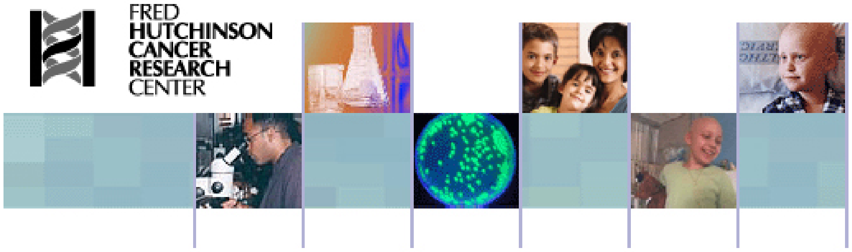
FRED HUTCHINSON CANCER RESEARCH CENTER



Advancing Knowledge, Saving Lives

BRIITE

Biomedical Research Institutions Information Technology Exchange

Semi-Annual Meeting
10-12 September 2003
Seattle, WA

Strategic Planning for IT Support of Grant-Funded Research

BRIITE: Biomedical Research Institutions Information Technology Exchange

10-12 SEP 2003, Seattle, WA

BRIITE Mission

BRIITE is an informal organization of research IT leaders from a number of biomedical research institutions. The mission of BRIITE is to facilitate the successful deployment of IT in support of research by:

- establishing personal contacts among those responsible for research computing activities at biomedical research institutions
- identifying and documenting common problems and interests
- seeking opportunities for partnership / consortium activities
- identifying common issues that should be brought to the attention of home institutions, of government agencies, and of funding or regulatory organizations

Meeting Goals

Information technology plays an essential role in almost all biomedical research. To be competitive, biomedical research organizations must provide access to key information technology as part of the institutional infrastructure available to their researchers. At the same time, individual investigators must be allowed maximum flexibility in the pursuit of their research. Determining how best to accomplish these sometimes conflicting goals will be the topic of this meeting – *Strategic Planning for IT Support of Grant-Funded Research*.

We will begin with a general discussion of the overall challenge, then turn our attention to a few individual topics which will be discussed in plenary session. Smaller working groups will then convene to elaborate on individual topics. We expect that interest in these issues will continue beyond the meeting itself, and that some working groups will continue their assessments so that the analyses begun at the meeting can ultimately be summarized in a collection of white papers.

Schedule – Day 0

September 10 (WED)

2:00 pm Steering Committee Meeting
Room J4-102, Yale Building, Fred Hutchinson Cancer Research Center

6:00 pm Reception & Dinner
Lowell-Hunt Catering, 1111 Fairview Ave, North

SCHEDULE – DAY 1

SEPTEMBER 11 (THU)

8:00 am	Continental breakfast *Sze Conference Room, Thomas Building, Fred Hutchinson Cancer Research Center*
8:15 am	Welcoming Comments – Introduction to BRIITE
8:30 am	IT support for grant funded research: Strategic issues *Robert Robbins, VP/IT, Fred Hutchinson Cancer Research Center*
9:30 am	PLENARY DISCUSSION: Who's using our systems: identity management, authorization, authentication, and usage logging *RL "Bob" Morgan, Sr. Technology Architect, C&C, University of Washington*
10:30am	BREAK
11:00 am	PLENARY DISCUSSION: Digital publishing support (web site development, data distribution, information for patients, study results dissemination) *Robert Robbins, VP/IT, Fred Hutchinson Cancer Research Center*
12:15 pm	LUNCH
1:30 pm	PLENARY DISCUSSION: Scientific data management – why is it so hard? *Nat Goodman, Sr. Research Scientist, Institute for Systems Biology*
2:30 pm	PLENARY DISCUSSION: Research access to clinical data *William B. Lober, M.D., Division of Biomedical and Health Informatics, UW*
3:30 pm	BREAK
4:00 pm	Working Group Sessions (self-organizing) examples: Identity Management Digital Publishing Support Scientific Data Management Research Access to Clinical Data
6:00 pm	Adjourn to hotel
6:30 pm	walk to dinner location
7:00 pm	Reception, Dinner *Daniel's Broiler, Seattle*
8:15 pm	PLENARY ADDRESS: caBIG, the Cancer Bioinformatics Informatics Grid *Ken Buetow, Director, NCICB, NCI, NIH*

Schedule – Day 2

September 12 (Fri)

8:00 am	Continental breakfast *Sze Conference Room, Thomas Building, Fred Hutchinson Cancer Research Center*
8:30 am	Working Group Sessions (continued) examples: Identity Management Digital Publishing Support Scientific Data Management Research Access to Clinical Data
10:00am	BREAK
10:30 am	Reports from the Working Group Sessions *speakers chosen by groups*
12:15 pm	LUNCH
1:30 pm	Information Technology Exchange *(Identify things that worked/things that didn't - brief presentations)*
3:00 pm	BREAK
3:30 pm	Brief tour of FHCRC core IT facilities *FHCRC staff*
5:00 pm	Adjourn

Schedule – Day 3

September 13 (SAT) – optional events

9:00 am OPTIONAL: Business meeting; planning future BRIITE activities; continuation of strategic planning
Location to be Announced

10:30 am OPTIONAL: Walking tour of FHCRC IT facilities; visit server rooms, wiring closets, interstitial floors of research buildings; inspect new research building – 350,000 square-foot Public Health Sciences building (to be opened Jan 2004)
FHCRC staff

12:00 pm OPTIONAL: LUNCH

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

© 2003, Robert J. Robbins

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.

