Institutional Standards: The Critical Missing Piece

http://www.esp.org/rjr/nist2003.pdf

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

rrobbins@fhcrc.org
(206) 667 2920

Institutional Standards: The Critical Missing Piece

http://www.esp.org/rjr/nist2003.pdf

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

rrobbins@fhcrc.org
(206) 667 2920

Institutional Standards: The Critical Missing Piece

http://www.esp.org/rjr/nist2003.pdf

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

rrobbins@fhcrc.org
(206) 667 2920

Abstract

In 1990, an NSF Invitational Workshop on Scientific Database Management brought together database experts and domain scientists to consider and document the challenges of scientific database management. Nearly fifteen years later, many of those challenges are still unmet. The problem is especially acute in biomedical research, where genome-project-driven technologies have unleashed a flood of data into a community (or rather a set of communities) with major sociological and structural impediments to effective large-scale data management.

Unlike "big-instrument, single-data-source" science (e.g., high-energy physics), most public-sector biomedical research occurs as "small-instrument, multi-data-source" science in small, investigator-initiated projects at universities or independent research organizations. Multi-source data from these smaller projects then, ideally, flow together into larger national or international resources (e.g., GenBank). The GenBank model, however, is only applicable to normal or paradigmatic science in the Kuhnian sense. In pre-paradigm fields or to fields undergoing paradigm-shifts, efforts to apply the GenBank model (by proposing national data standards and repositories) will fail. Even in some normal science fields (e.g., functional tomography of the brain), efforts to apply the GenBank model will experience difficulties because of the limitations of the current scientific publishing model (e.g., total transfer of copyright to the journal publisher).

In "small-instrument, multi-data-source" science budgets are small and the allocation for local data management is usually inadequate. Resources for extending the local project to include support for participation in a national data repository are usually non-existent.

As noted in the NSF Workshop, the relational data model is an inadequate abstraction for representing many kinds of biological data (e.g., pedigrees, taxonomies, maps, metabolic networks, food chains). Efforts to deploy object-oriented DBMS have not met with widespread success. Compromise efforts to force complex biological data structures into relational models have resulted in locally effective kludges that do not admit ready integration into larger data collections. The effective use of taxonomies in bio databases quickly results in the need for tri-state logic, something not easily implemented with commercial RDBMS.

National efforts to close many of these gaps in effective bio data management will founder on problems of scale. How will the development of a national data standard help a small-RO1 PI who can barely afford any information infrastructure, much less generic systems that interoperate well with large communities? How can an individual researcher hope to address problems resulting from the current science publishing model? What systems are readily available to help a researcher comply with government requirements to share data while also complying with other government requirements to protect human-subjects privacy? The answer to these, and other challenges, will lie in the development of institutional standards for IT support of grant-funded research. These institutional standards are indeed the critical missing piece.

Meeting Overview

Goal:

To identify opportunities for information science (IS) standards and standards development to facilitate bioscience and biomedical research.

Meeting Overview

Purposes:

To define the current and emerging state of information science (IS) standards related to bioscience and biomedical research, and

To identify barriers and gaps to, and opportunities and pathways for, IS standards development and implementation to enhance bioscience and biomedical research.

Meeting Overview

Scope:

- Biomedical Data Integration Standards

  (e.g., ontology, data format, nomenclature)

- Networked Science

  (e.g., IS standards to harness teragrid-scale computing)

- Quantitative Computational Biology

  (e.g., standards required to improve today's environment for quantitative computational biology, especially modelling of complex systems)

Meeting Overview

Scope:

What barriers and gaps might prevent us from achieving these goals?

(e.g., standards required to improve today's environment for quantitative computational biology, especially modelling of complex systems)

Caution from the Present

Resource Inadequacy:

- Current government spending on bio-medical information infrastructure is far too low to achieve the solutions many have envisioned.

Caution from the Past

Scientific Database Management

Final Report

edited by

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

Report of the Invitational NSF Workshop on

Scientific Database Management

12–13 March 1990

Charlottesville, Virginia

Anita K. Jones, Chairperson

Technical Report 90-21

August 1990

Caution from the Past

U Va Tech Reports:

- CS-90-21

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

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

- CS-90-22

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

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

Caution from the Past

Two major conclusions:

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

Caution from the Past

Two major conclusions:

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

Topics

Problems:

- Resource-adequacy problems
- Database Problems
- Data-source Problems

Solutions:

- More Resources
- Better Database Products
- Institutional Support for Biomedical IT

Resource Problems

Topics

- Resource-adequacy problems

  Current levels of government spending are woefully inadequate to meet the needs of public-sector biomedical research.

Rhetorical Question

Which is likely to be more complex:

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

Rhetorical Question

Which is likely to be more complex:

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

Business Factoids

Five years ago, United Parcel Service:

- used redundant multi-terabyte databases to track all packages in transit
- had 4,000 full-time employees dedicated to IT
- spent one billion dollars per year on IT
- had an income of 1.1 billion dollars, against revenues of 22.4 billion dollars

Business Comparisons

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

Federal Funding of Biomedical-IT

Appropriate funding level:

- approx. 5-15% of research funding
- *i.e.*, 1 - 3 **billion** dollars per year

Federal Funding of Biomedical-IT

Appropriate funding level:

- approx. 5-15% of research funding
- *i.e.*, 1 - 3 **billion** dollars per year

Source of estimate:

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

Federal Funding of Biomedical-IT

A

Warning:

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

Source of estimate:

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

Resource Solutions

Federal Funding of Biomedical-IT

Solutions might occur at many levels:

- Industry partnerships?
- Agency initiatives, like BISTI or caBIG?
- Agency infrastructure support, like CCSGs?
- Leverage investments by working at the INSTITUTIONAL level (e.g., caBIG)

Database Problems

Topics

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

Database I

Basics

Relational Databases

Business Databases:

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

Relational Databases

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

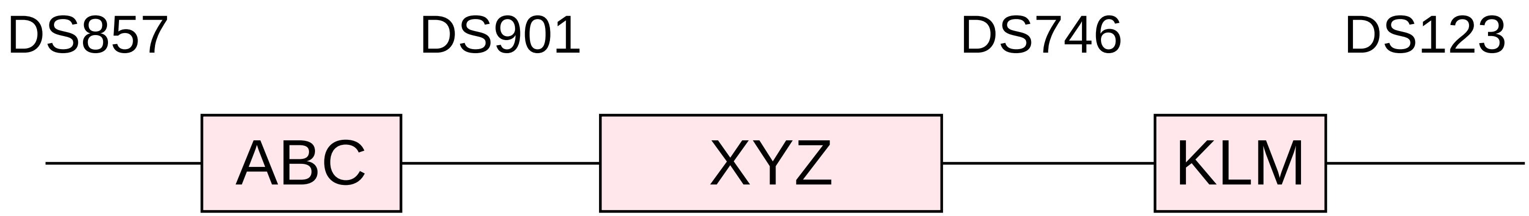
Relational Databases

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

Relational Databases

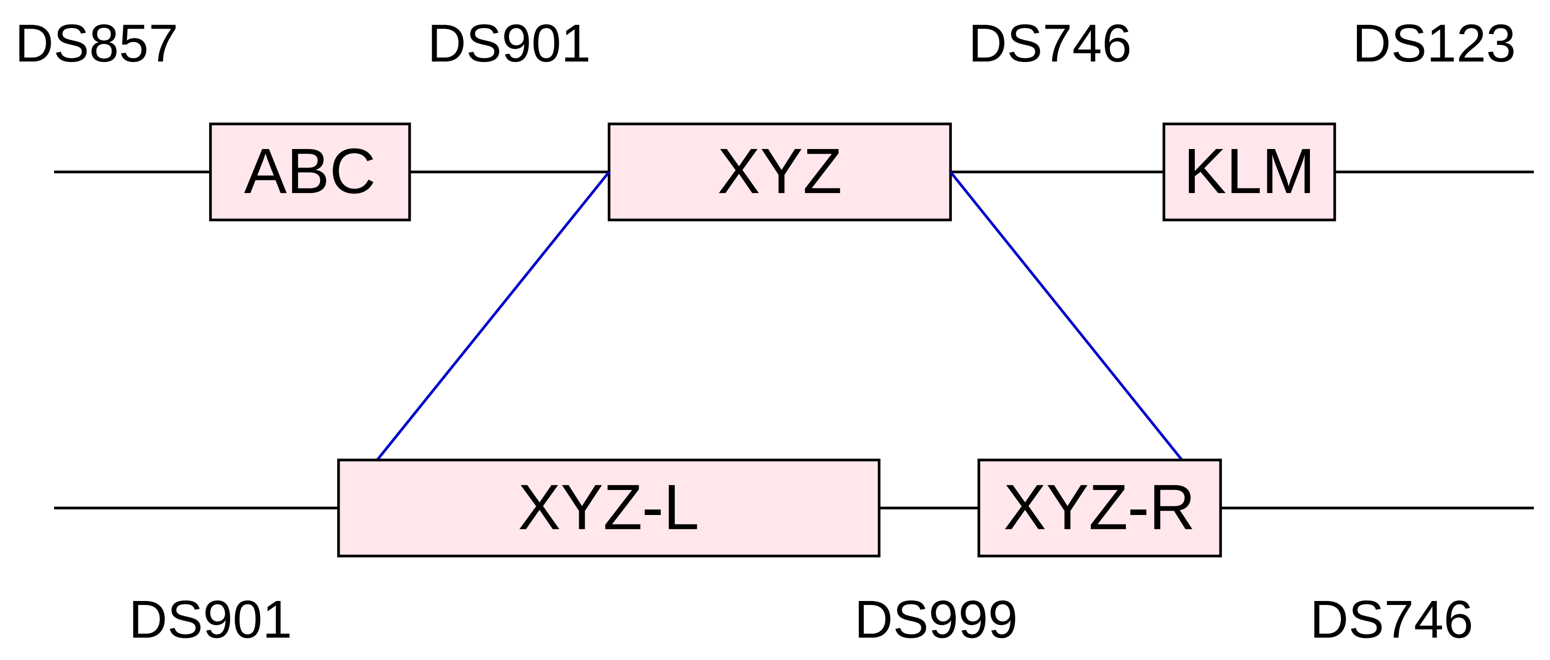
Real Objects:	Hypothetical Objects:
• CONCRETE • STABLE (or known instability) • IMMUTABLE (more or less)	• INSUBSTANTIAL • UNSTABLE • HIGHLY MUTABLE (lumping and splitting)

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

Data Models

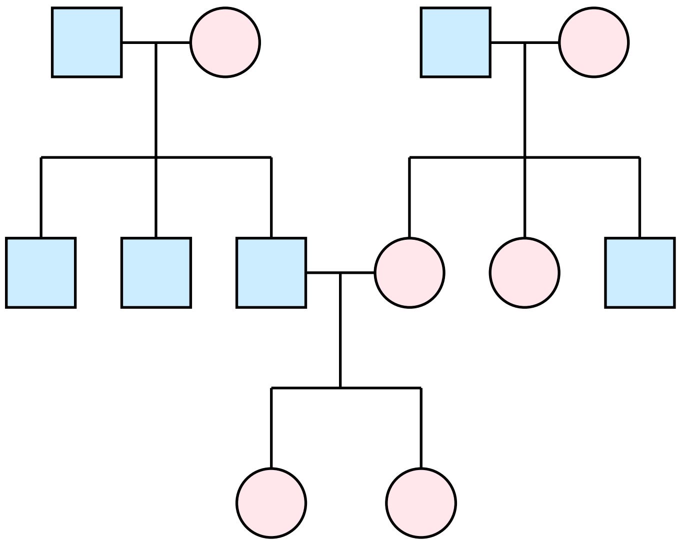
Data-model Challenges

Many bio-data problems involve:

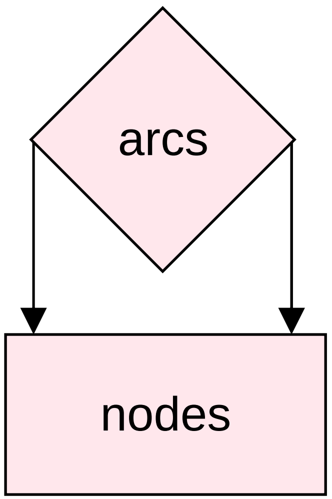
- Graphs: pedigrees, taxonomies, partial orderings, etc…
- Repeat time series observations, with inconsistent results
- Provisional conclusions
- Universal linking tables

Graph Challenges

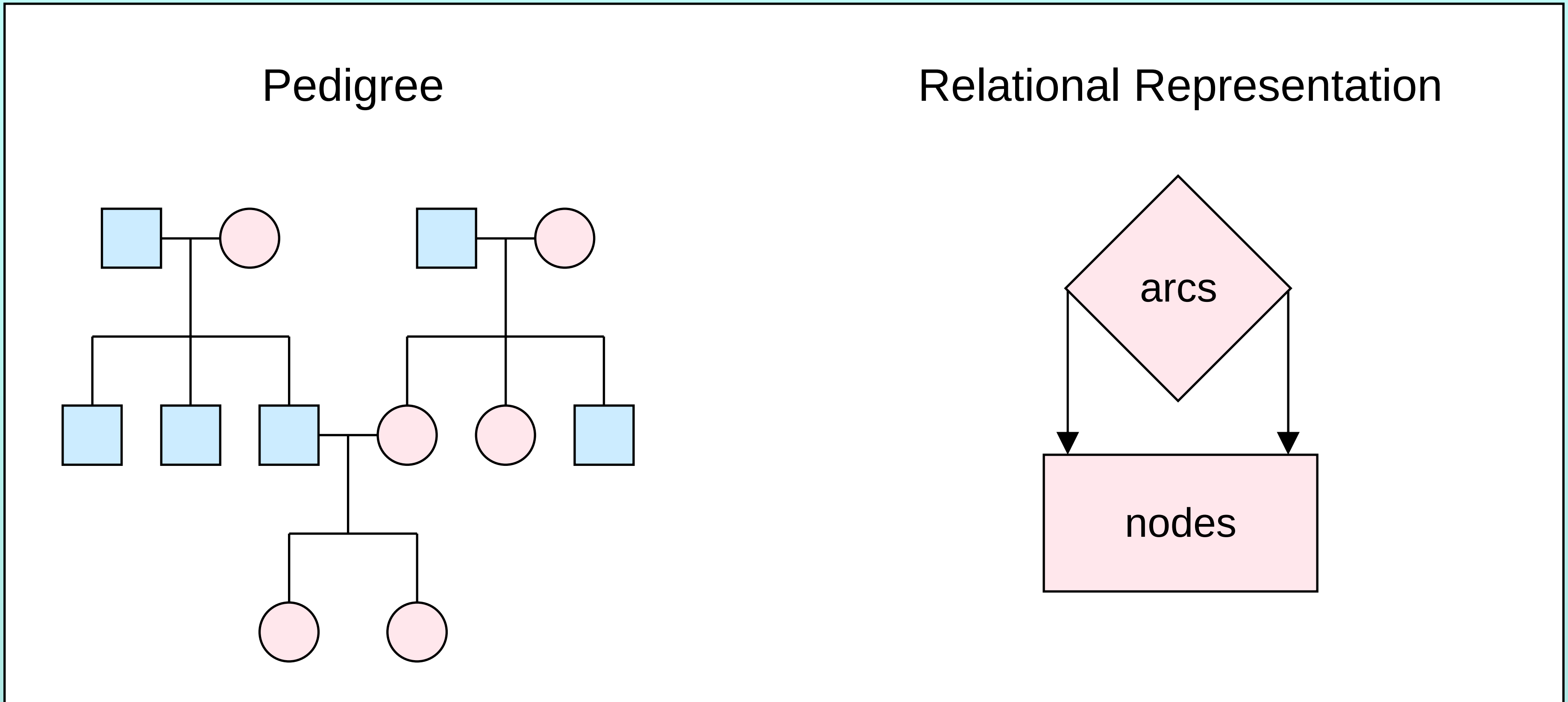
Pedigree



Relational Representation

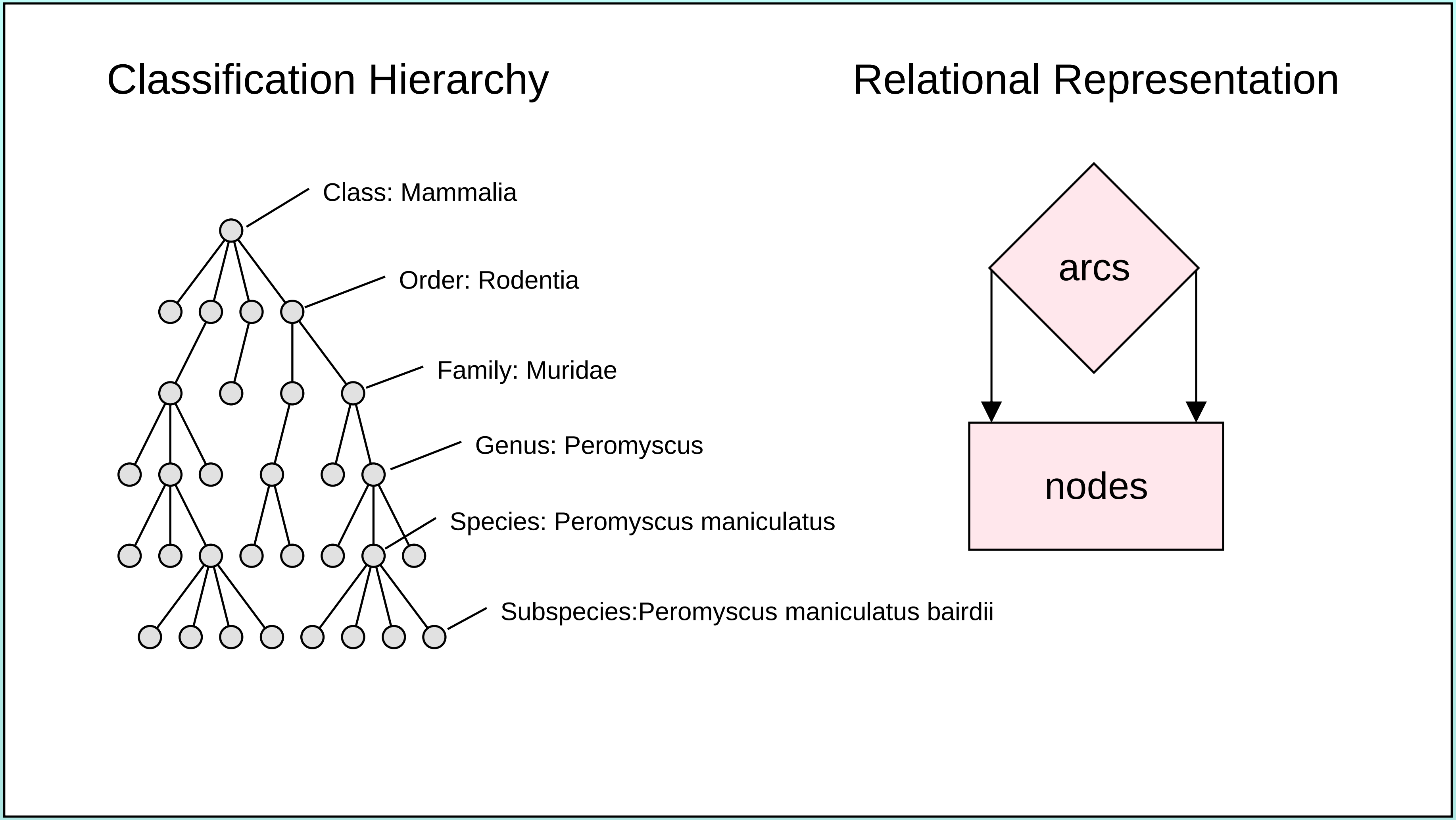


Graph Challenges



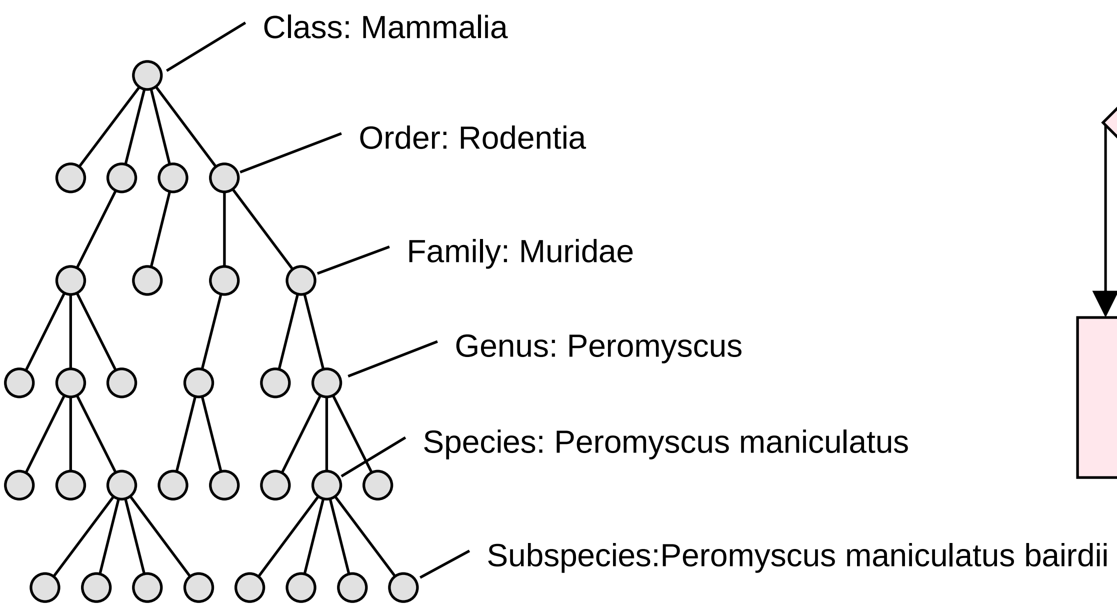
A biological pedigree can be represented as a directed graph structure relating two classes of nodes (males and females) with specific constraints: all nodes have two and only two parents, one male and one female. In a relational database, this graph can be represented as a pair of tables.

Graph Challenges

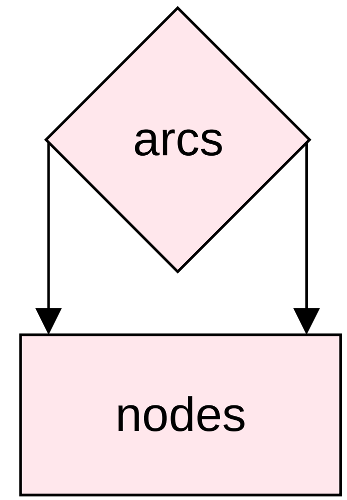


Graph Challenges

Classification Hierarchy



Relational Representation



A simple organismal classification hierarchy can be represented as a single-rooted, connected, directed graph structure with the specific constraint: all nodes have one and only one parent. In a relational database, this graph can be represented as a pair of tables.

Graph Challenges

Classification Hierarchy

Relational Representation

Graph problem:

Any graph can be represented in a relational database as a pair of tables. Enforcing the constraints for a particular graph, however, requires complex procedural code.

A simple organismal classification hierarchy can be represented as a single-rooted, connected directional graph structure with the specific constraint: all nodes have one and only one parent. In a relational database, this graph can be represented as a pair of tables.

Graph Challenges

Classification Hierarchy

Relational Representation

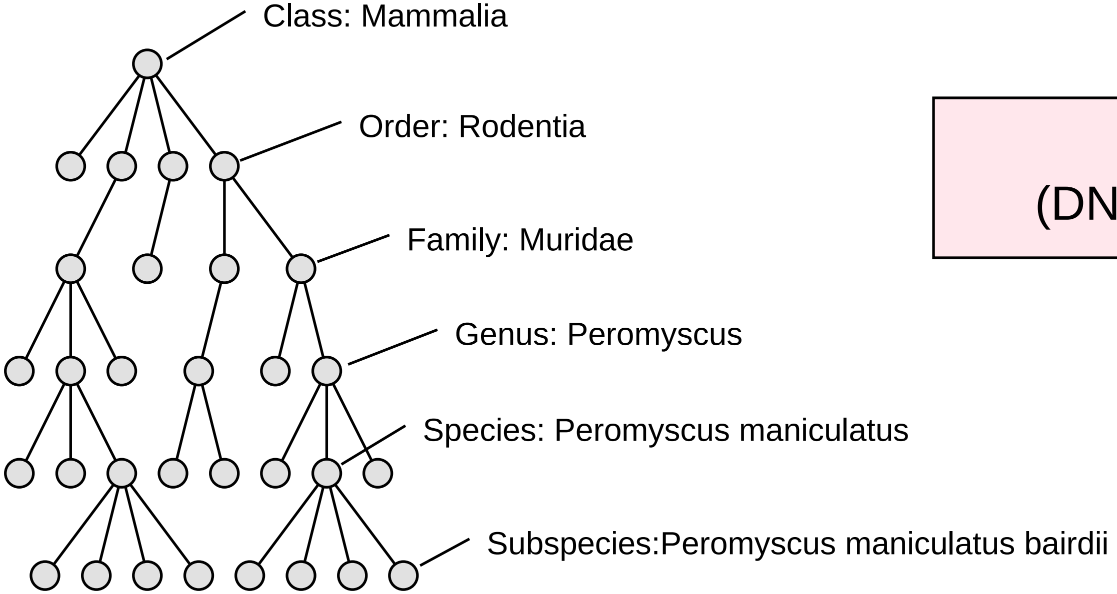
Graph solutions needed:

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)

A simple organismal classification hierarchy can be represented as a single-rooted, connected directional graph structure with the specific constraint: all nodes have one and only one parent. In a relational database, this graph can be represented as a pair of tables.

