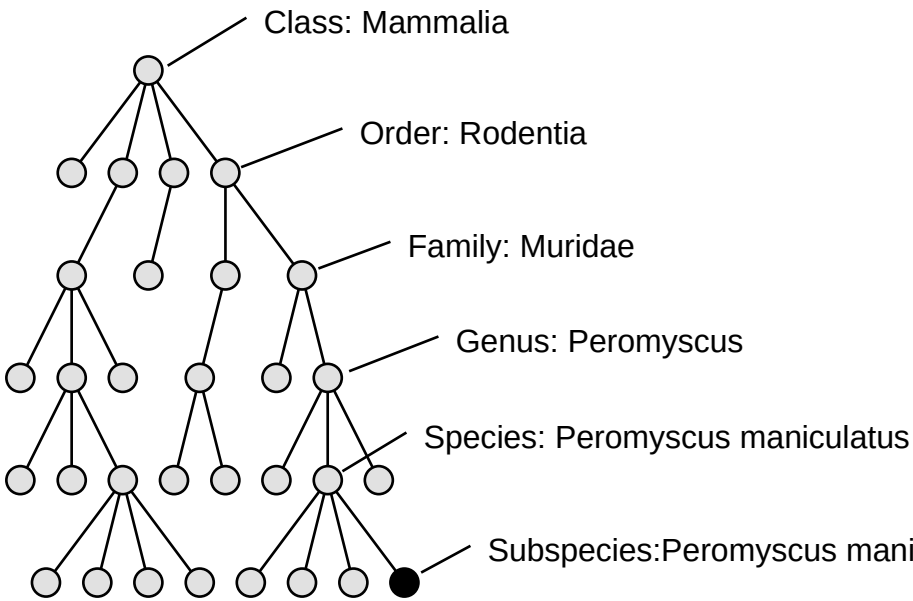


Data Objects to be Classified

Data object
(expression arrays?)

Classification Challenges

Classification Hierarchy



Data Objects to be Classified

Classified as:

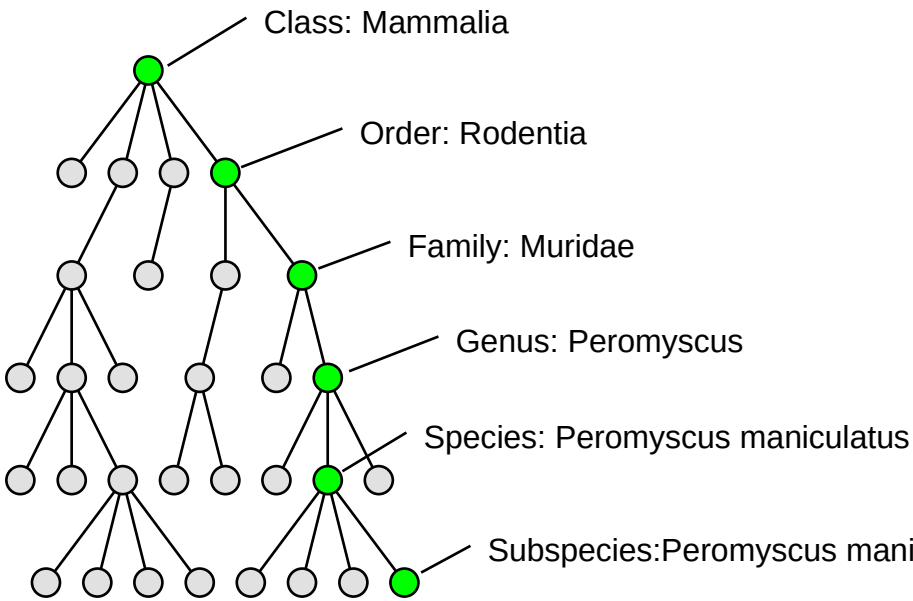
Peromyscus maniculatus bairdii

Data object
(expression arrays?)

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:

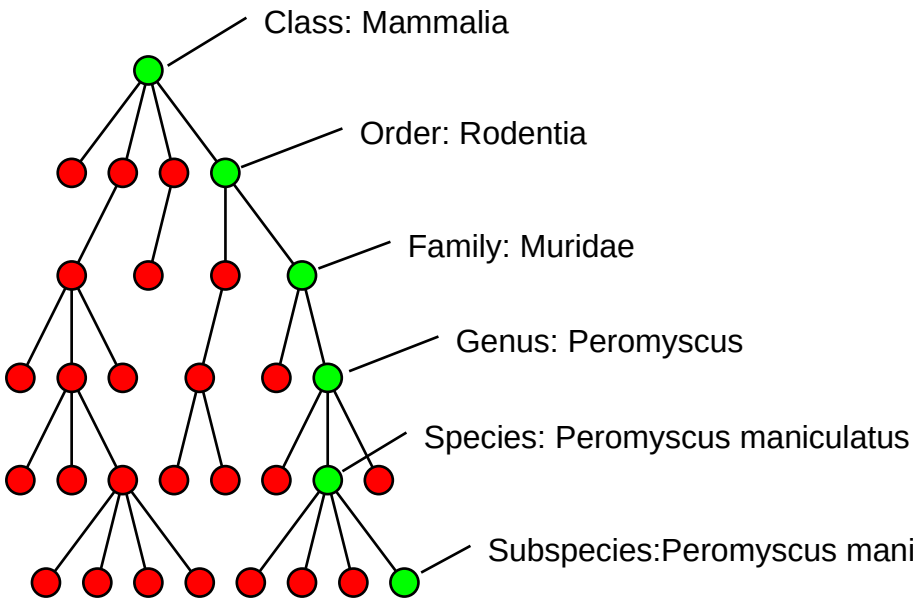
Peromyscus maniculatus bairdii

Data object
(expression arrays?)

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:

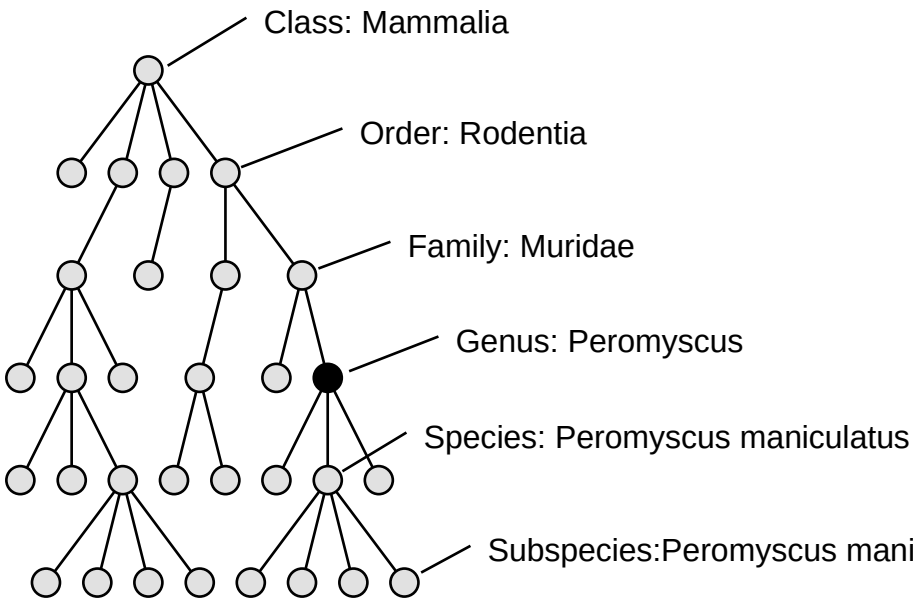
Peromyscus maniculatus bairdii

Data object
(expression arrays?)

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:

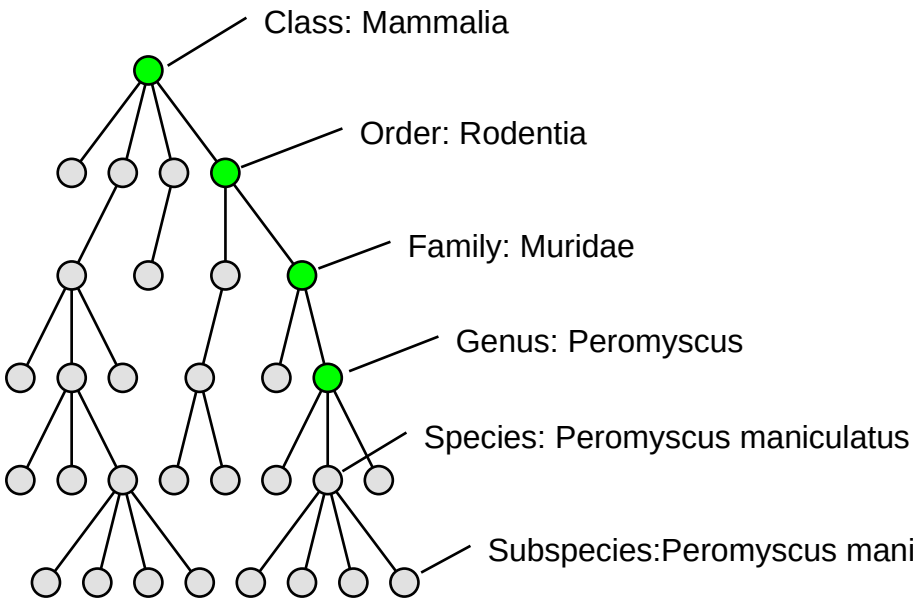
Peromyscus

Data object
(expression arrays?)

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:

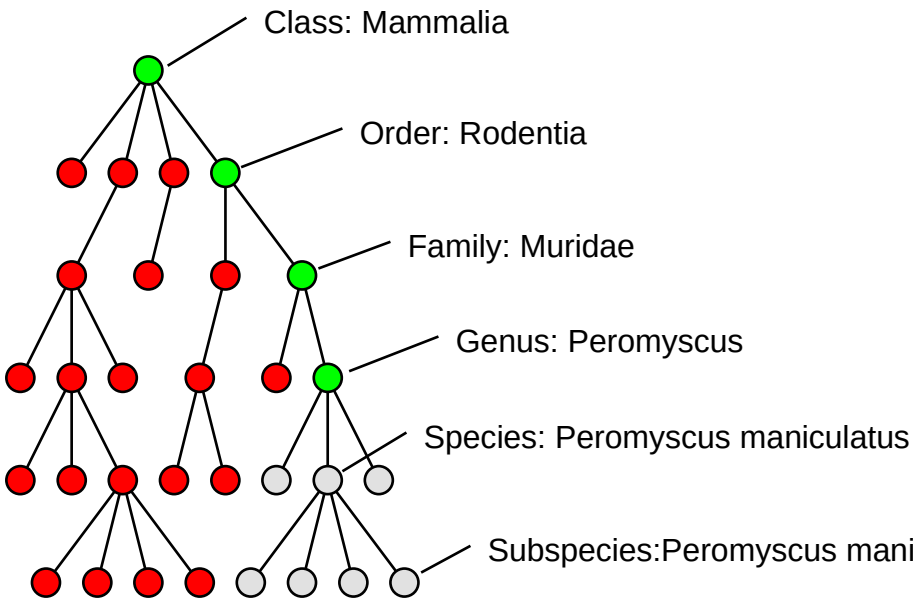
Peromyscus

Data object
(expression arrays?)

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:

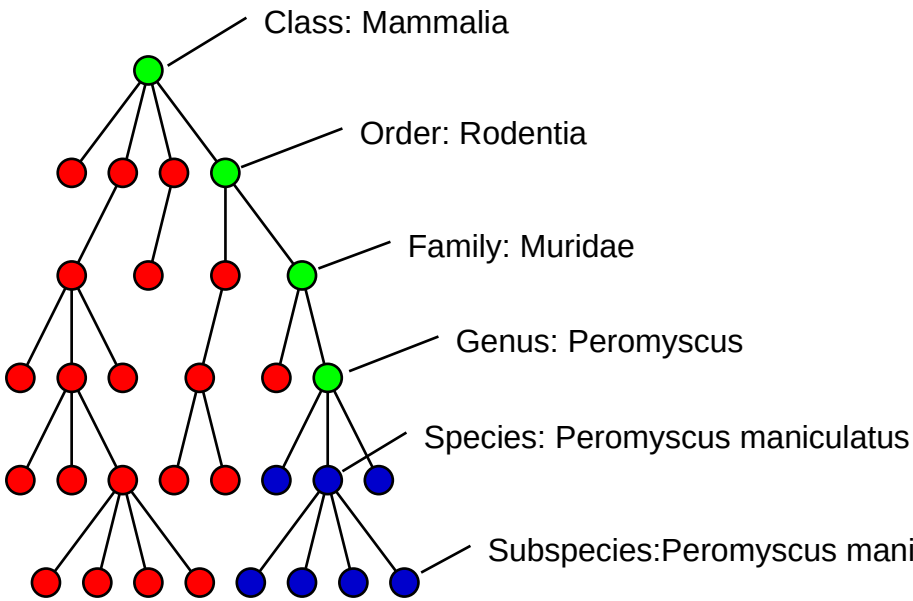
Peromyscus

Data object
(expression arrays?)

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
(expression arrays?)

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 IV

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.

The Vision

We must begin to think of the computational infrastructure of genome research as a federated information infrastructure of interlocking pieces.

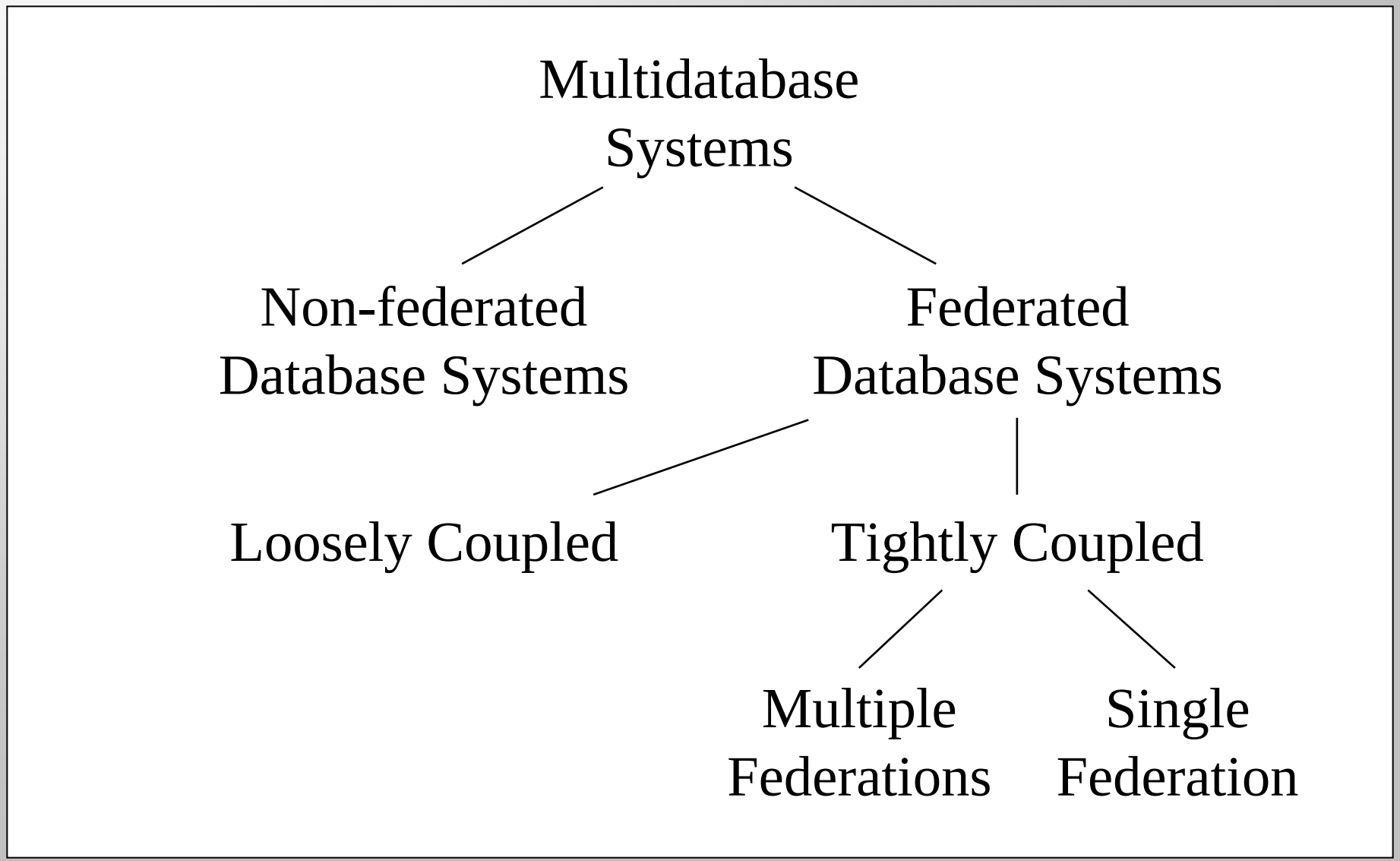
Report of the Invitational DOE Workshop on Genome Informatics, 26-27 April 1993, Baltimore, Maryland

http://www.esp.org/foundations/bioinformatics/holdings/doe-white-paper.pdf

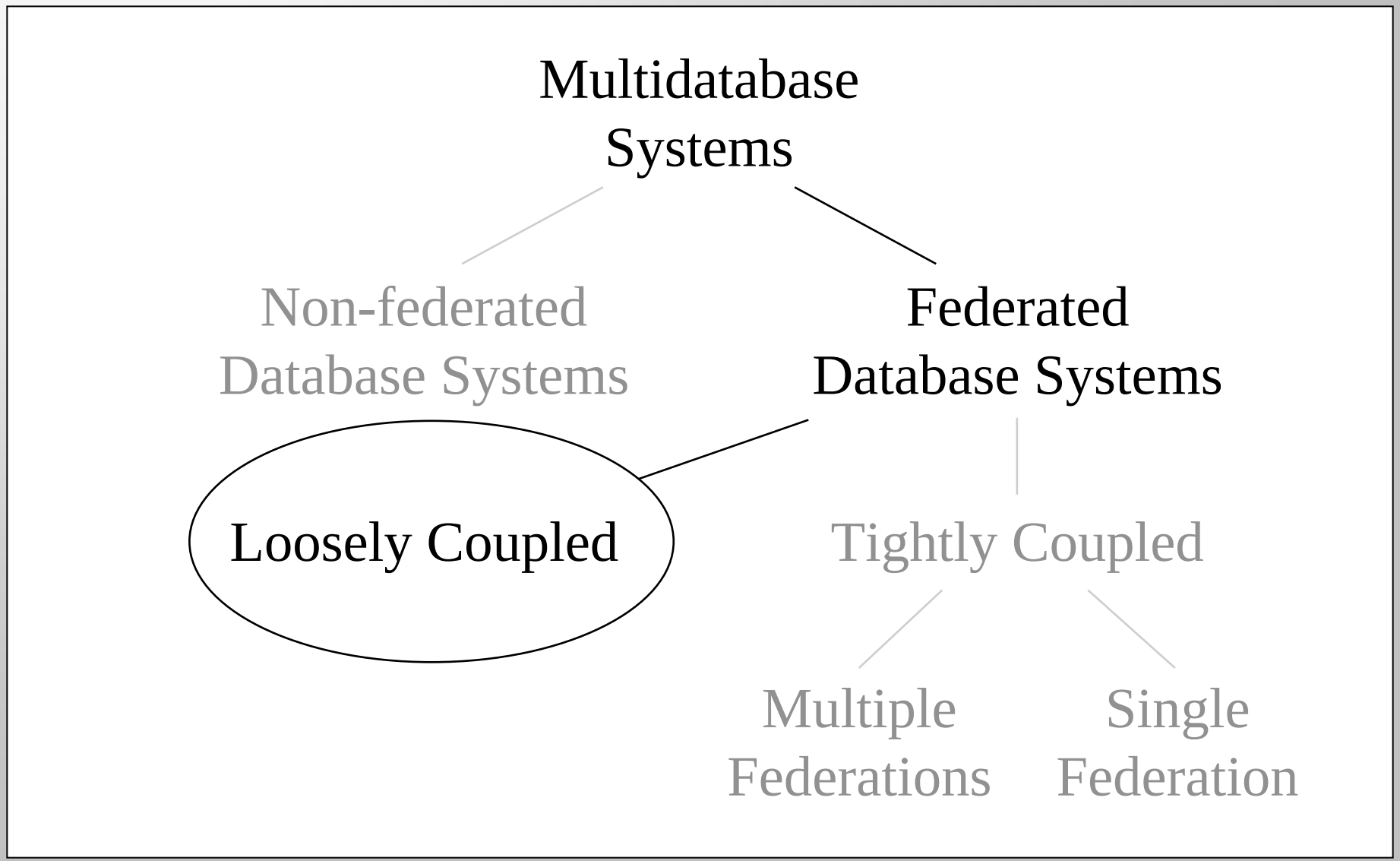
Multidatabase Taxonomy

- A ***multidatabase system*** (MDBS) supports simultaneous operations on multiple (perhaps different) component databases.
- A ***federated database system*** (FDBS) has autonomous components, whereas ***non-federated database systems*** are unitary.
- A federated system with no strong central federation management is considered ***loosely coupled***.
- One with strong central management and with federation database administrators controlling access to the components is ***tightly coupled***.
- A ***single federation*** allows only one centrally managed federated schema; a ***multiple federation*** allows multiple centrally managed schemas.

Multidatabase Taxonomy



Multidatabase Taxonomy



Multidatabase Challenges

- The coordinated updating of loosely coupled databases is still an unsolved problem.
- Maintaining inter-database referential integrity across loosely coupled databases is still an unaddressed problem.

Multidatabase Challenges

- The coordinated updating of loosely coupled databases is still an unsolved problem.
- Maintaining inter-database referential integrity across loosely coupled databases is stll an unaddressed problem.

Both of these challenges must be addressed and SOLVED before a truly effective semantic web of shared digital data can be achieved.