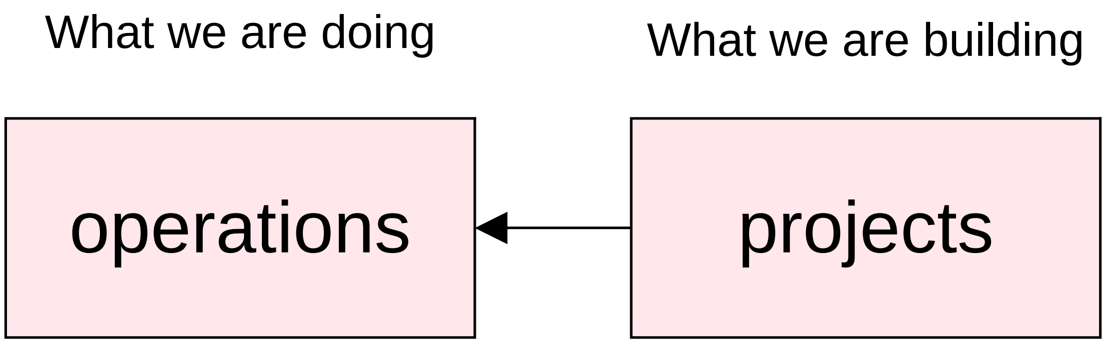
Strategic Planning

What we are doing

operations

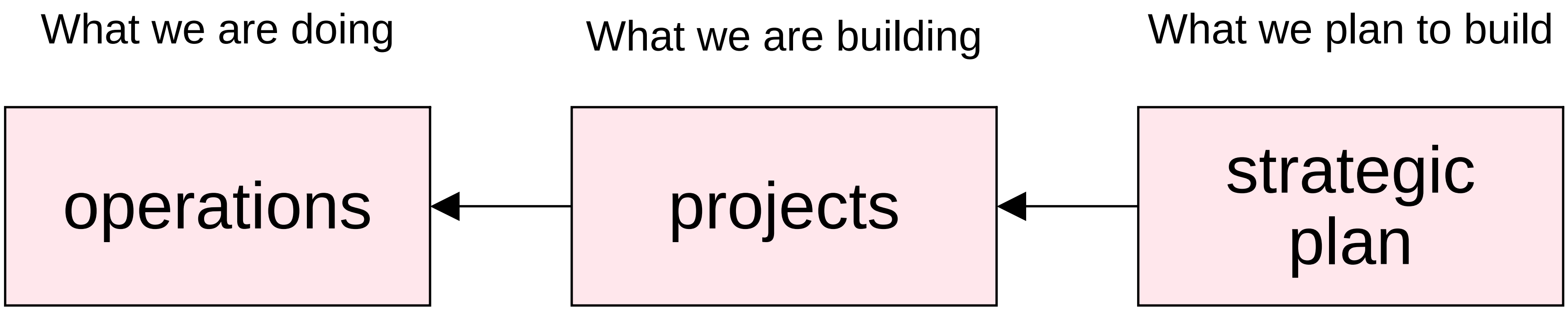
© 2003, Robert J. Robbins

Strategic Planning



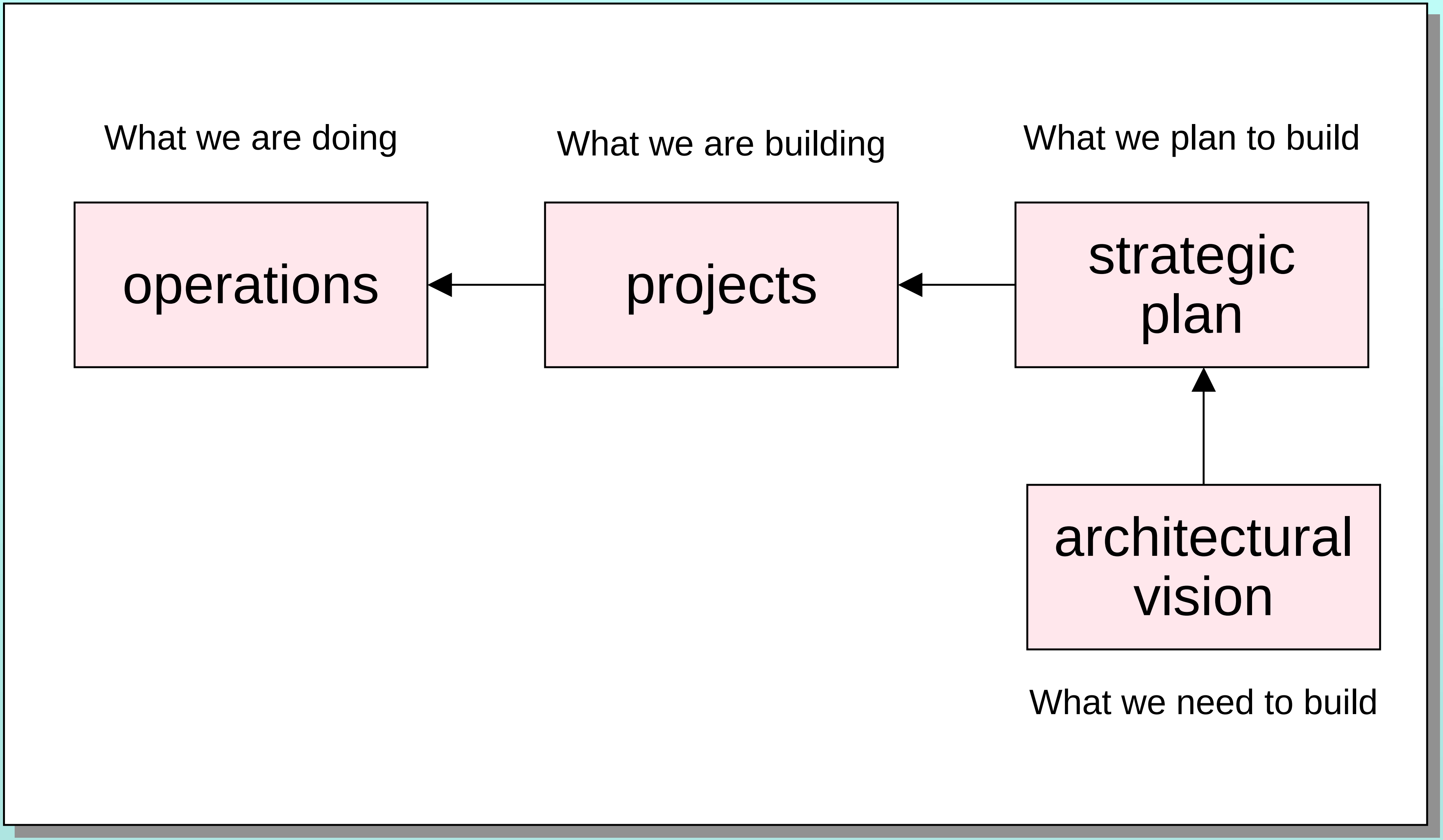
© 2003, Robert J. Robbins