Classification Challenges

Classification Hierarchy

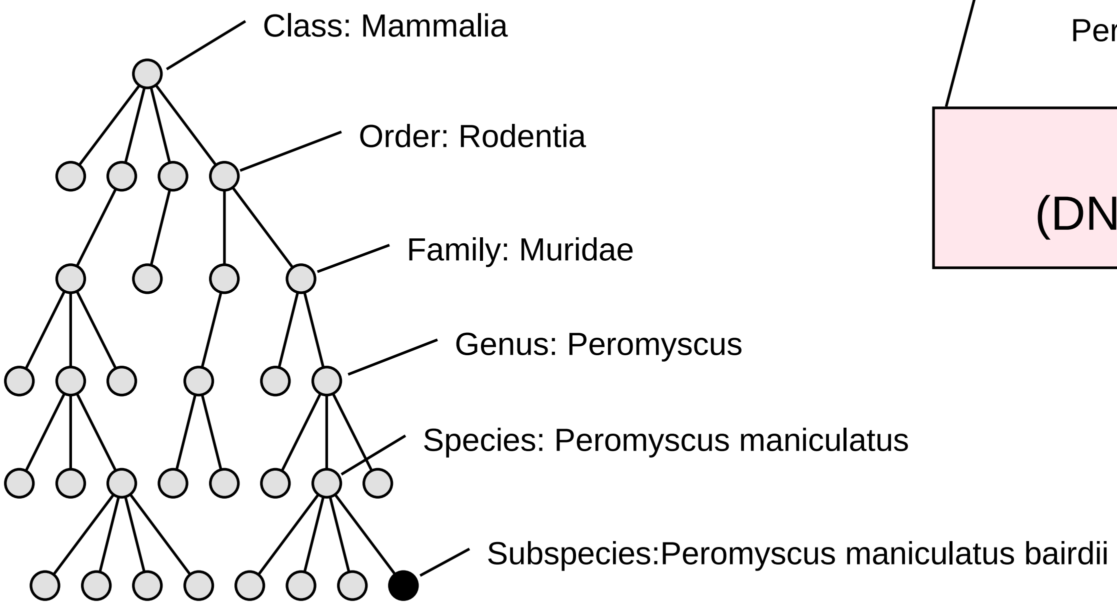


Data Objects to be Classified

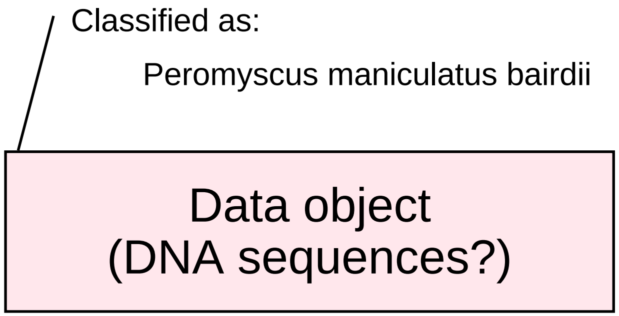
Data object
(DNA sequences?)

Classification Challenges

Classification Hierarchy

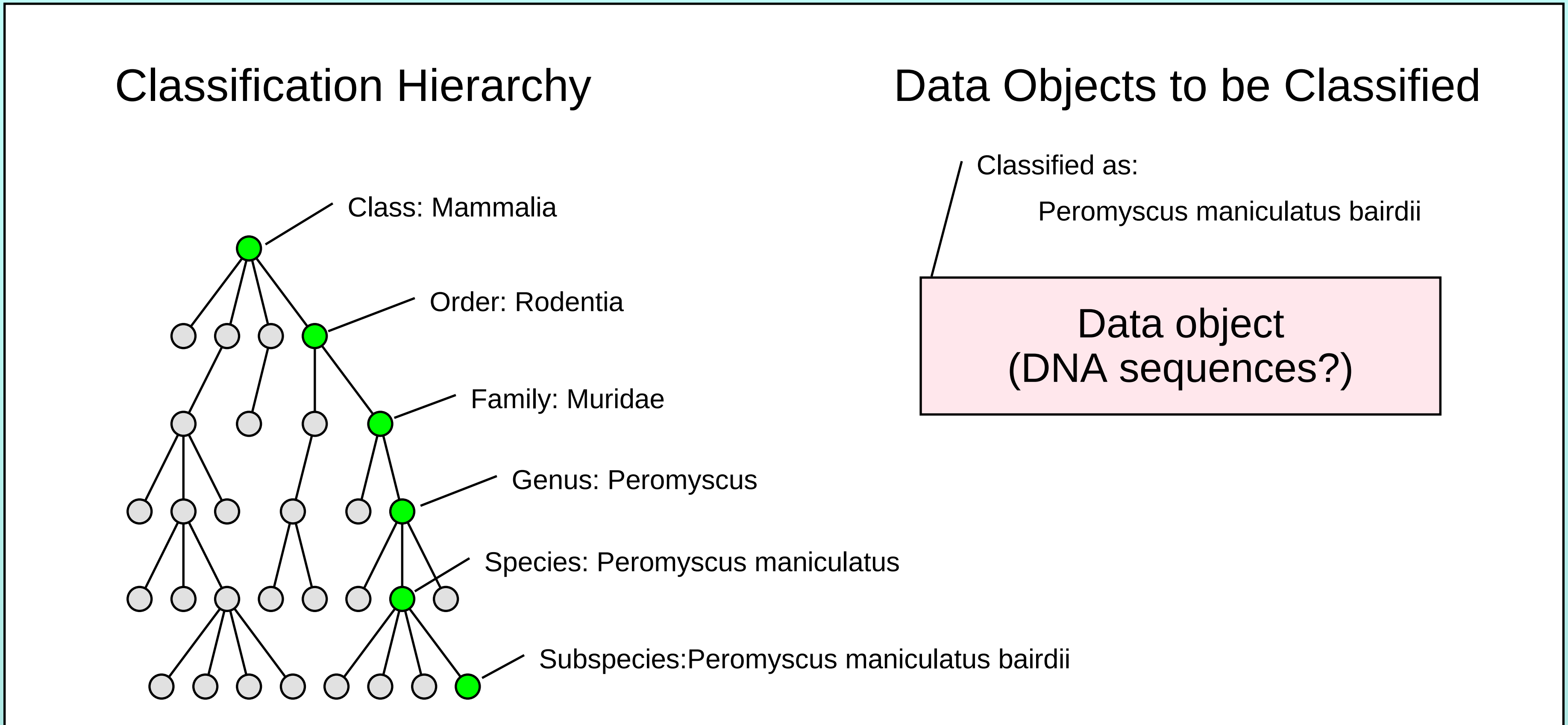


Data Objects to be Classified



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

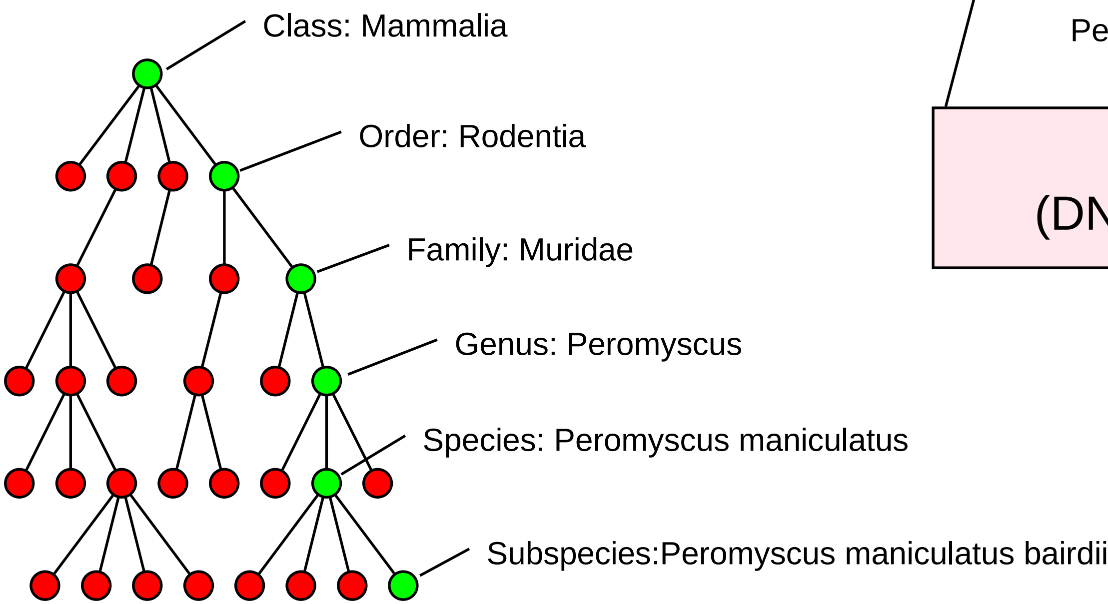
Classification Challenges



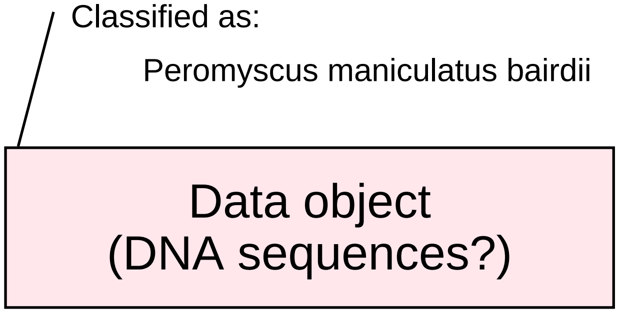
Suppose we permit querying at any level, but require classification of objects at leaf level. Then all questions referring to nodes on the path from the classification point to the top return **TRUE**,

Classification Challenges

Classification Hierarchy



Data Objects to be Classified

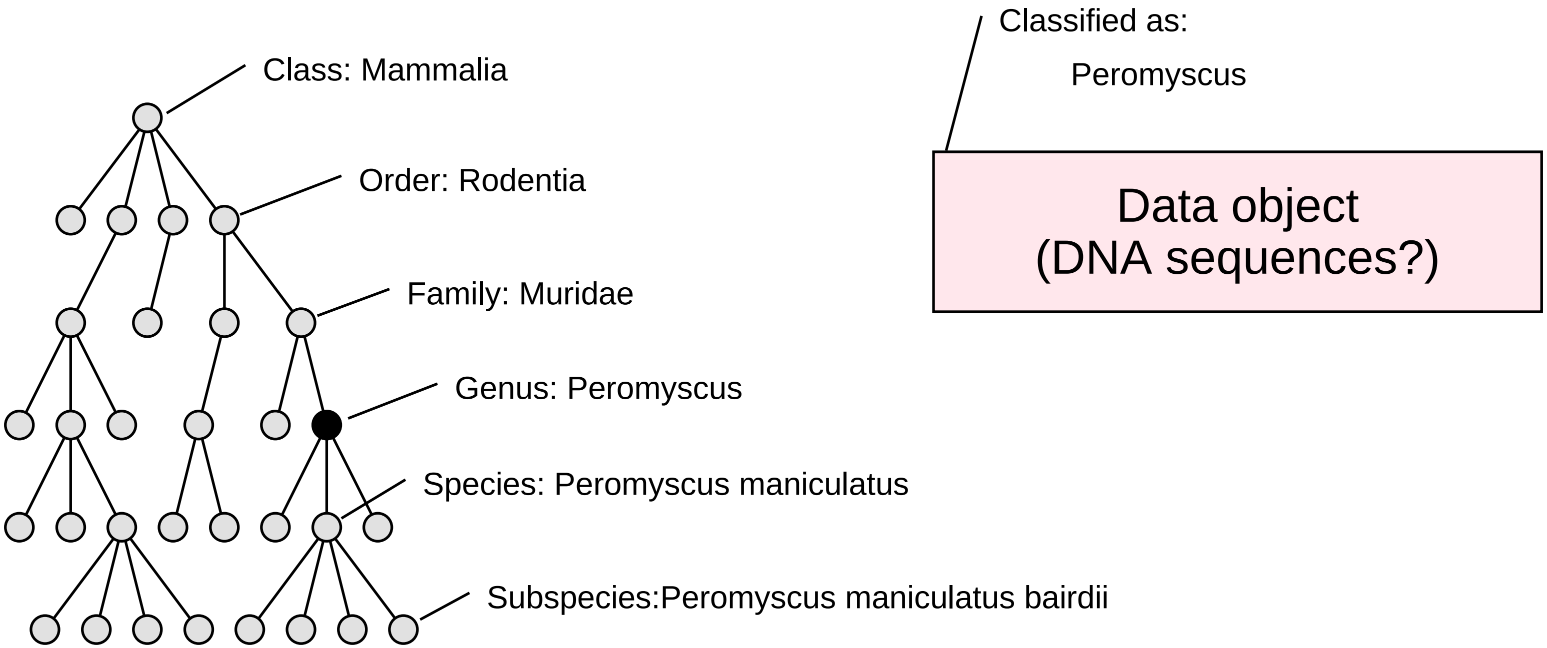


Suppose we permit querying at any level, but require classification of objects at leaf level. Then all questions referring to nodes on the path from the classification point to the top return **TRUE**, all others **FALSE**.

Classification Challenges

Classification Hierarchy

Data Objects to be Classified

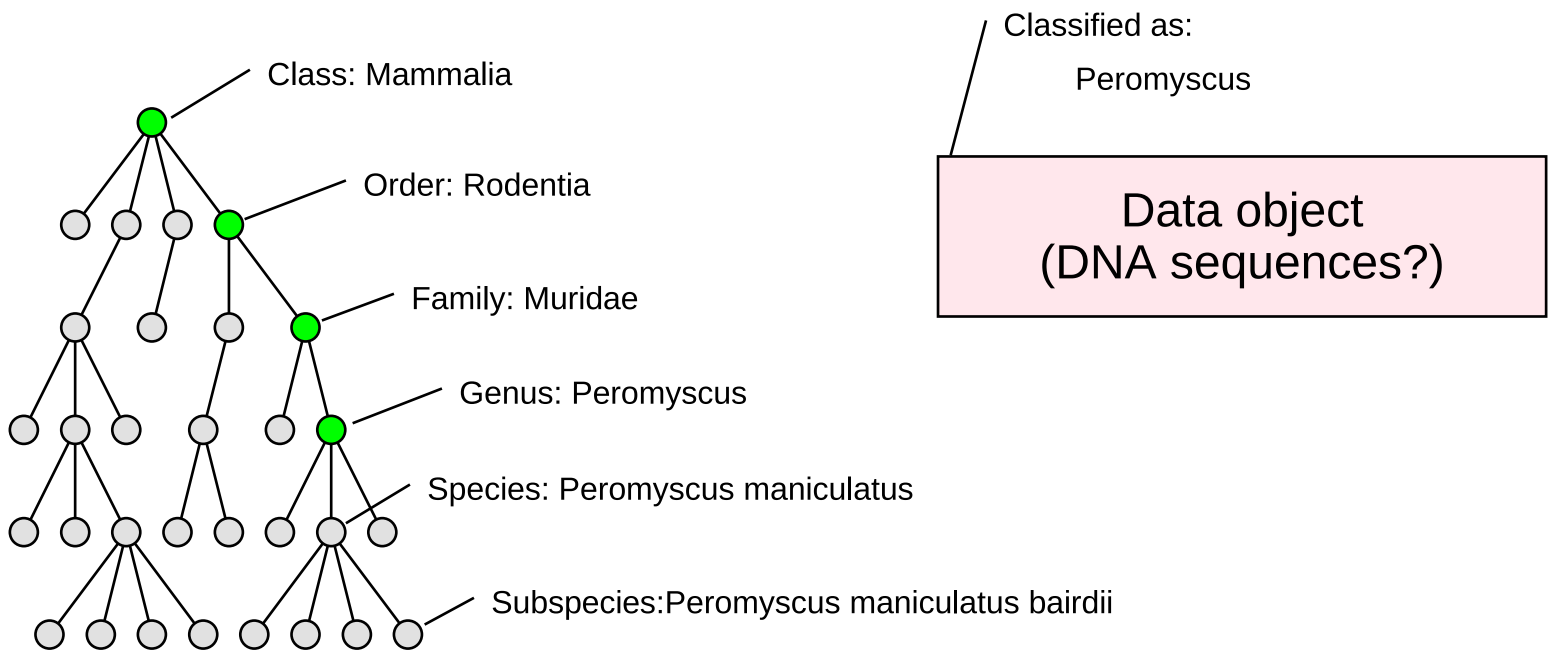


Now, suppose the we permit querying at any level, and also that we allow classification of objects at any level.

Classification Challenges

Classification Hierarchy

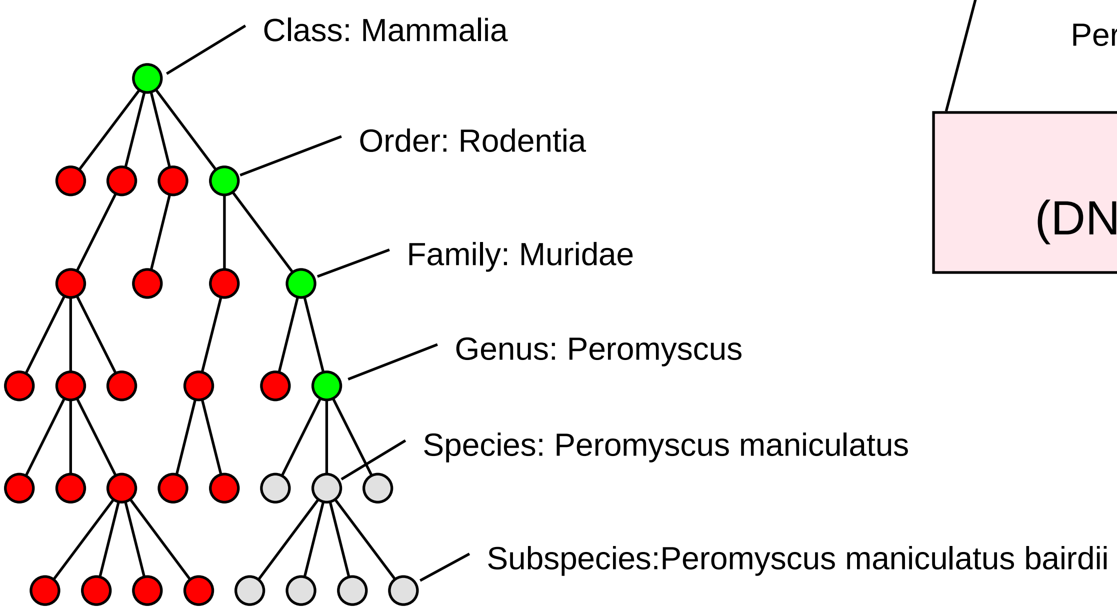
Data Objects to be Classified



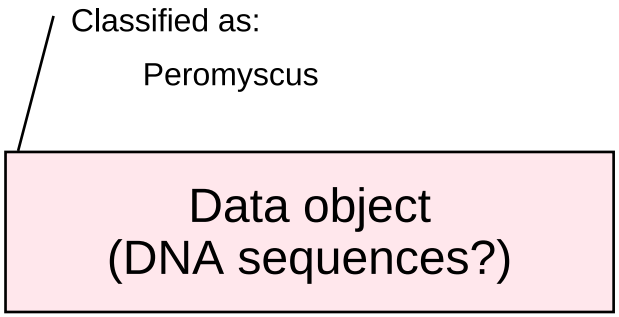
Now, suppose the we permit querying at any level, and also that we allow classification of objects at any level. Then all questions referring to nodes on the path from the classification point to the top return **TRUE**,

Classification Challenges

Classification Hierarchy

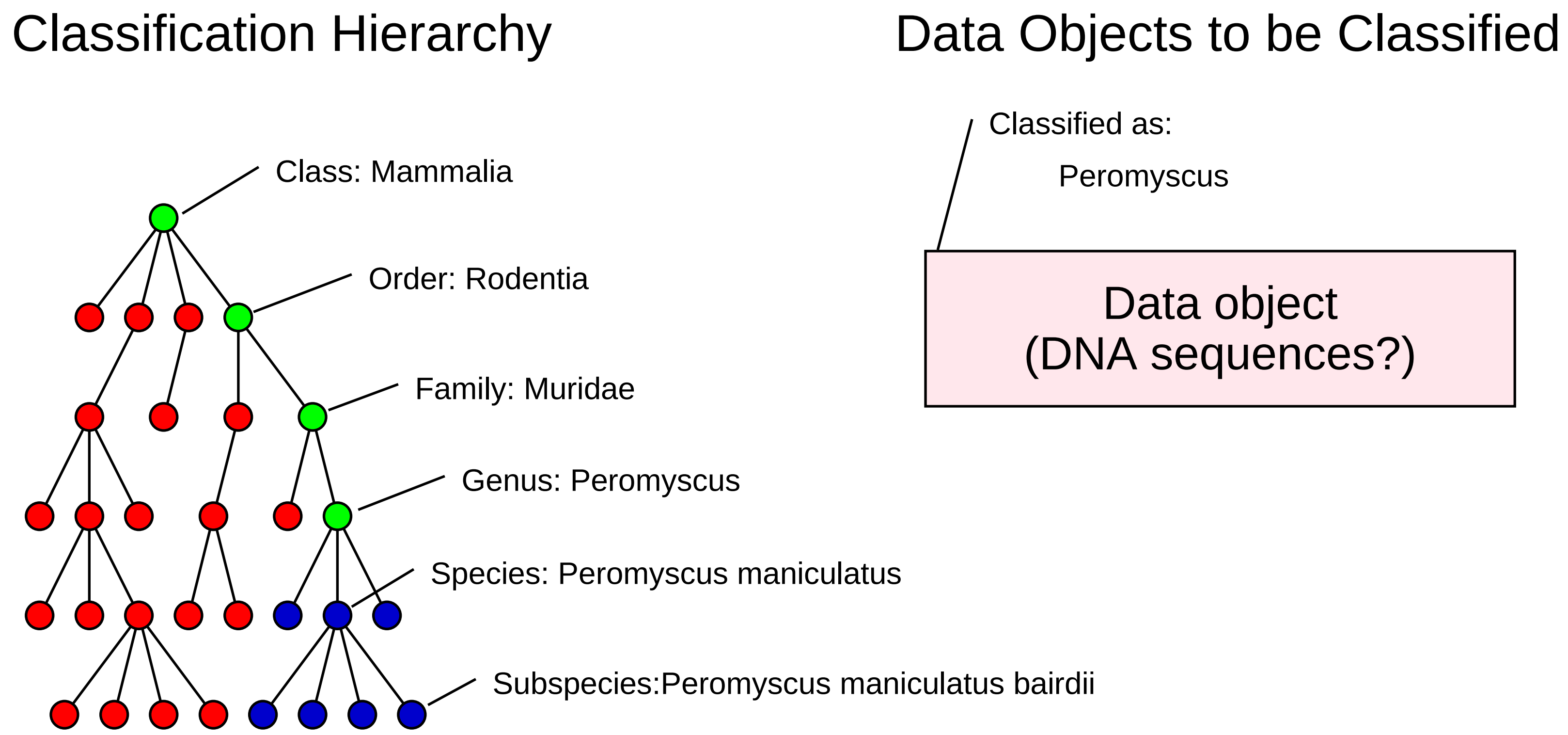


Data Objects to be Classified



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



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

Classification Challenges

Classification Hierarchy

Data Objects to be Classified

Tri-state logic required:

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

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

Database III

Data Integration

Data Integration Crisis

Adequate connections among data objects in different databases do not exist.

Without adequate connectivity, much of the value of the data will be lost.

Data Integration Goals

Achieve conceptual integration of biomedical data.