Data Source Problems

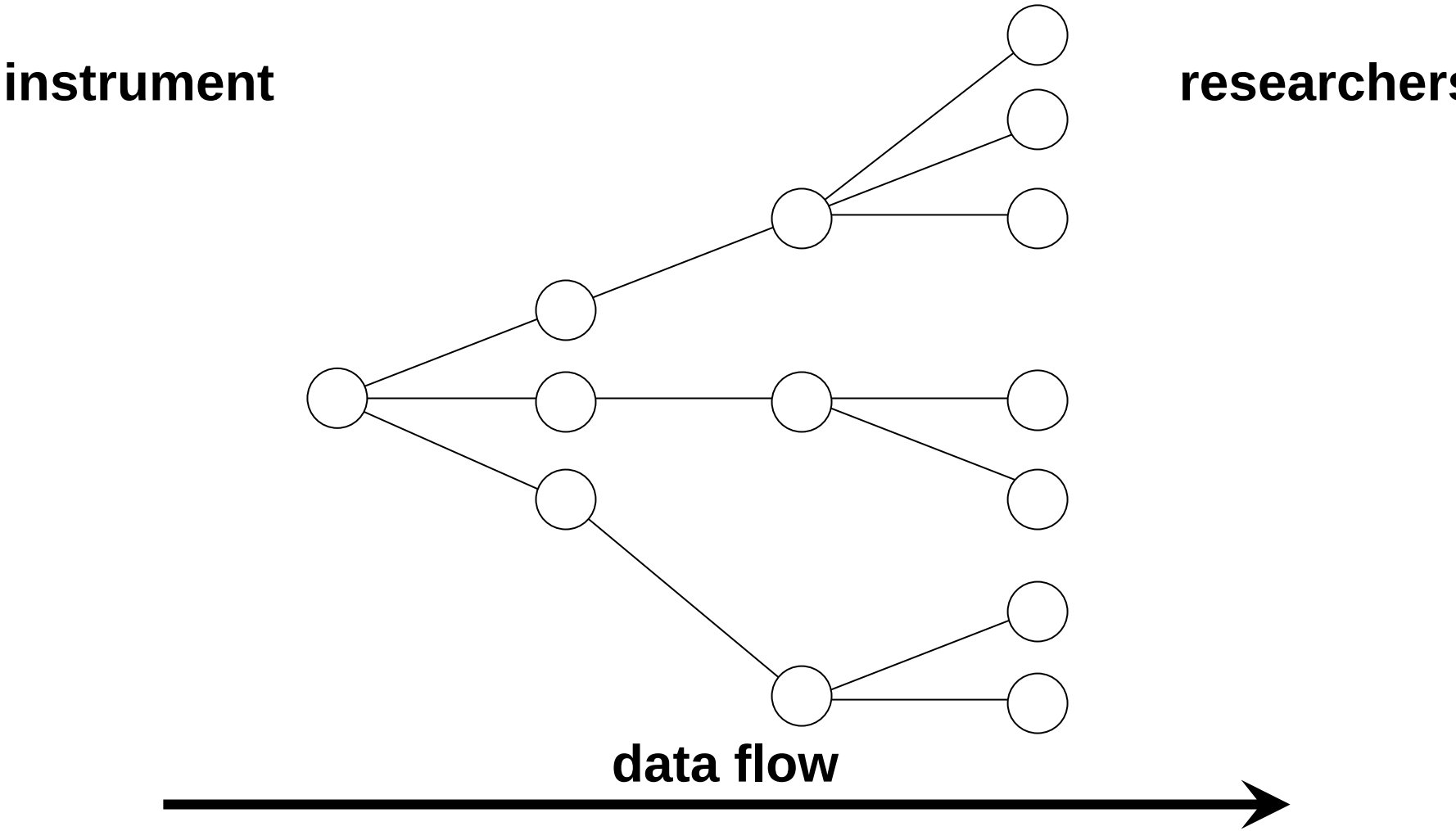
Topics

- Data-source problems
  - Biology is a small-instrument, multi-source science.
  - Integrating multi-source data is hard.
  - Consistency flows in the wrong direction.