Strategic Planning



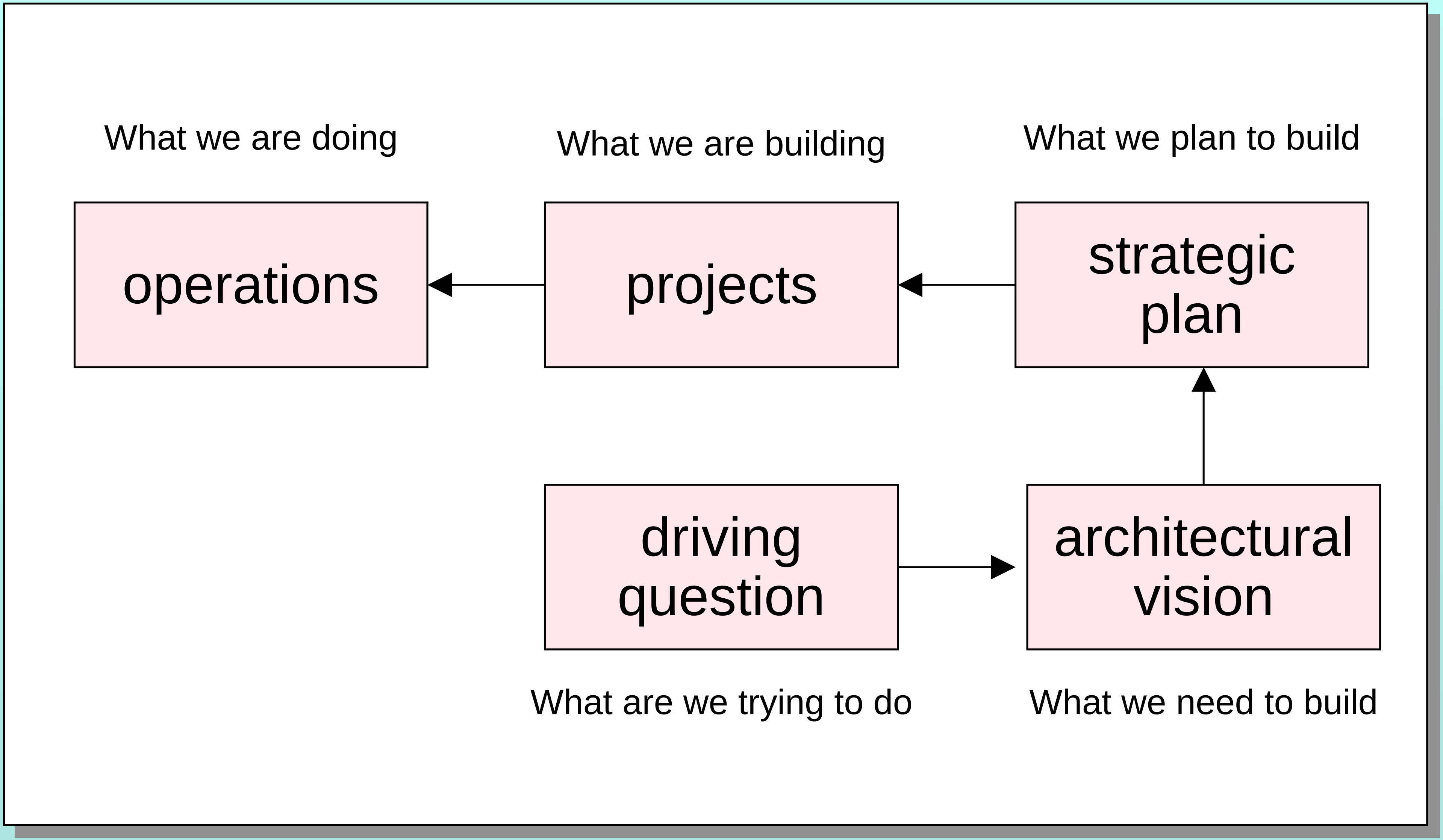
© 2003, Robert J. Robbins

Strategic Planning



© 2003, Robert J. Robbins

Strategic Planning



© 2003, Robert J. Robbins

Strategic Planning

W uild

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?