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

Data Integration Impediments

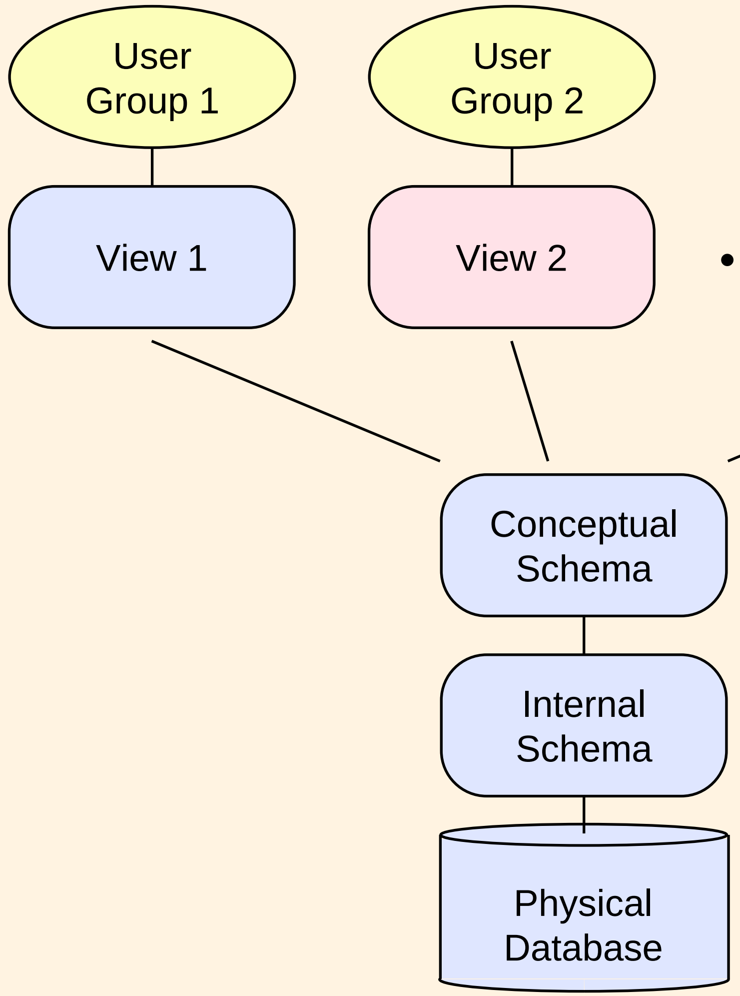
Technical: Integrating distributed, hetero-geneous databases is not easy.

Sociological: Local incentives encourage competition, not cooperation.

Conceptual: Semantic mismatches exist among databases.

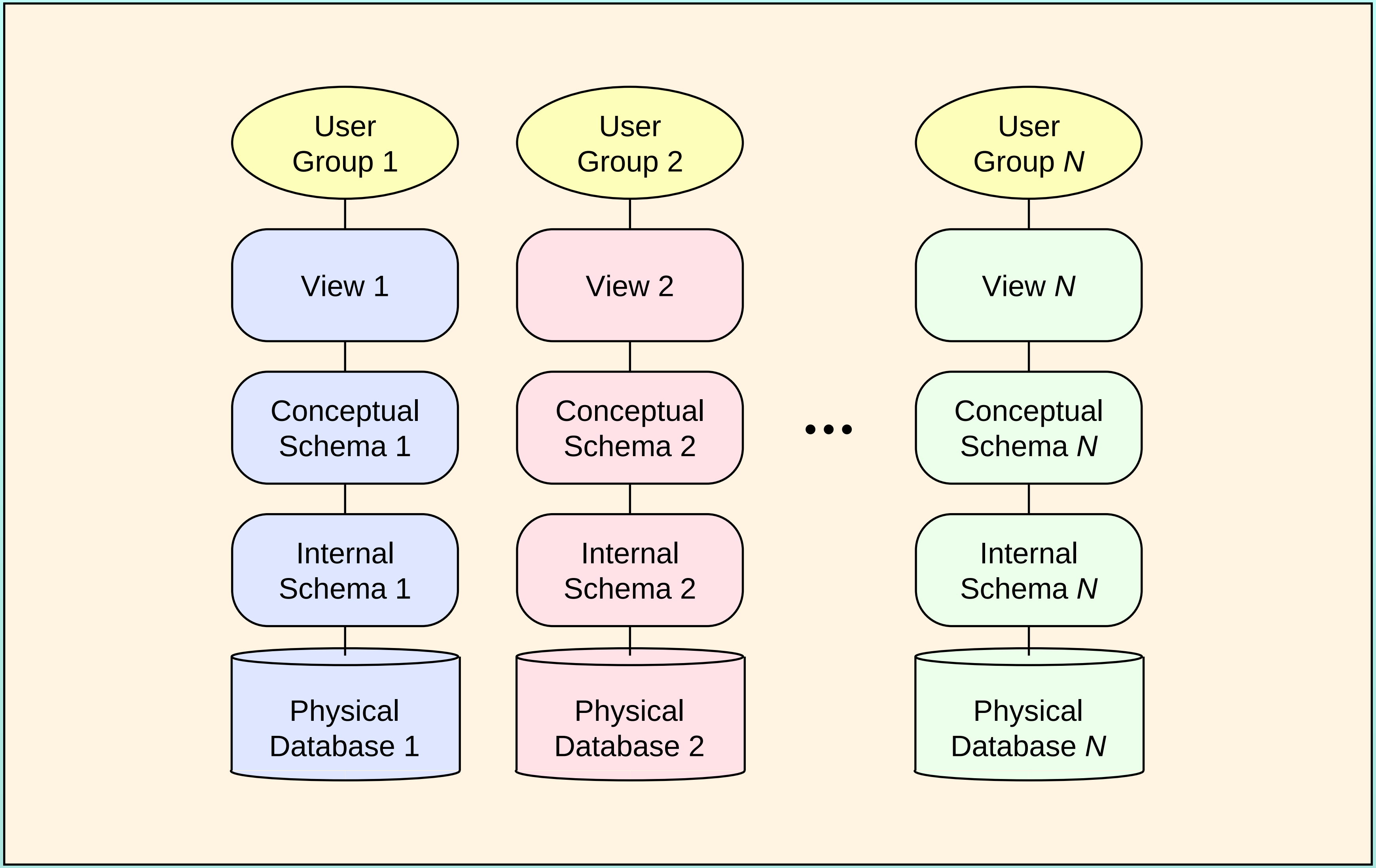
Technical Impediments

Multiple Views

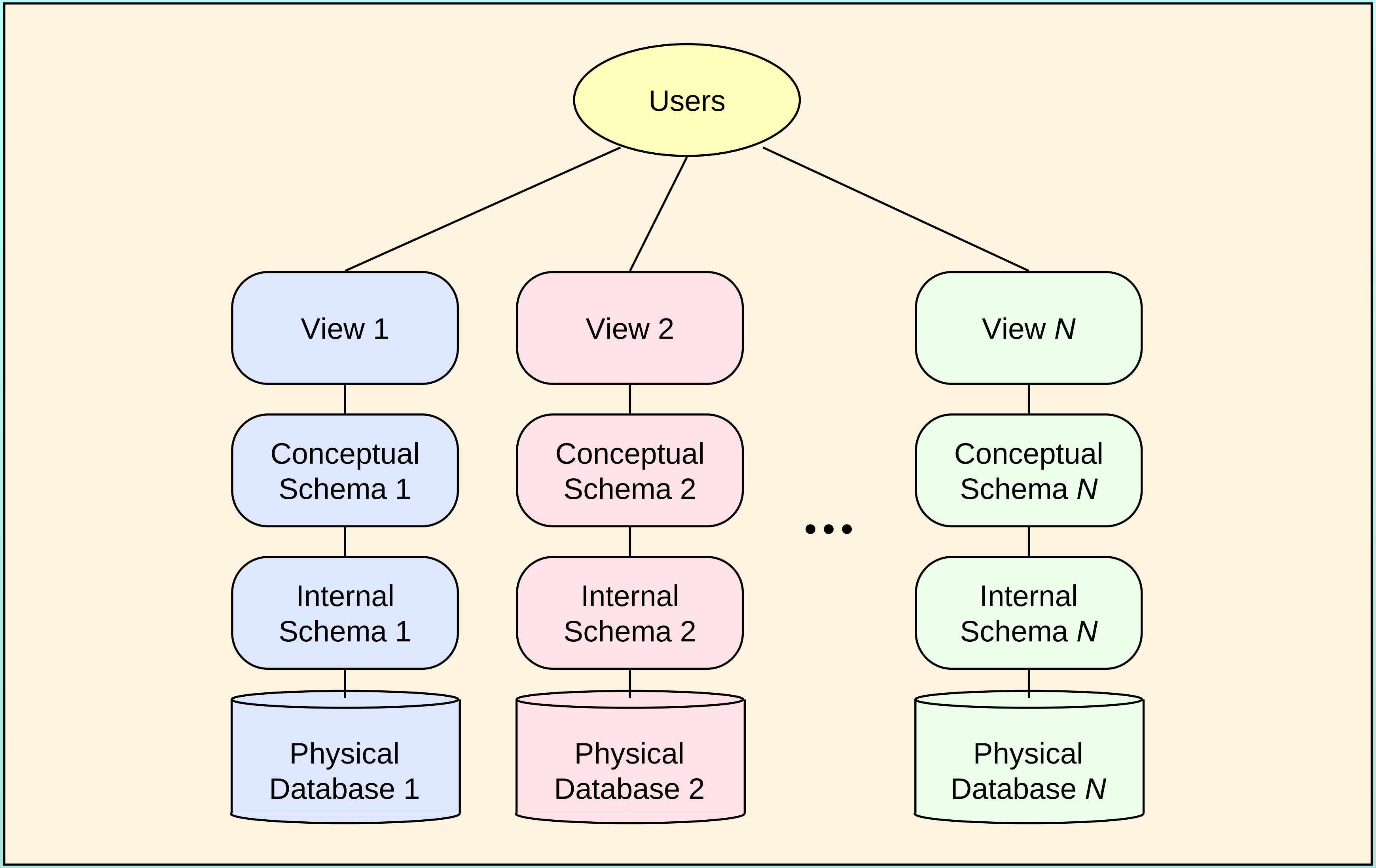


Database designs are layered, with each layer at a different level of abstraction.

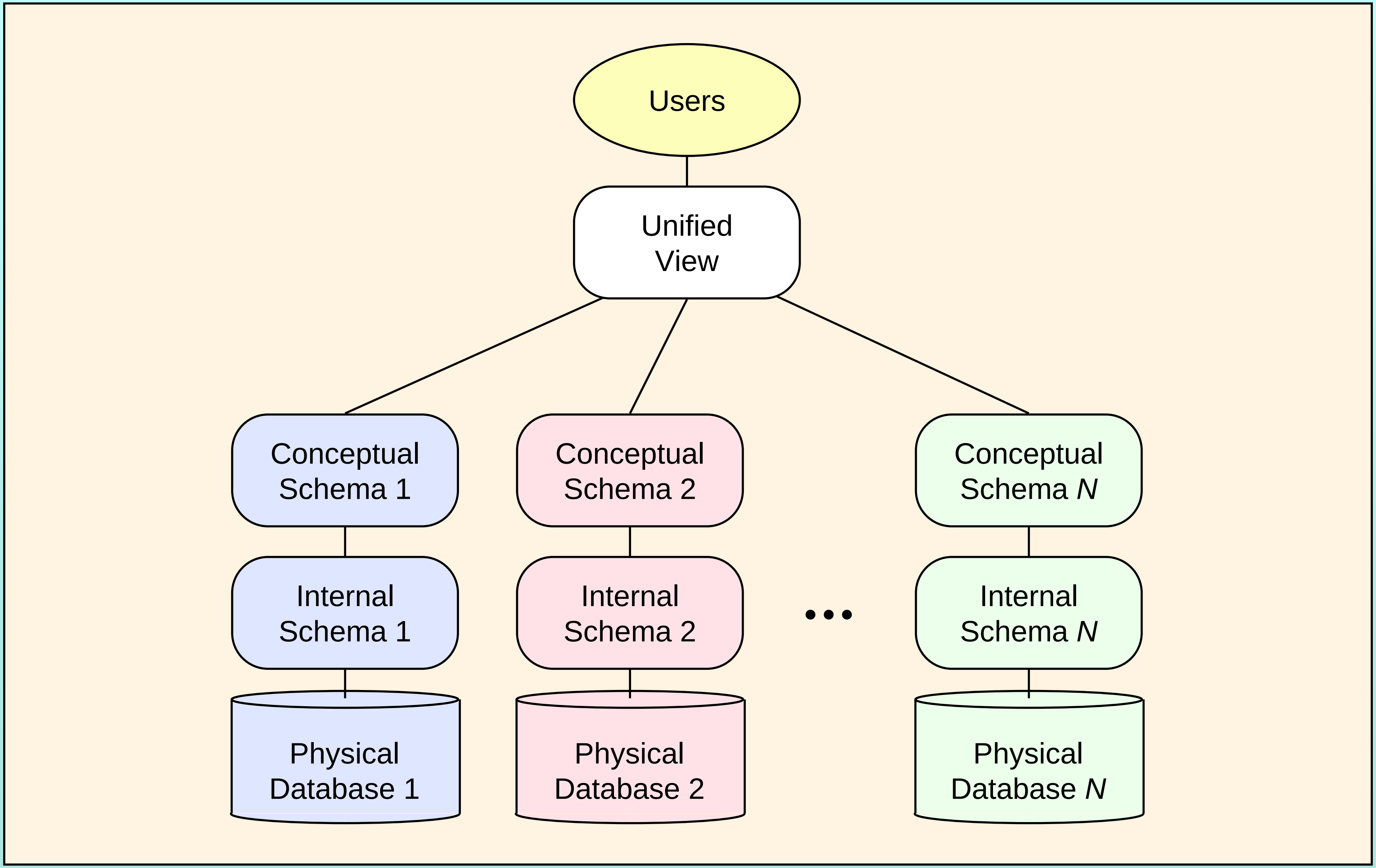
Multiple Databases



Current Situation



Desired Situation



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

Multidatabase Taxonomy

UNFEASIBLE	COMMON GLOBAL SCHEMA
	MEDIUM TO LONGER TERM SOLUTION
	SHORT TERM SOLUTION
UNACCEPTABLE	DO NOTHING IN ASSURING INTEROPERABILITY

Options for integrating networked databases (adapted from Chorafas and Steinmann, 1993).

Multidatabase Taxonomy

Tightly Coupled: single organizational entity overseeing information resources relevant to biomedical research

•
•
•

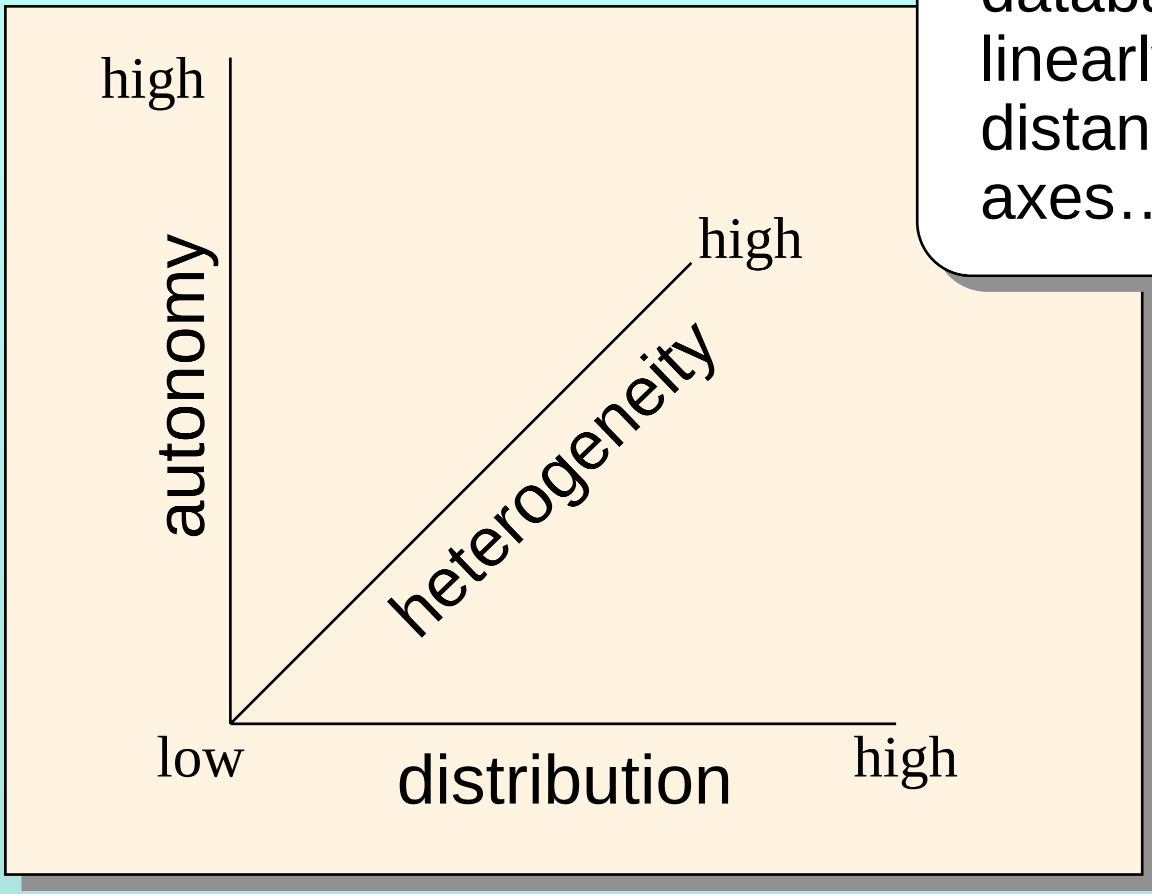
adoption of common DBMSs at participating sites

shared data model across participating sites

common semantics for data publishing

Loosely Coupled: common syntax for data publishing

Difficulty Dimensions

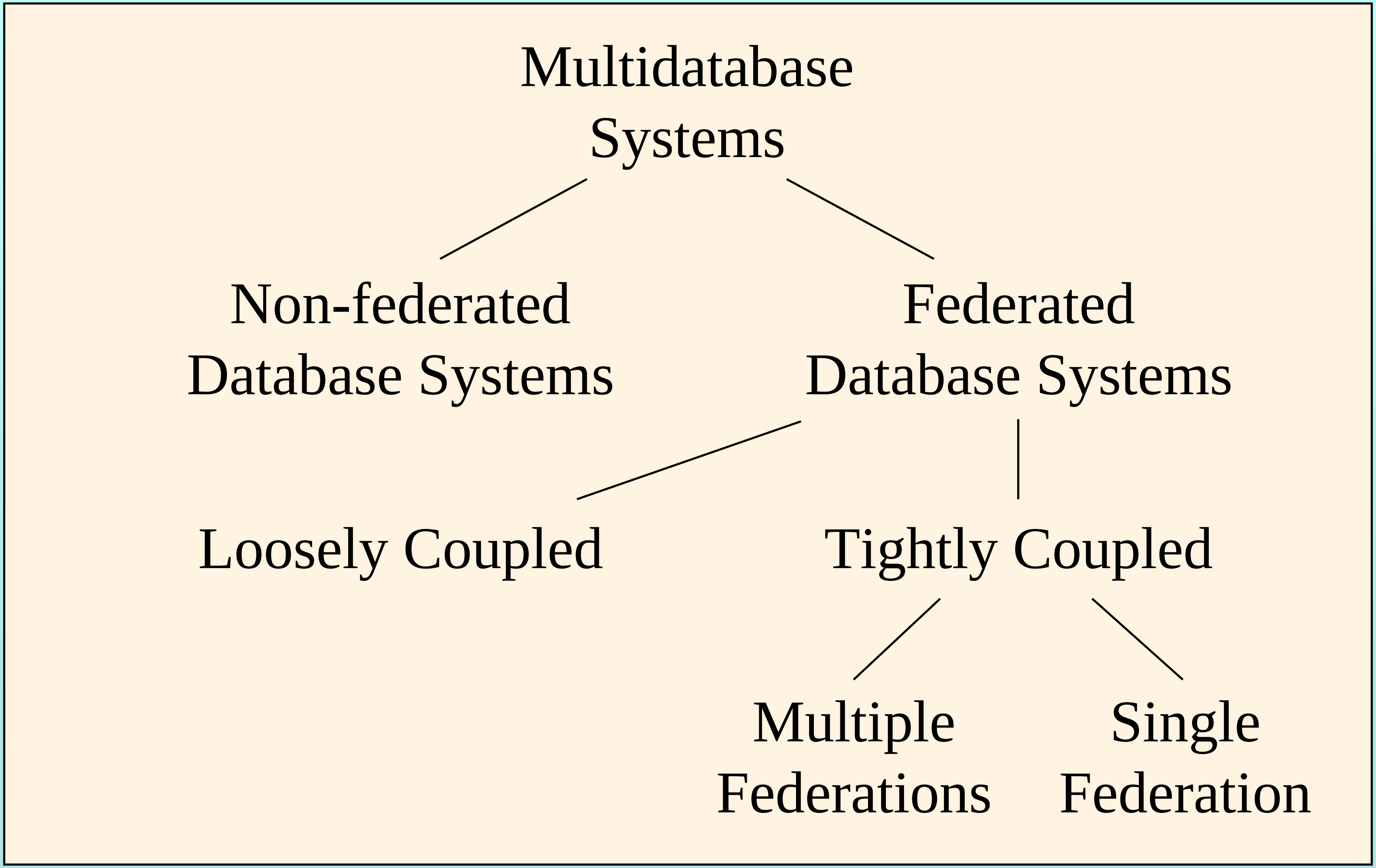


Difficulty in connecting databases scales non-linearly as a function of distance along all three axes…

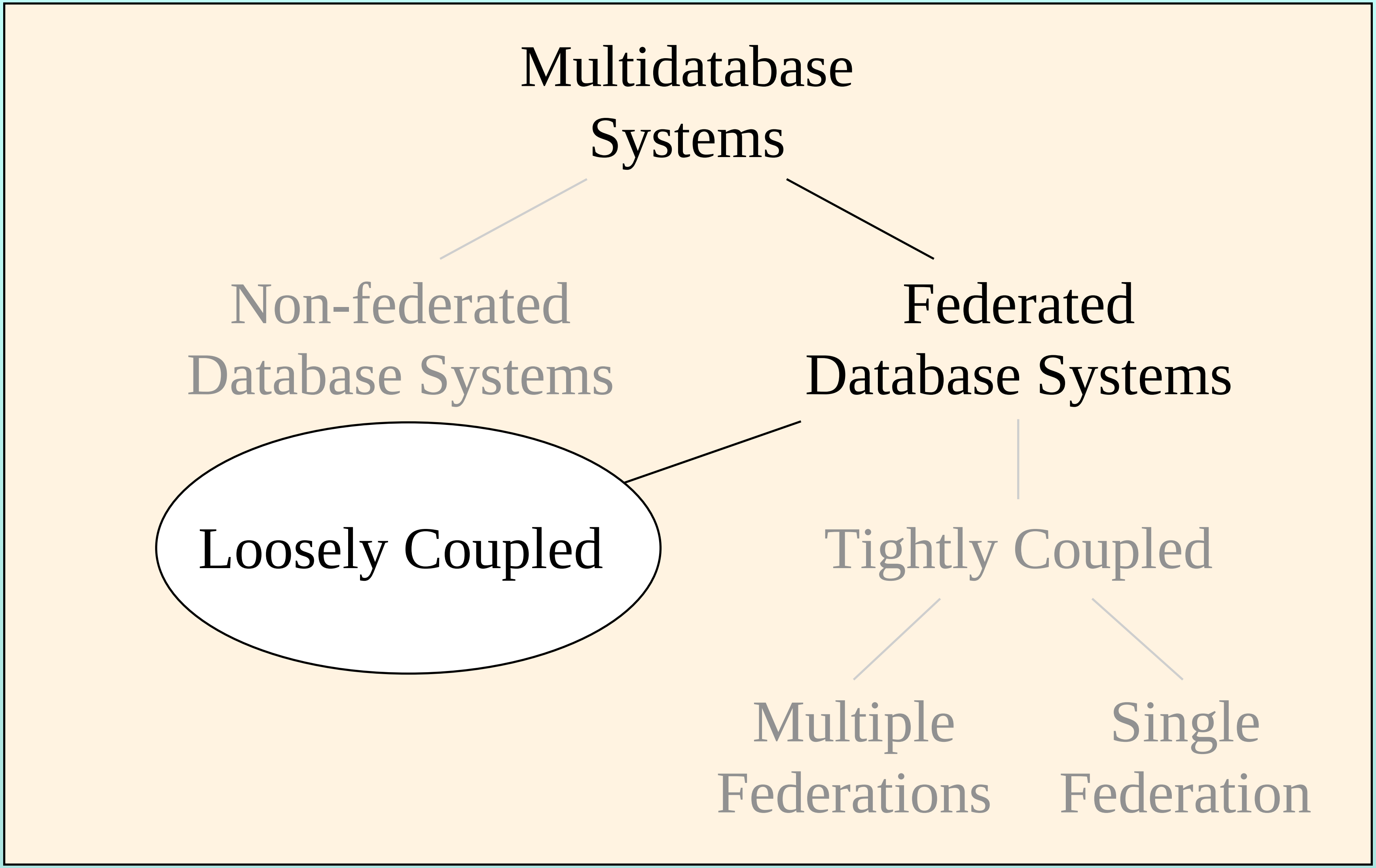
Multidatabase Taxonomy

- 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.