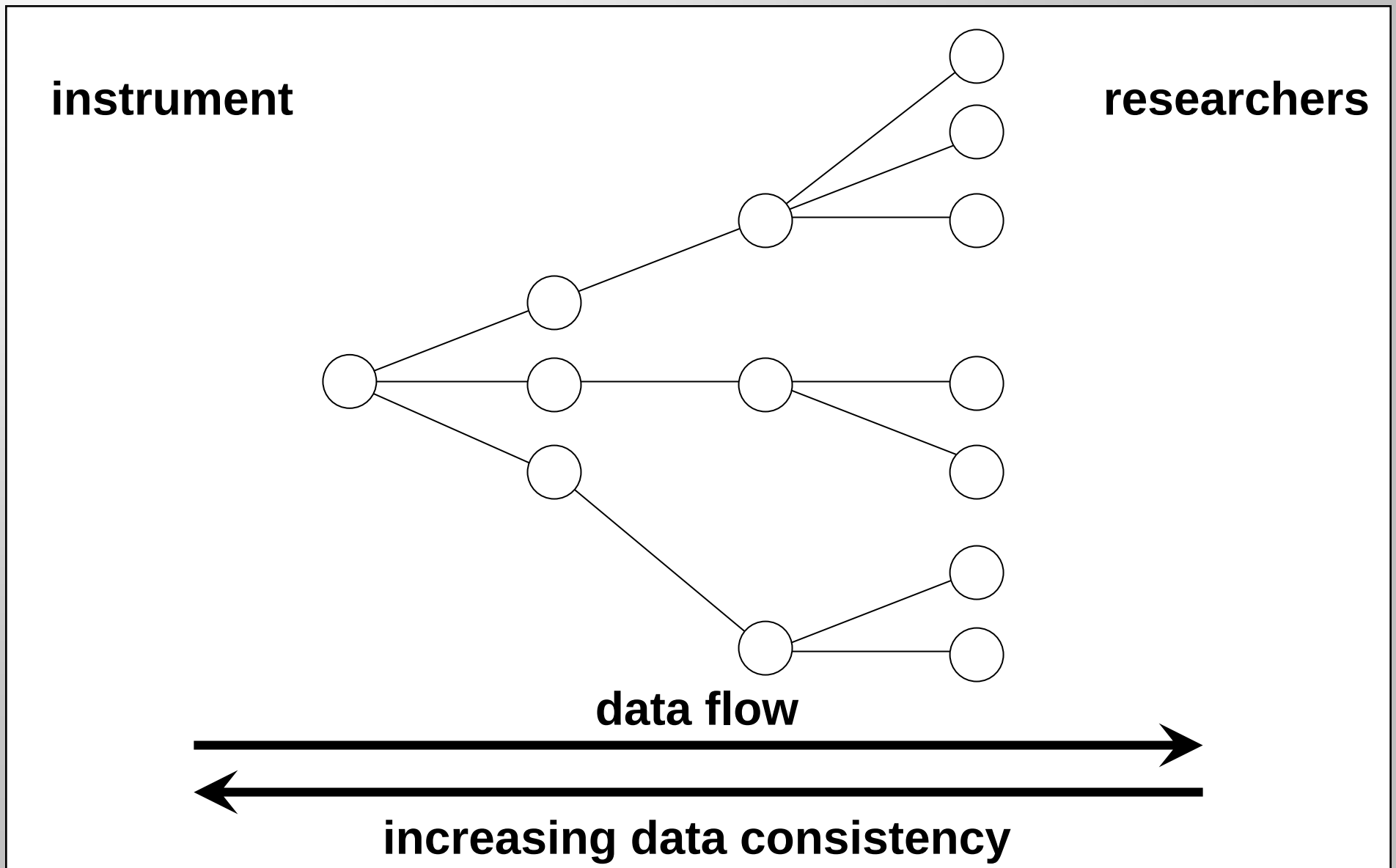
Source I

Basics

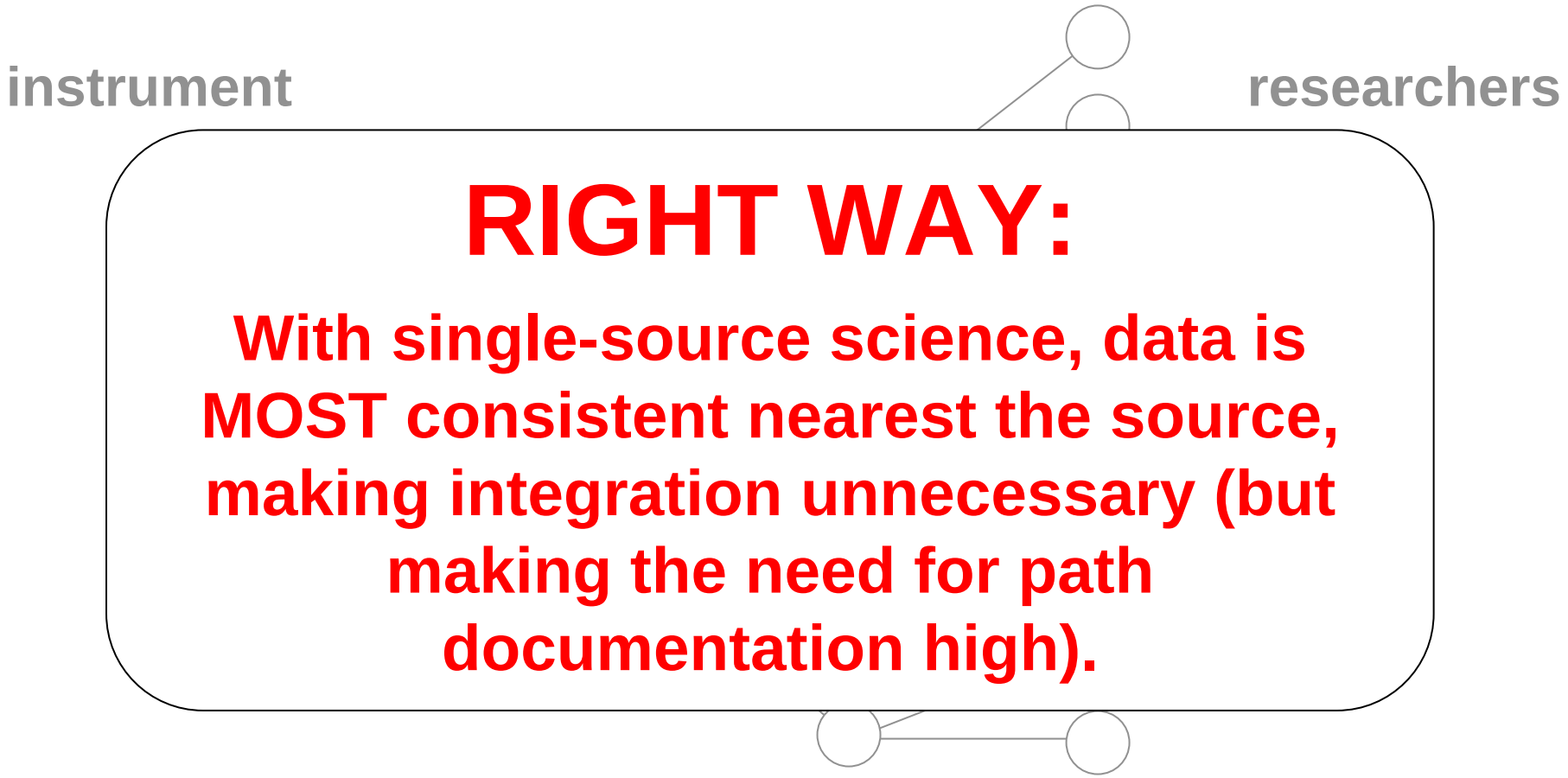
Single-instrument Science



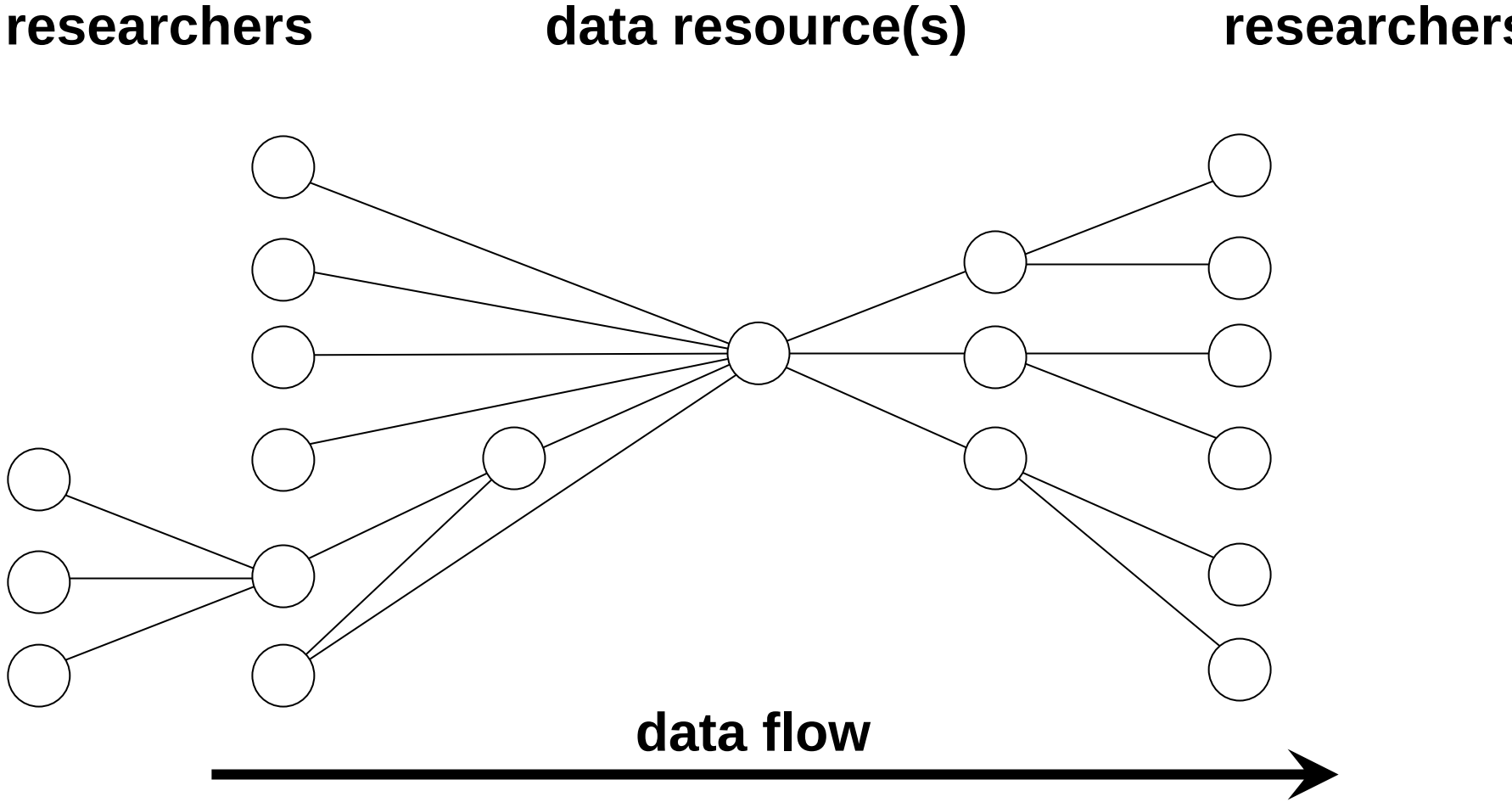
Single-instrument Science



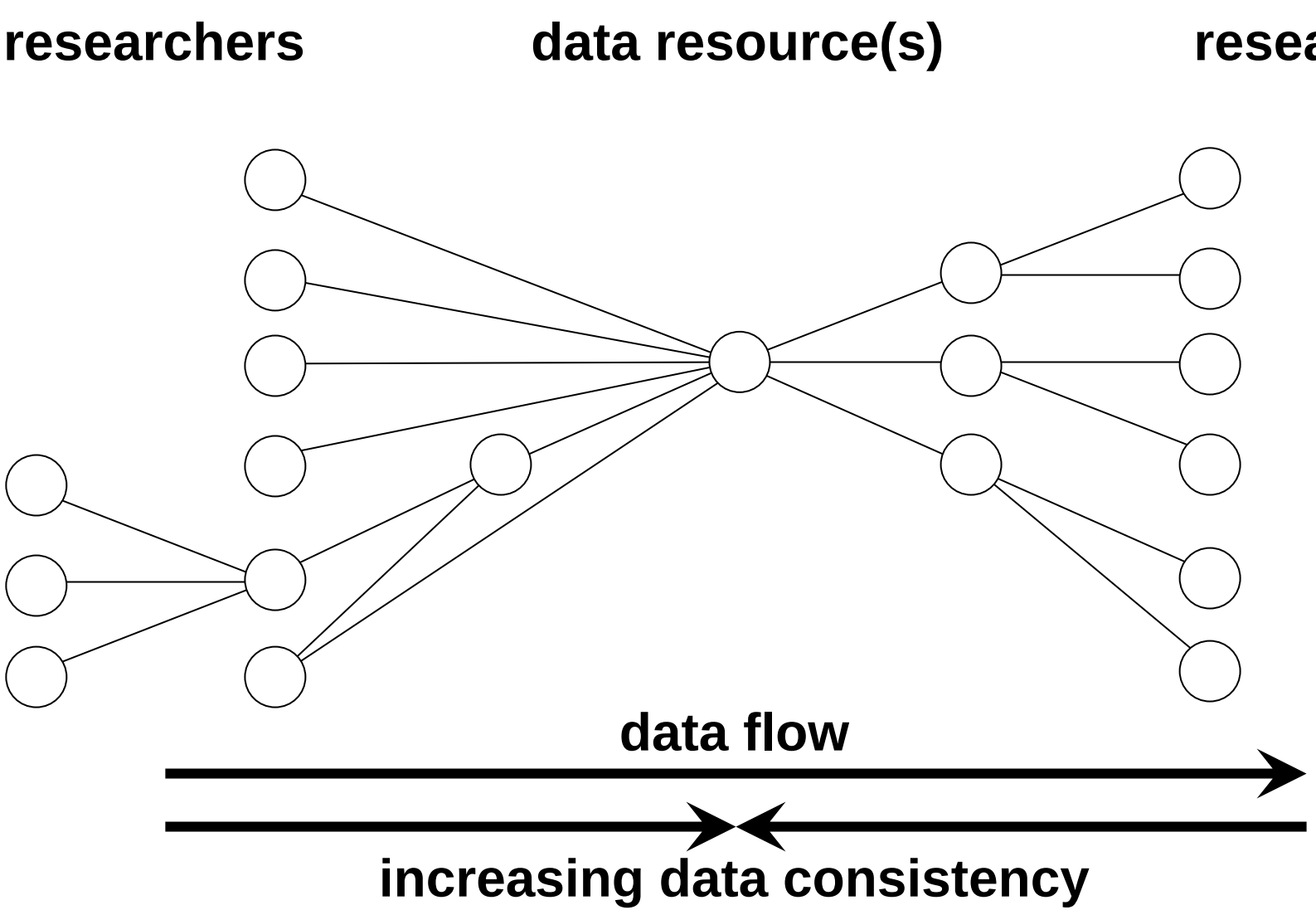
Single-instrument Science



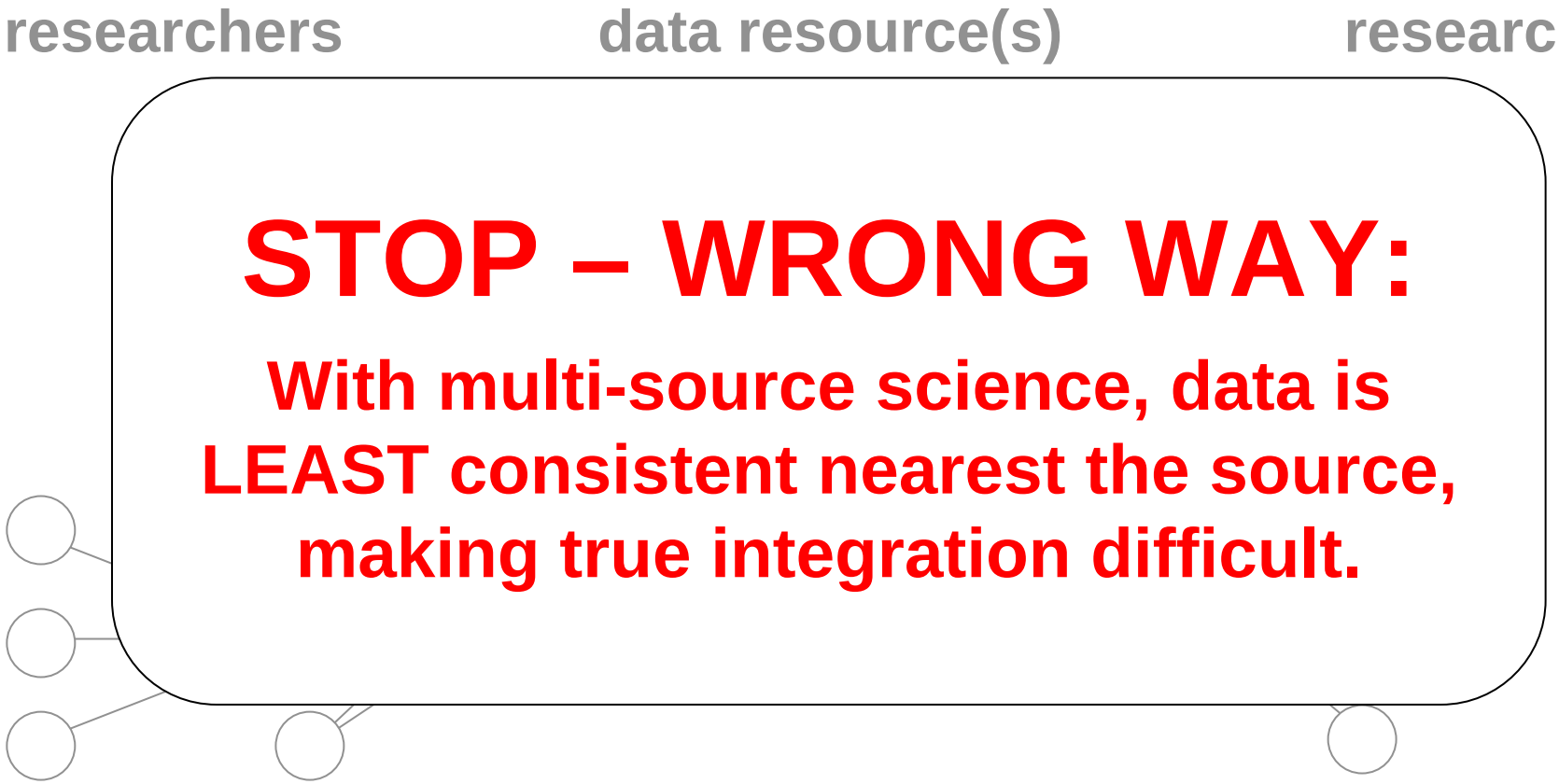
Multi-instrument Science



Multi-instrument Science

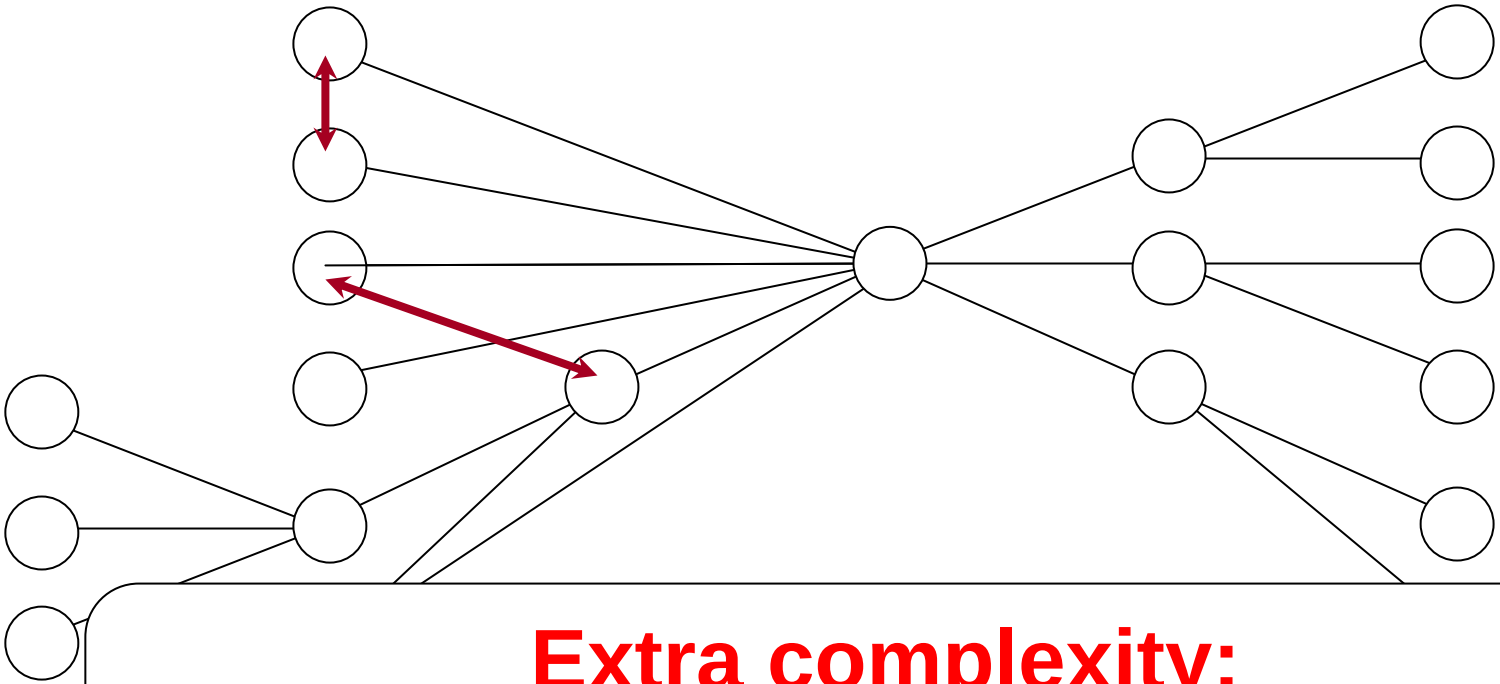


Multi-instrument Science



Multi-instrument Science

researchers data resource(s) researchers

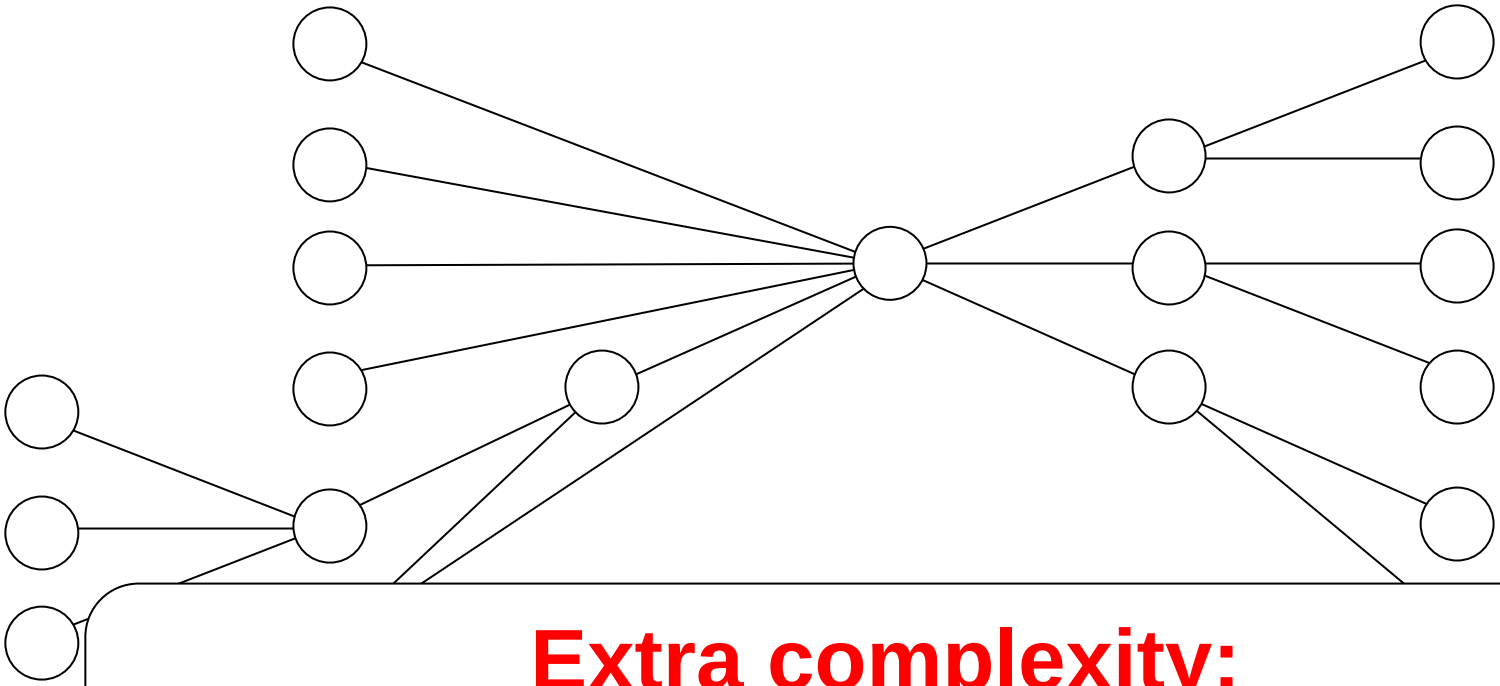


Extra complexity:

Undocumented, uncoordinated local data exchange

Multi-instrument Science

researchers data resource(s) researchers

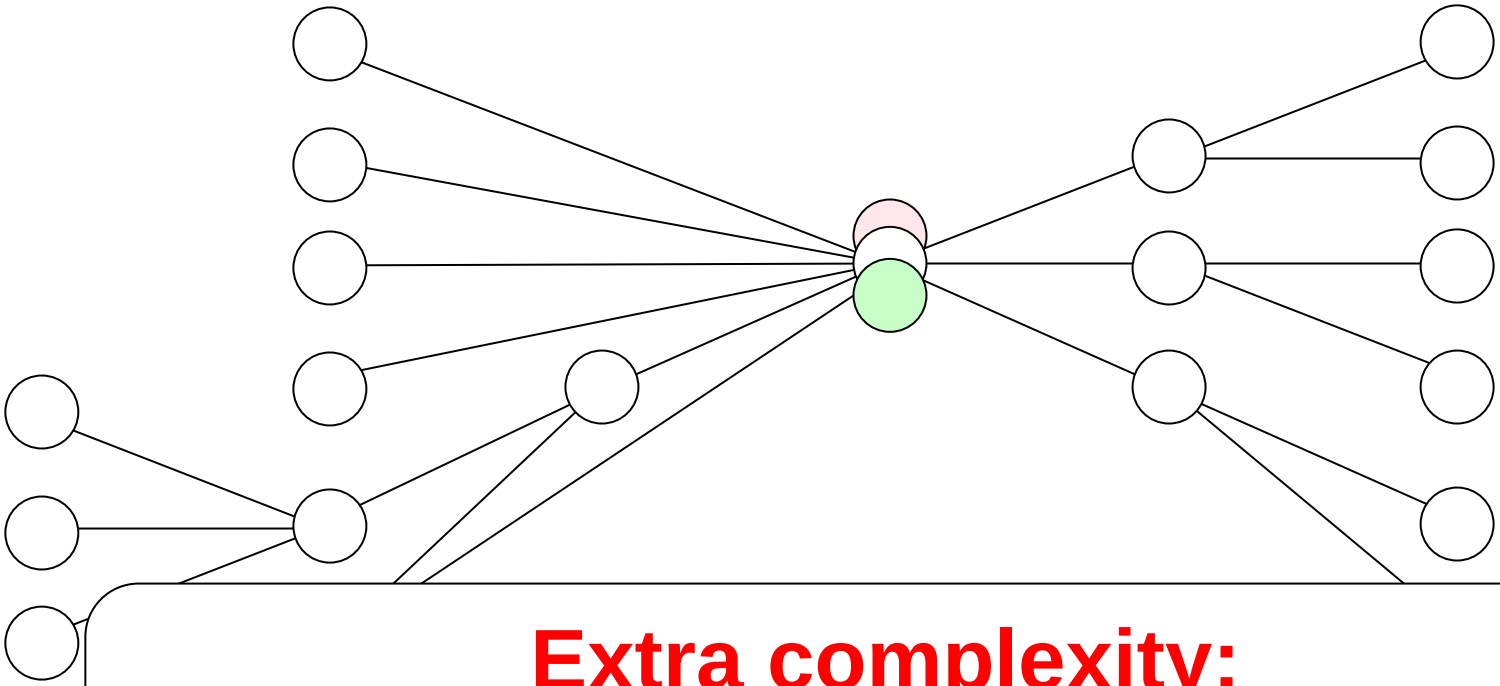


Extra complexity:

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.

Challenges Due To Problems In Metaphysics

Semantic Web

Issues

Object Identity

- **The CHALLENGE:**
  - **A semantic web requires inter-database referential integrity.**
  - **Inter-database referential integrity requires reliable and stable primary keys.**
  - **Primary keys provide for the persistent maintenance of identity.**
  - **If the concept of identity cannot be agreed upon, the proper use of primary keys cannot be agreed upon.**
  - **Without common, persistent primary keys, inter-database referential integrity is impossible.**

Object Identity

- **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

- **In any semantic web for the life sciences, no matter what technology is used, several constraints must be addressed:**
  - **RETAIL VS WHOLESALE CUSTOMERS: The semantic web must support the retail needs for coherence and the wholesale need for variation and disagreement.**
  - **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

- **In any semantic web for the life sciences, no matter what technology is used, several constraints must be addressed:**
  - **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

Semantic Web

Background Issues

Philosophical Issues: Identity

- **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 can be many "different" books associated with the same ISBN identifier; there can also be several different ISBN identifiers associated with the "same" book.**

Philosophical Issues: Properties

- **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)**

Philosophical Issues: Properties

- 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 a semantic web of biology.

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

Philosophical Issues: Properties

- **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**

Philosophical Issues: Properties

- 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.

Philosophical Issues: Identification

- **"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 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*"**

Philosophical Issues: Identification

- "Identification" is a process that reduces ambiguity. Ambiguity-reducing identification can occur in a number of different ways:
  - thout
    .,
  - er of
    may
    s
  - 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*"

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.

Practical Issues: Identifying What?

- **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)**

Practical Issues: Identifying What?

- 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?

Identification vs Specification

- **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**

Identification vs Specification

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).

Practical Issues: Identity Claims

- **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**

Practical Issues: Identity Claims

- Different methods exist for answering the question whether or not two objects are the same :
  - DEMONSTRATED IDENTITY: the identifiers are the same and the
  - ntial
  - e
  - ate of the essential properties are unknown

With checksums, LSIDs are an instance of DEMONSTRATED identity.

Without checksums, LSIDs are an instance of ASSERTED identity.

Object Identity: Open Issues

- **Several open issues must be addressed as a semantic web is deployed:**
  - **Context-free semantics are hard**
  - **Funding models support local optimization**
  - **Data degradation and time limited transactions**
  - **Sociology of cutting edge science**

Challenges Due to Science Itself

Challenges/Limits

- **Science is constantly changing**
- **Scientific “facts” are never globally consistent**
- **Scientific databases are never perfect**
- **Resources are always limiting**
- **Needs are constantly changing**
- **Technology keeps evolving**

Challenges/Limits

- Science is constantly changing

THE REAL CHALLENGE:

Doing something genuinely useful anyway.

Challenges/Limits

Data Inconsistency

Logic 101

- **If premise “A” is false, then the statement “IF A then B” is always true, regardless of the truth value of “B”.**

Logic 101

- If premise “A” is false, then the statement “IF A then B” is always true, regardless of the truth value of “B”.
- **That is, with a false antecedent you can prove anything.**

Logic 101

- If premise “A” is false, then the statement “IF A then B” is always true, regardless of the truth value of “B”.
- That is, with a false antecedent you can prove anything.
- **“A and not A” is always false.**

Logic 101

- If premise “A” is false, then the statement “IF A then B” is always true, regardless of the truth value of “B”.
- That is, with a false antecedent you can prove anything.
- “A and not A” is always false.
- **Feeding inconsistent premises into a logical calculator yields nonsense.**

Logic 101

- If premise “A” is false, then the statement “IF A then B” is always true, regardless of the truth value of “B”.
- Feeding inconsistent premises into a logical calculator yields nonsense.

Seamless access to inconsistent data is a bad idea.

Challenges/Limits

Errors Accrete

GOAL: A Caution

In parallel to the molecular database GenBank (but operating on completely different principles), GBIF envisions a future in which all sorts of information about any species (gene sequences, occurrence in ecosystems, specific locality data, ecological relationships, physiological requirements and so on) would be compiled on demand from many, disparate, continuously updated databases.

SpeciesBANK would effectively be an encyclopedia of species that is continuously filling in missing or supplanting outdated information.

GOAL: A Caution

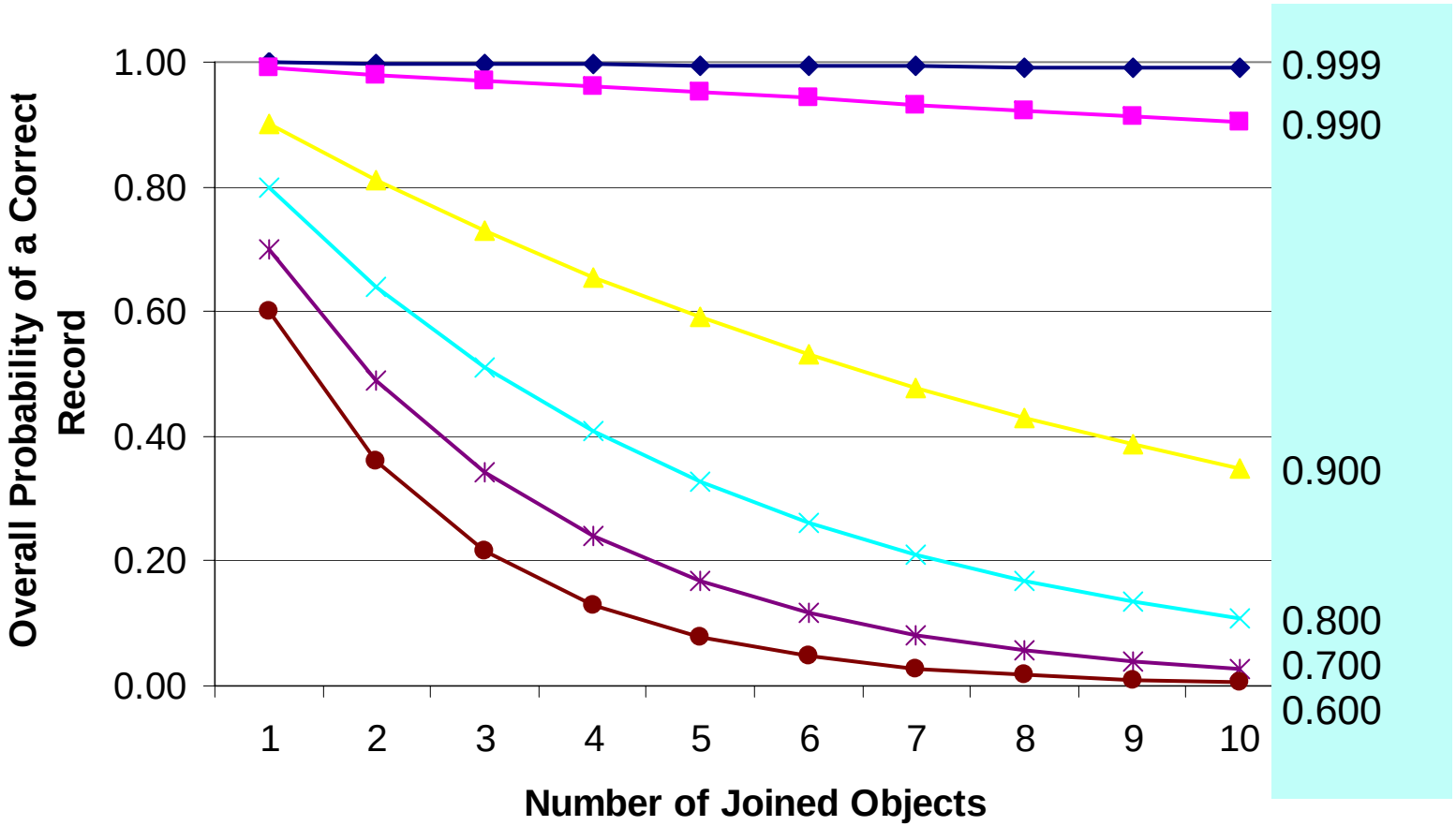
In parallel to the molecular database GenBank (but operating on completely different principles), GBIF envisions a future in which all sorts of information about any species (gene sequences, occurrence in ecosystems, specific locality data, ecological relationships, physiological requirements and so on) would be compiled on demand from many, disparate, continuously updated databases.

SpeciesBANK would effectively be an encyclopedia of species that is continuously filling in missing or supplanting outdated information.

Declining Overall Probabilities

- If a “record” in SpeciesBANK is assembled (joined) from data components maintained independently, and
- If the component data collections are not perfect (e.g., the probability of correct = p),
- Then the proportion of completely correct SpeciesBANK records in a query will be given by p^n, where n is the number of elements joined in the query.

Declining Overall Probabilities



As p goes down, p^n goes down a lot faster.

Declining Overall Probabilities

What kinds of error rates (or inconsistency rates) occur in real data sets?

A recent study of human genome data (chromosome band location of genes), in two large, curated databases, showed an average error rate of 0.1, giving p = 0.9.

What about some species data?