al

o build

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

W uild

al

o build

Example of important driving question:

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

© 2003, Robert J. Robbins

Strategic Planning

W… …uild

…al

… build

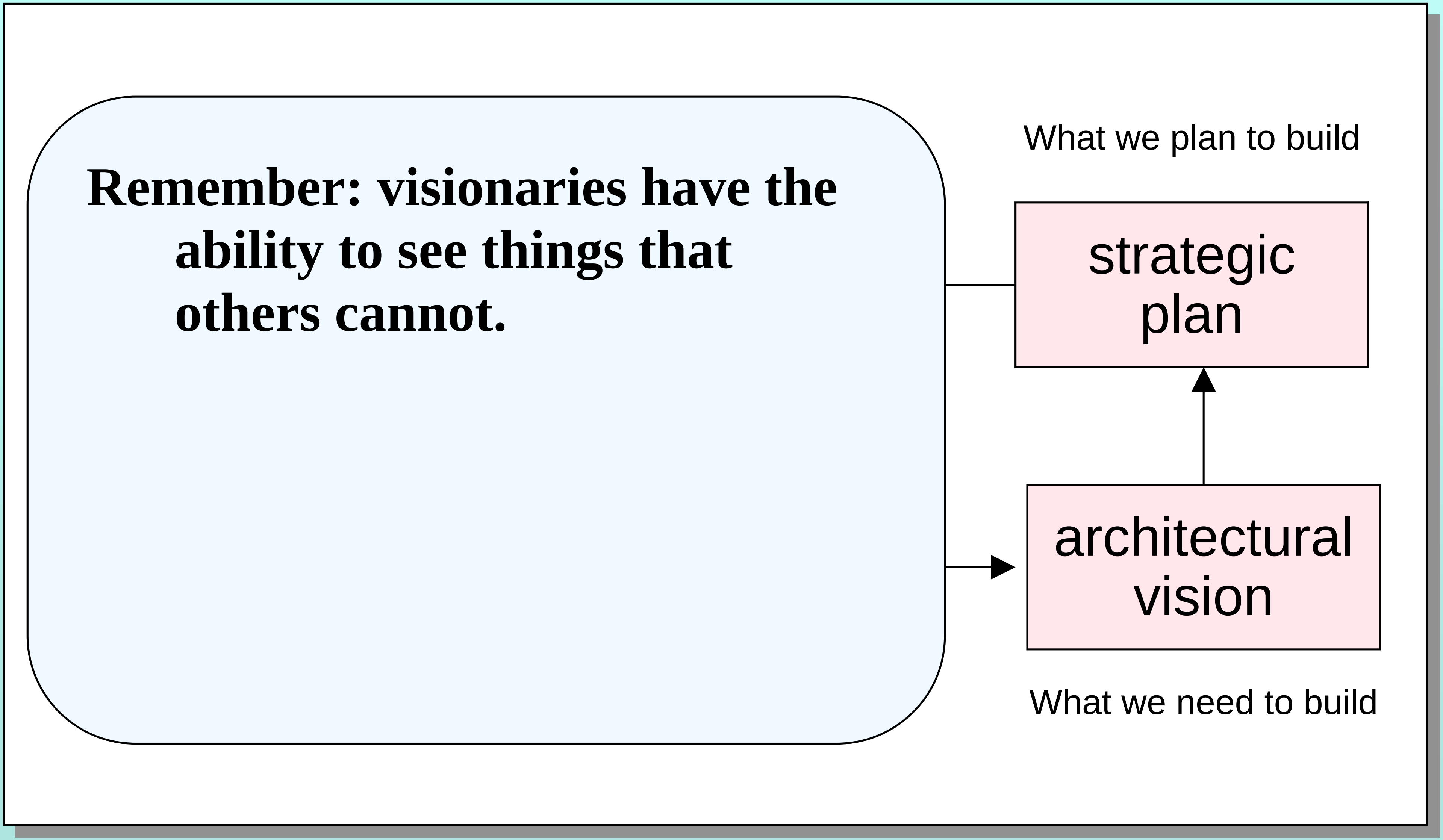
Example of important driving question:

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.