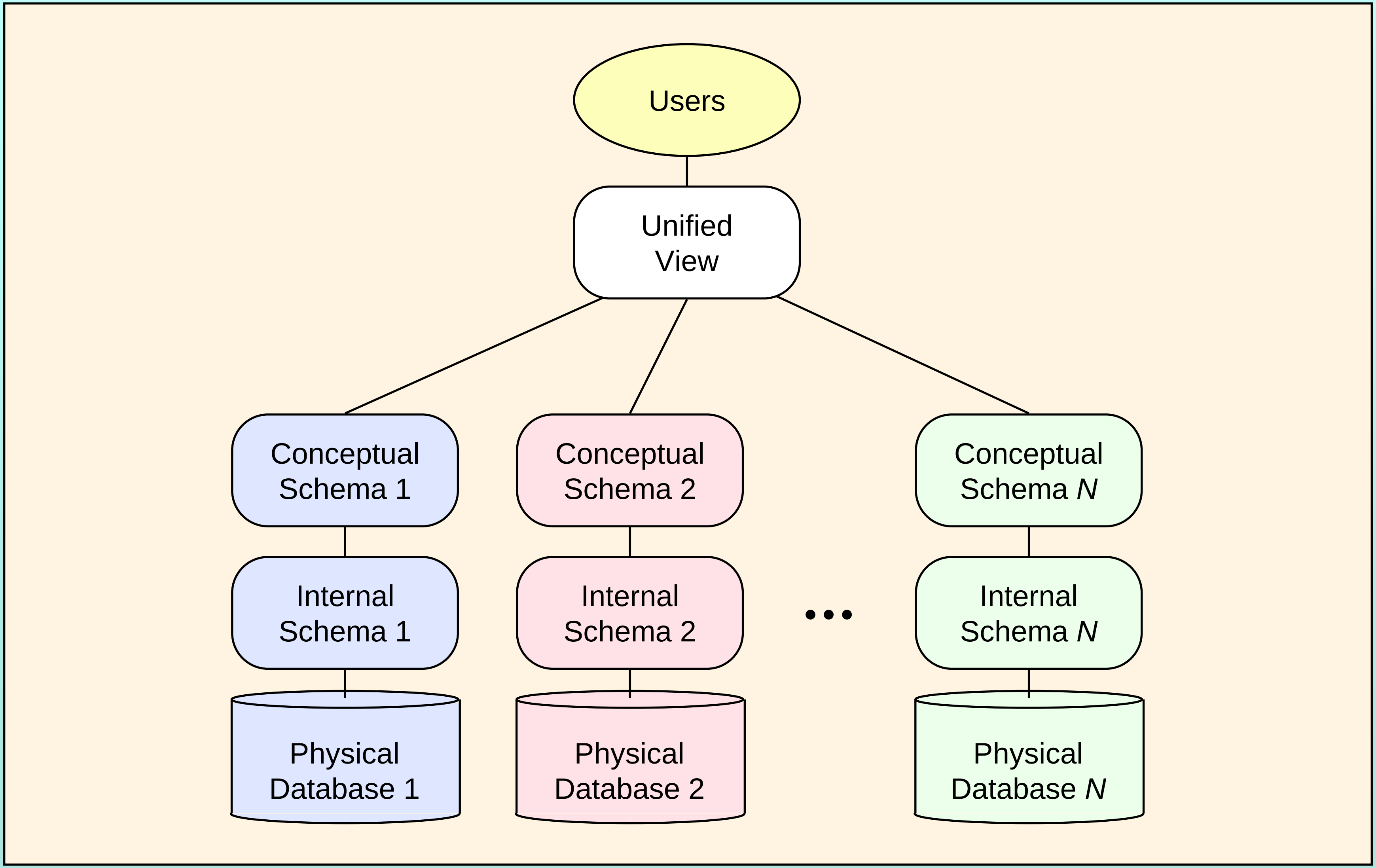
Multidatabase Taxonomy



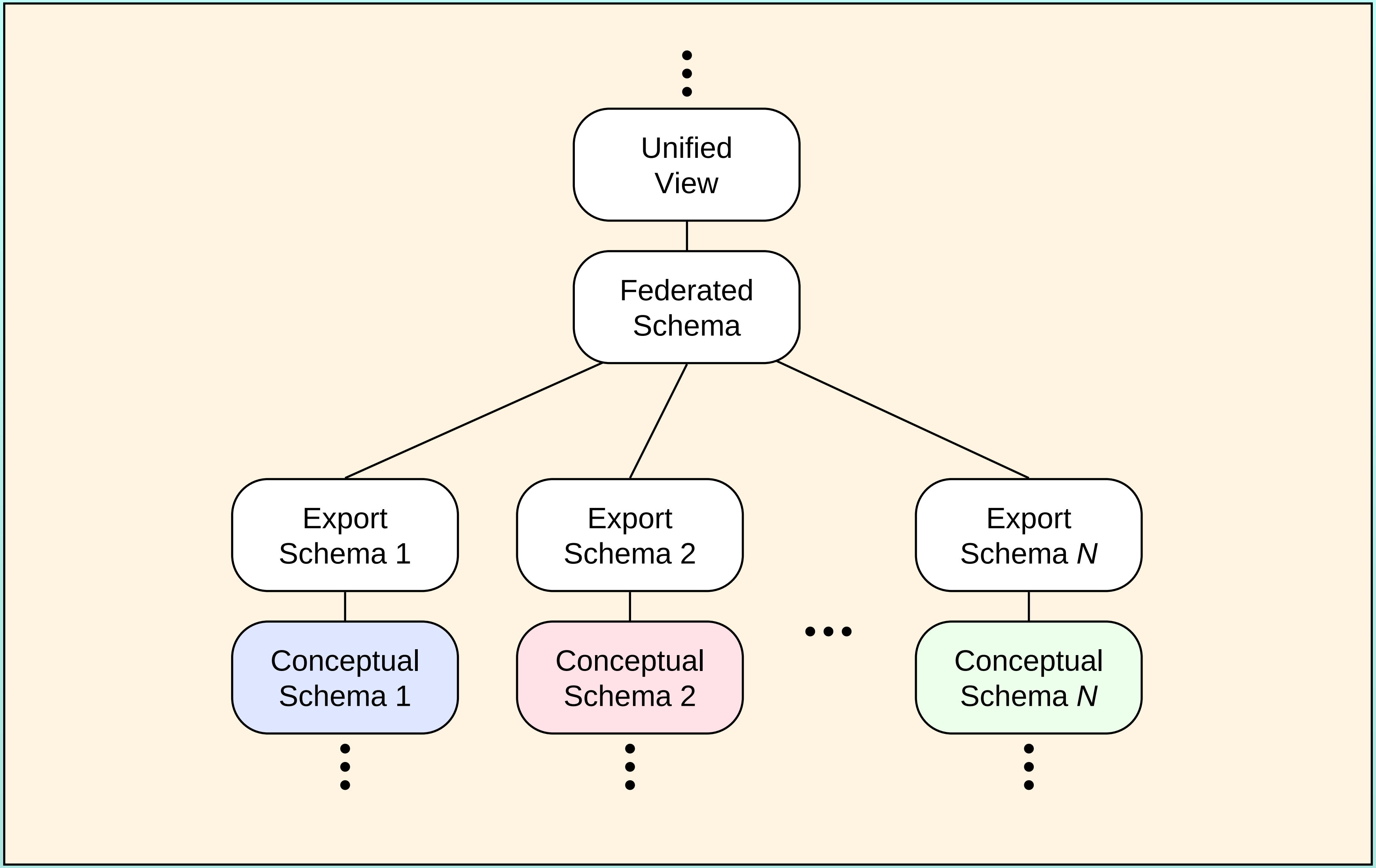
Multidatabase Taxonomy



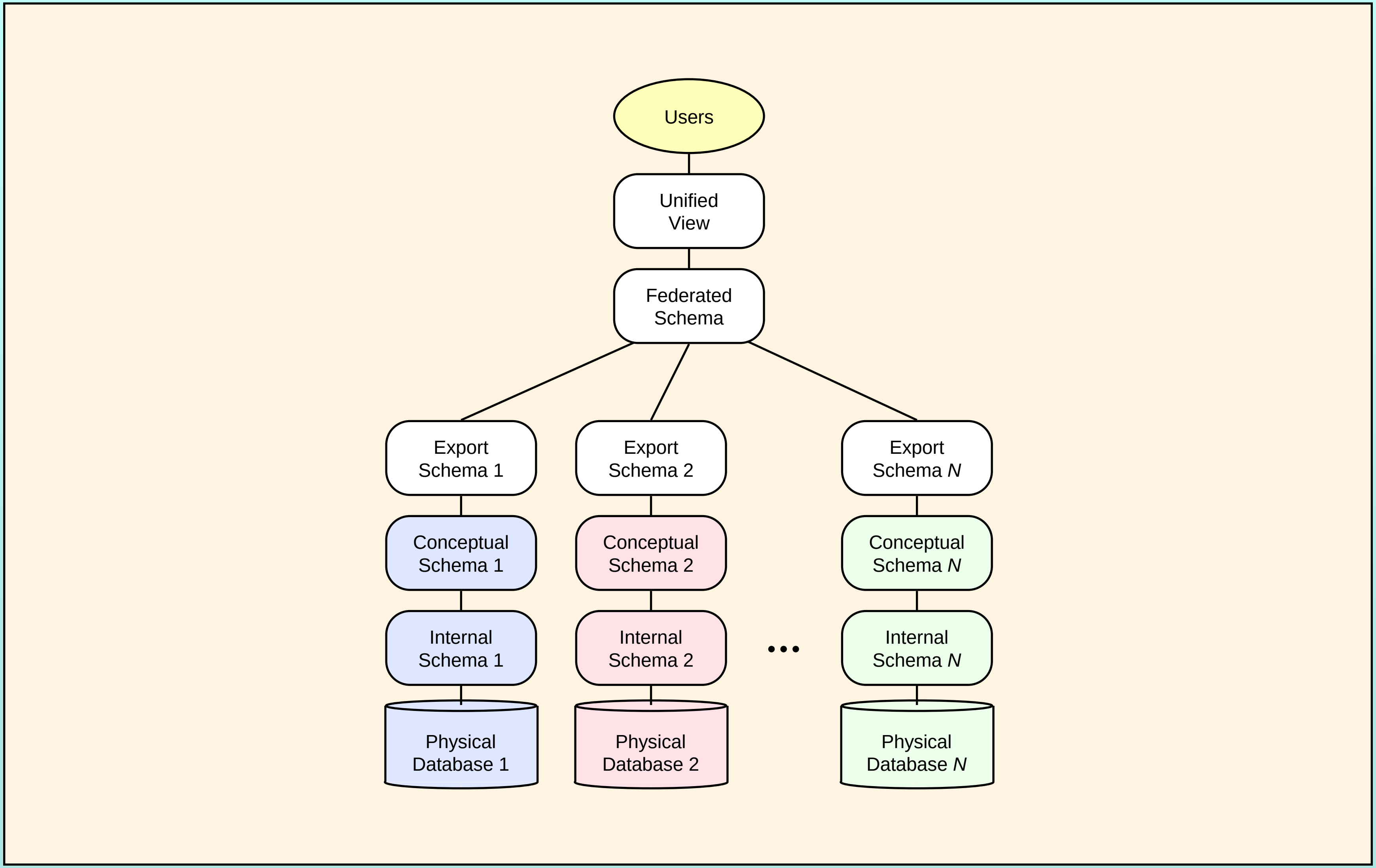
Desired Situation



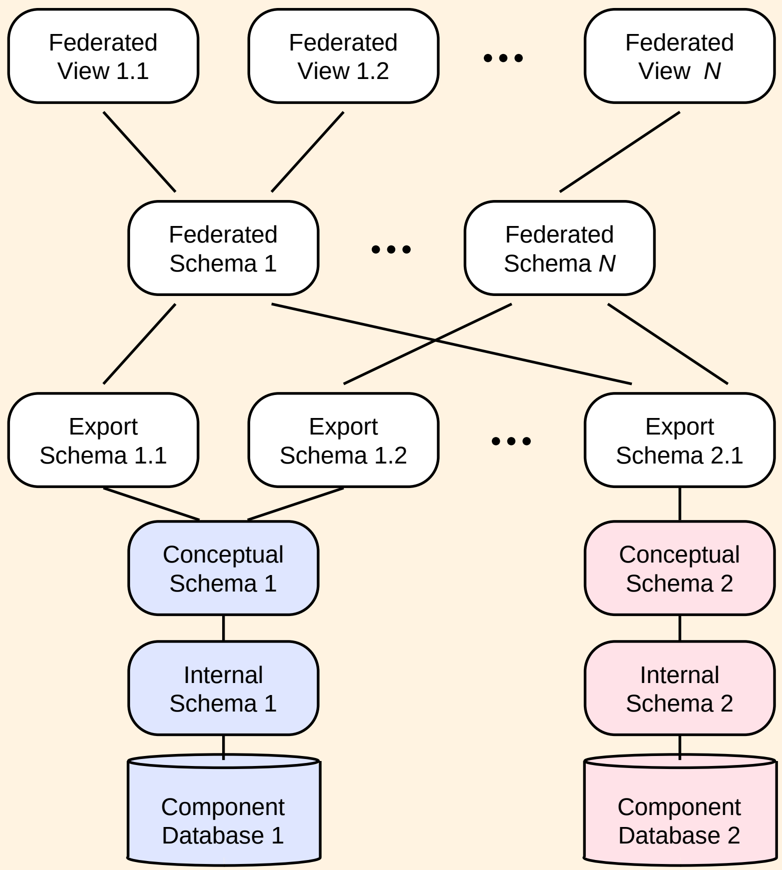
More Layers


Unified
View
Federated
Schema
Export
Schema 1
Export
Schema 2
Export
Schema *N*
Conceptual
Schema 1
Conceptual
Schema 2
Conceptual
Schema *N*

Federated Schema



Multiple Federations



Schema Change

Schema-change Issues

Problems occur at many levels:

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

Schema-change Issues

Problems occur at many levels:

Schema Evolution:

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

- More issues…

Database Solutions

Database Solutions

Solutions might occur at many levels:

- Development of more sophisticated products by vendors.
- Adoption of consistent (if inadequate) methods in the meanwhile.
- Facilitate equivalent solutions across grants by providing equivalent infrastructure support at the institutional level.

Data Source Problems

Topics

- Data-source problems

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

  Integrating multi-source data is hard.

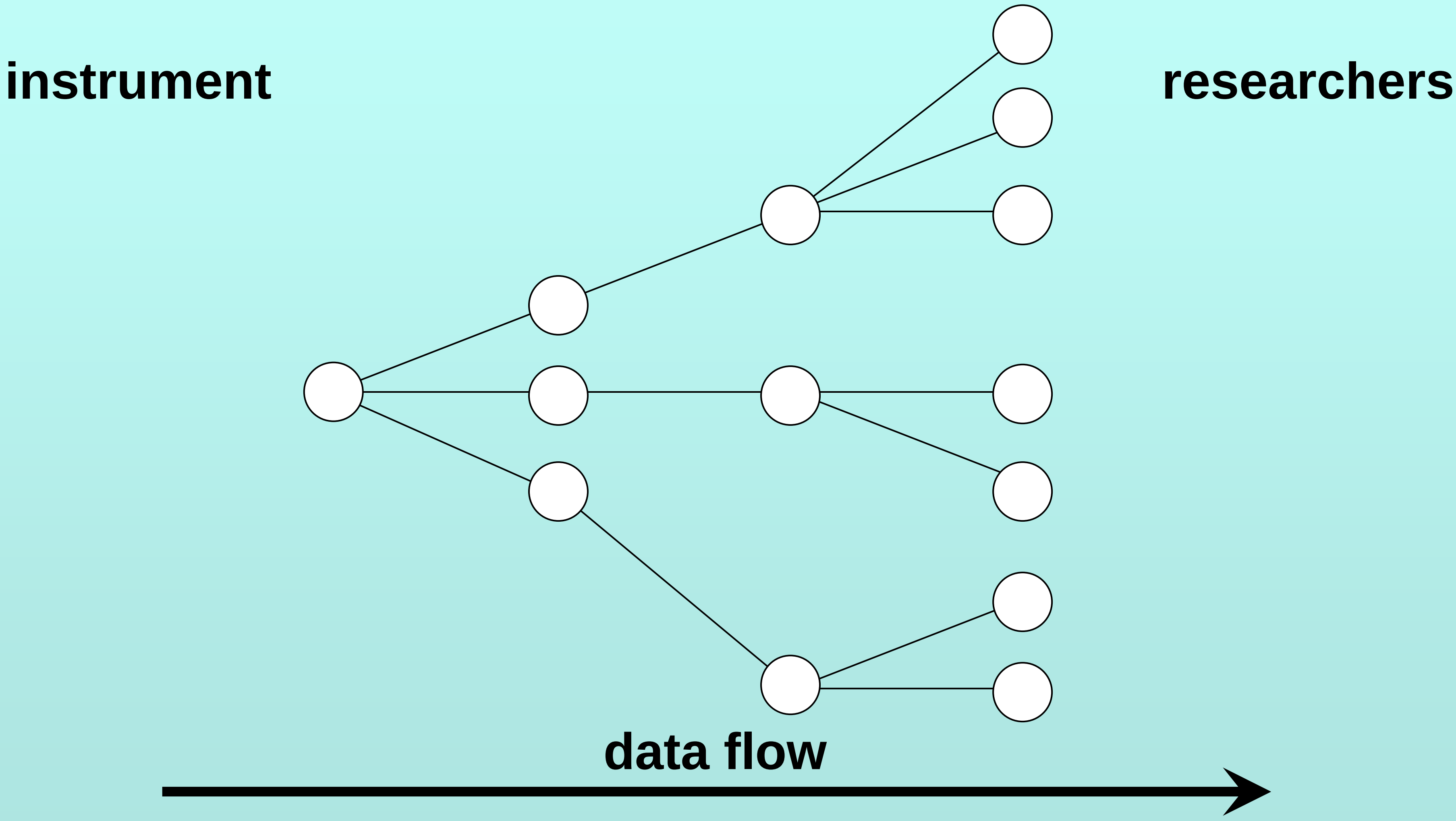
  Consistency flows in the wrong direction.

  GenBank is a false model.