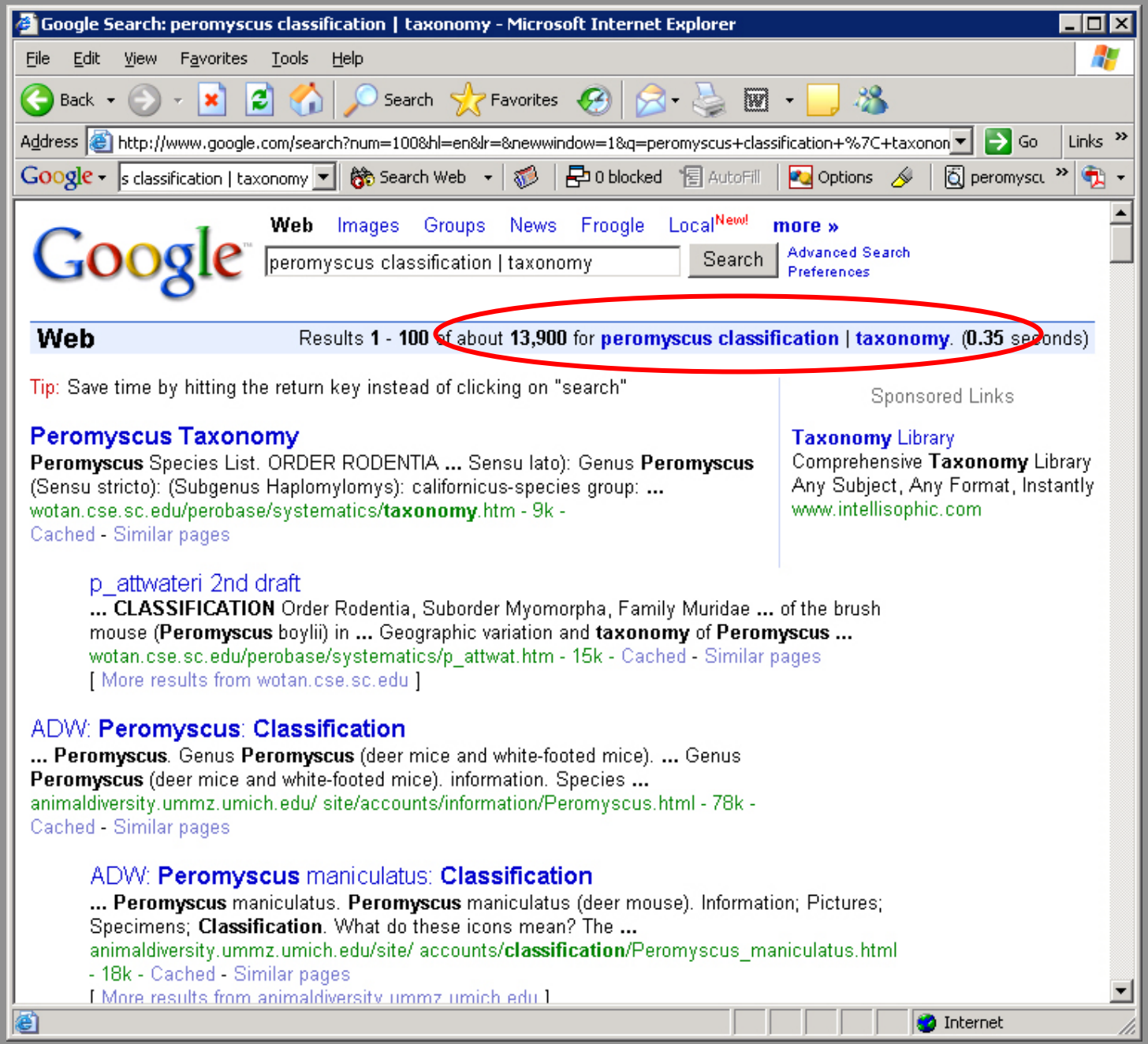
Challenges/Limits

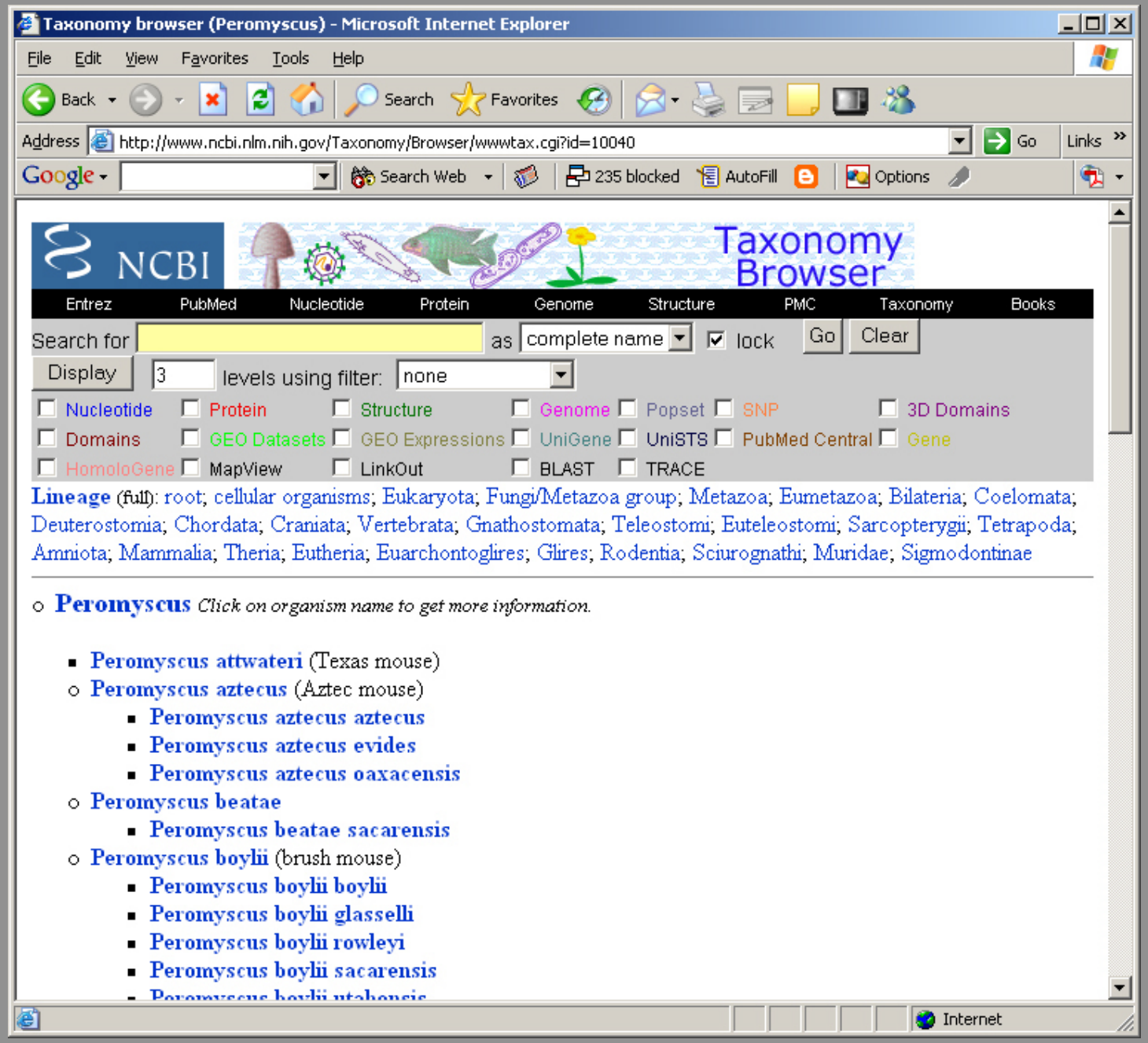
An Example

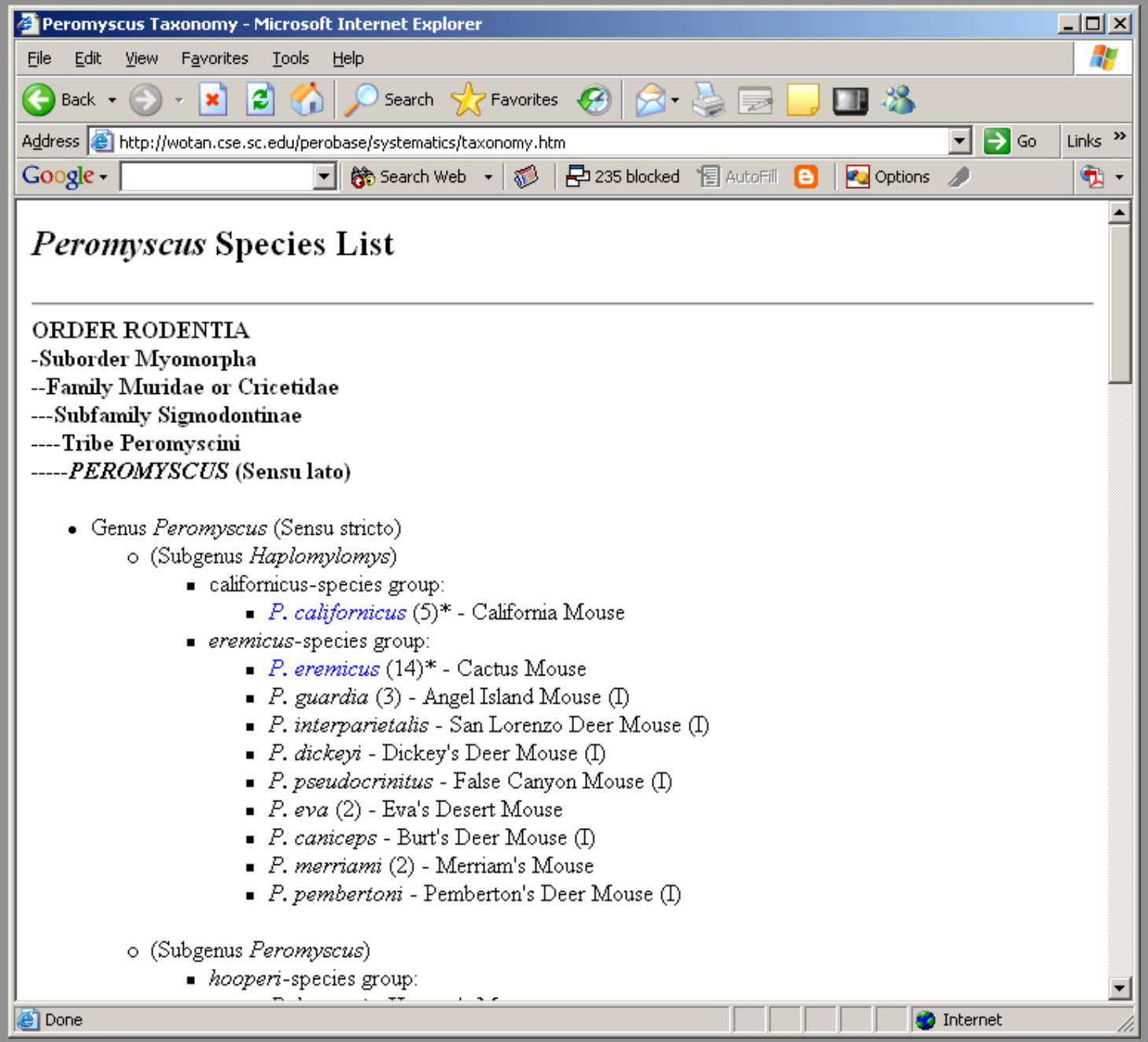
Peromyscus: example

Source: http://cedarcreek.umn.edu/mammals/cricetidae.html

Peromyscus maniculatus


Google Search: peromyscus classification | taxonomy - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://www.google.com/search?num=100&hl=en&lr=&newwindow=1&q=peromyscus+classification+%7C+taxonon
Go Links
Google s classification | taxonomy Search Web 0 blocked AutoFill Options peromyscu
Google
Web Images Groups News Froogle Local New! more »
peromyscus classification | taxonomy
Search
Advanced Search
Preferences
Web
Results 1 - 100 of about 13,900 for peromyscus classification | taxonomy. (0.35 seconds)
Tip: Save time by hitting the return key instead of clicking on "search"
Peromyscus Taxonomy
Peromyscus Species List. ORDER RODENTIA ... Sensu lato): Genus Peromyscus (Sensu stricto): (Subgenus Haplomylomys): californicus-species group: ...
wotan.cse.sc.edu/perobase/systematics/taxonomy.htm - 9k -
Cached - Similar pages
Sponsored Links
Taxonomy Library
Comprehensive Taxonomy Library
Any Subject, Any Format, Instantly
www.intellisophic.com
p_attwateri 2nd draft
... CLASSIFICATION Order Rodentia, Suborder Myomorpha, Family Muridae ... of the brush mouse (Peromyscus boylii) in ... Geographic variation and taxonomy of Peromyscus ...
wotan.cse.sc.edu/perobase/systematics/p_attwat.htm - 15k - Cached - Similar pages
[More results from wotan.cse.sc.edu]
ADW: Peromyscus: Classification
... Peromyscus. Genus Peromyscus (deer mice and white-footed mice). ... Genus Peromyscus (deer mice and white-footed mice). information. Species ...
animaldiversity.ummz.umich.edu/ site/accounts/information/Peromyscus.html - 78k -
Cached - Similar pages
ADW: Peromyscus maniculatus: Classification
... Peromyscus maniculatus. Peromyscus maniculatus (deer mouse). Information; Pictures; Specimens; Classification. What do these icons mean? The ...
animaldiversity.ummz.umich.edu/site/ accounts/classification/Peromyscus_maniculatus.html
- 18k - Cached - Similar pages
[More results from animaldiversity.ummz.umich.edu]
Internet

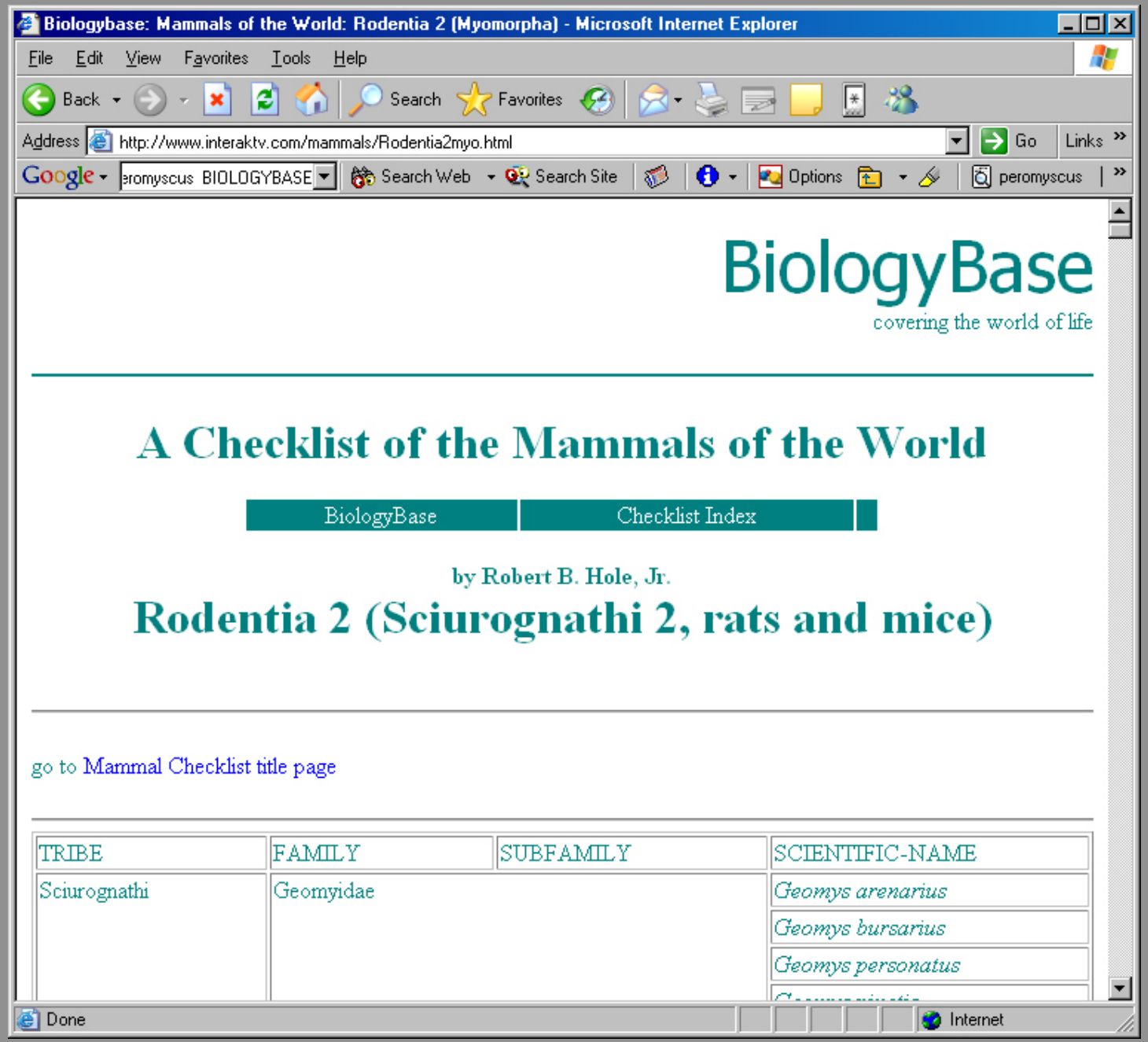

Taxonomy browser (Peromyscus) - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=10040
Go Links
Google Search Web 235 blocked AutoFill Options
NCBI
Taxonomy Browser
Entrez PubMed Nucleotide Protein Genome Structure PMC Taxonomy Books
Search for as complete name lock Go Clear
Display 3 levels using filter: none
Nucleotide Protein Structure Genome Popset SNP 3D Domains
Domains GEO Datasets GEO Expressions UniGene UniSTS PubMed Central Gene
HomoloGene MapView LinkOut BLAST TRACE
Lineage (full): root; cellular organisms; Eukaryota; Fungi/Metazoa group; Metazoa; Eumetazoa; Bilateria; Coelomata; Deuterostomia; Chordata; Craniata; Vertebrata; Gnathostomata; Teleostomi; Euteleostomi; Sarcopterygii; Tetrapoda; Amniota; Mammalia; Theria; Eutheria; Euarchontoglires; Glires; Rodentia; Sciurognathi; Muridae; Sigmodontinae
Peromyscus Click on organism name to get more information.
Peromyscus attwateri (Texas mouse)
Peromyscus aztecus (Aztec mouse)
Peromyscus aztecus aztecus
Peromyscus aztecus evides
Peromyscus aztecus oaxacensis
Peromyscus beatae
Peromyscus beatae sacarensis
Peromyscus boylii (brush mouse)
Peromyscus boylii boylii
Peromyscus boylii glasselli
Peromyscus boylii rowleyi
Peromyscus boylii sacarensis
Internet



Peromyscus Species List

ORDER RODENTIA
-Suborder Myomorpha
--Family Muridae or Cricetidae
---Subfamily Sigmodontinae
----Tribe Peromyscini
-----_PEROMYSCUS_ (Sensu lato)

- Genus *Peromyscus* (Sensu stricto)
  - (Subgenus *Haplomylomys*)
    - californicus-species group:
      - *P. californicus* (5)* - California Mouse
    - *eremicus*-species group:
      - *P. eremicus* (14)* - Cactus Mouse
      - *P. guardia* (3) - Angel Island Mouse (I)
      - *P. interparietalis* - San Lorenzo Deer Mouse (I)
      - *P. dickeyi* - Dickey's Deer Mouse (I)
      - *P. pseudocrinitus* - False Canyon Mouse (I)
      - *P. eva* (2) - Eva's Desert Mouse
      - *P. caniceps* - Burt's Deer Mouse (I)
      - *P. merriami* (2) - Merriam's Mouse
      - *P. pembertoni* - Pemberton's Deer Mouse (I)
  - (Subgenus *Peromyscus*)
    - *hooperi*-species group:


Biologybase: Mammals of the World: Rodentia 2 (Myomorpha) - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://www.interaktv.com/mammals/Rodentia2myo.html
Go Links
Google eromyscus BIOLOGYBASE Search Web Search Site Options peromyscus
BiologyBase
covering the world of life
A Checklist of the Mammals of the World
BiologyBase
Checklist Index
by Robert B. Hole, Jr.
Rodentia 2 (Sciurognathi 2, rats and mice)
go to Mammal Checklist title page
TRIBE
FAMILY
SUBFAMILY
SCIENTIFIC-NAME
Sciurognathi
Geomyidae
Geomys arenarius
Geomys bursarius
Geomys personatus
Done
Internet

Peromyscus: number of species

NCBI:	42
Perobase:	55
BiologyBASE:	53
Total:	64
In common:	32

Peromyscus: number of species

NCBI:	42
Perobase:	55
BiologyBASE:	53
Total:	64
In common:	32

Hmmm. Fifty percent concordance across only three resources.

Not so hot…

Challenges/Limits

Constant Revision

GOAL: Another Caution

In parallel to the molecular database GenBank (but operating on completely different principles), GBIF envisions a future in which all sorts of information about any species (gene sequences, occurrence in ecosystems, specific locality data, ecological relationships, physiological requirements and so on) would be compiled on demand from many, disparate, continuously updated databases.

SpeciesBANK would effectively be an encyclopedia of species that is continuously filling in missing or supplanting outdated information.

GOAL: Another Caution

In parallel to the molecular database GenBank (but operating on completely different principles), GBIF envisions a future in which all sorts of information about any species (gene sequences, occurrence in ecosystems, specific locality data, ecological relationships, physiological requirements and so on) would be compiled on demand from many, disparate, continuously updated databases.

SpeciesBANK would effectively be an encyclopedia of species that is continuously filling in missing or supplanting outdated information.

Primary Literature

- Each contribution to the primary literature is an original contribution. It may be based on prior findings, or it may completely overturn prior findings.
- NO REQUIREMENT OF CONSISTENCY exists between any two documents in the primary literature.

Encyclopedia of Science

Should a biological database be a compilation of scientific facts, or should it be a collection of scientific observations?

A compilation of facts is appealing, but…

Scientific “facts” have a way of changing with more scientific observations, and the growing burden of constant editing to achieve accuracy and internal consistency would be difficult.

Encyclopedia of Science

Science continually evolves. Scientific knowledge is under constant revision in the light of new evidence. From a practical point of view, it is not the ultimate truth of the scientific world picture that matters, but the [current] scientific answers to particular questions...

The concept of an archive of reliable scientific knowledge is much too schematic. There is no *Encyclopaedia* where *all* well-established science, and only well-established science, may be consulted. If such an institution existed, it would be in constant agitation, as new information was being added, and old facts and assertions struck out.

Ziman, J. 1978. *Reliable Knowledge: An Exploration of the Grounds for Belief in Science*. London: Cambridge University press.

Encyclopedia of Science

Science continually evolves. Scientific knowledge is under constant revision in the light of new evidence. From a practical point of view, it is not the ultimate truth of the scientific world picture that matters, but the [current] scientific answers to particular questions...

The concept of an archive of reliable scientific knowledge is much too schematic. There is no *Encyclopaedia* where *all* well-established science, and only well-established science, may be consulted. If such an institution existed, it would be in constant agitation, as new information was being added, and old facts and assertions struck out.

Ziman, J. 1978. *Reliable Knowledge: An Exploration of the Grounds for Belief in Science*. London: Cambridge University press.

Limits to Global Integration

OBSERVATION:

It is easy to imagine “global integration of biodiversity data” as a goal for a future, successful SpeciesBANK program.

Limits to Global Integration

ASSERTION:

The notion of final “global integration” is simply inconsistent with the actual practice of science and the notion of temporary global integration is nonsensical.

Databases as Primary Literature

Most scientists hold primary literature in high regard, while giving less credence to secondary and tertiary sources.

but

Databases as Primary Literature

[The] layman who attempts to consult all the [primary literature] relevant to a particular scientific question is soon wearied and appalled by the confusion and diversity of fact and opinion that he will find.

Ziman, J. 1978. *Reliable Knowledge: An Exploration of the Grounds for Belief in Science*. London: Cambridge University press.

Databases as Primary Literature

[The] layman who attempts to consult all the [primary literature] relevant to a particular scientific question is soon wearied and appalled by the confusion and diversity of fact and opinion that he will find. At the research frontier, scientific knowledge is untested, unselected, contradictory and outwardly chaotic.