© 2003, Robert J. Robbins

Strategic Planning

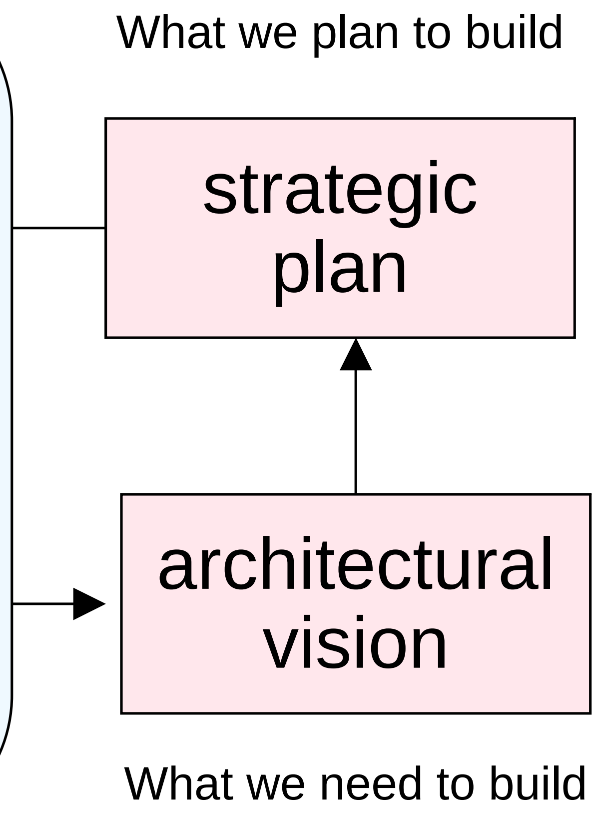


Strategic Planning

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?

© 2003, Robert J. Robbins

Conclusions (Inferences)

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

© 2003, Robert J. Robbins

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.

© 2003, Robert J. Robbins

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.

© 2003, Robert J. Robbins

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.

© 2003, Robert J. Robbins

TCP/IP Network Model

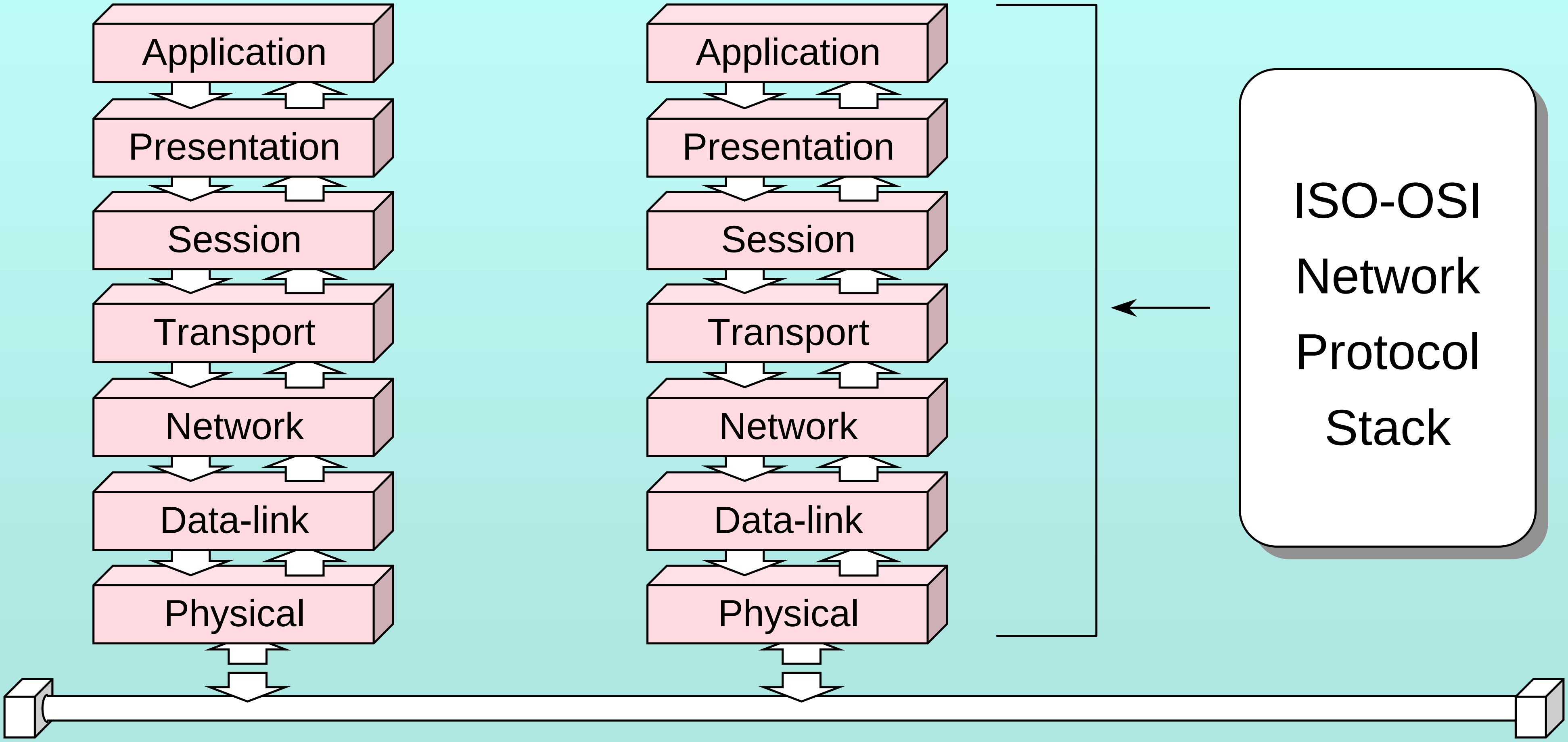
Technology

Technical Attributes

- Highly abstracted components
- Layered architecture
- Modular construction
- Clearly defined interfaces
- No interactions except through interfaces
- Declarative user interface