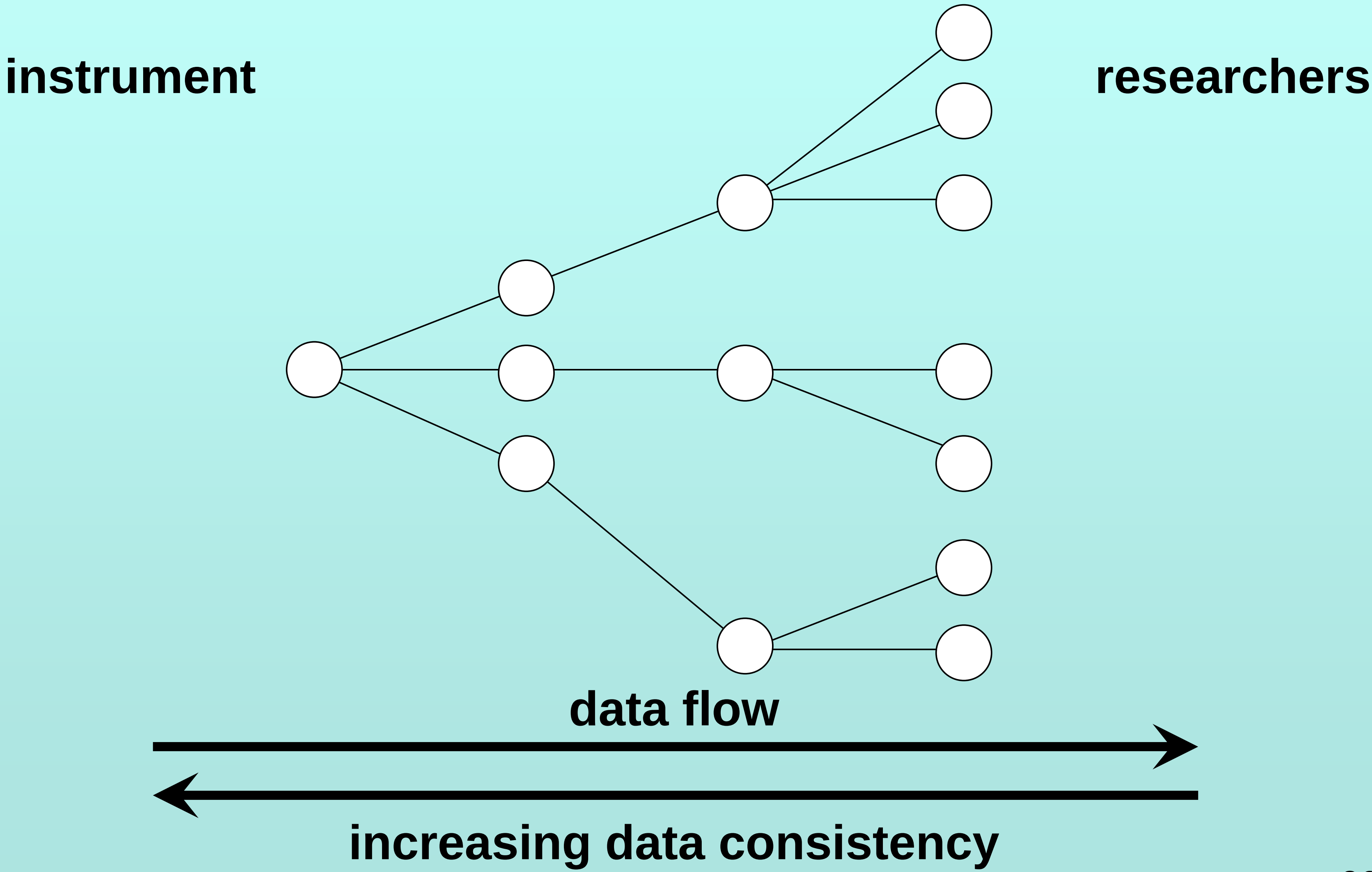
Source I

Basics

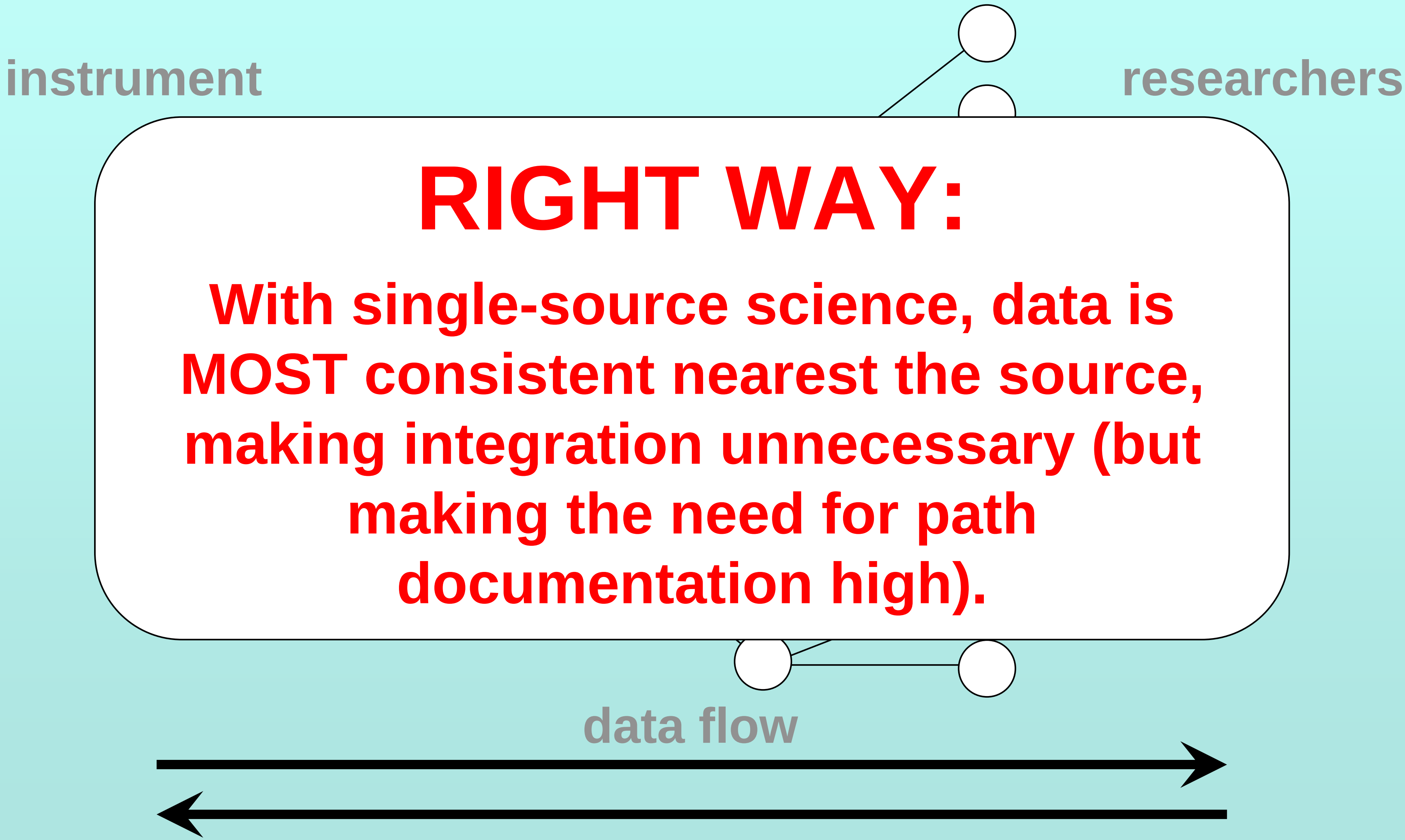
Single-instrument Science



Single-instrument Science



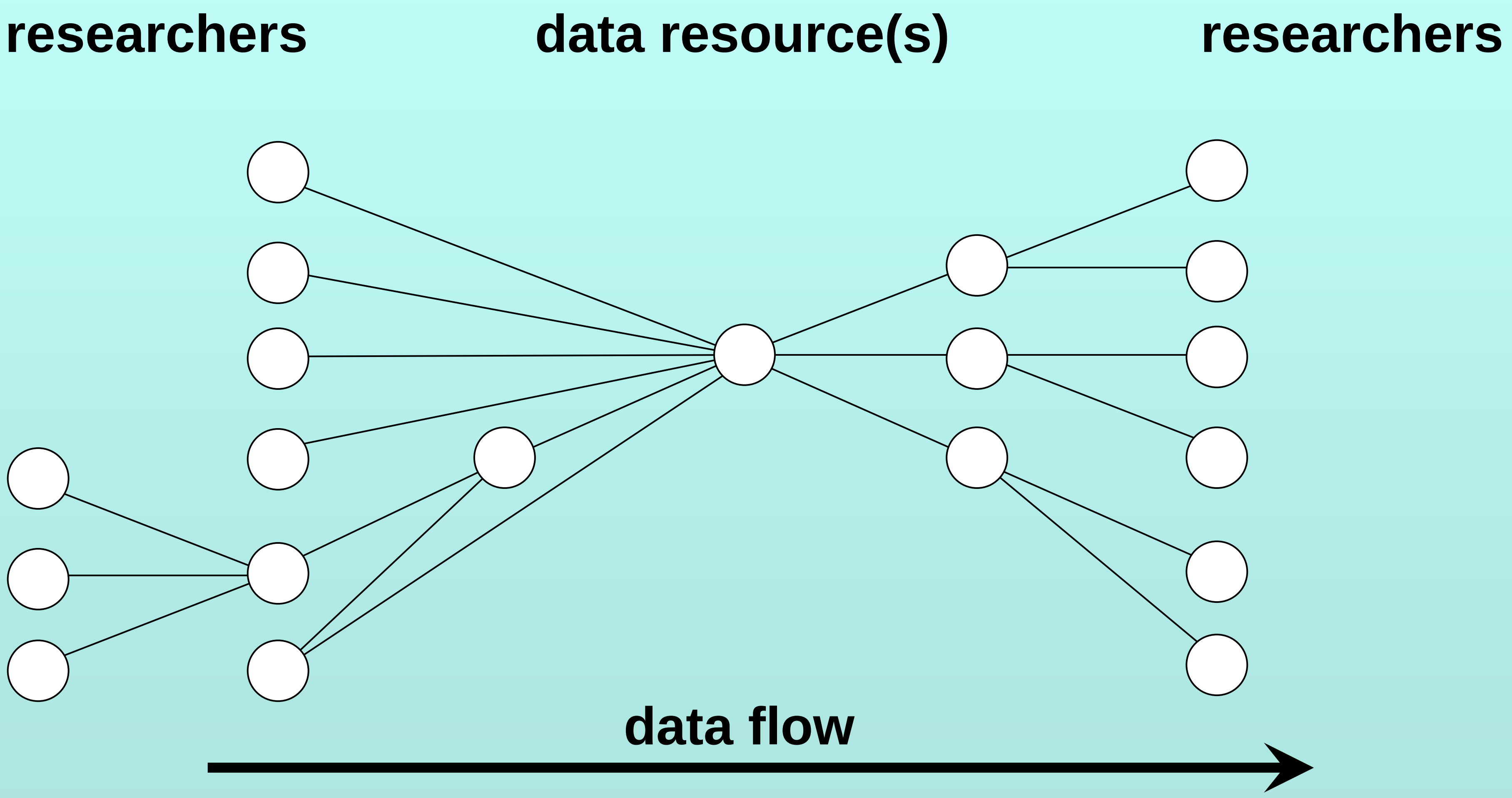
Single-instrument Science



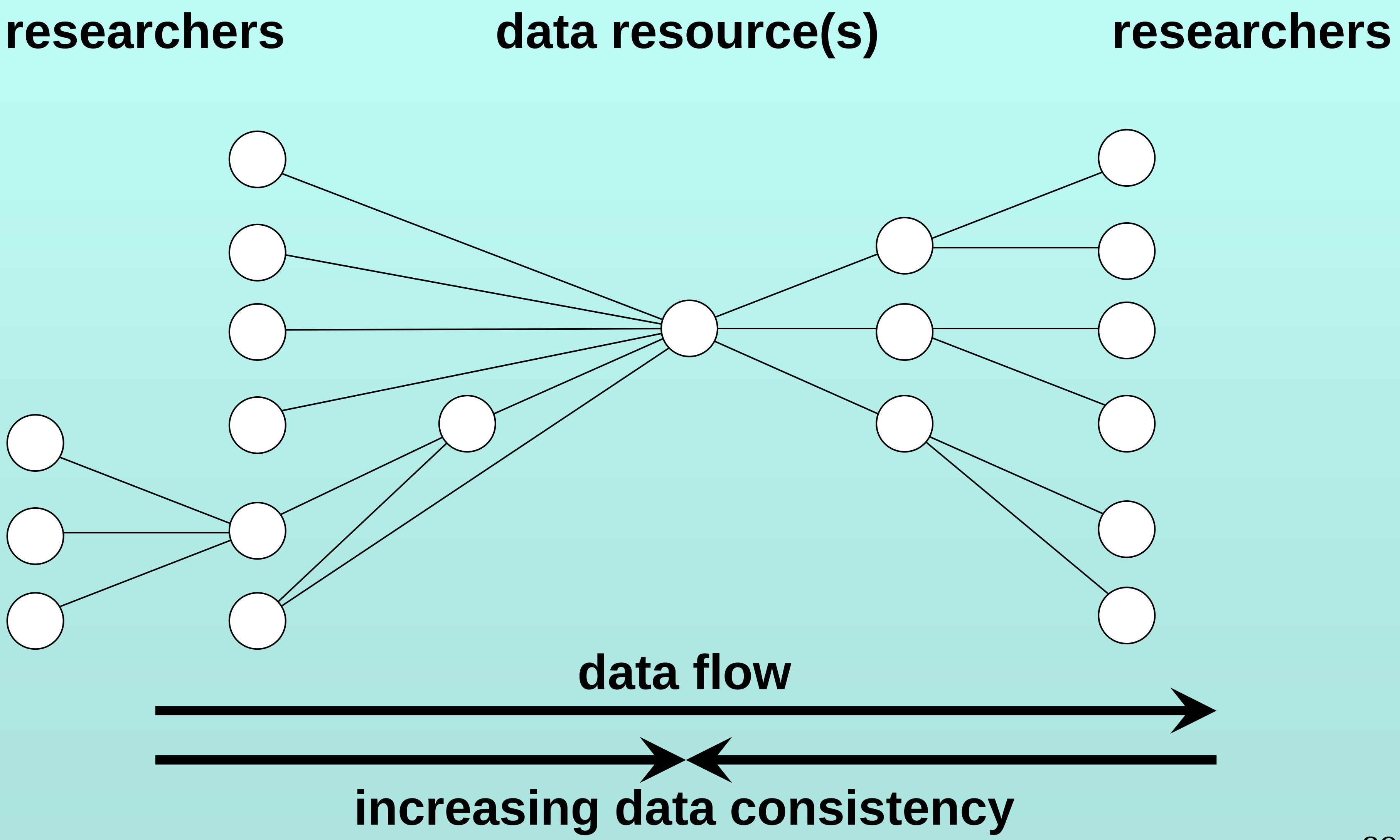
RIGHT WAY:

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

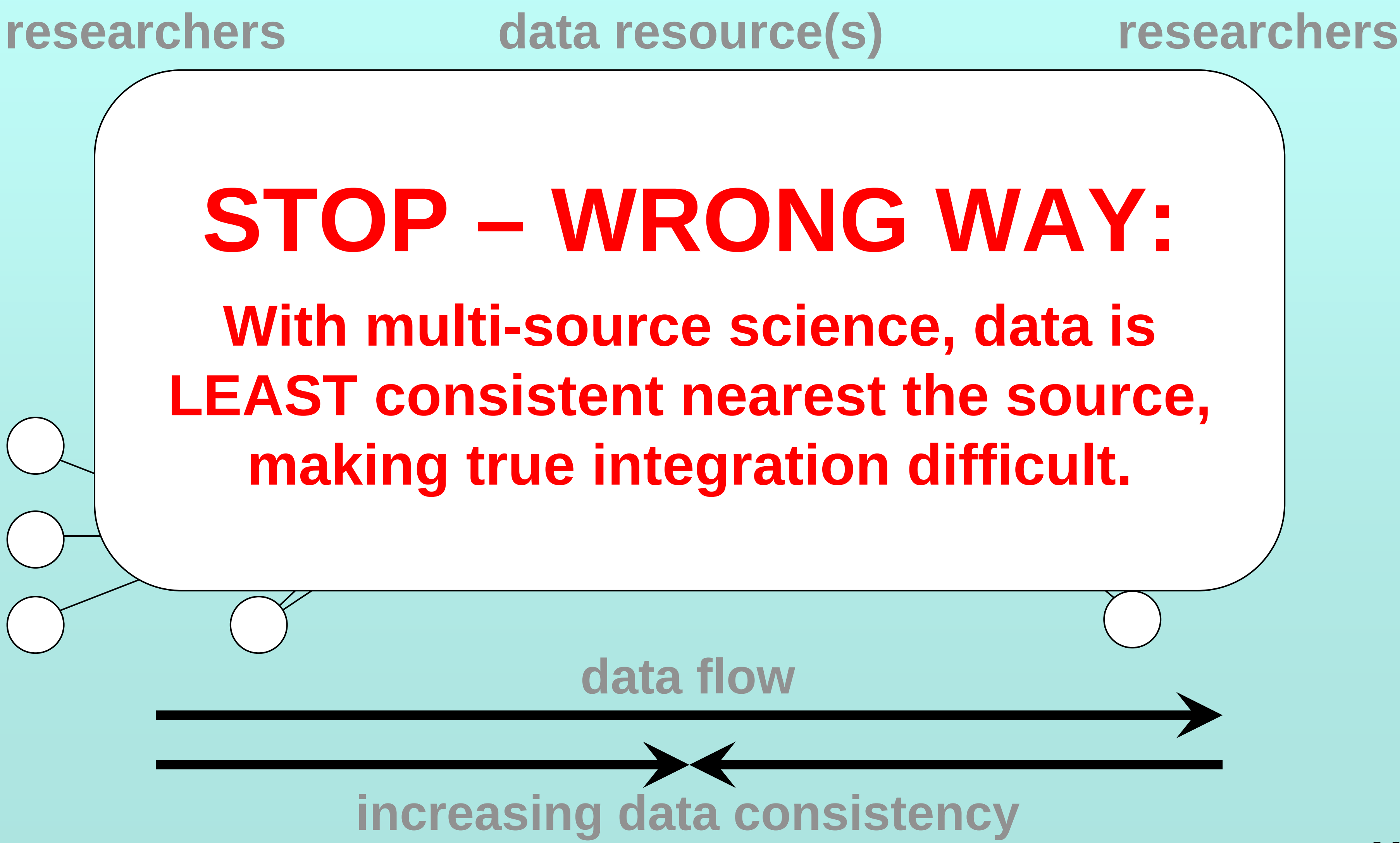
Multi-instrument Science



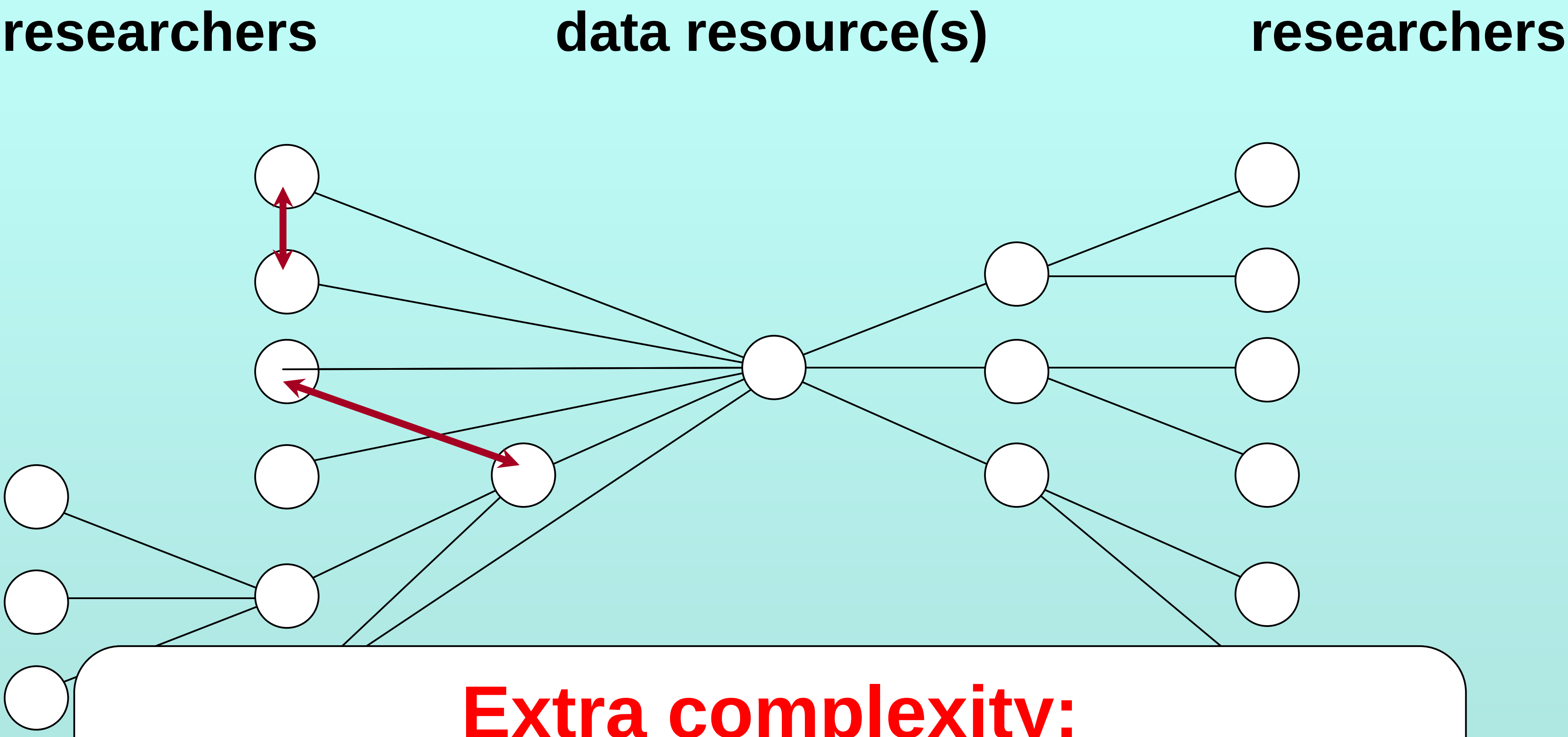
Multi-instrument Science



Multi-instrument Science



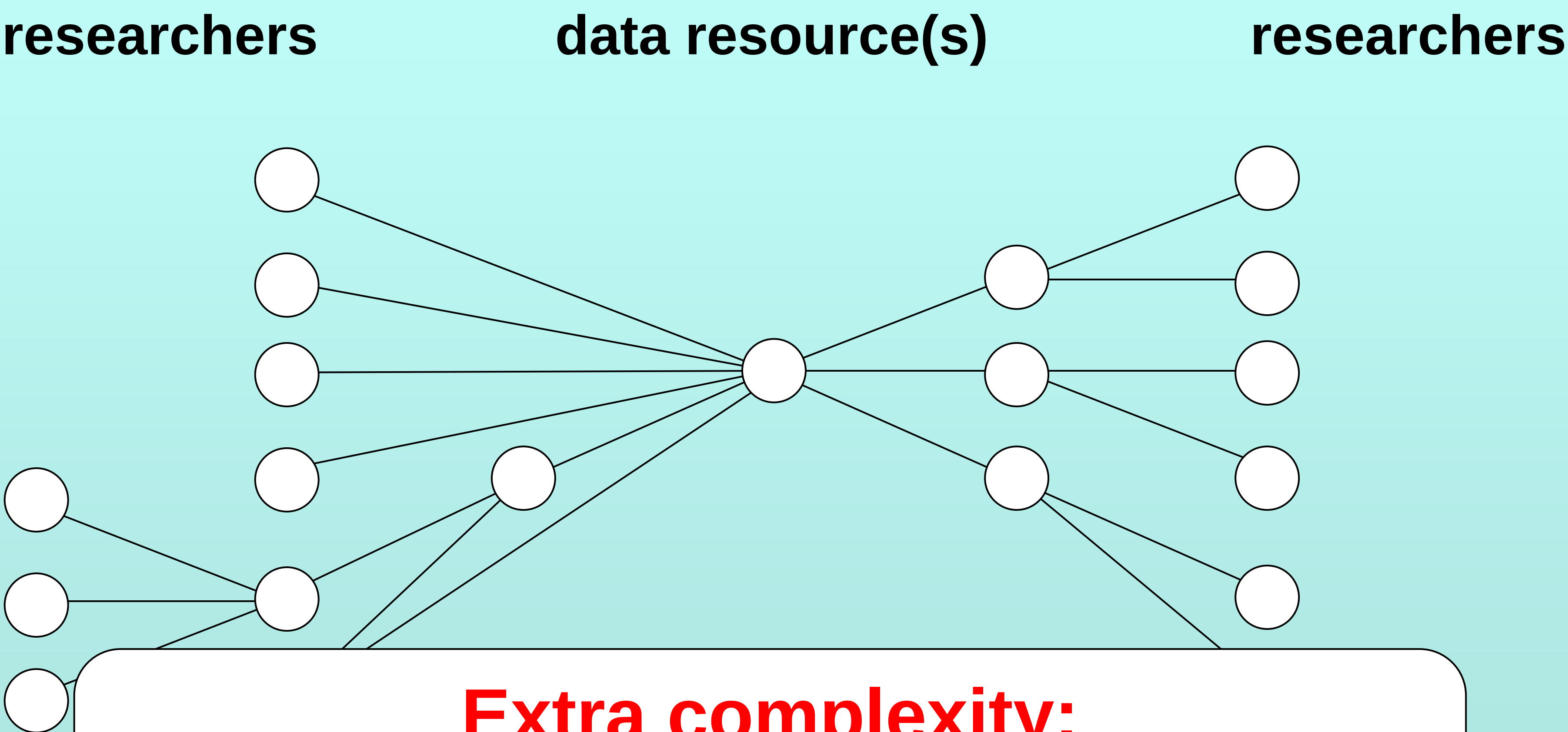
Multi-instrument Science



Extra complexity:

Undocumented, uncoordinated local data exchange

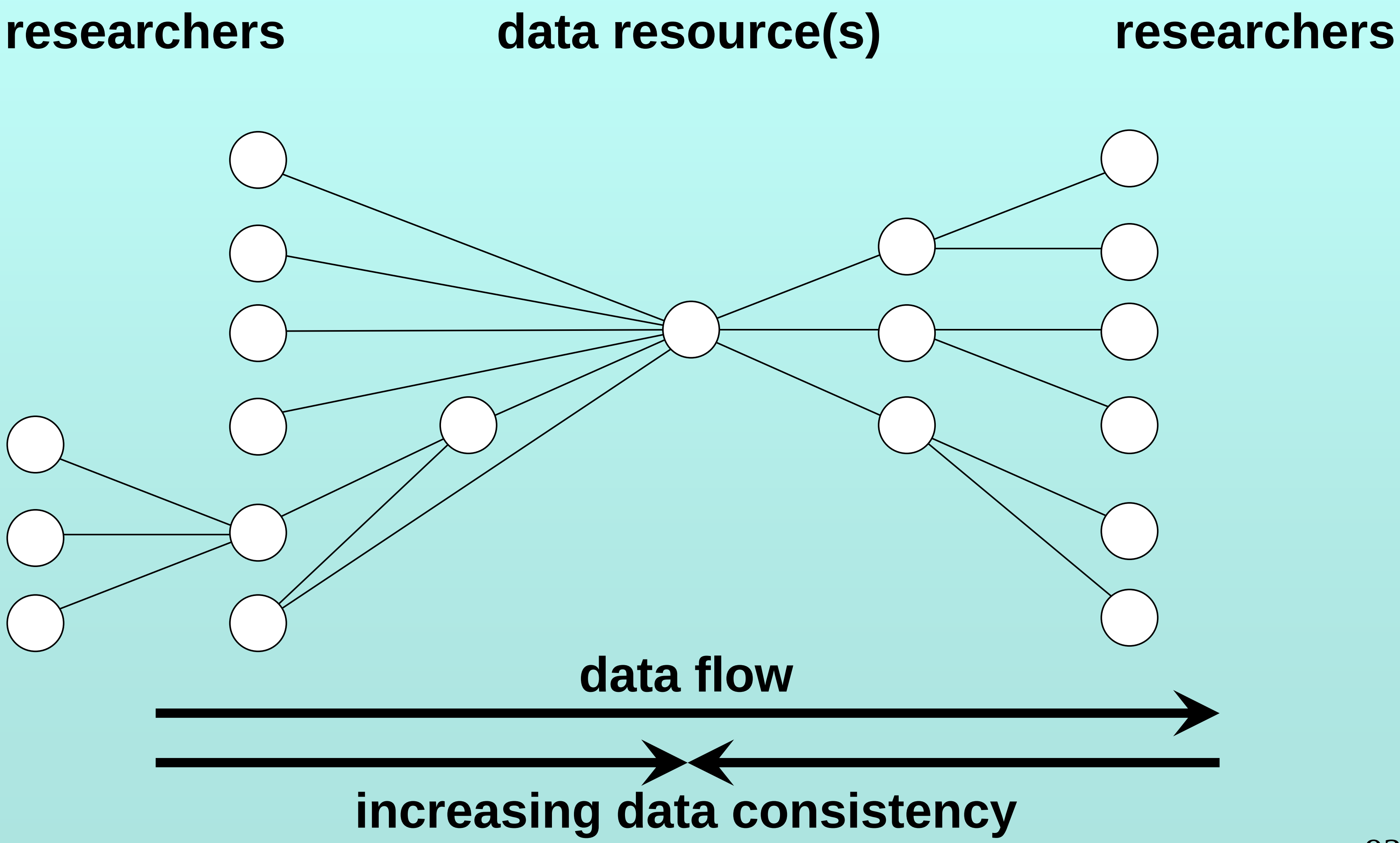
Multi-instrument Science



Extra complexity:

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

Multi-instrument Science



Source II

Scope

Data-source Scope Issues

Problems occur at many levels:

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

Source III

Solution: GenBank

GenBank as a False Model

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

Source IV

Real Solutions

Data-source Solutions

Institutional Solutions:

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

Institutional Support

Strategic Planning for IT Support of Grant-funded Research

(http://www.esp.org/rjr/briite-01.pdf)

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

rrobbins@fhcrc.org
(206) 667 2920

Strategic Planning
for IT Support of
Grant-funded Research

Eh?

Strategic Planning: >= 5 years

Grant-funded: <= 5 years

rrobbins@fhcrc.org
(206) 667 2920

How can you do strategic planning for supporting grants not yet in existence at the time of planning?

How can you do strategic planning for supporting grants not yet in existence at the time of planning?

Clearly, this can be done only in a generic sense.

How can you do strategic planning for supporting grants not yet in existence at the time of planning?

Clearly, this can be done only in a generic sense.

But what is the essence of generic support for IT support of grant-funded research?

How can you do strategic planning

Is it perhaps,

CENTRALIZED SUPPORT FOR DISTRIBUTED COMPUTING

support for IT support of grant-funded research?

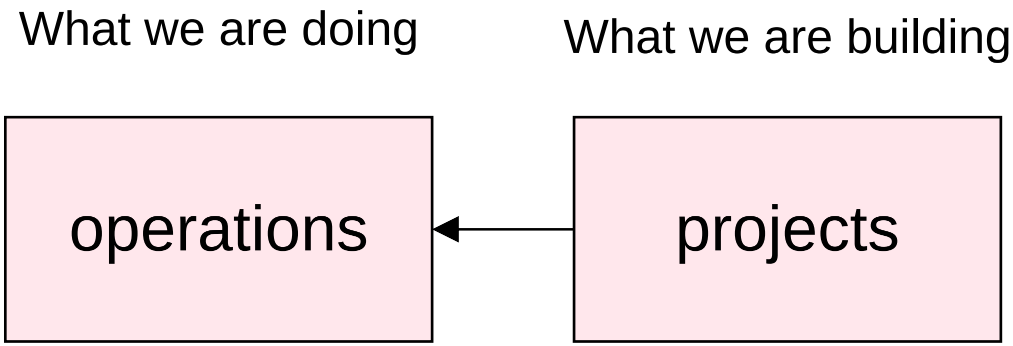
Strategic Planning for grant-funded research requires ***fourth-box*** thinking: a strategic architectural vision in response to some driving question.