Ziman, J. 1978. *Reliable Knowledge: An Exploration of the Grounds for Belief in Science*. London: Cambridge University press.

Databases as Primary Literature

No amount of magic can integrate “data” that are

untested,

unselected,

contradictory, and

outwardly chaotic

into anything resembling a coherent whole.

Databases as Primary Literature

Can there ever be a biological database of everything?

In a word: NO

Constant Revision

An Example

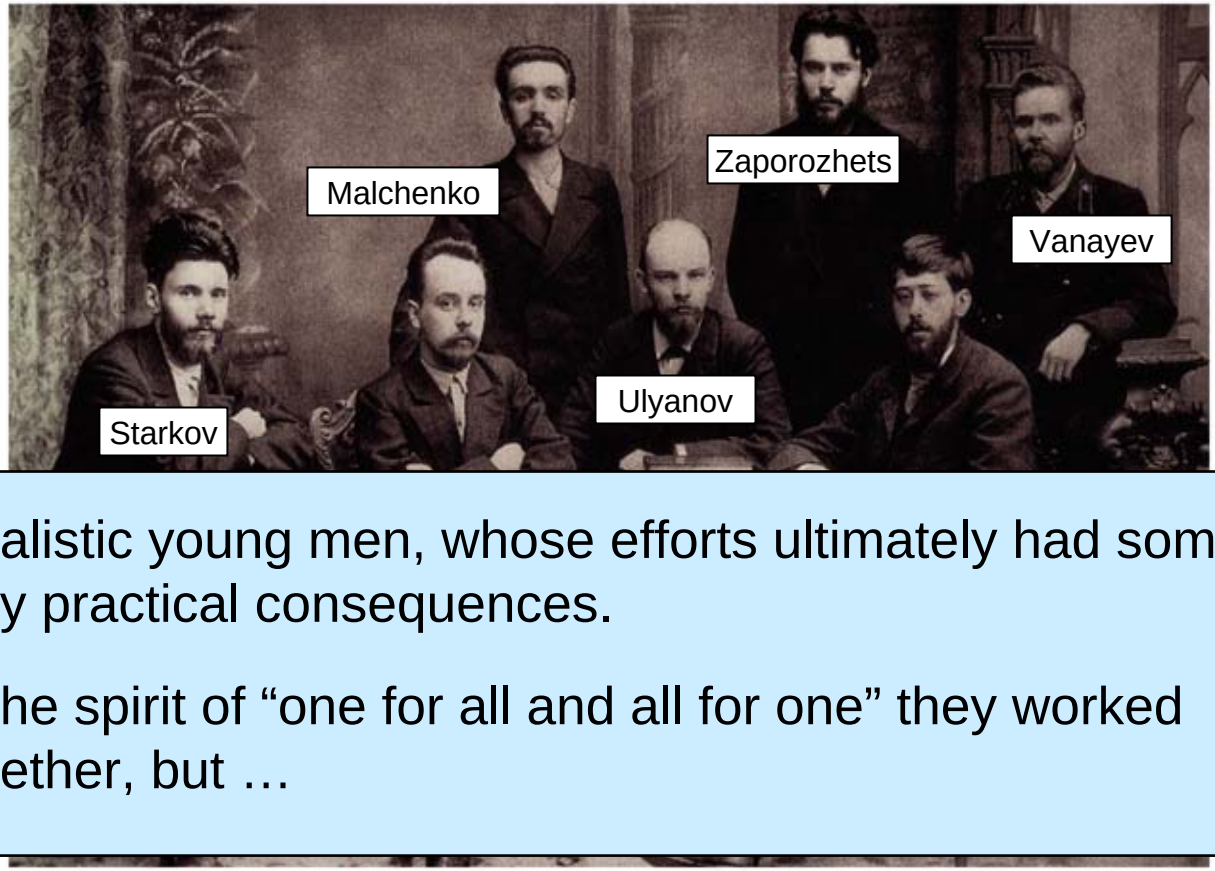
The Perils of Constant Revision

St. Petersburg Union of Struggle for the Liberation of the Working Class
Photograph taken in 1897

The Perils of Constant Revision

St. Petersburg Union of Struggle for the Liberation of the Working Class
Photograph taken in 1897

The Perils of Constant Revision



Idealistic young men, whose efforts ultimately had some very practical consequences.

In the spirit of “one for all and all for one” they worked together, but …

St. Petersburg Union of Struggle for the Liberation of the Working Class
Photograph taken in 1897

The Perils of Constant Revision

In 1929, Malchenko was arrested and accused of being a "wrecker". He was executed 18 November 1930.

As a counter-revolutionary wrecker of the party, he could hardly have been a participant in its early creation, so…

St. Petersburg Union of Struggle for the Liberation of the Working Class
Photograph taken in 1897

The Perils of Constant Revision

History required some correction.

Thus, when the picture was next published…

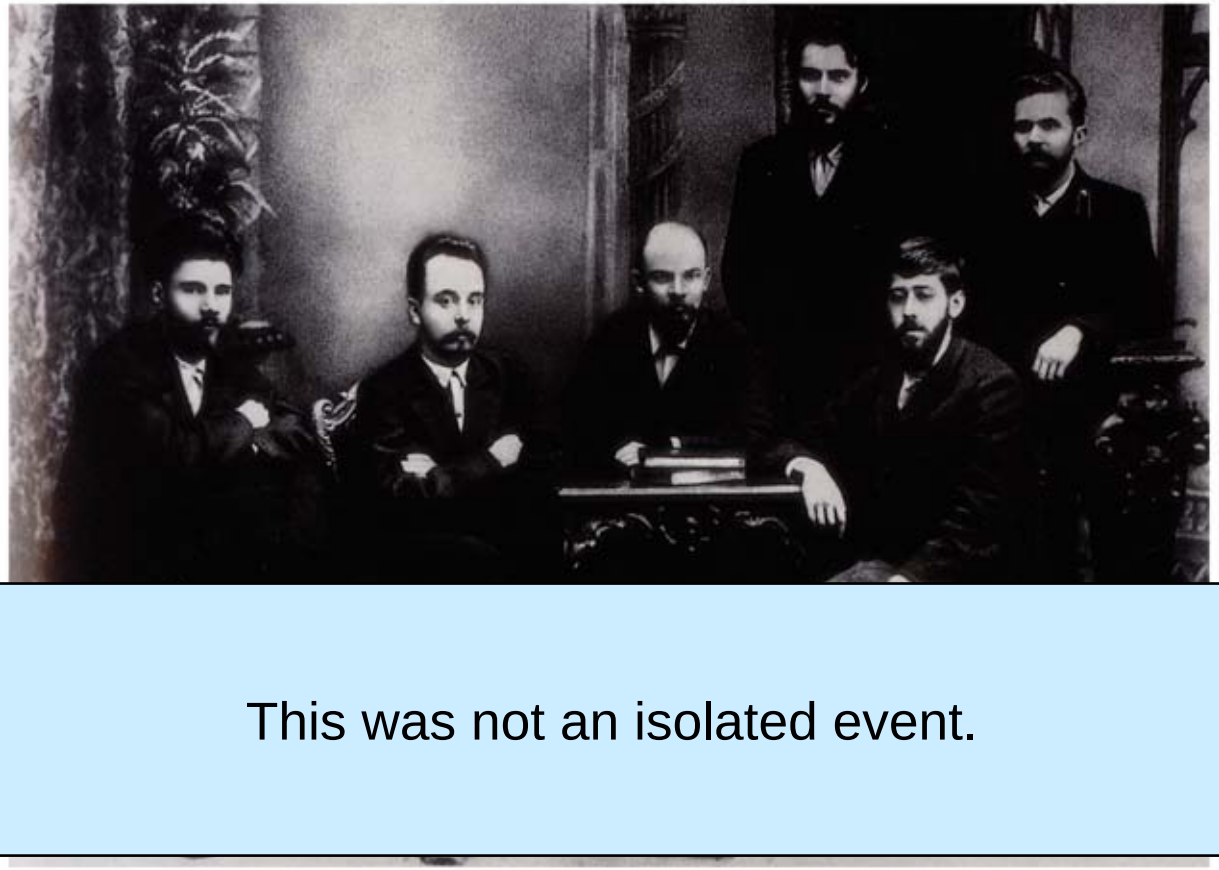
St. Petersburg Union of Struggle for the Liberation of the Working Class
Photograph taken in 1897

The Perils of Constant Revision

Malchenko was gone.

St. Petersburg Union of Struggle for the Liberation of the Working Class
Photograph published in 1939

The Perils of Constant Revision



This was not an isolated event.

St. Petersburg Union of Struggle for the Liberation of the Working Class
Photograph published in 1939

The Perils of Constant Revision

Stalin, with comrades

The Perils of Constant Revision

Stalin, with fewer comrades

The Perils of Constant Revision

Photograph from 1934 Russian edition of *Ten Years of Uzbekistan*

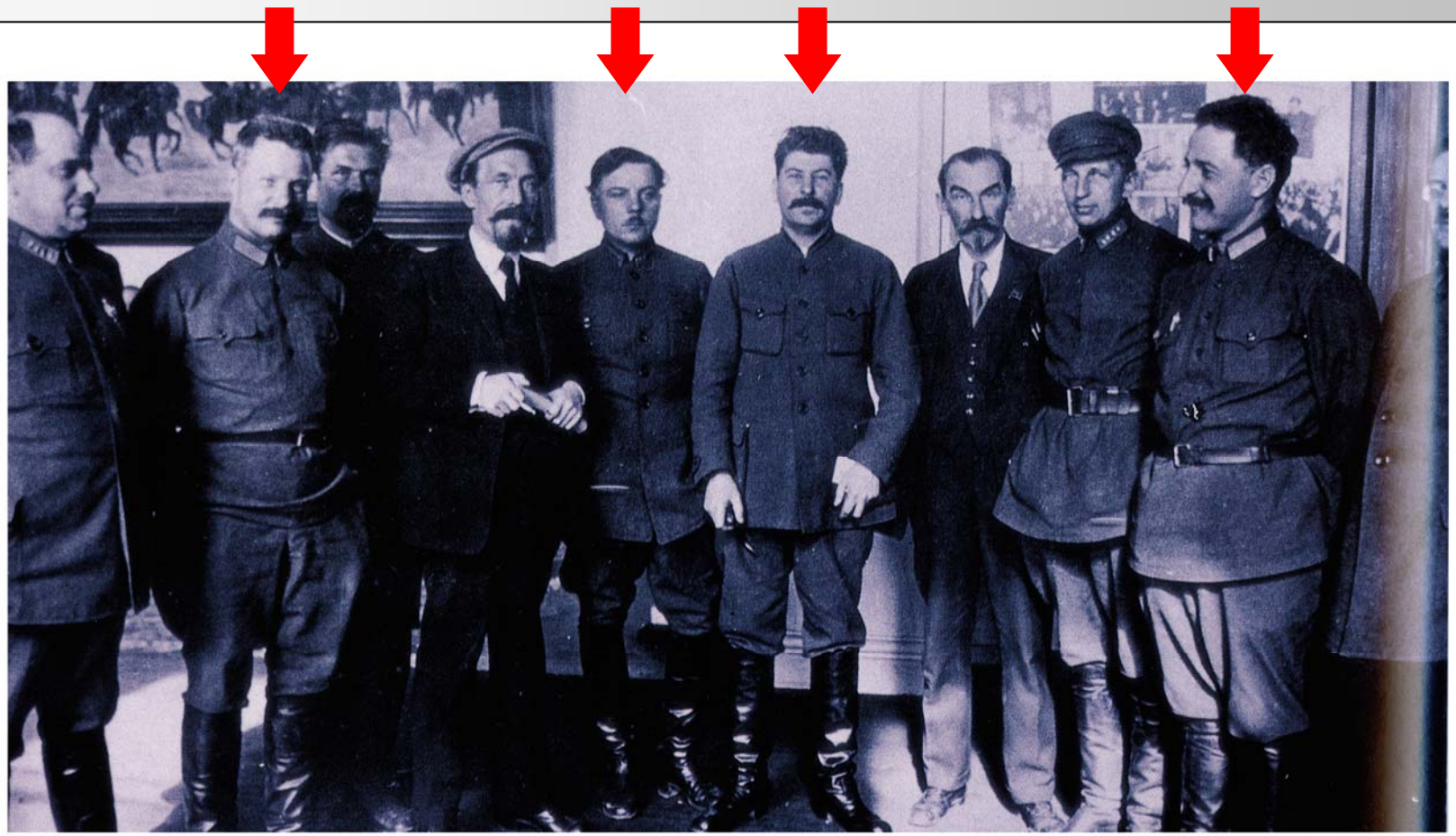
The Perils of Constant Revision

Photograph from 1935 Uzbek edition of *Ten Years of Uzbekistan*

The Perils of Constant Revision

Ten Comrades at the 14th Party Congress in 1925

The Perils of Constant Revision



Ten Comrades at the 14th Party Congress in 1925

The Perils of Constant Revision

In 1939 there were four.

The Perils of Constant Revision

Four

The Perils of Constant Revision

Three

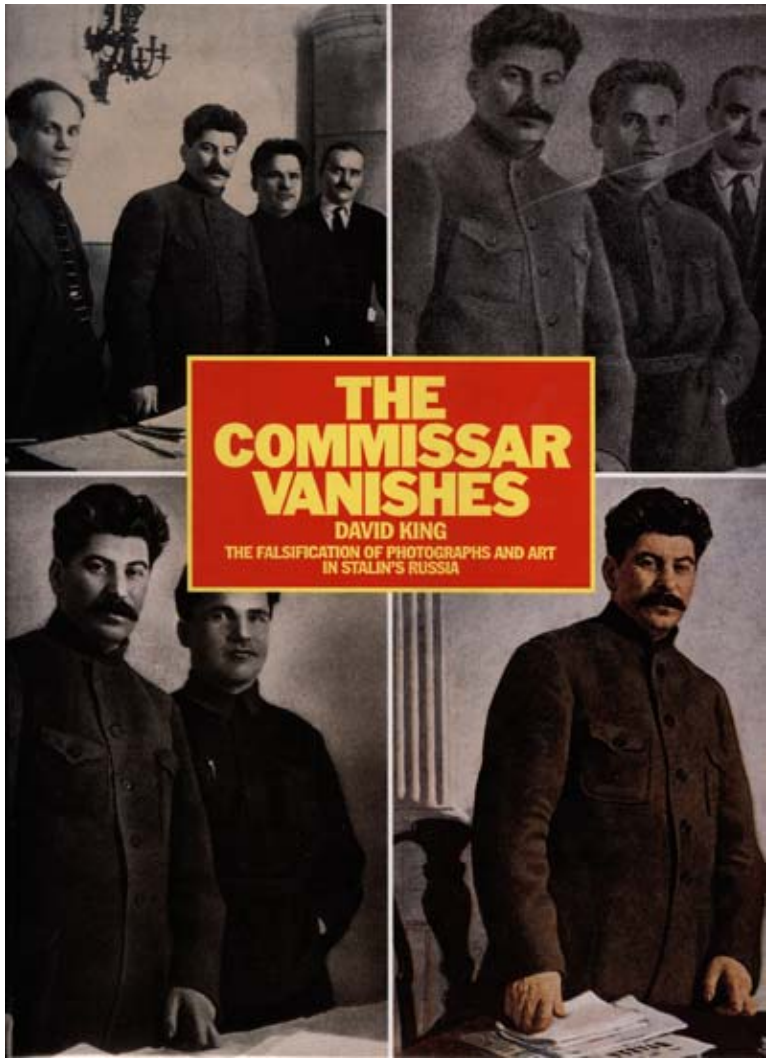
The Perils of Constant Revision

Two

The Perils of Constant Revision

One

This book documents the efforts of the communist party to edit the historical record so that it always reflected current party dogma.



It provides a lesson in the fundamental impossibility of such a task.

Challenges Due to Inappropriate Standards

Standards

- Using standards always seems like a good idea, but

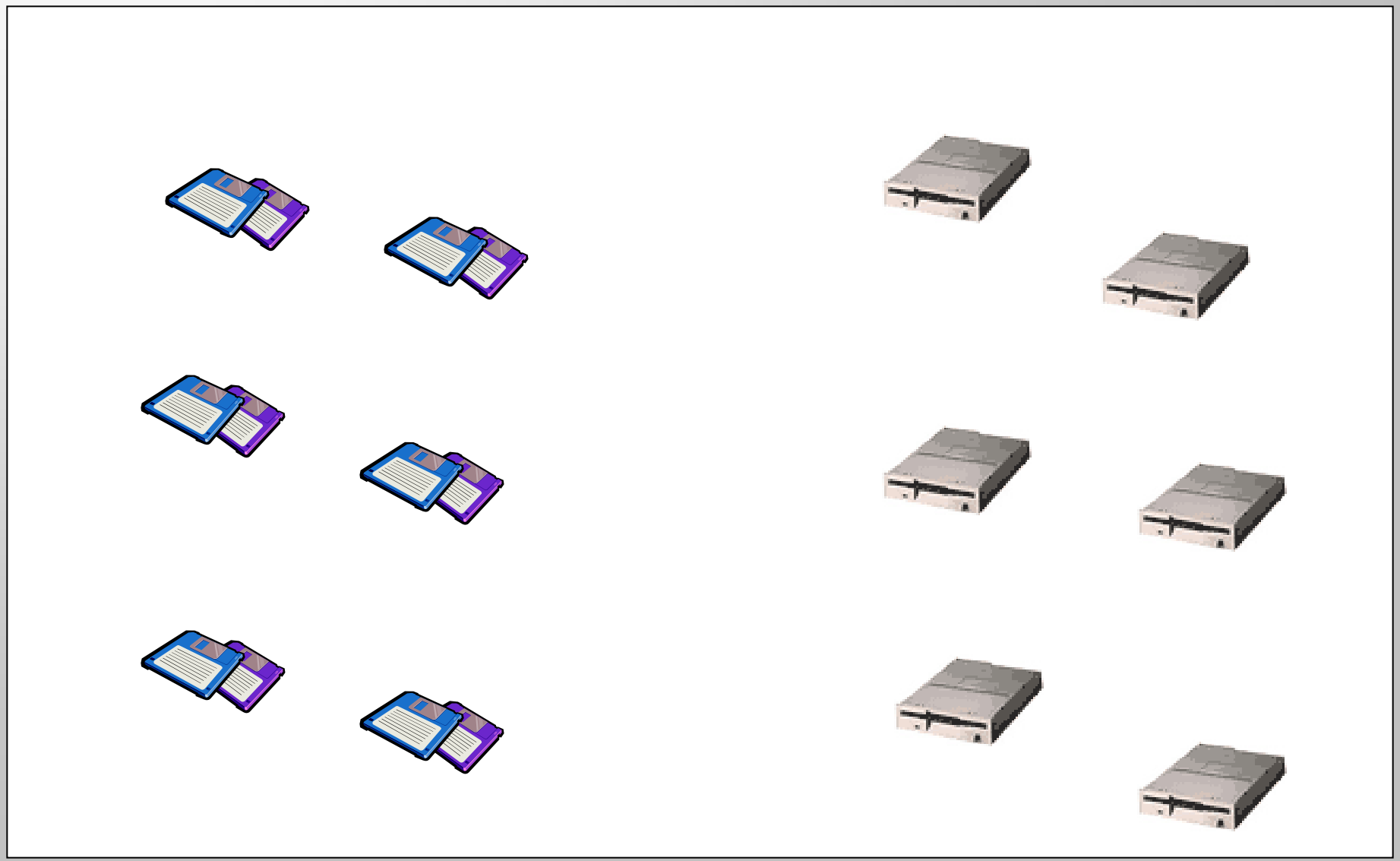
Standards

- Using standards always seems like a good idea, but
- avoiding premature standards is important, and

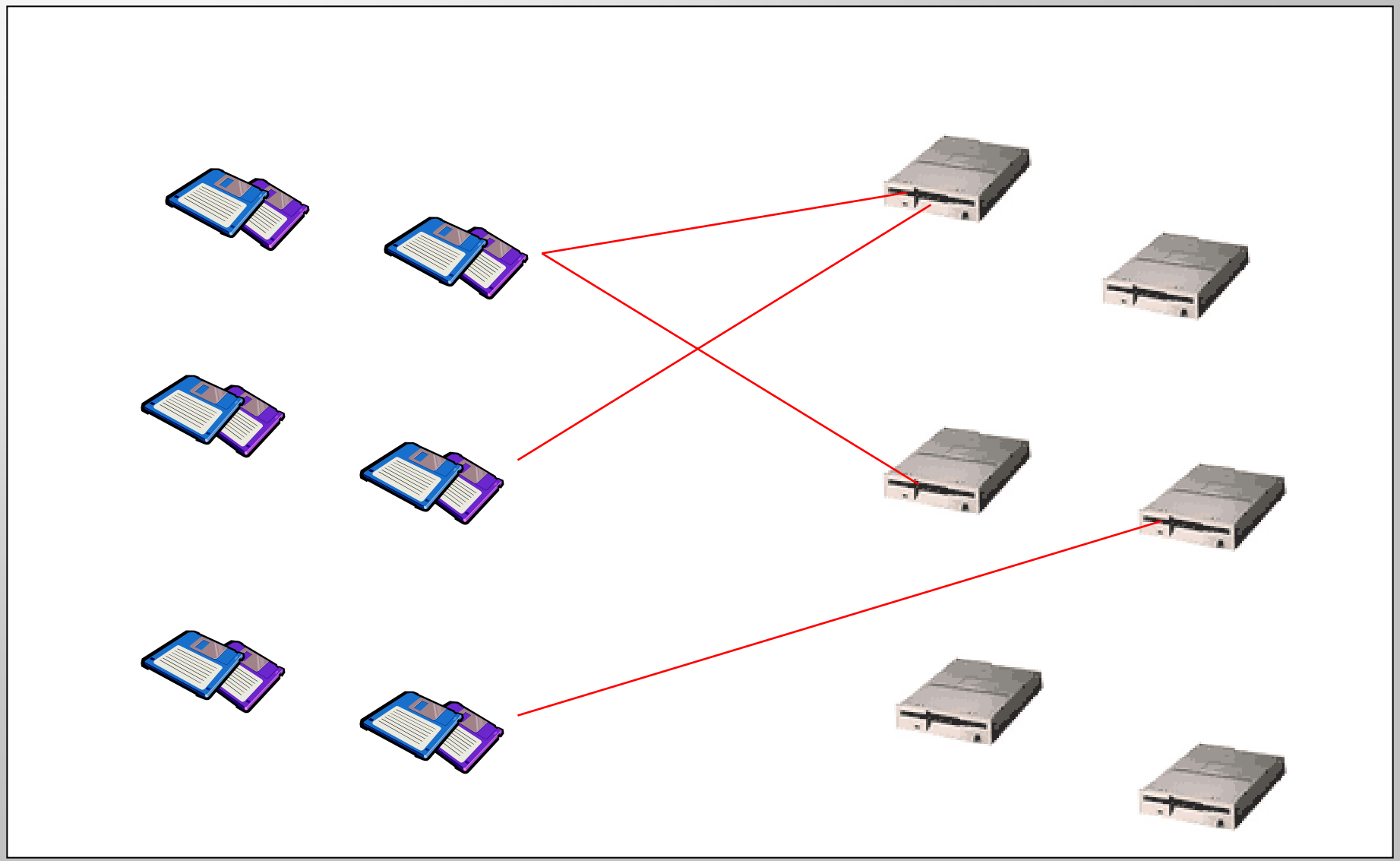
Standards

- Using standards always seems like a good idea, but
- avoiding premature standards is important, and
- adopting bad standards can cripple an IT endeavor, especially one with global ambitions.