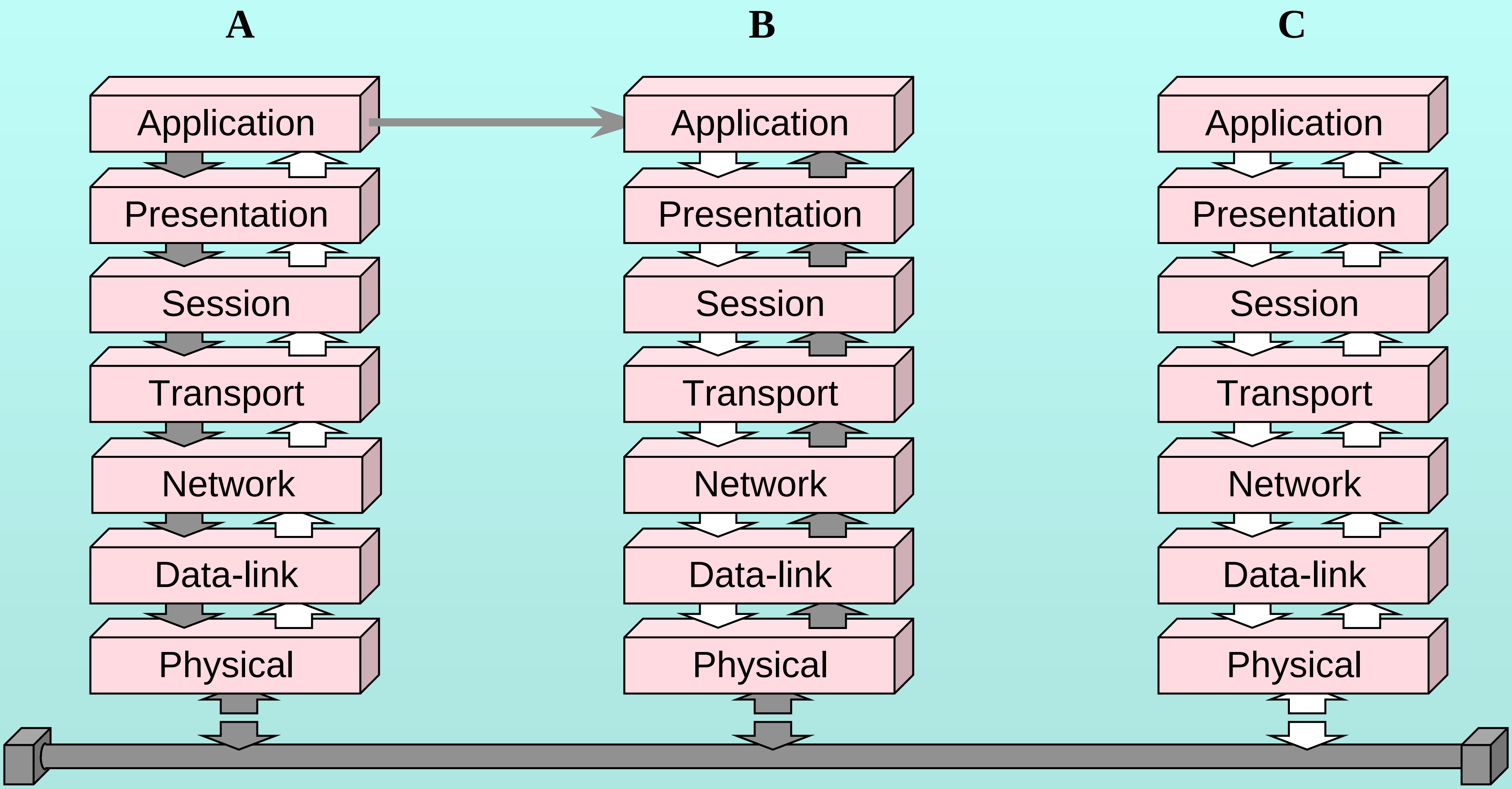
© 2003, Robert J. Robbins

ISO-OSI Network Model



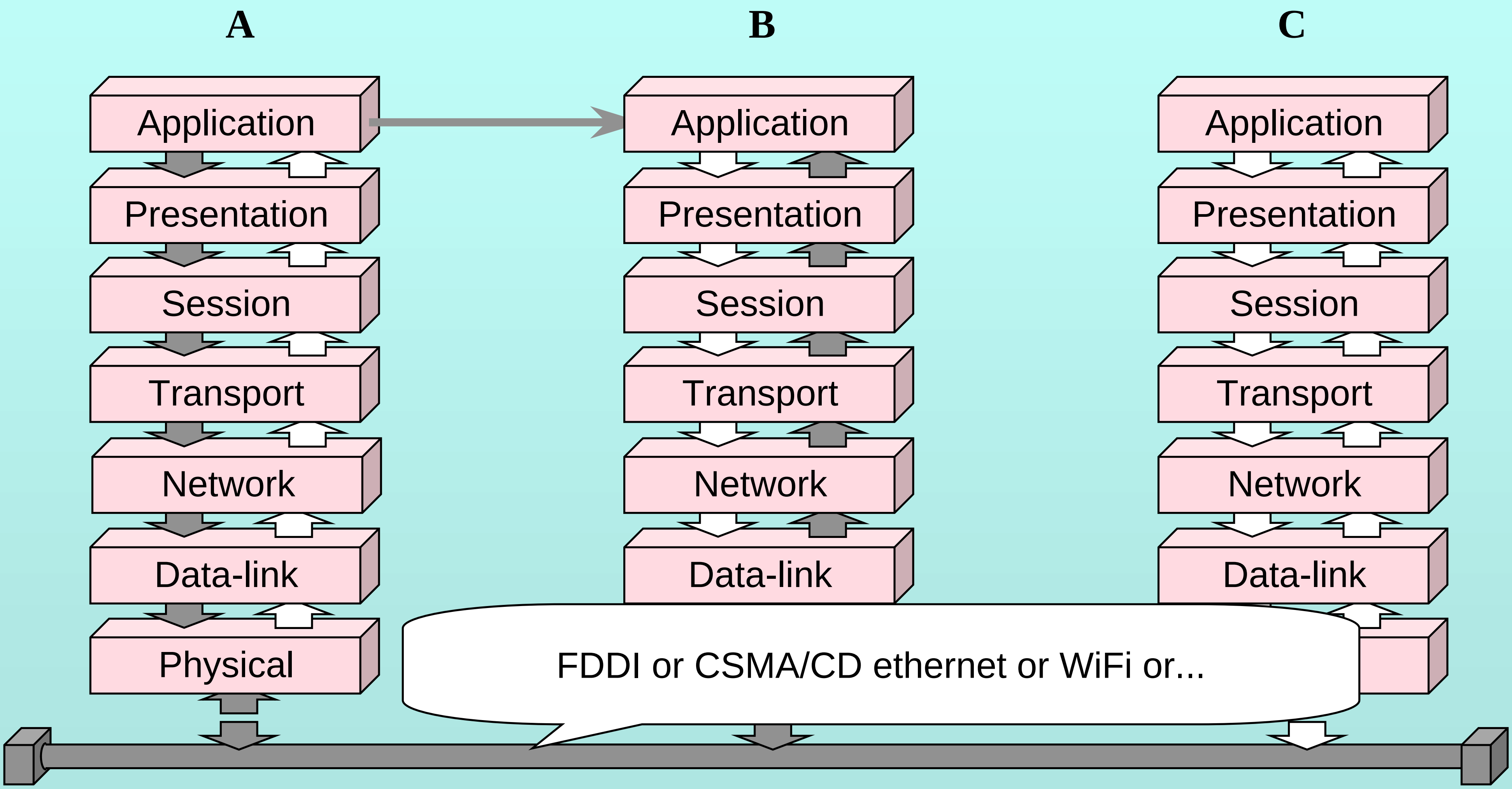