Strategic Planning

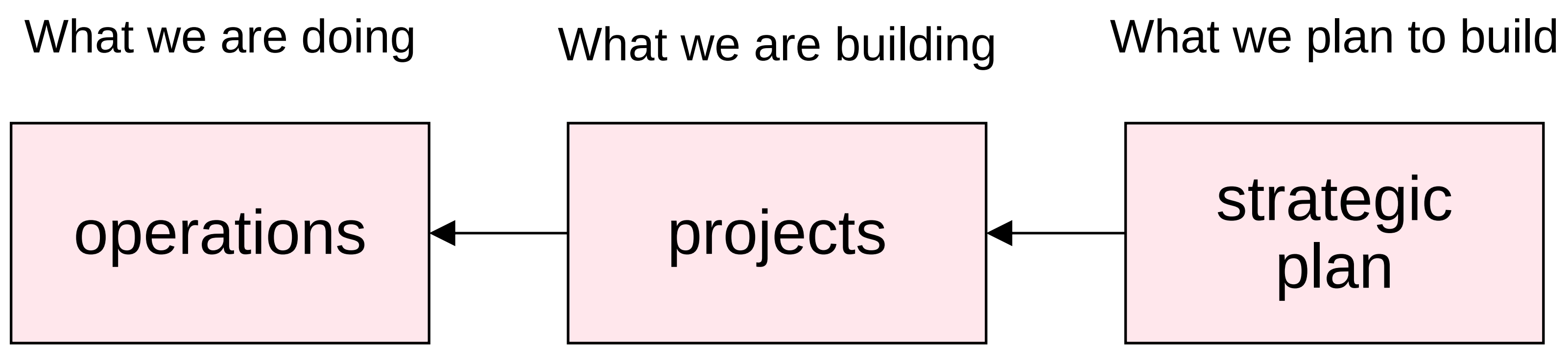
What we are doing

operations

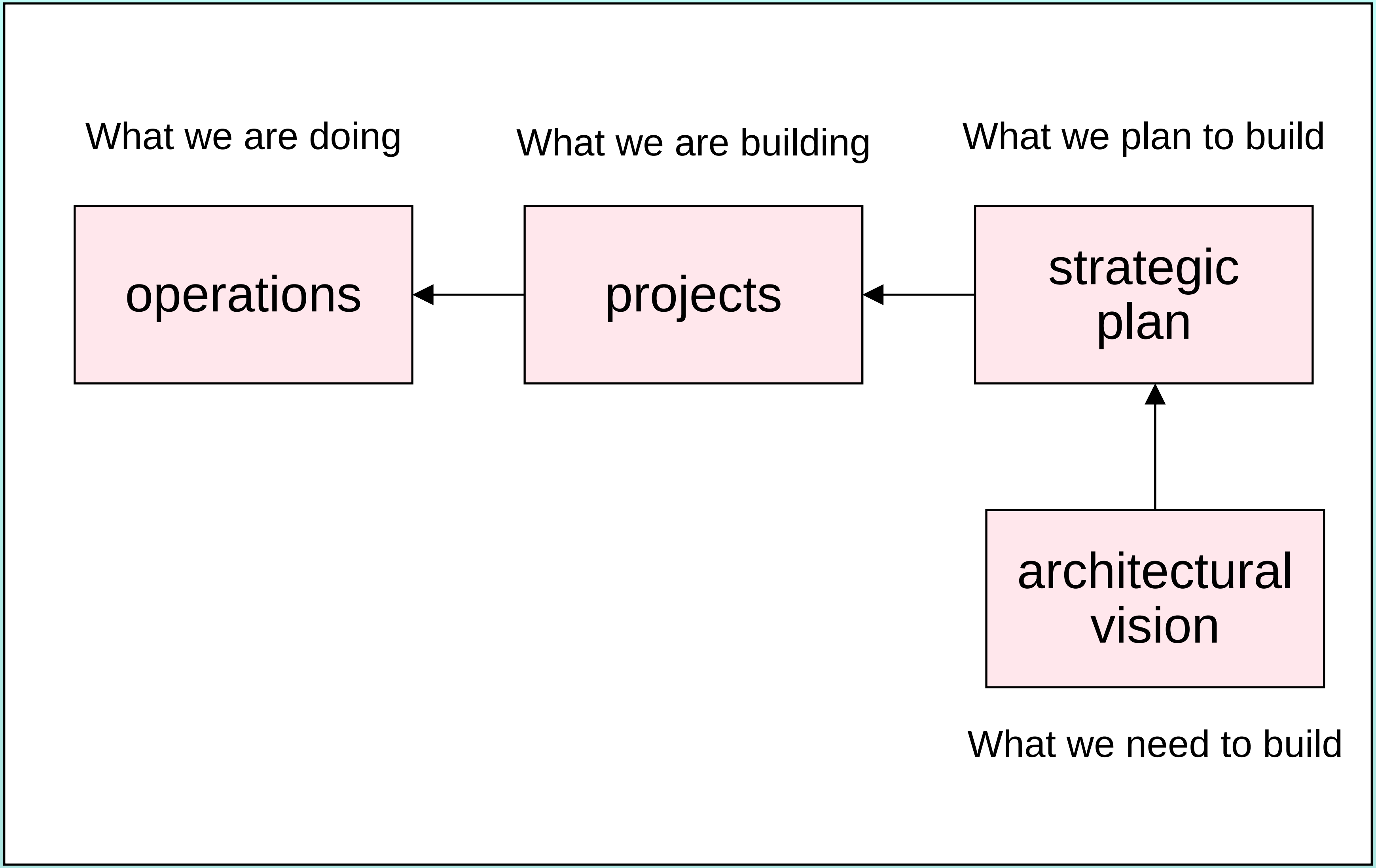
Strategic Planning



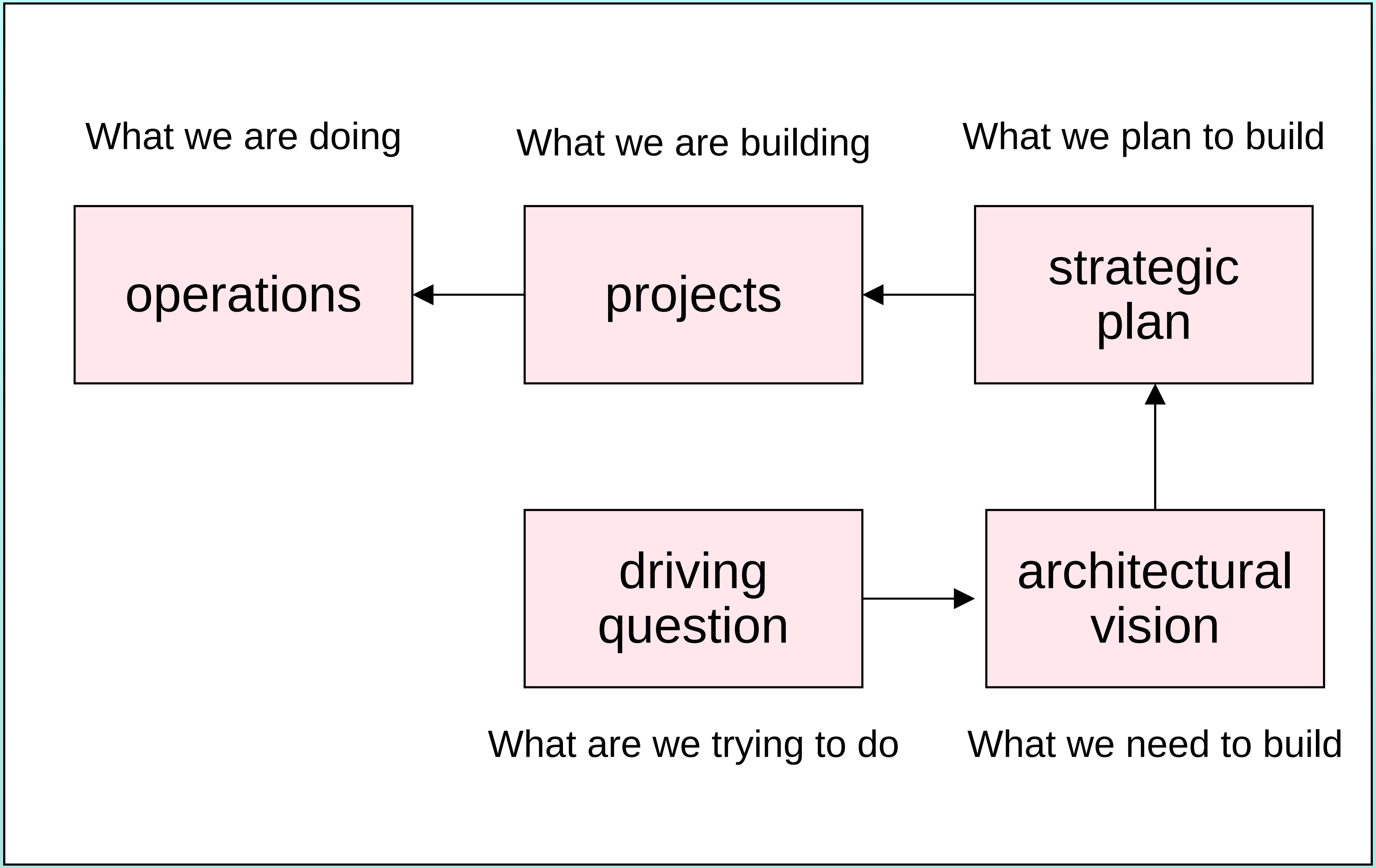
Strategic Planning



Strategic Planning



Strategic Planning



Strategic Planning

Example of important driving question:

Q: How could you design a communication system that will continue to function, even when pieces have been totally destroyed?

Strategic Planning

Example of important driving question:

Q: How could you design a communication system that will continue to function, even when pieces have been totally destroyed?

A: ARPANET packet-switched network

Strategic Planning

Example of important driving question:

Q: How can you get different networks, using different computers and different operating systems and different network protocols to interoperate?

Strategic Planning

Example of important driving question:

Q: How can you get different networks, using different computers and different operating systems and different network protocols to interoperate?

A: TCP / IP (the INTERNET)

Strategic Planning

Example of important driving question:

Q: How could you separate business logic from the technical manipulation of the contents of databases?

Strategic Planning

Example of important driving question:

Q: How could you separate business logic from the technical manipulation of the contents of databases?

A: The RELATIONAL MODEL of databases.

Strategic Planning

Example of important driving question:

Q: What can a biomedical institution do to maximize the effectiveness of IT at the level of individual grants?

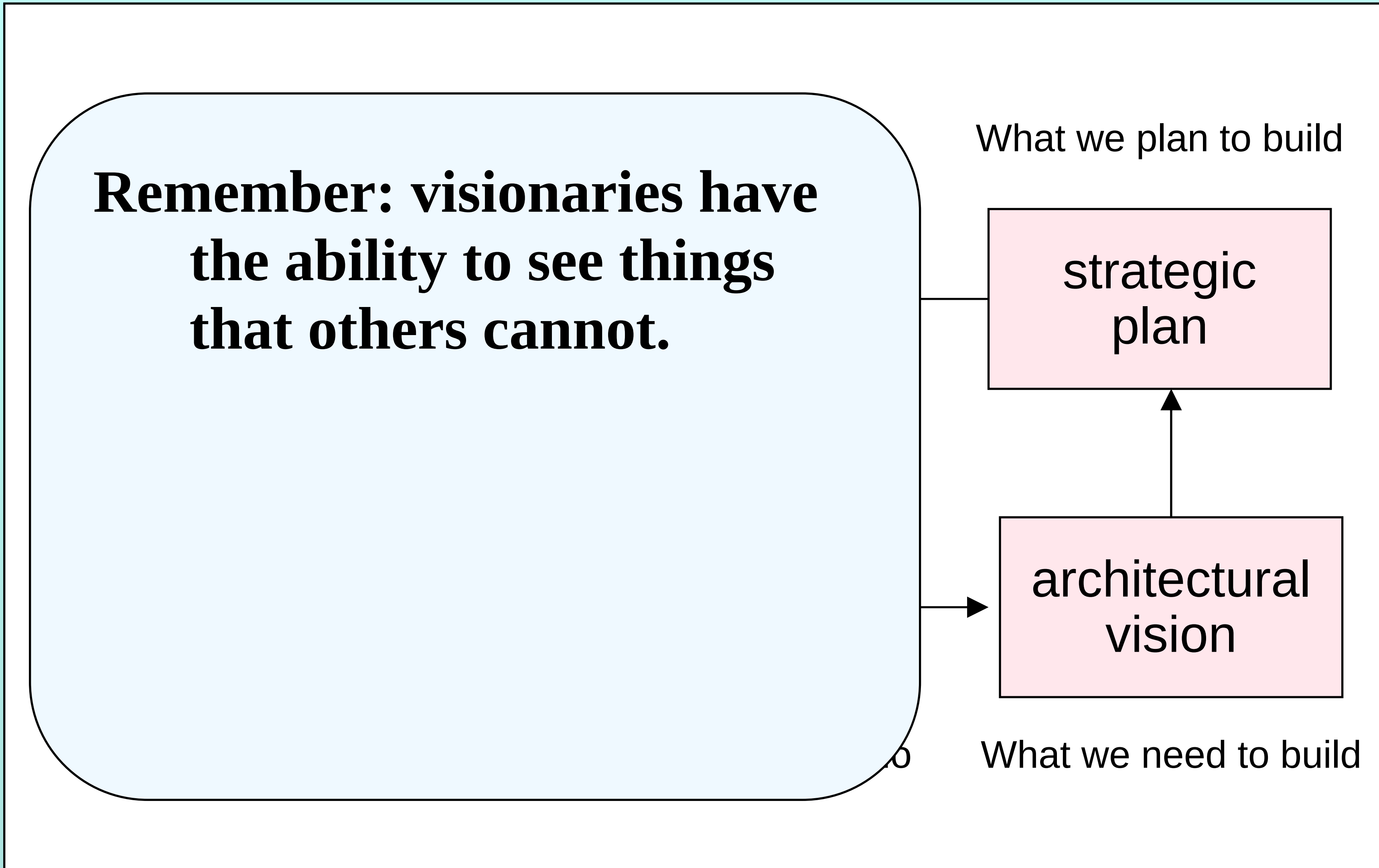
Strategic Planning

Example of important driving question:

Q: What can a biomedical institution do to maximize the effectiveness of IT at the level of individual grants?

A: That’s the question for this meeting. A strong case can be made for centralized support of distributed computing.

Strategic Planning

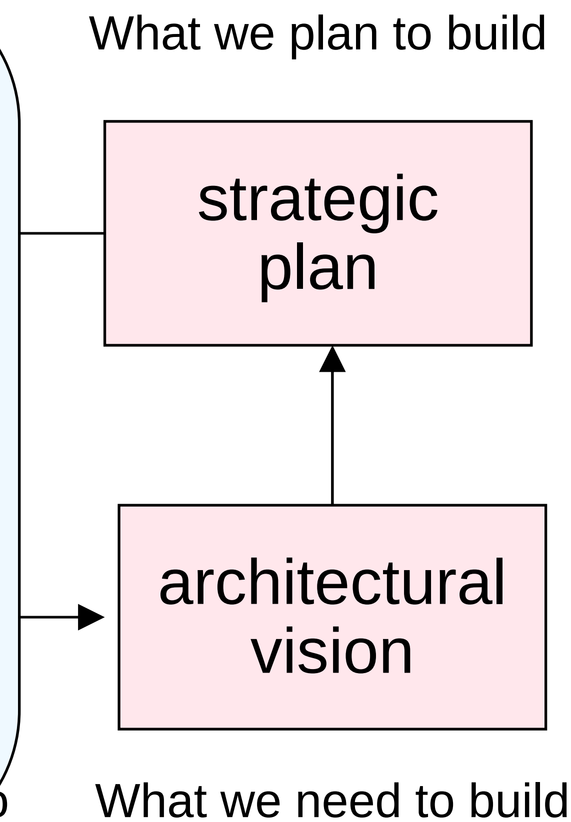


Strategic Planning

Remember: visionaries have the ability to see things that others cannot.

This is also true of those with various forms of dementia.

Expect some skepticism along the way…



Strategic Planning

TCP / IP networking and RDBMS are two of the most useful tools in the history of IT.

What can we learn from the history of their development?

Conclusions (Inferences)

- Truly valuable IT comes from a driving question, informing an architectural vision.

Conclusions (Inferences)

- Truly valuable IT comes from a driving question, informing an architectural vision.
- You must know your GOAL and handle the trade-offs accordingly.

Conclusions (Inferences)

- Truly valuable IT comes from a driving question, informing an architectural vision.
- You must know your GOAL and handle the trade-offs accordingly.
- The resulting architectural vision may have a NEWSPEAK flavor.

Conclusions (Inferences)

- Truly valuable IT comes from a driving question, informing an architectural vision.
- You must know your GOAL and handle the trade-offs accordingly.
- The resulting architectural vision may have a NEWSPEAK flavor.
- Ultimately, the results are stunning in their power, flexibility, and extensibility.

TCP/IP & RDBMS Pattern

- Formulate driving question
- Develop vision of what might be
- Explore logical consequences of vision
- Prototype
- Expand/extend/revise vision
- Prototype
- Repeat…

Patience is a Virtue

Internet Time:

- A sustained explosion of growth and technical innovation…

Patience is a Virtue

Internet Time:

- A sustained explosion of growth and technical innovation…
- after 35 years of patient, painstaking planning, testing, and development.

Patience is a Virtue

Internet Time:

- A sustained explosion of growth and technical innovation…
- after 35 years of patient, painstaking planning, testing, and development.

Conceptually, packet-switched networking began in 1960; the idea of internetworking was created in the 1970s; the whole thing took off in 1995…

BRIITE Challenge

BRIITE Challenge

- Confirm driving question
- Begin to plan architectural vision
- Identify possible components
- Describe ideal functions of components
- Imagine how functions might be achieved
- Assess how design might affect function
- Consider how components might interact
- Repeat as necessary

Possible Modules

Possible Modules

- Basic Infrastructure
- Authorization, Authentication, Auditing
- Digital Publishing Support
- Scientific Database I: Data Models & Design
- Scientific Database II: Data Integration
- Scientific Database Support III: Community Databases
- Scientific Database Support IV: Public dB Integration

Possible Modules

- Clinical Research I: Research Access to Clinical Data
- Clinical Research II: Research Trials
- Clinical Research III: Controlled Vocabularies
- Clinical Research IV: Specimen Management
- Clinical Research V: Tumor / Disease Registries
- Laboratory Information Management Systems
- Shared Resource Support

Possible Methods

- Top down: ideal solutions
- Bottom up: current problems
- Iterative: both, back and forth…

Top-down Example

Authorization, Authentication, etc.

Every administrator of a computer resource needs some way to identify users, to authorize them to access the resource, to authenticate them when they access the resource, and to log and audit them when they use the resource. In a typical academic environment, there are many, many different approaches to handling these tasks.

What if, once upon a time in the future, there were to be a system called GLAAAS…

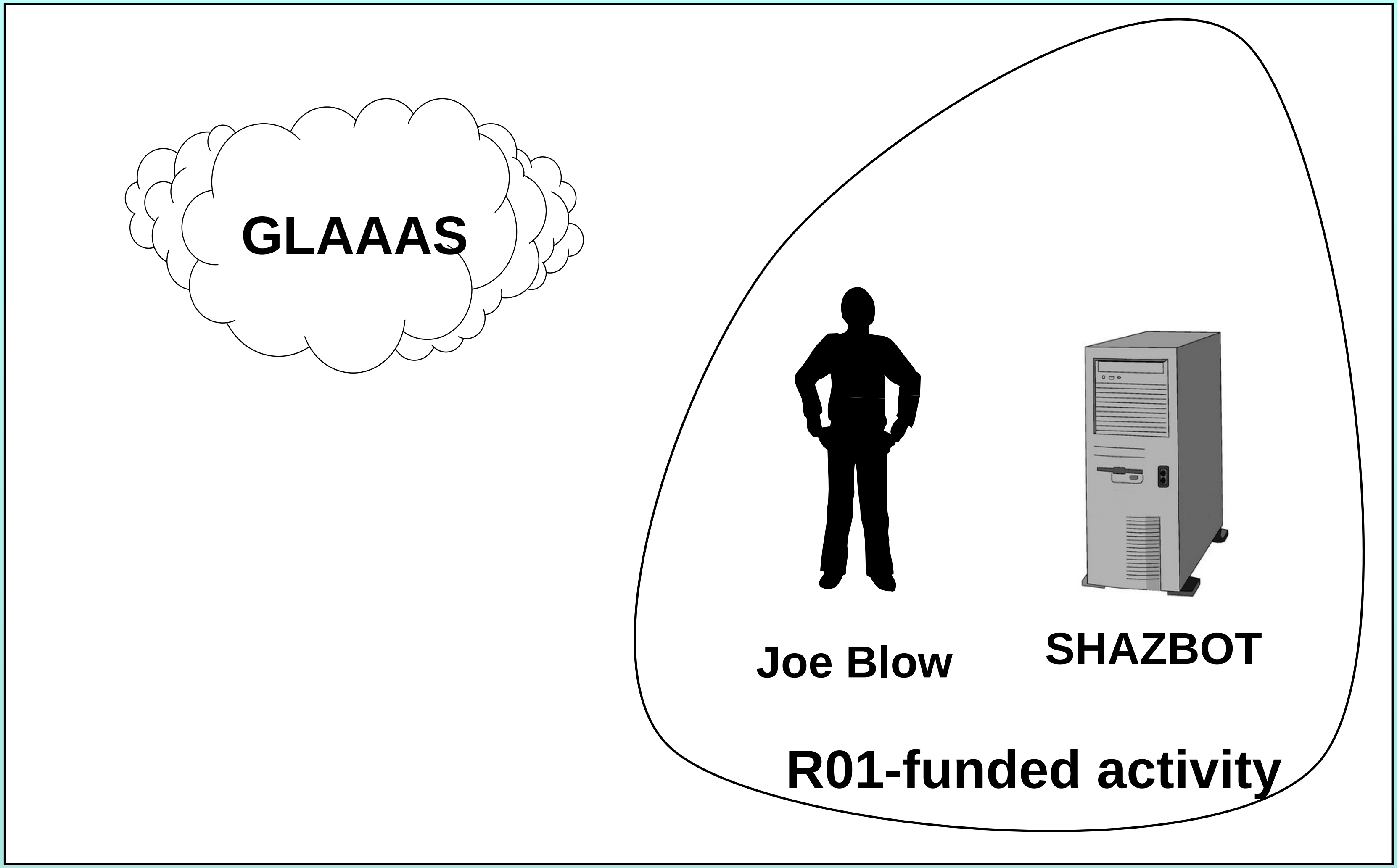
GLAAAS

GLAAAS is a GLobal Authorization, Authentication, and Auditing System that can be used to assign, track, and audit permissions to use IT resources on any server that participates in GLAAAS.