Bad Data-exchange Standard

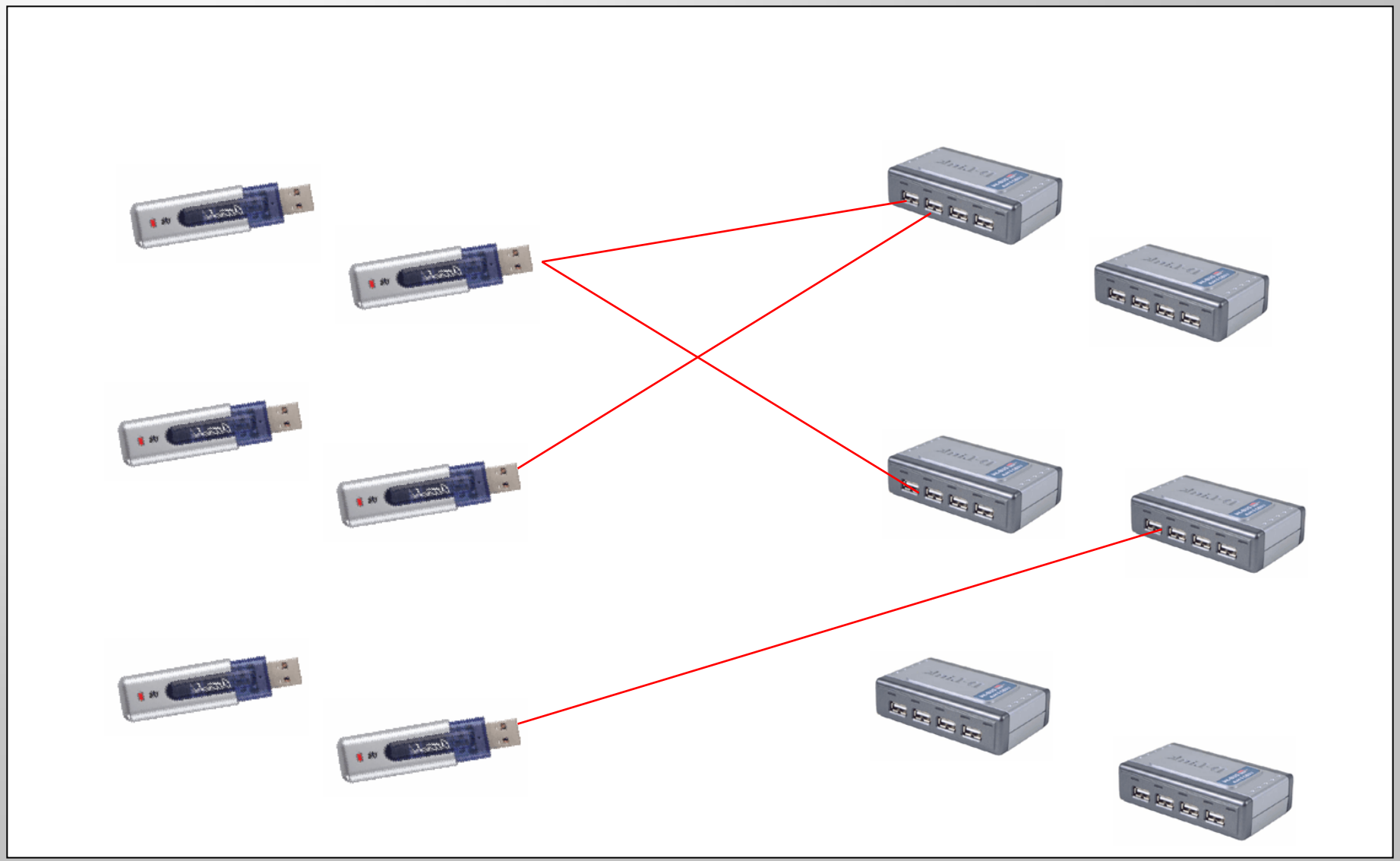


Bad Data-exchange Standard

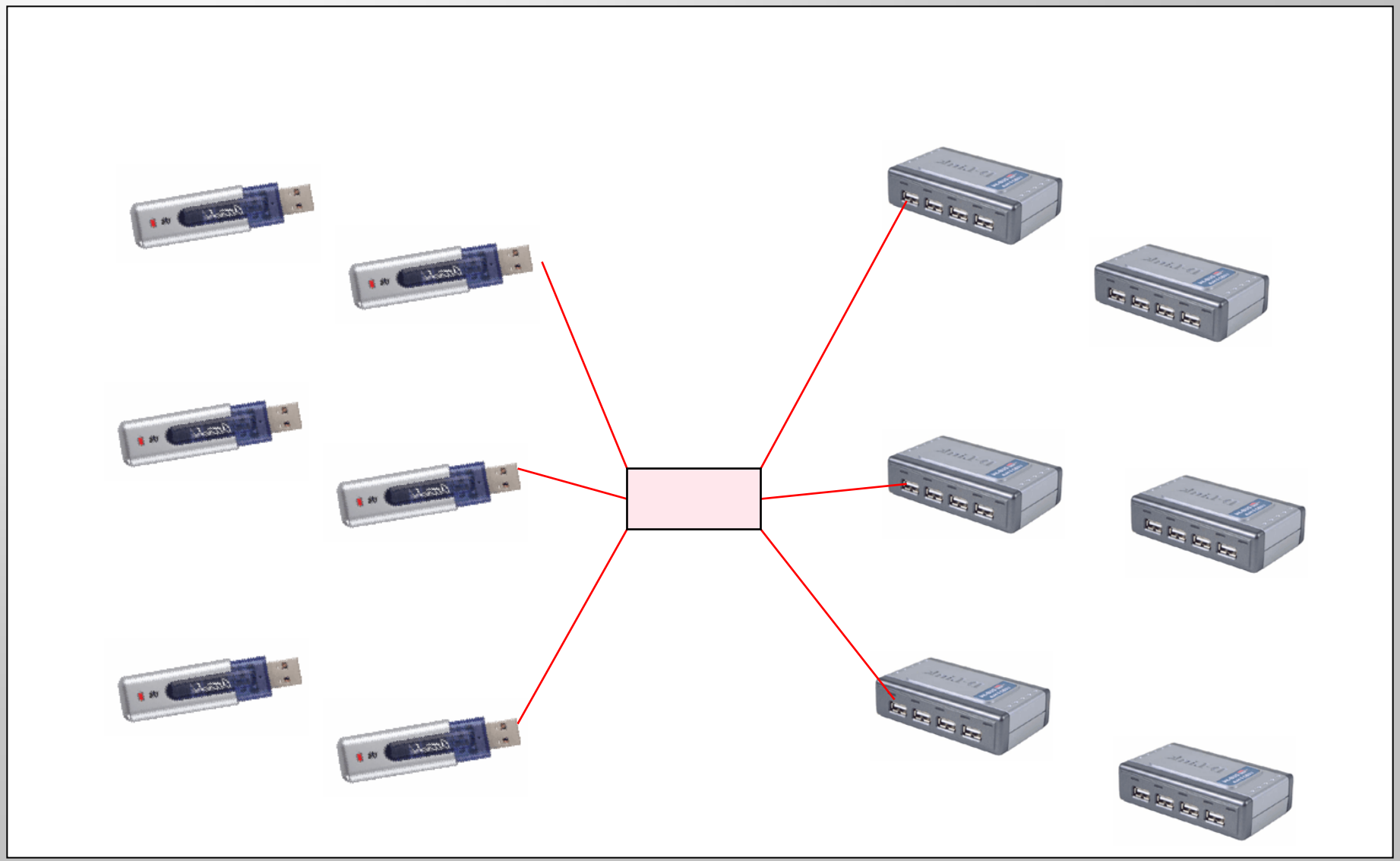


Good Data-exchange Standard

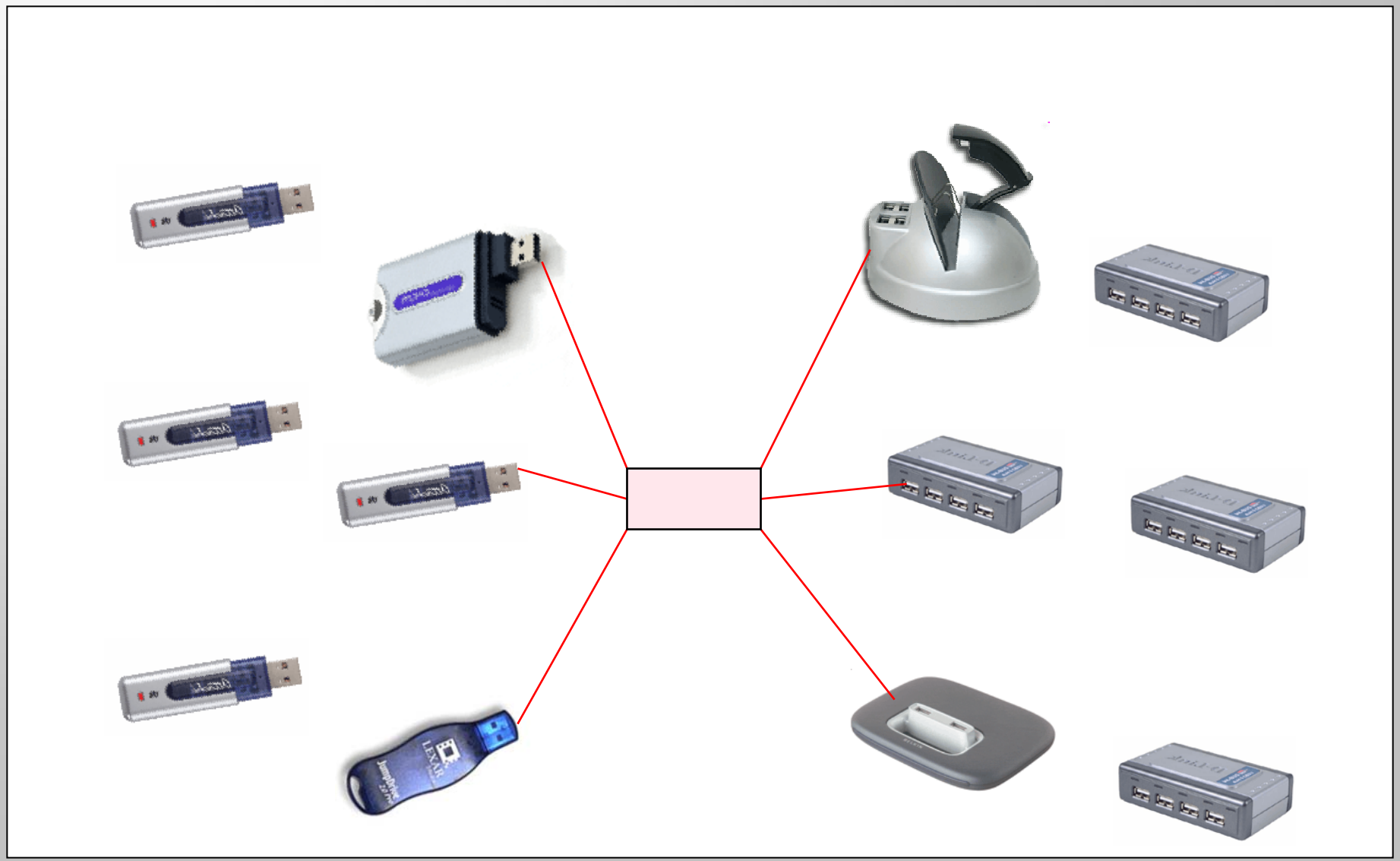
Good Data-exchange Standard



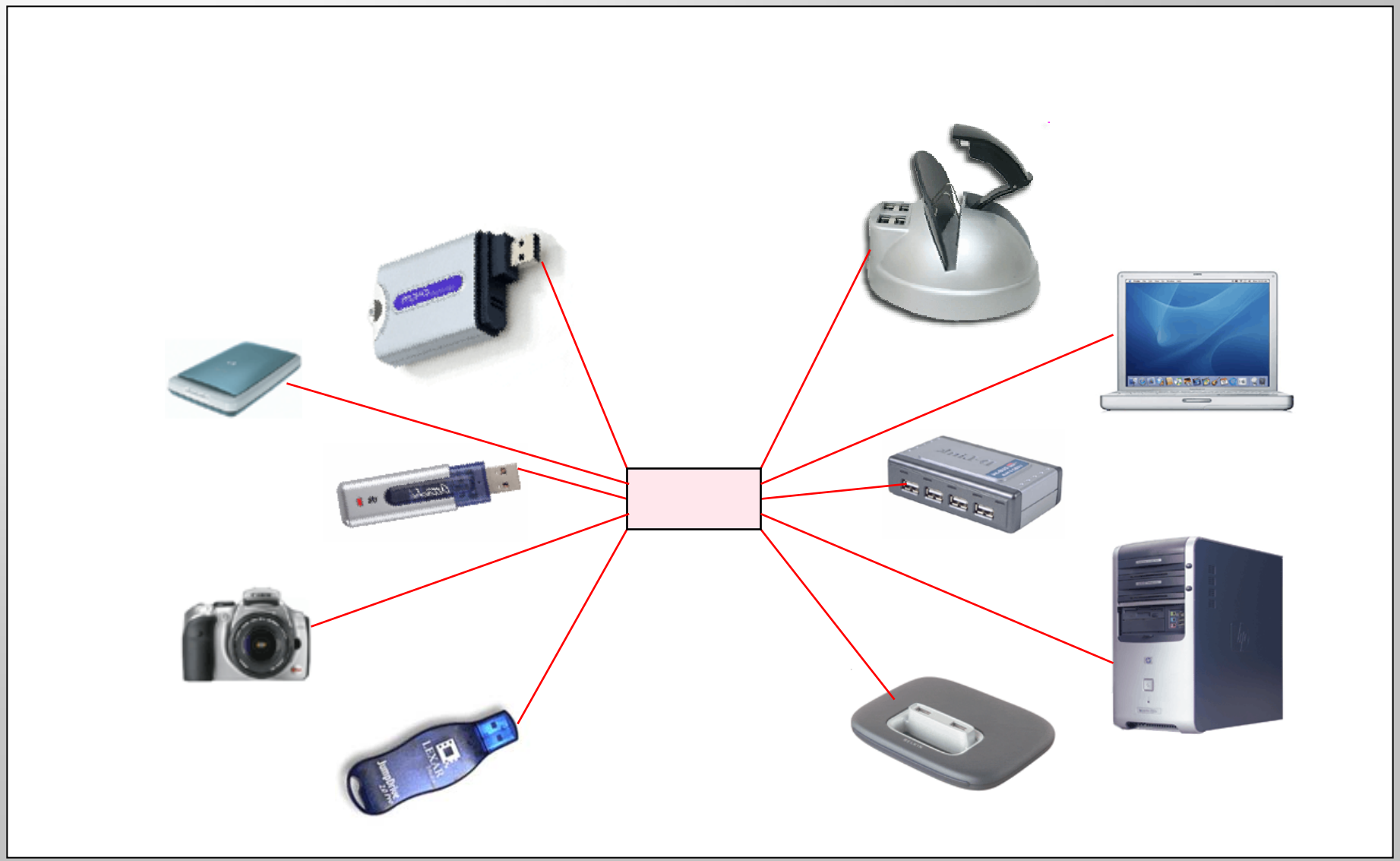
Good Data-exchange Standard



Good Data-exchange Standard



Good Data-exchange Standard



Challenges Due to IT Industry Trends

Industry Trends, I

- The advance of technology is relentless.

Industry Trends, I

- The advance of technology is relentless.
- New technology, new standards, new capabilities are constantly appearing.

Industry Trends, I

- The advance of technology is relentless.
- New technology, new standards, new capabilities are constantly appearing.
- Challenges once thought to be impossible yield to new solutions.

Industry Trends, I

- The advance of technology is relentless.
- New technology, new standards, new capabilities are constantly appearing.
- Challenges once thought to be impossible yield to new solutions.
- Newly developed technologies, like web-services and XML-schema data systems make digital data-sharing systems a real possibility.

Industry Trends, I

But always remember,

In fifteen years, today's technology will seem as hopelessly dim and inadequate as 1990s technology seems today.

To build digital data-sharing systems, we must USE current technology but we must be careful not to DEPEND on that technology.

Industry Trends, II

- As technology matures ease of use become more and more important.

Industry Trends, II

- As technology matures ease of use become more and more important.
- Real user value occurs when technology is engineered away to invisibility.

Industry Trends, II

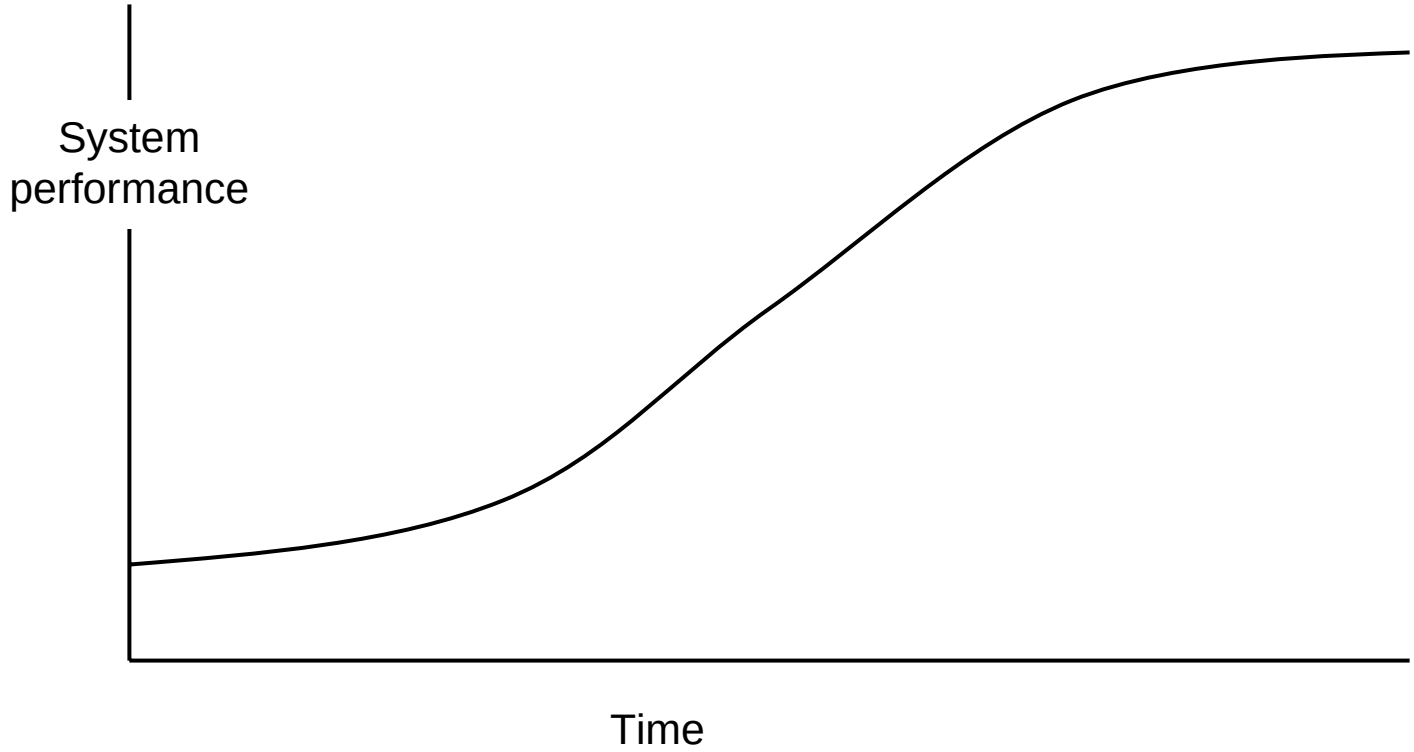
To build truly useful SpeciesBANK systems,

We must appreciate and effectively use advanced technology.