© 2003, Robert J. Robbins

ISO-OSI Network Model



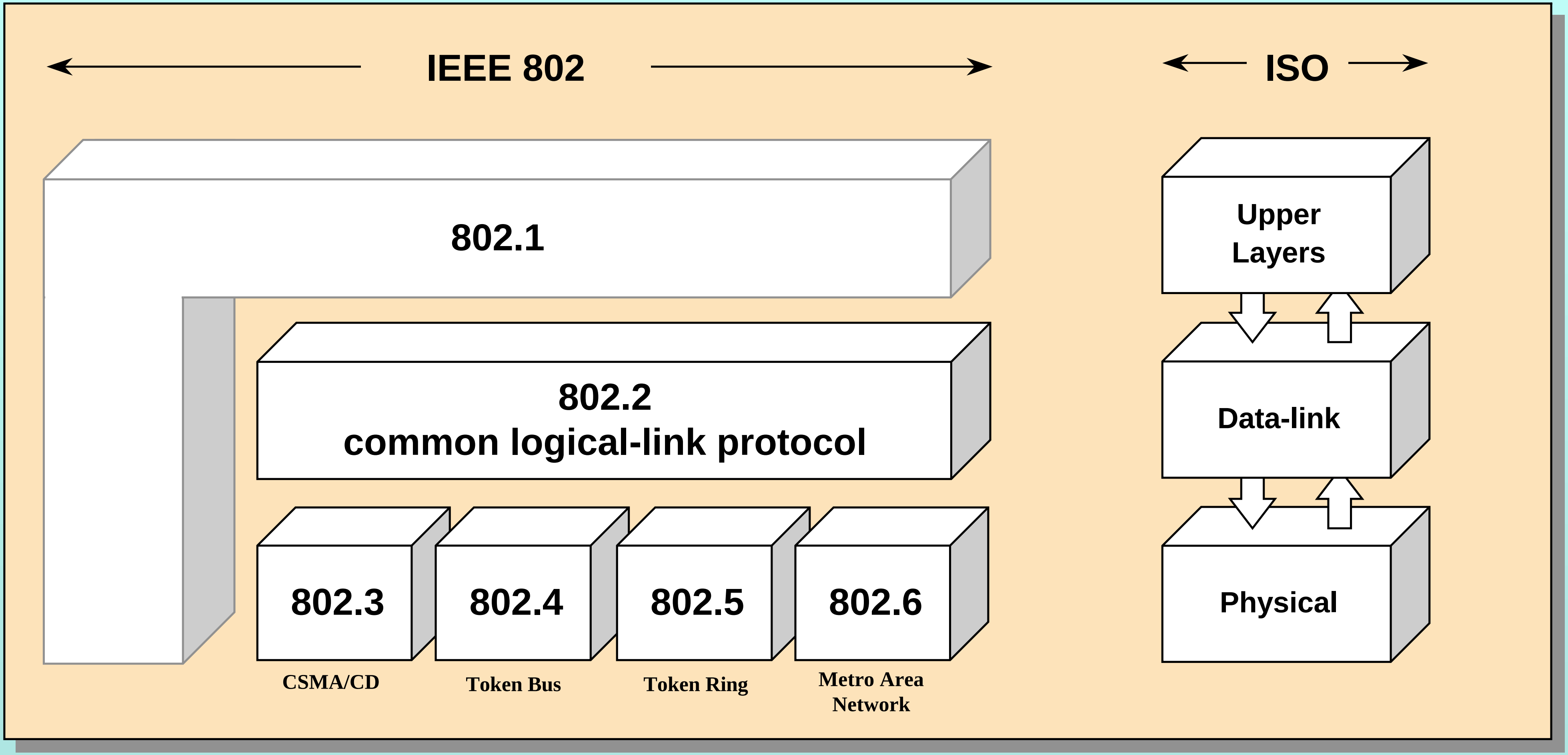