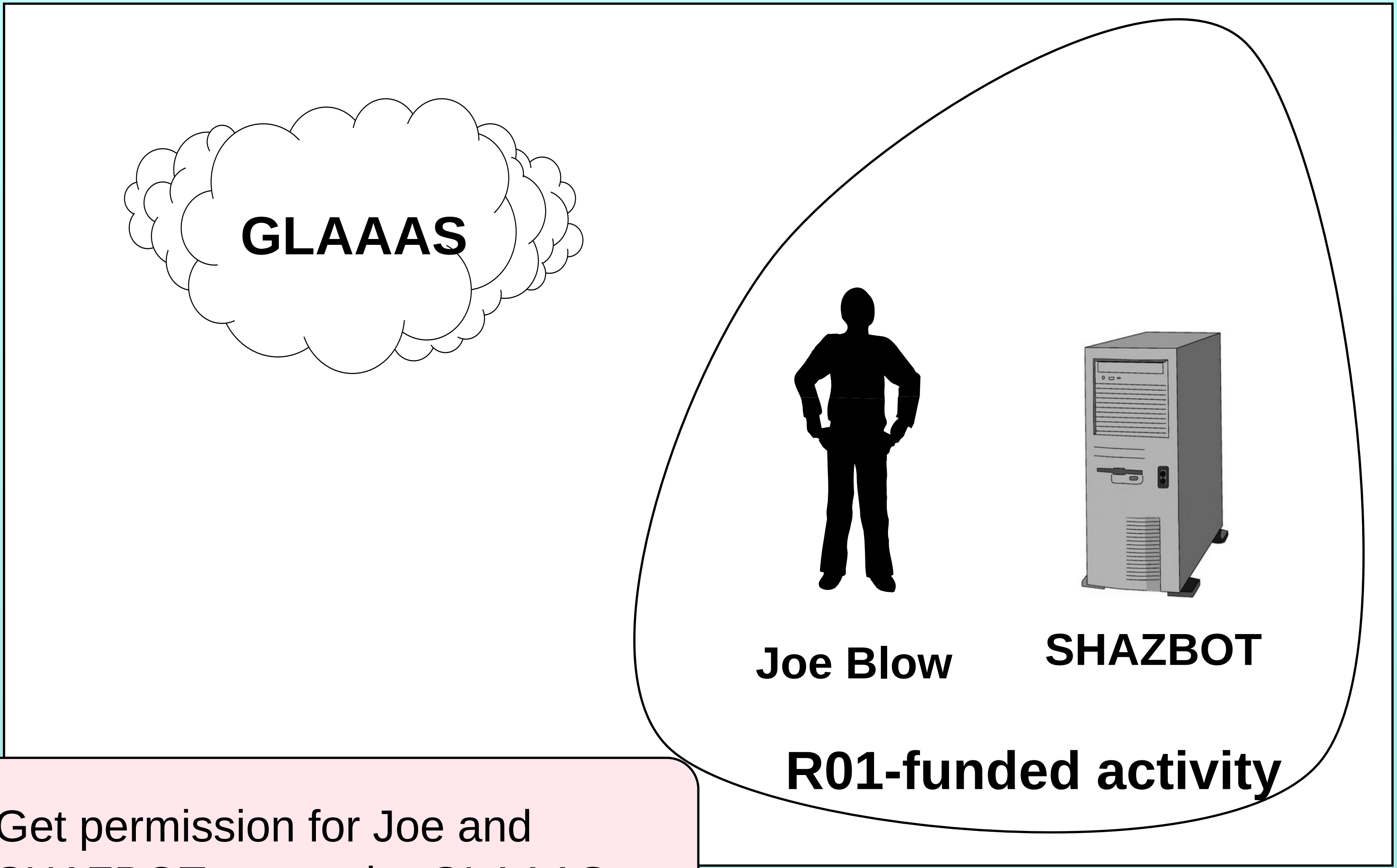
GLAAAS works with any operating system and makes almost no demands on the configuration of any participating server.

GLAAAS provides gPAMs (general pluggable authentication modules) and gPLMs (general pluggable logging modules) to all participating servers.

GLAAAS

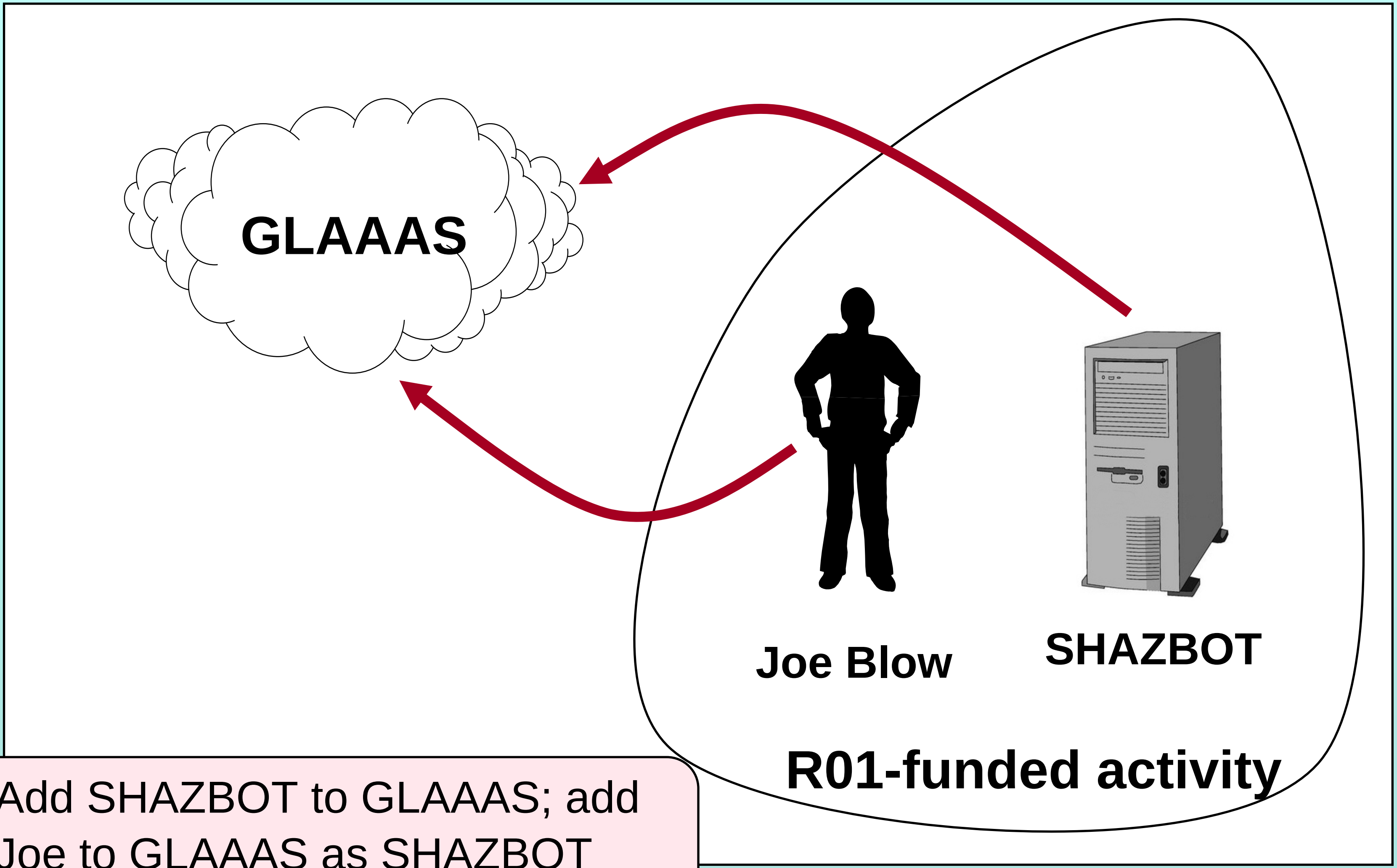


GLAAAS



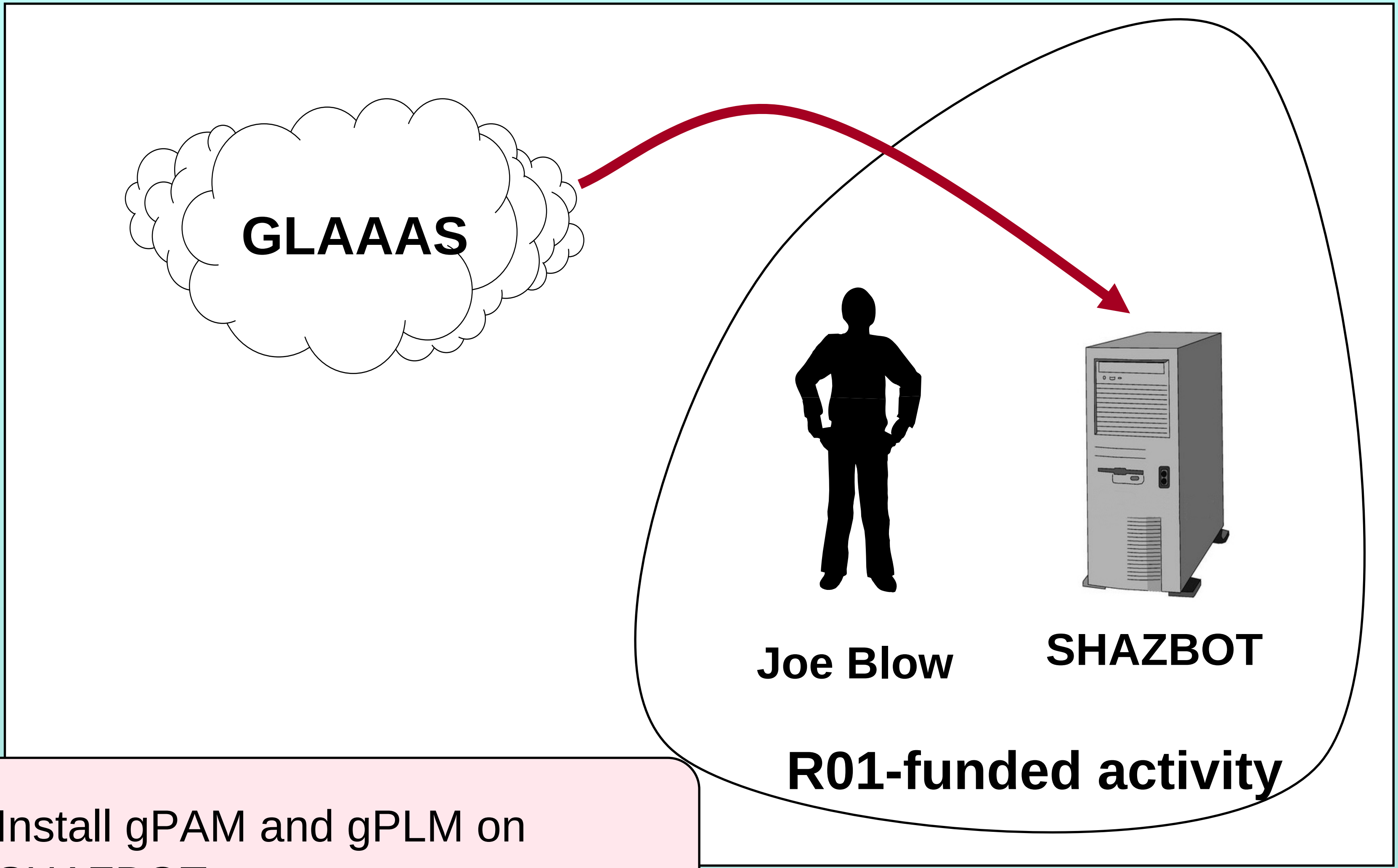
Get permission for Joe and SHAZBOT to use the GLAAAS.

GLAAAS



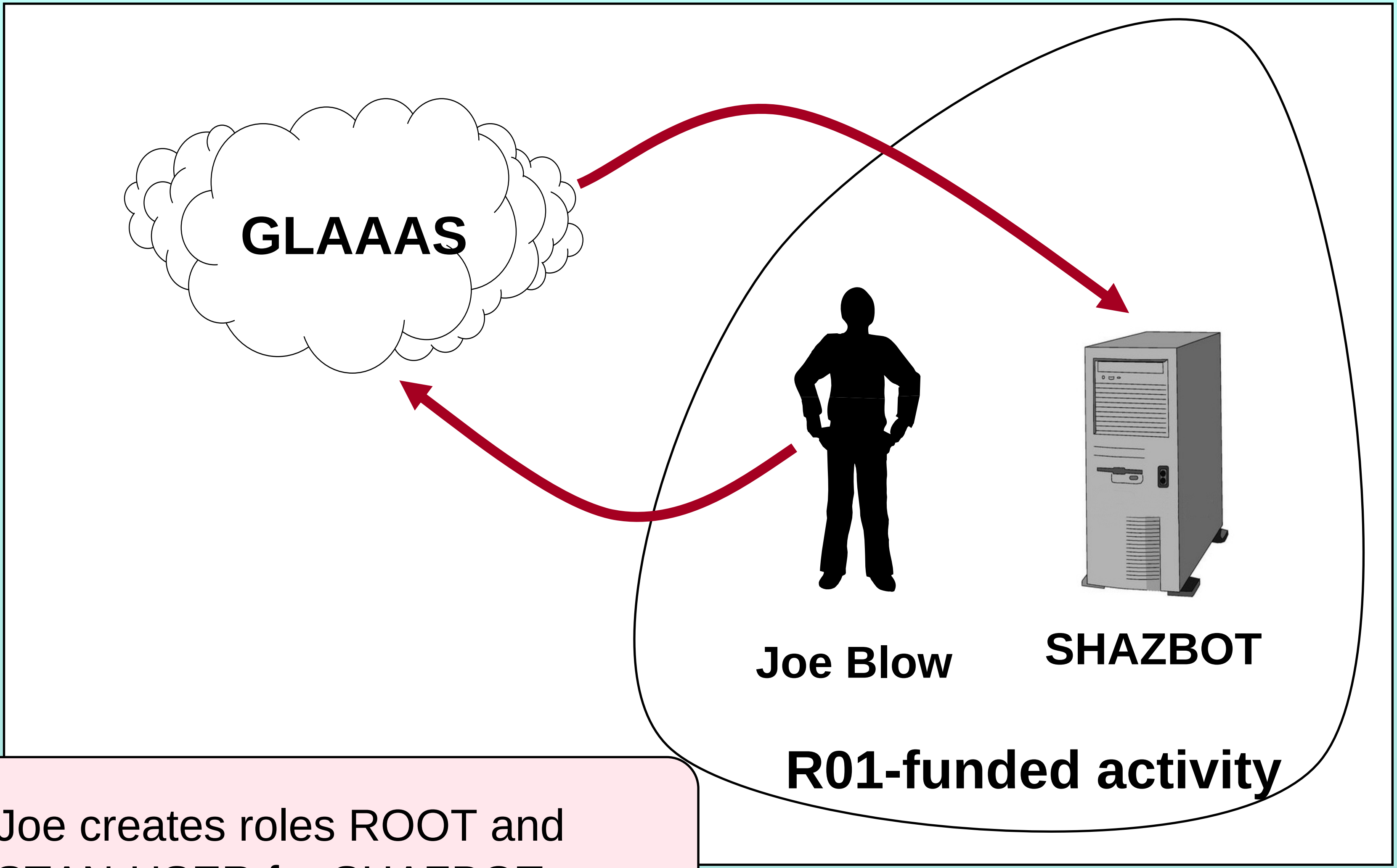
Add SHAZBOT to GLAAAS; add Joe to GLAAAS as SHAZBOT admin.

GLAAAS



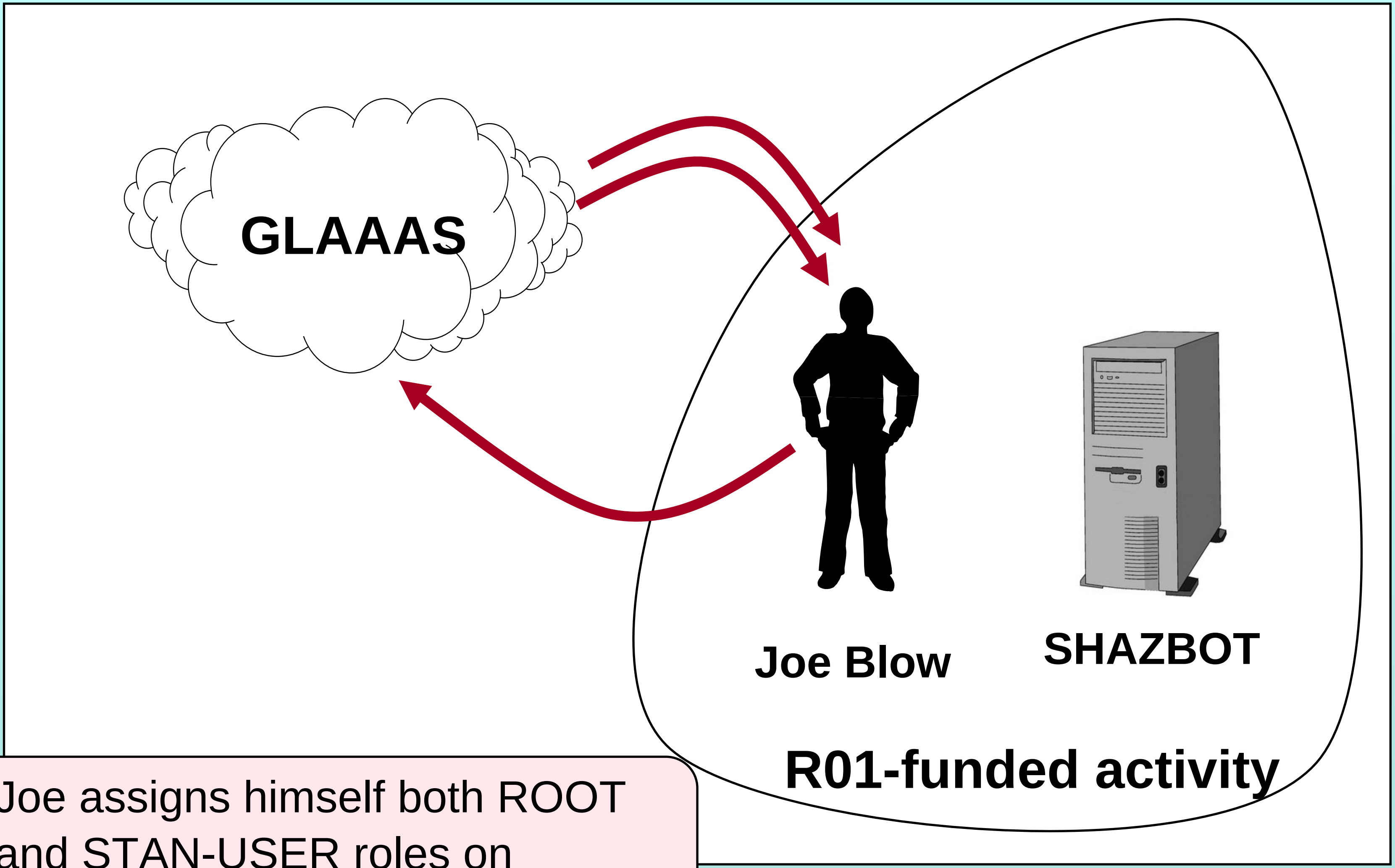
Install gPAM and gPLM on SHAZBOT.

GLAAAS



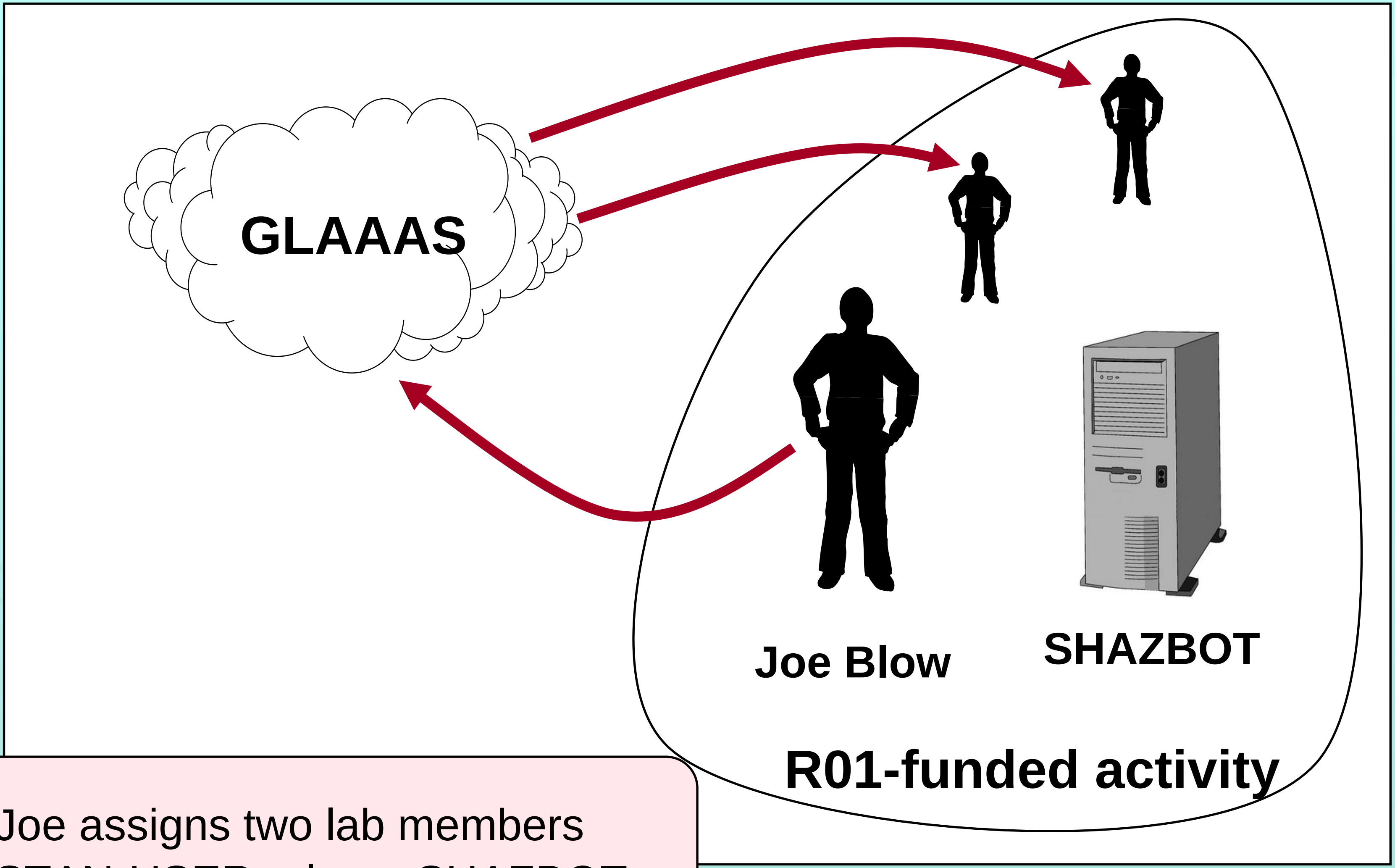
Joe creates roles ROOT and STAN-USER for SHAZBOT.

GLAAAS



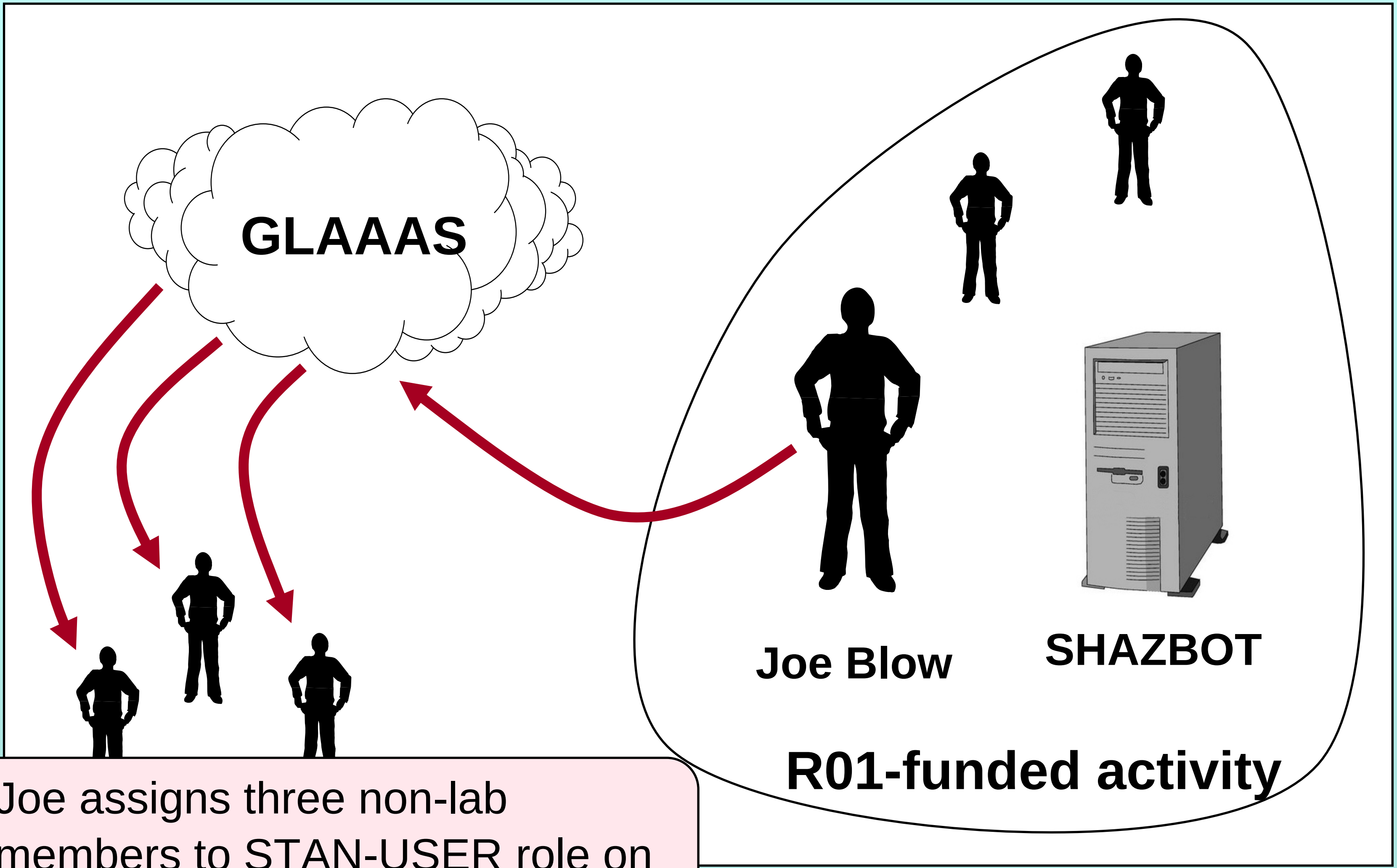
Joe assigns himself both ROOT and STAN-USER roles on SHAZBOT.