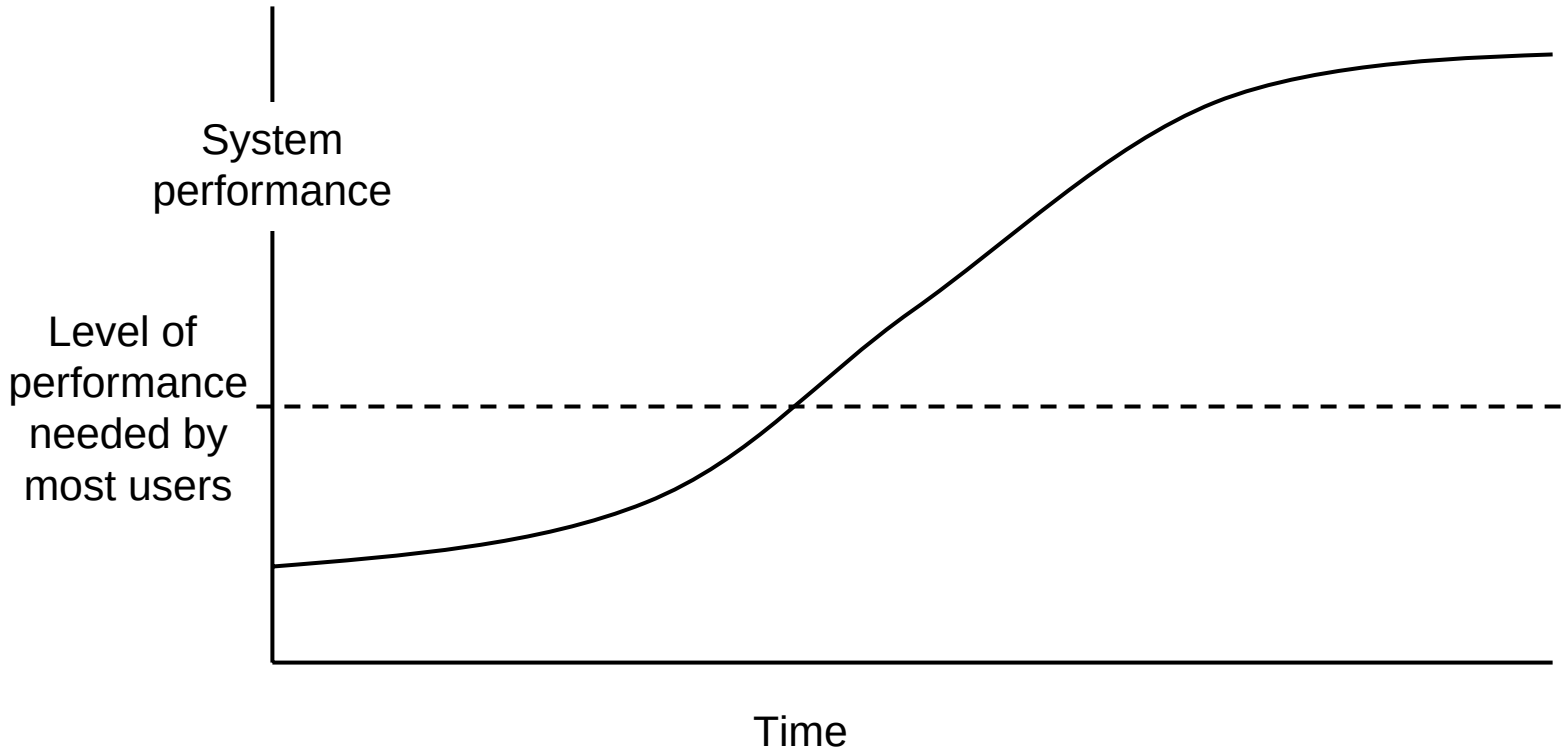
But, we must never allow ourselves to become enamored of that technology.

Our success will depend on our knowledge of the process and practice of science than on our expertise with information technology.

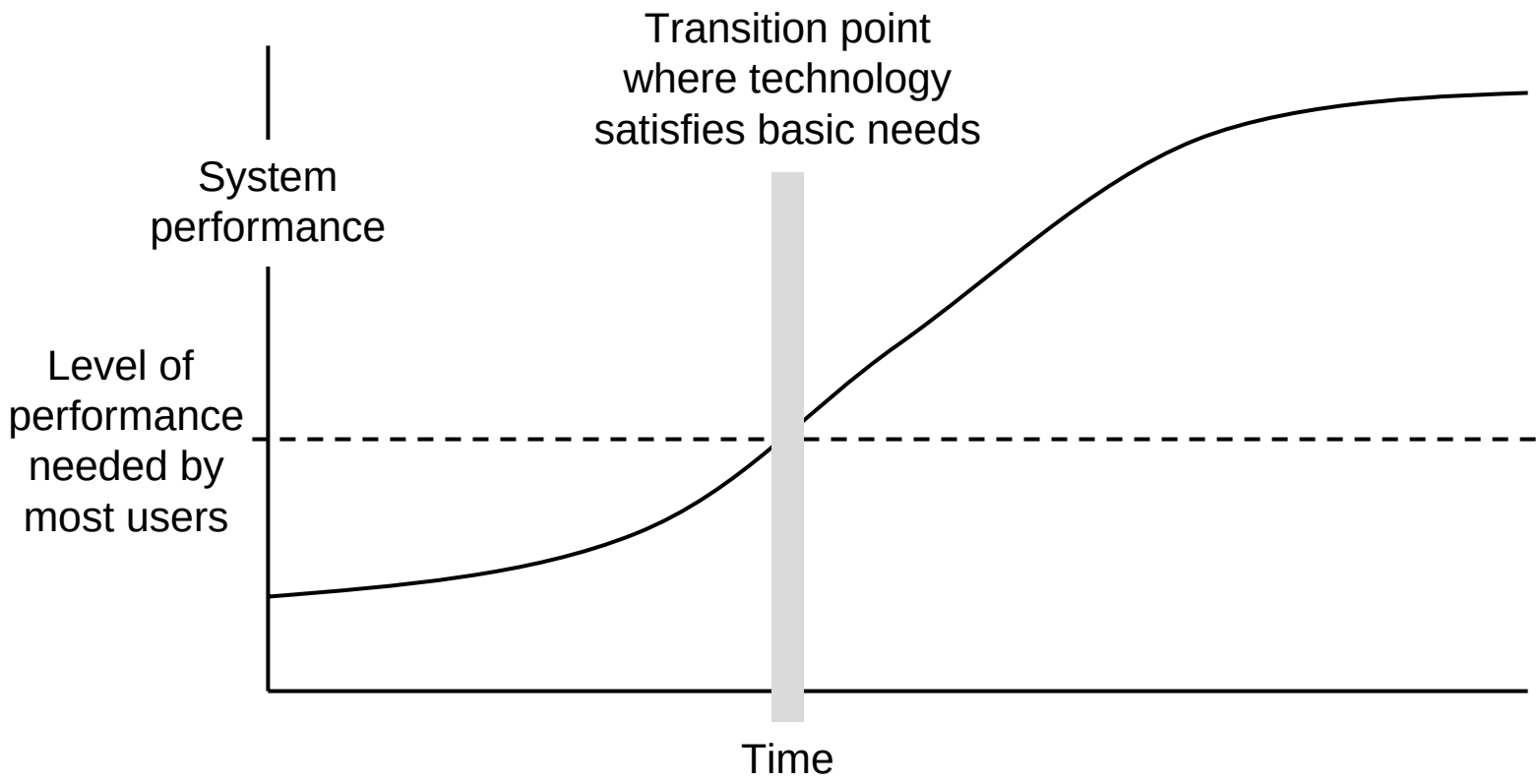
Industry Trends



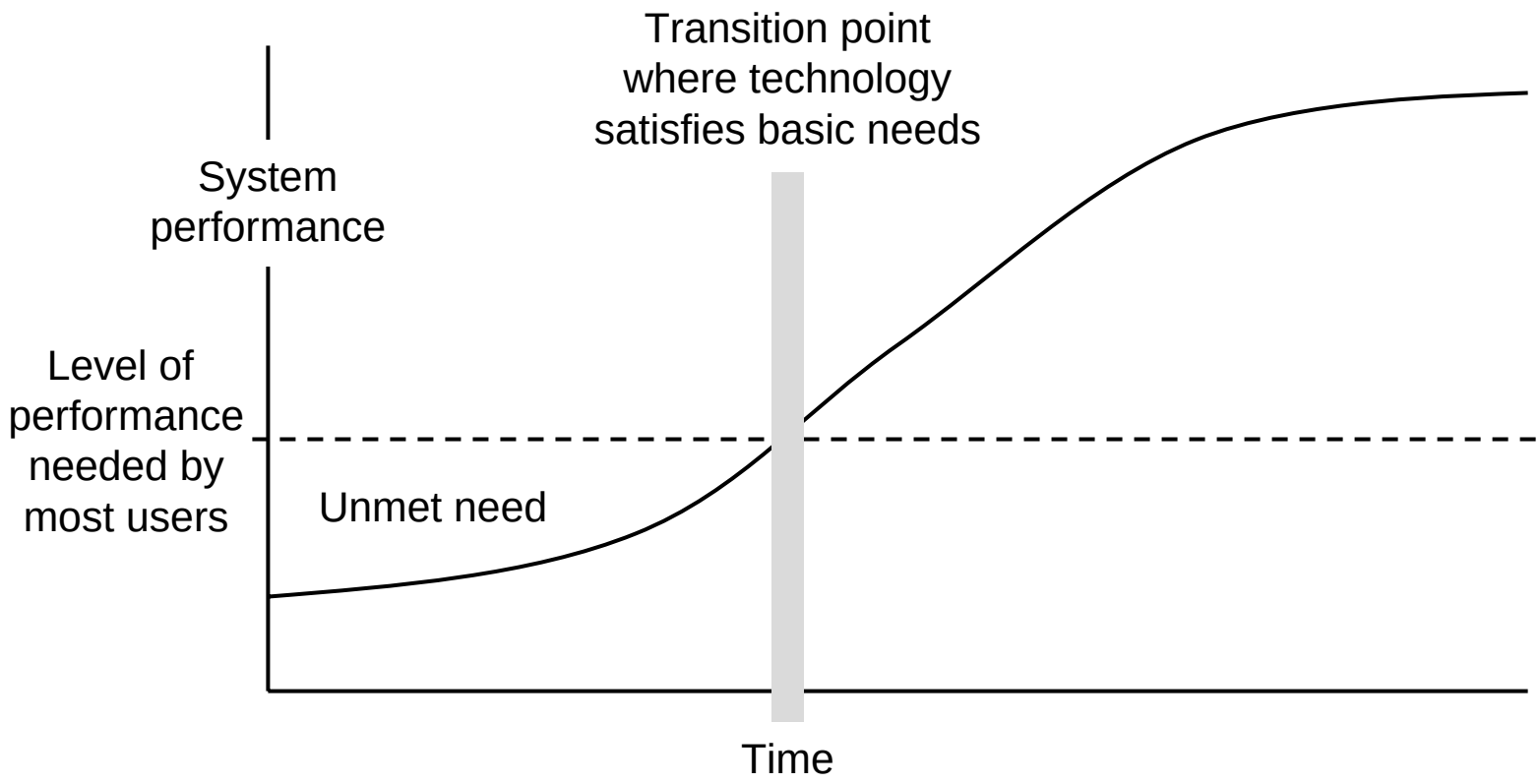
Industry Trends



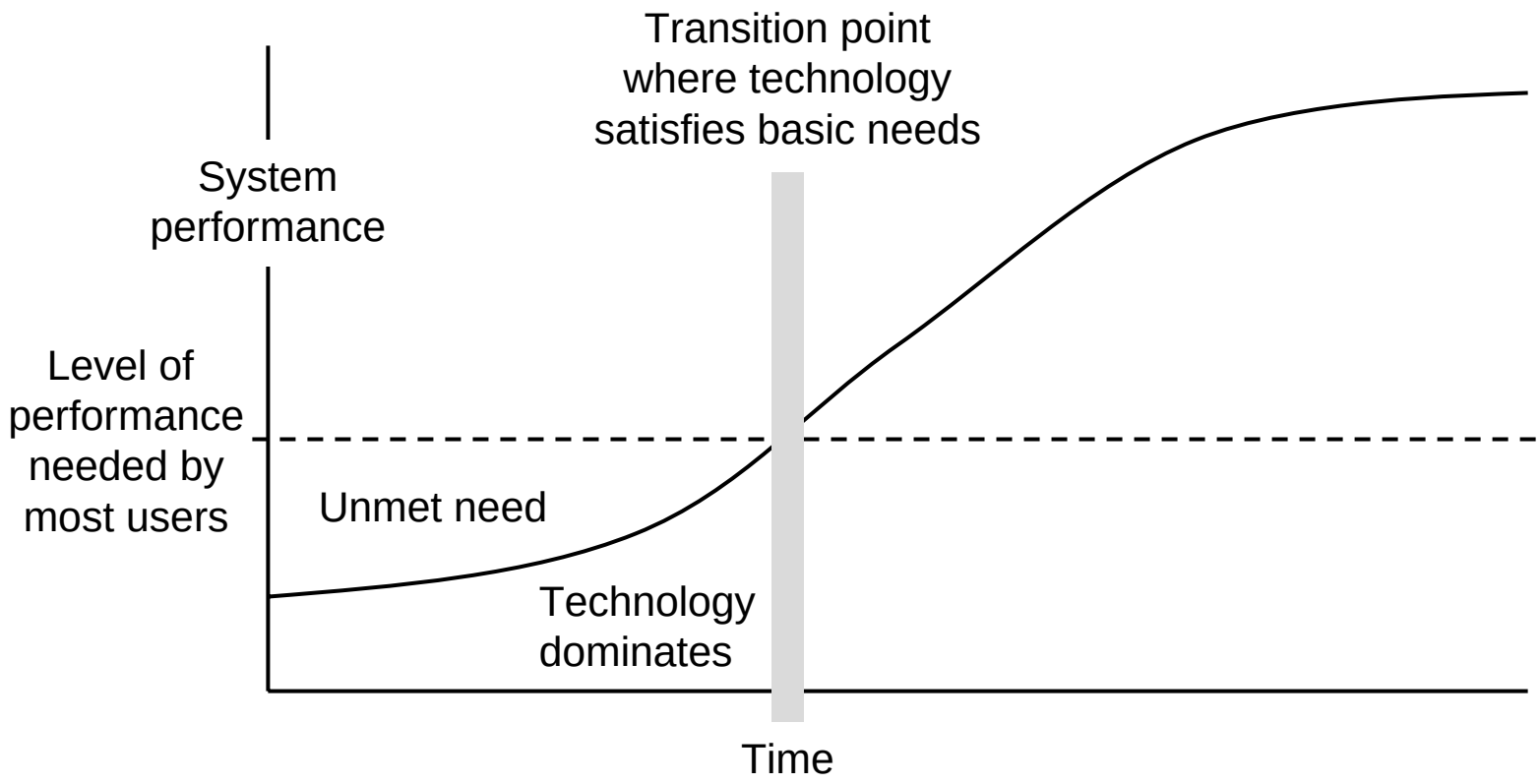
Industry Trends



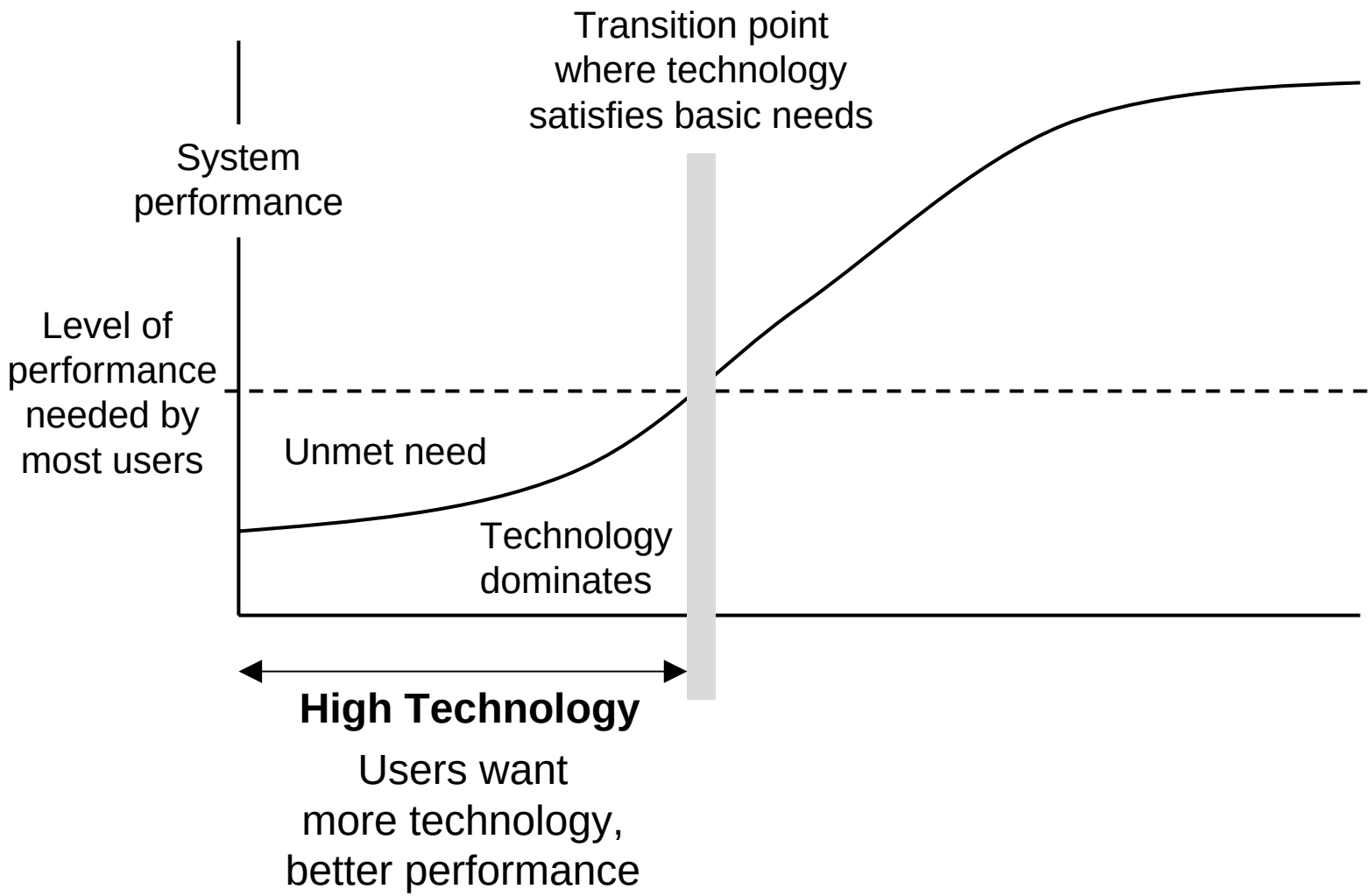
Industry Trends



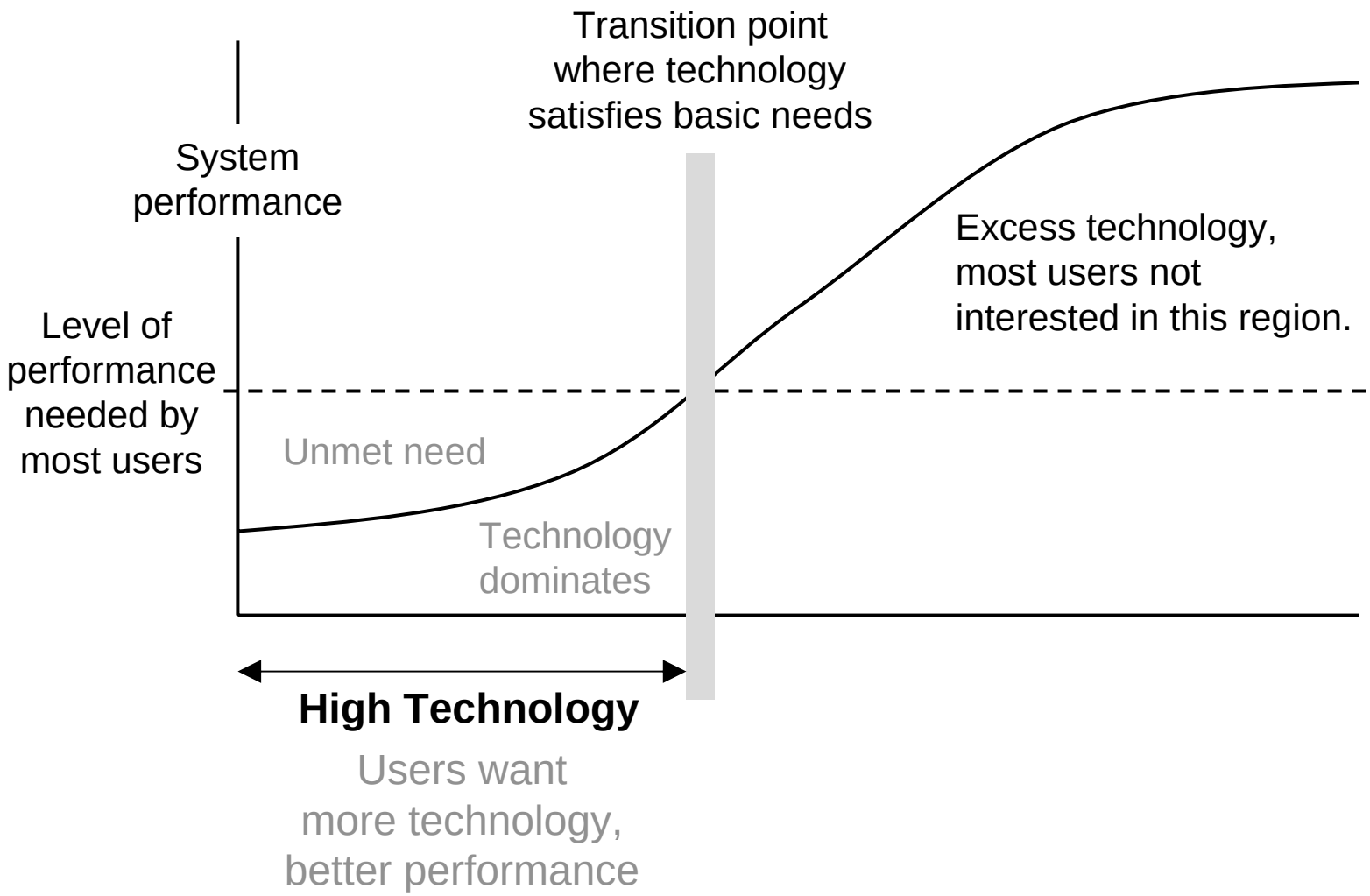
Industry Trends



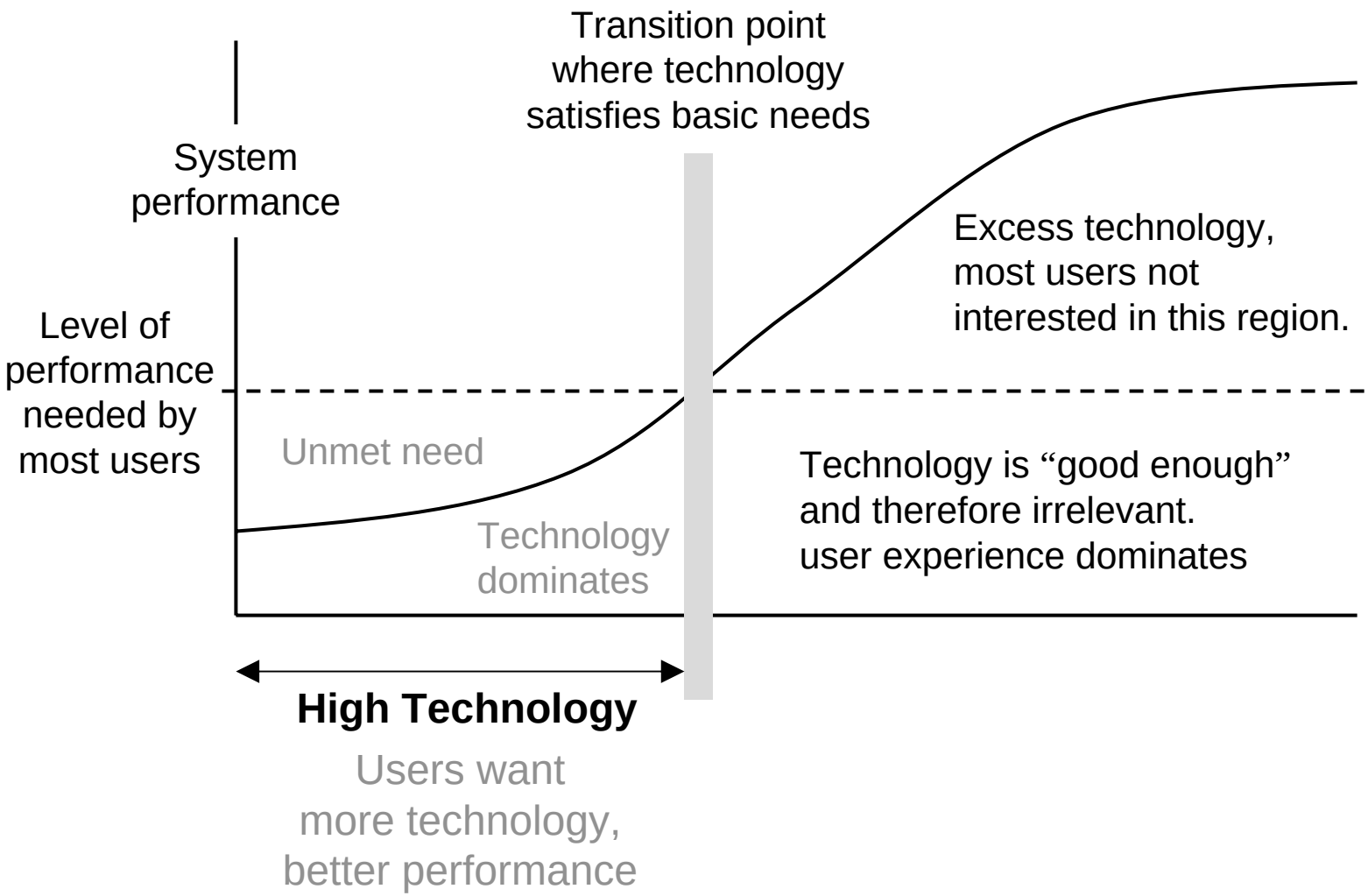
Industry Trends



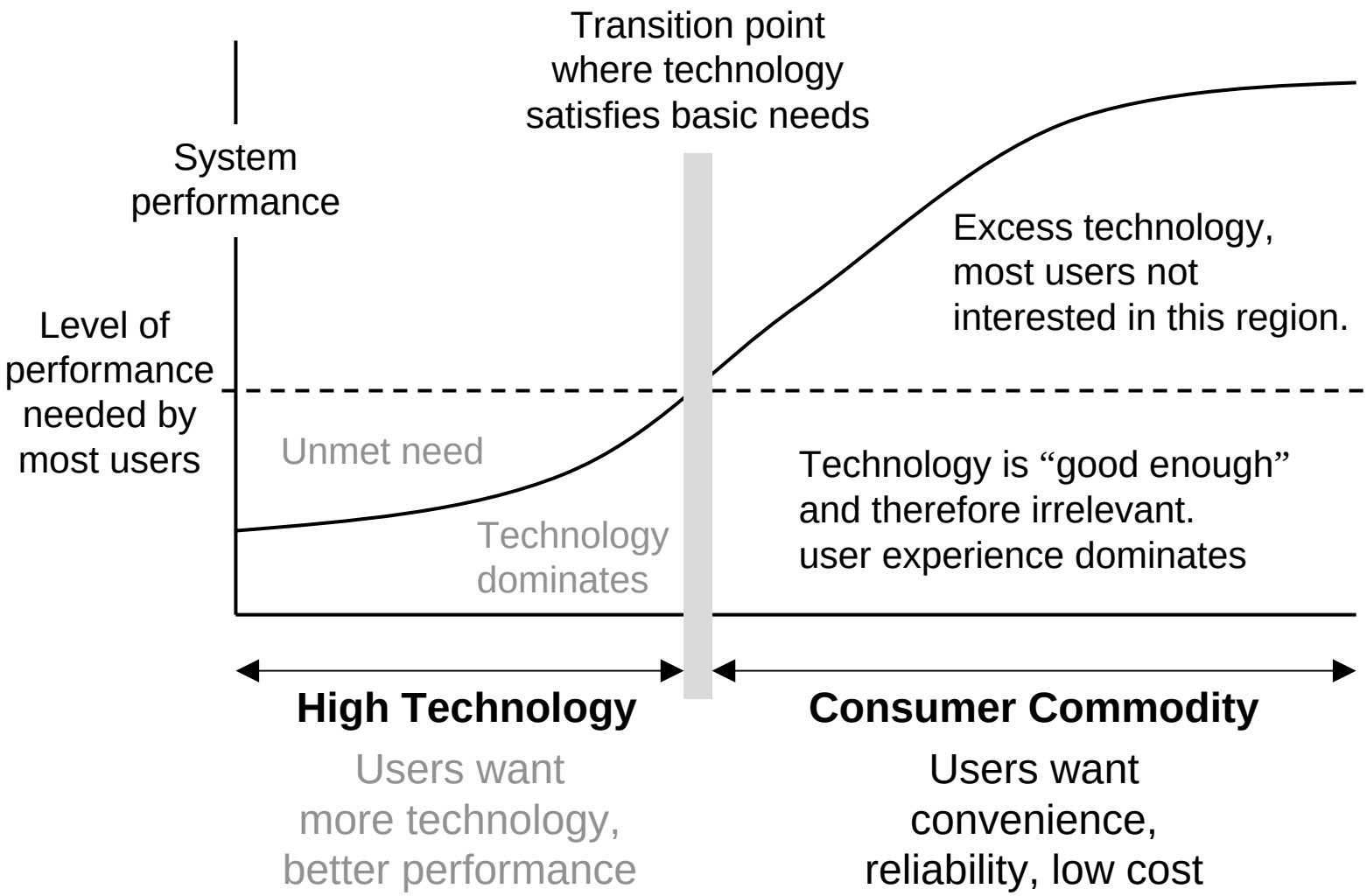
Industry Trends



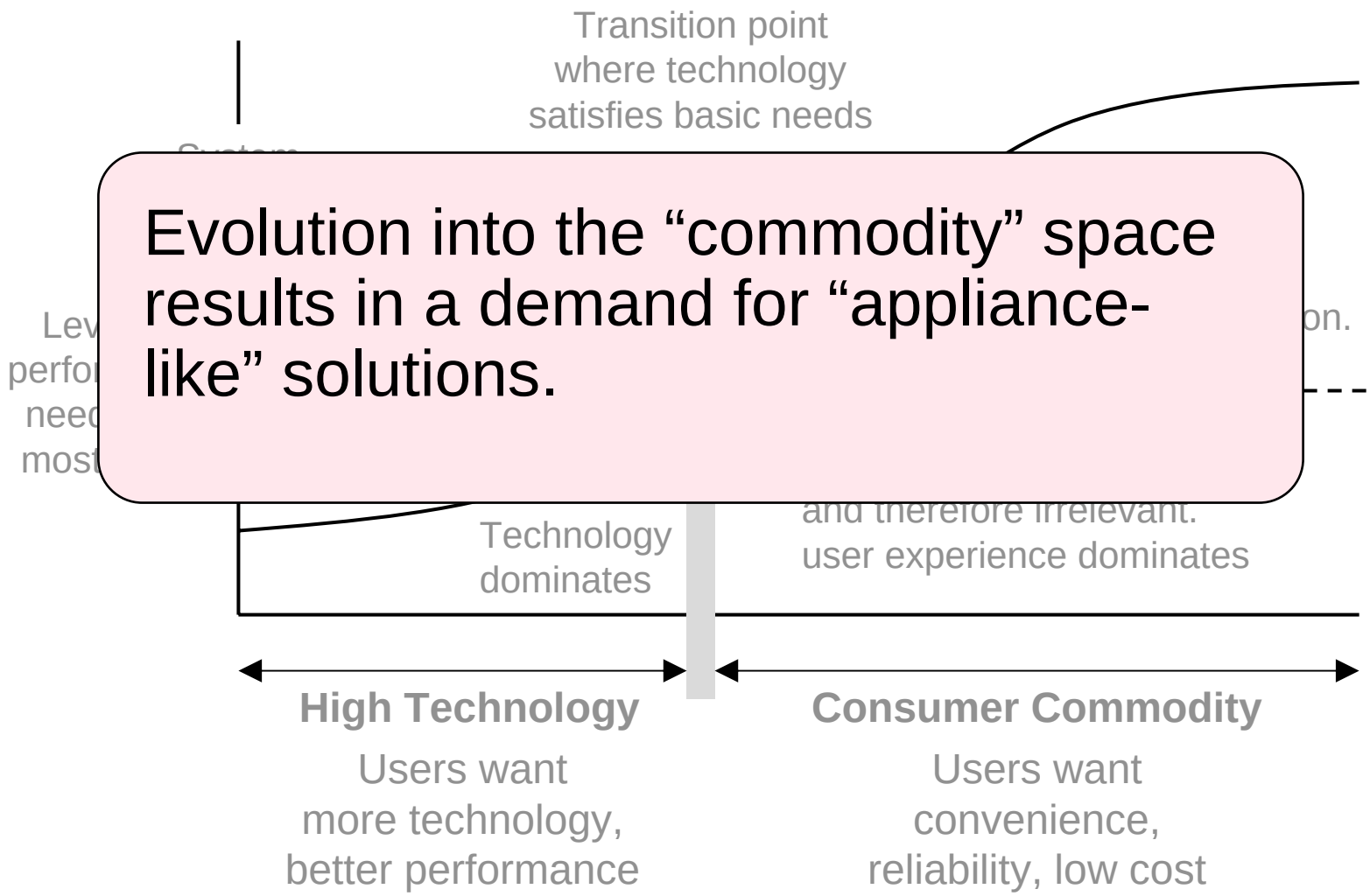
Industry Trends



Industry Trends

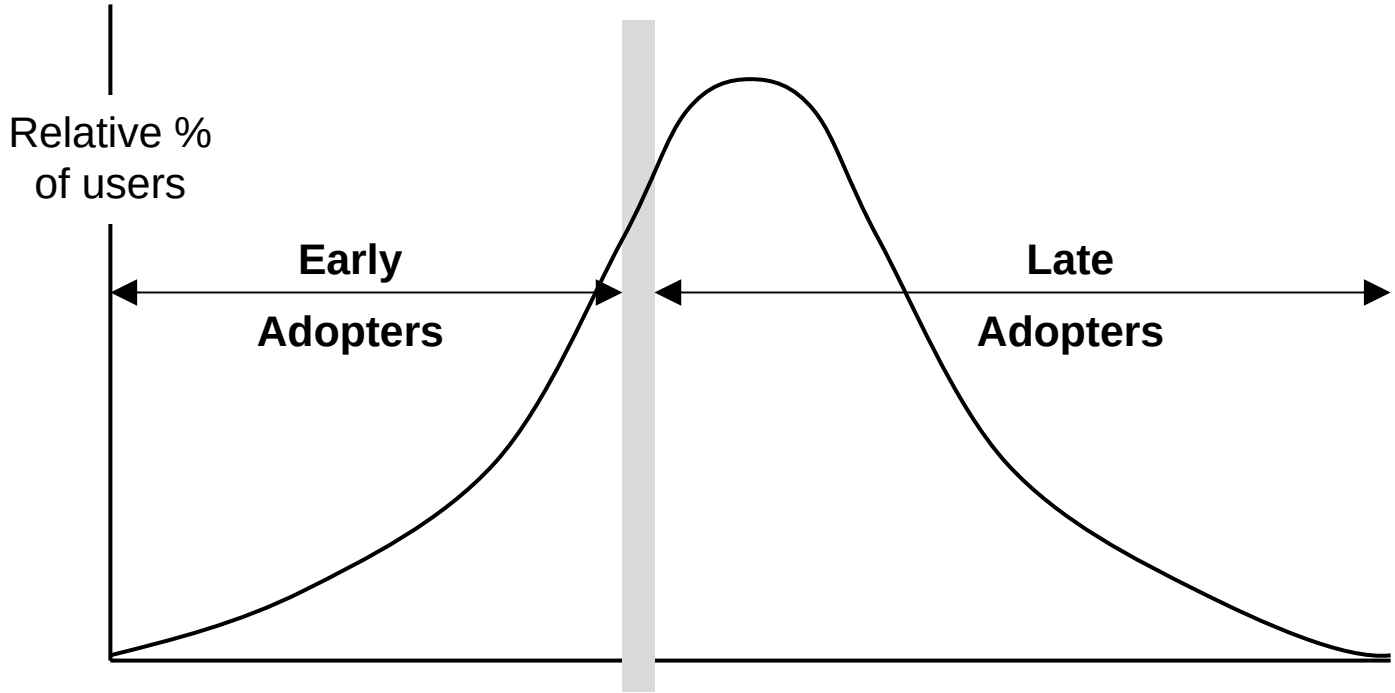


Industry Trends



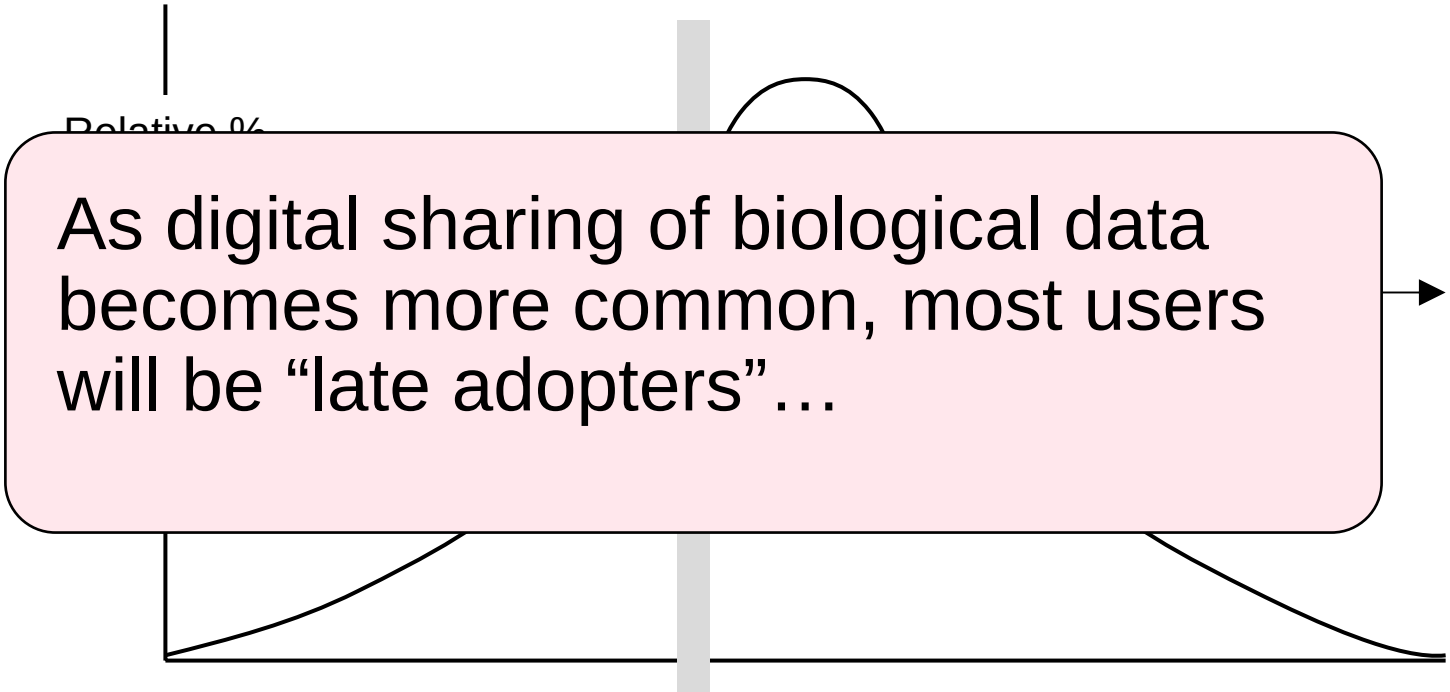
Evolution into the “commodity” space results in a demand for “appliance-like” solutions.

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.

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.

Challenges Due to the Inevitability of Change

Universal Interoperability

- **Hard…**

Logical Simplicity

- In a federated, component-based environment, the biggest challenge is managing complexity.
- This requires a commitment to simplicity.
- Components must be entirely self-contained.
- All inter-component communication occurs only through well defined interfaces.
- Systems must be designed to accommodate change.

Logical Simplicity

- In a federated, component-based environment, the biggest challenge is managing complexity.
- This requires a commitment to simplicity.
- Components must be entirely self-contained.
- All inter-component communication occurs only through well defined interfaces.
- Systems must be designed to accommodate change.

Driving Assumption

- Many use case requirements across the federation will be inconsistent and some will be genuinely contradictory.

Driving Assumption

- Many use case requirements across the federation will be inconsistent and some will be genuinely contradictory.
- The federation must work anyway.

Challenges Due to Limits of Social Scalability

Social Scalability

- In a truly federated environment, long term success for a federated information infrastructure will depend upon social scalability.
- Social scalability CANNOT be achieved through normative pronouncements.
- Experience suggests that social scalability is best achieved through a combination of pure laissez faire individualism and social consequences – i.e., social contracts.

Social Scalability

- In a truly federated environment, long term success for a federated security model will d
- S gh n
- E b laissez faire individualism and social consequences – i.e., social contracts.

Negotiated social contracts – not mandated technical solutions – drive the emergence of standards in a federation.

Social Consequences

- Every individual is free to do whatever he/she chooses.

Social Consequences

- Every individual is free to do whatever he/she chooses.
- Every other individual is free to respond however he/she chooses.

Social Consequences

- Every individual is free to do whatever he/she chooses.
- Every other individual is free to respond however he/she chooses.
- Interactive relationships then sort things out.

Social Consequences

- Every individual is free to do whatever he/she chooses.
- Every other individual is free to respond however he/she chooses.
- Interactive relationships then sort things out.
- Examples:

  One cuts, the other chooses.

Social Consequences

- Every individual is free to do whatever he/she chooses.
- Every other individual is free to respond however he/she chooses.
- Interactive relationships then sort things out.
- Examples:

  I am free to suppress my caller ID; if I do, you are free to refuse to answer my calls.

Social Consequences

- Every individual is free to do whatever he/she chooses.
- Every other individual is free to respond however he/she chooses.
- Interactive relationships then sort things out.
- Examples:

  You are free to run your systems in as stupid and incoherent manner as you choose; if you do, I am free to refuse to have anything to do with your systems.

Logical Issues

- Rules governing behavior can be permissions or prohibitions.

Logical Issues

- Rules governing behavior can be permissions or prohibitions.
- The union set of contradictory permissions is a very flexible environment.

Logical Issues

- Rules governing behavior can be permissions or prohibitions.
- The union set of contradictory permissions is a very flexible environment.
- The union set of contradictory prohibitions is the null set.

Logical Issues

- Rules governing behavior can be permissions or prohibitions.
- The union set of contradictory permissions is a very flexible environment.
- The union set of contradictory prohibitions is the null set.
- Use case requirements across a federation will be contradictory.

Logical Issues

- Rules governing behavior can be permissions or pr
- T
- T
- U

If a federated information infrastructure is to deliver services greater than the null set, it must be technically implemented on the aggregation of permissions, not prohibitions.

Behavioral constraints should be achieved on a virtual organization basis, through negotiated social contracts.

Logical Issues

- Rules governing behavior can be permissions or pr
- T
- T
- U

For example, the components of a federated information system should make it easy for users to behave according to common standards, but it should not mandate that they do so.

Social Scalability: Required Reading

James Madison
Alexander Hamilton
John Jay

The Federalist Papers

Social Scalability: Required Reading

James Madison
Alexander Hamilton
John Jay

The Federalist Papers

There is no better source of ideas on how to build systems that work in a decentralized social environment.

Remember, you can't change human nature, so you must design systems that work **despite** human nature.

Social Scalability: Required Reading

Alexander Hamilton
Jame
John

THEOREM:

When there is no authority to **compel** participation in standard systems, then one must **entice** participation in standard systems.

Social Scalability: Required Reading

Alexander Hamilton

Jame

John

OUR TASK:

To devise an infrastructure for effective and enticing data-sharing systems with semantic-web-like properties that will work despite all of the challenges we have considered.

END