GLAAAS



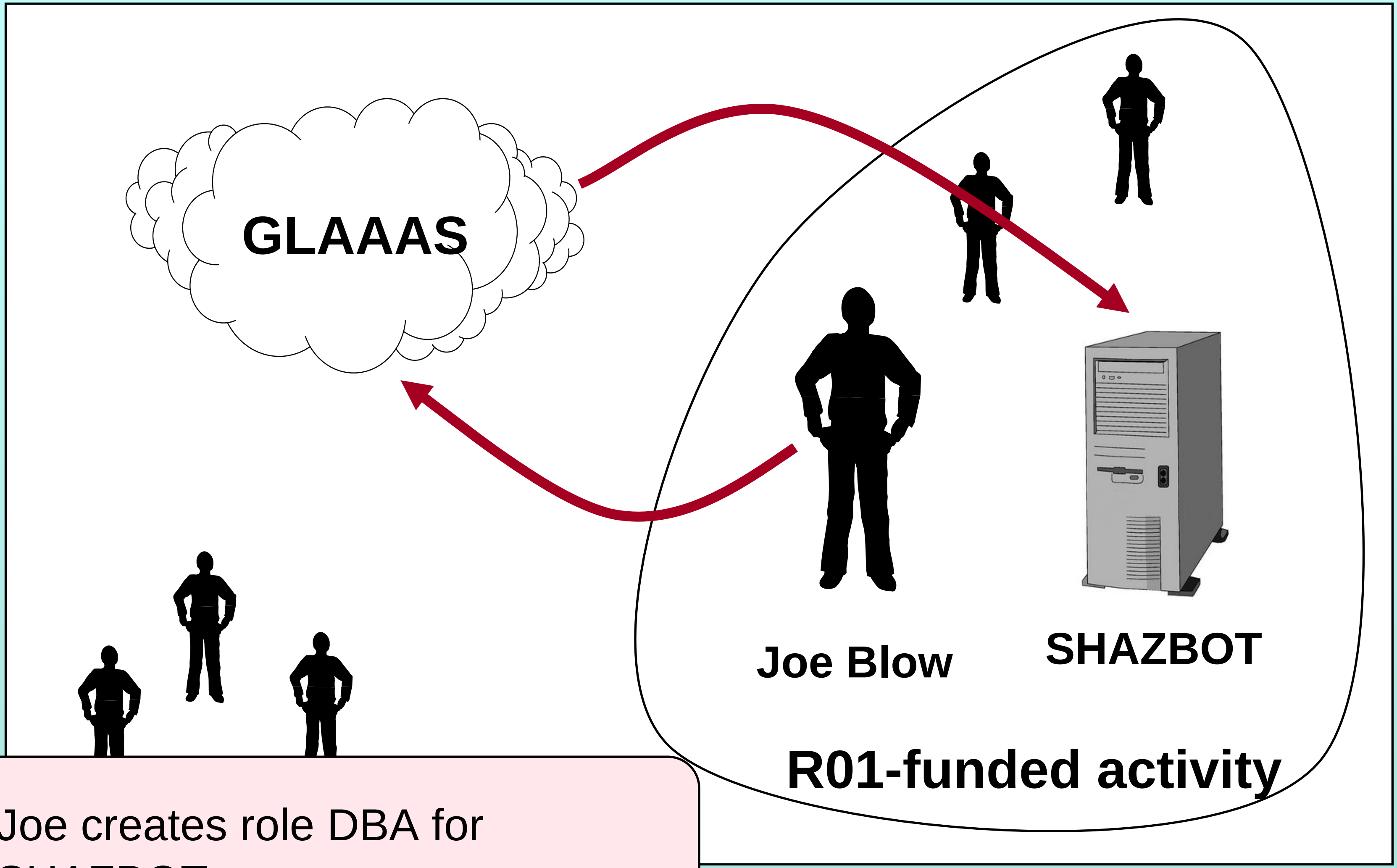
Joe assigns two lab members STAN-USER role on SHAZBOT.

GLAAAS



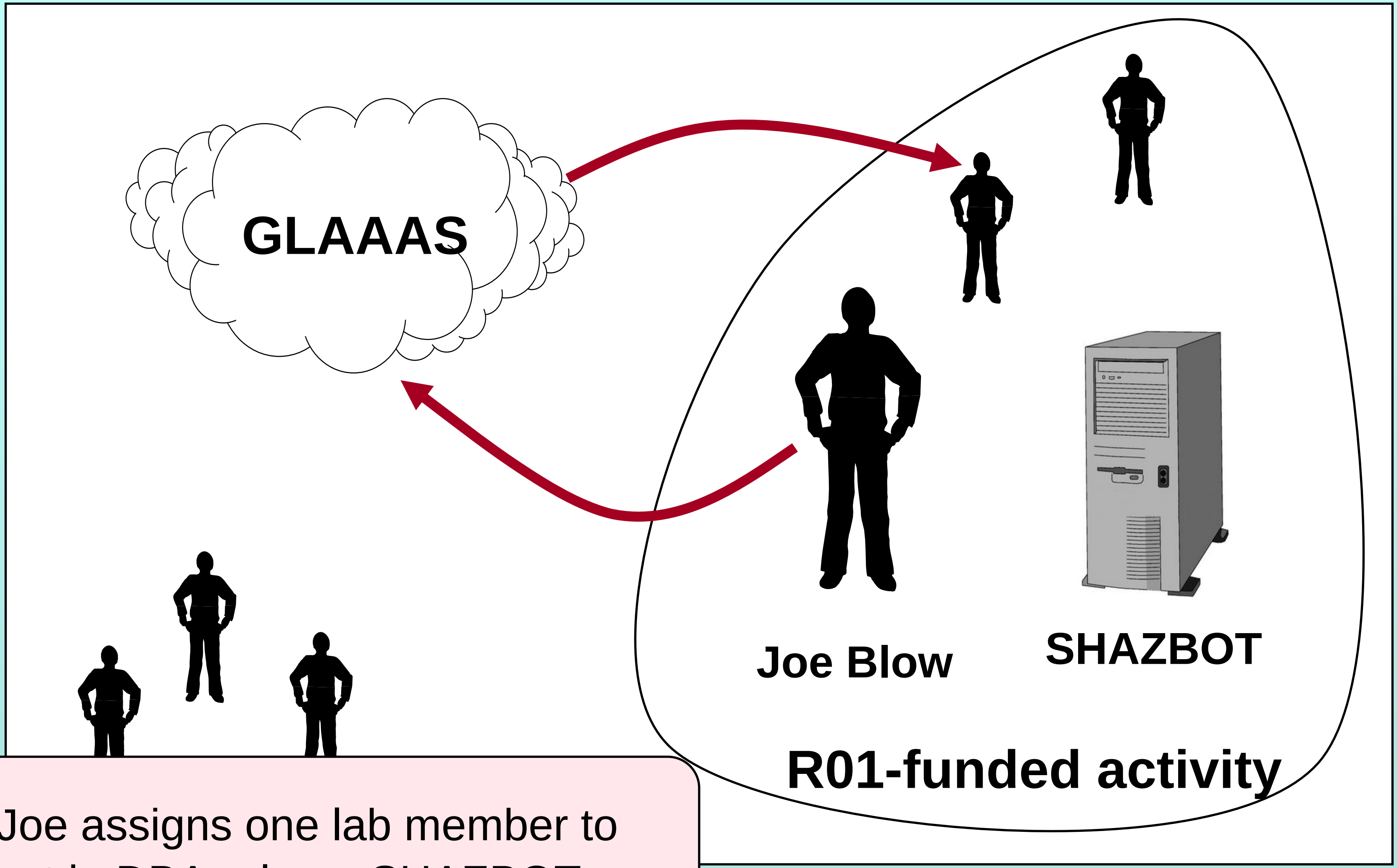
Joe assigns three non-lab members to STAN-USER role on SHAZBOT.

GLAAAS



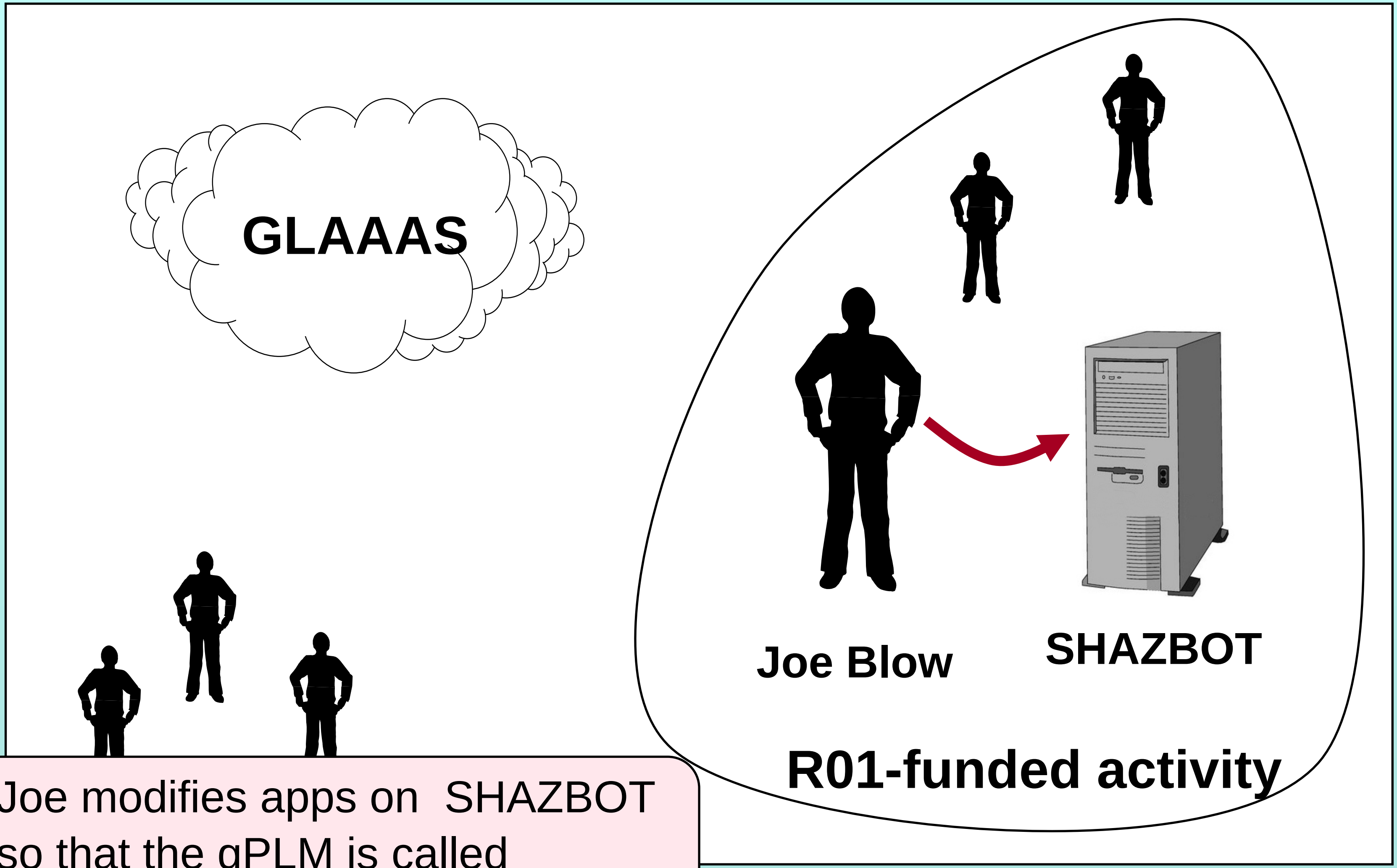
Joe creates role DBA for SHAZBOT.

GLAAAS



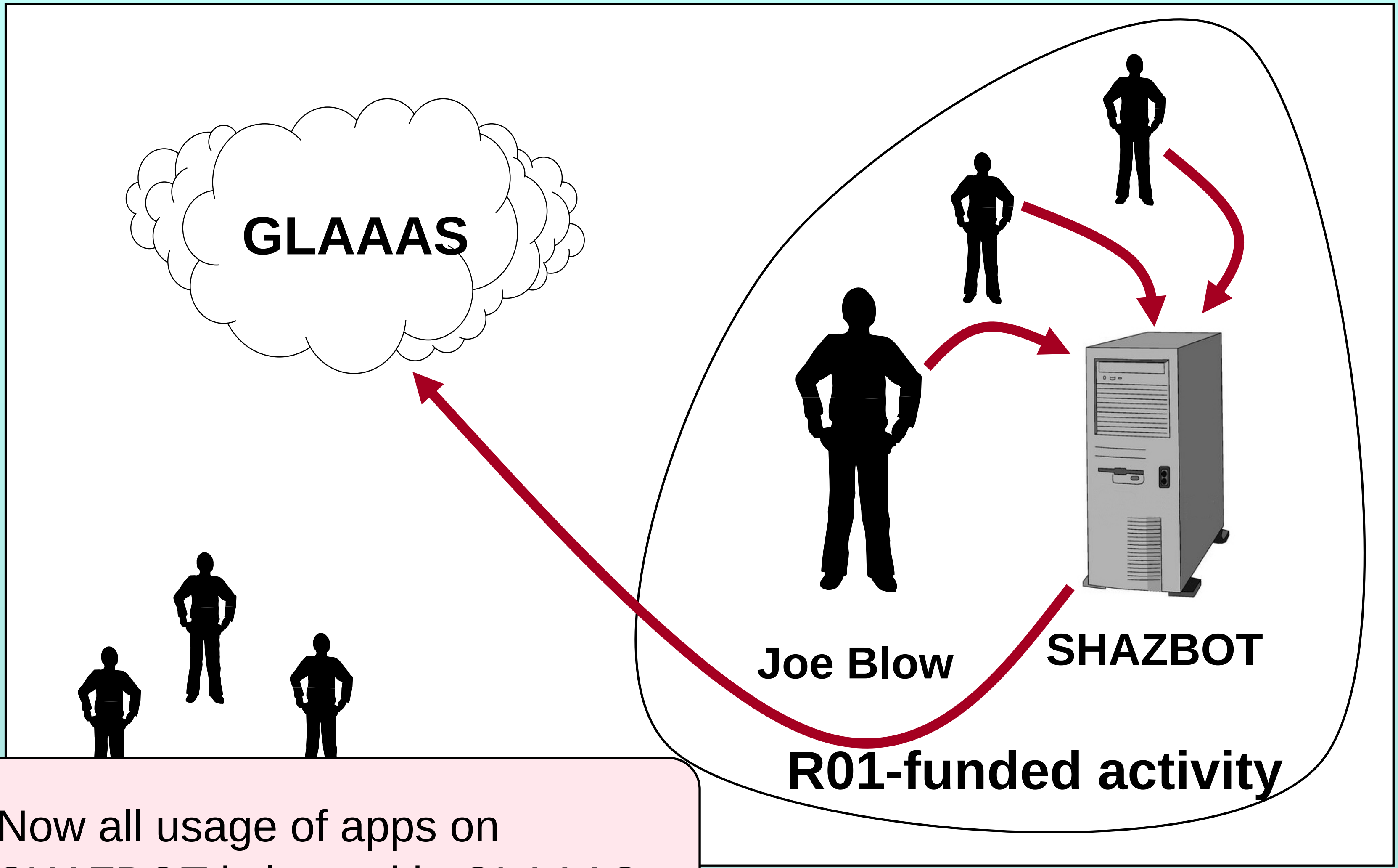
Joe assigns one lab member to act in DBA role on SHAZBOT.

GLAAAS



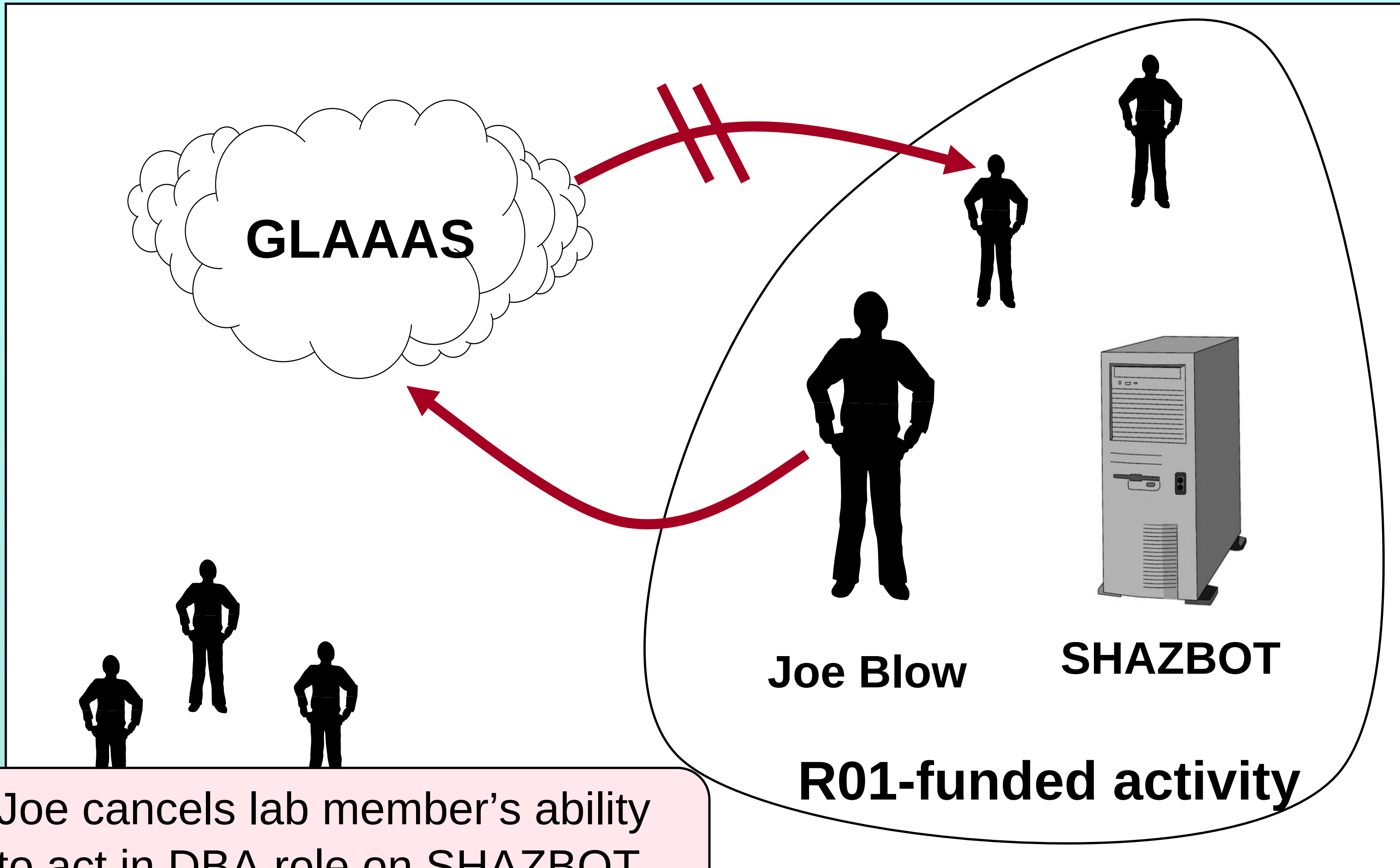
Joe modifies apps on SHAZBOT so that the gPLM is called frequently.

GLAAAS



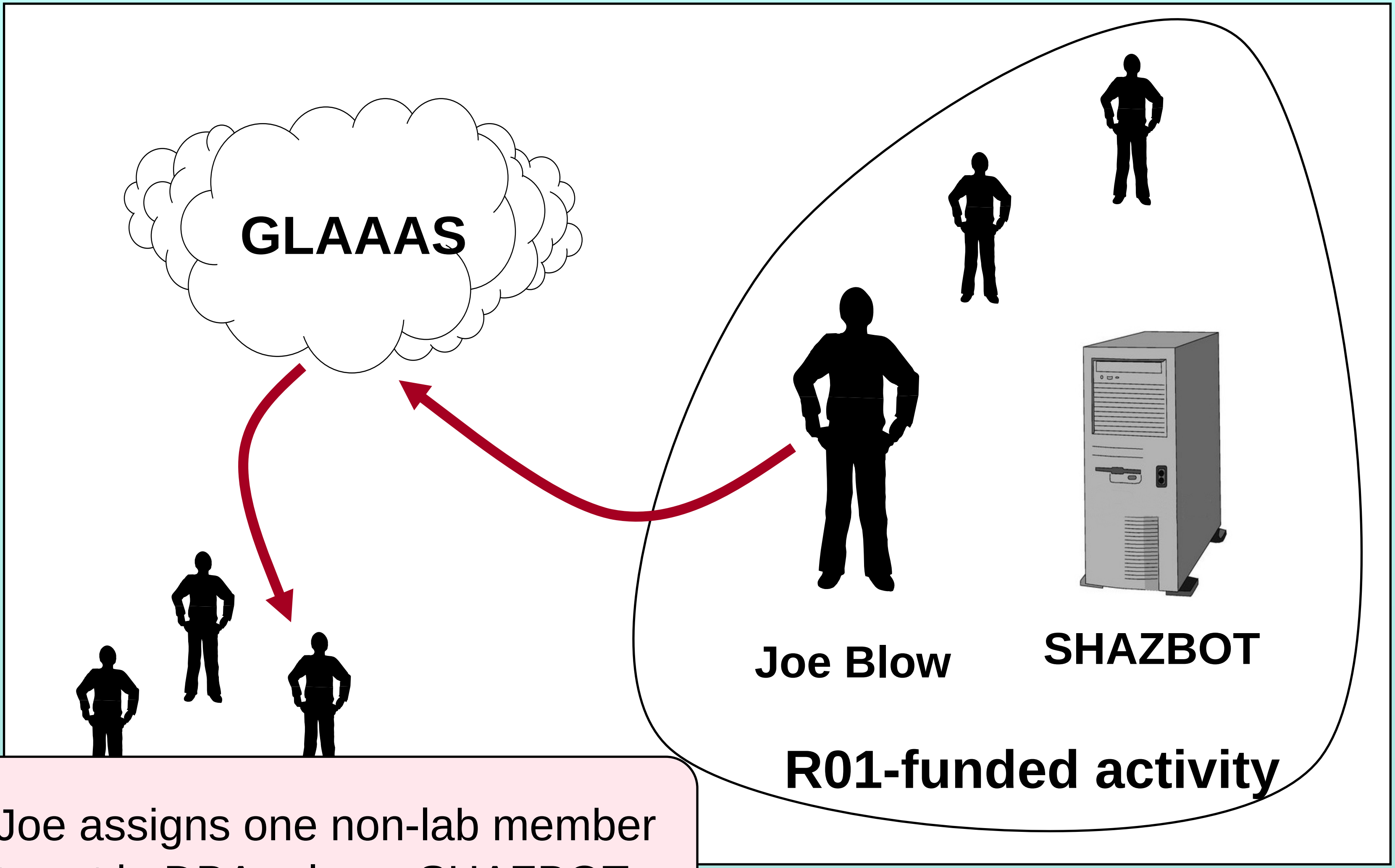
Now all usage of apps on SHAZBOT is logged in GLAAAS.

GLAAAS



Joe cancels lab member's ability to act in DBA role on SHAZBOT, then …

GLAAAS



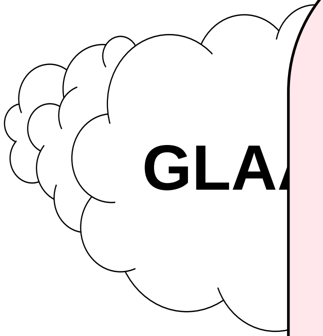
Joe assigns one non-lab member to act in DBA role on SHAZBOT.

GLAAAS

All of these changes in authorization, authentication, and logging for SHAZBOT occur without any USER having to make any changes to his/her account and without any effect on the user's permissions or access on any other system.

USERs assigned multiple roles on a machine can request a change to a different authorized role at any time, without having to reauthenticate. USERs can be simultaneously connected in multiple roles, if needed.

GLAAAS



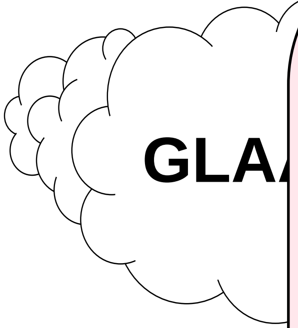
What else might GLASS do?

Provide truly GLOBAL support, by working with similar systems at other campuses?

Support the management of GROUPS of people, so that permission could be granted to the right group, but the responsibility for maintaining the group is no longer the system administrator’s?

….?

GLAAAS



Technically, how might GLASS actually work?

....?

Slides:

http://www.esp.org/rjr/nist2003.pdf