© 2003, Robert J. Robbins

ISO-OSI Network Model

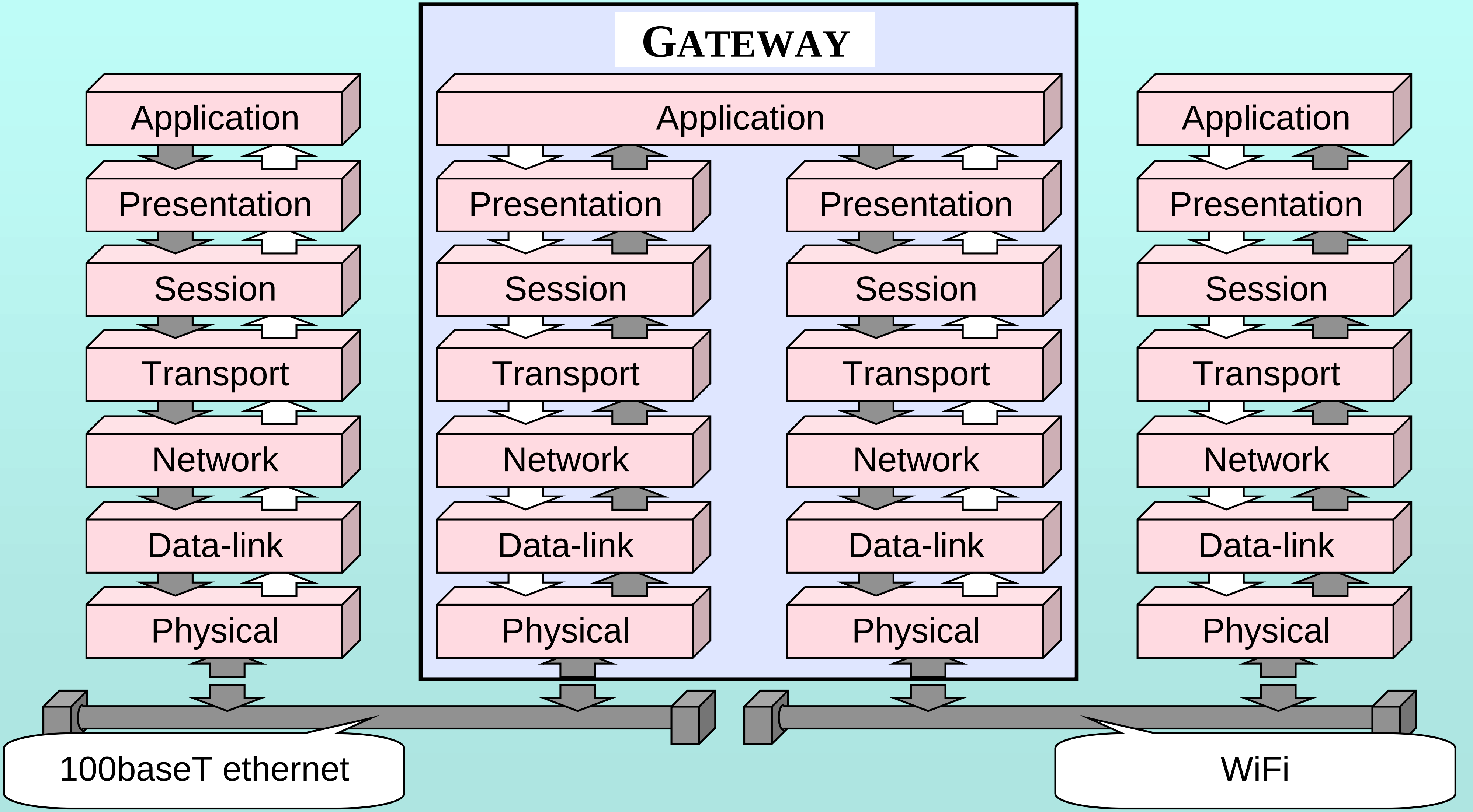


© 2003, Robert J. Robbins

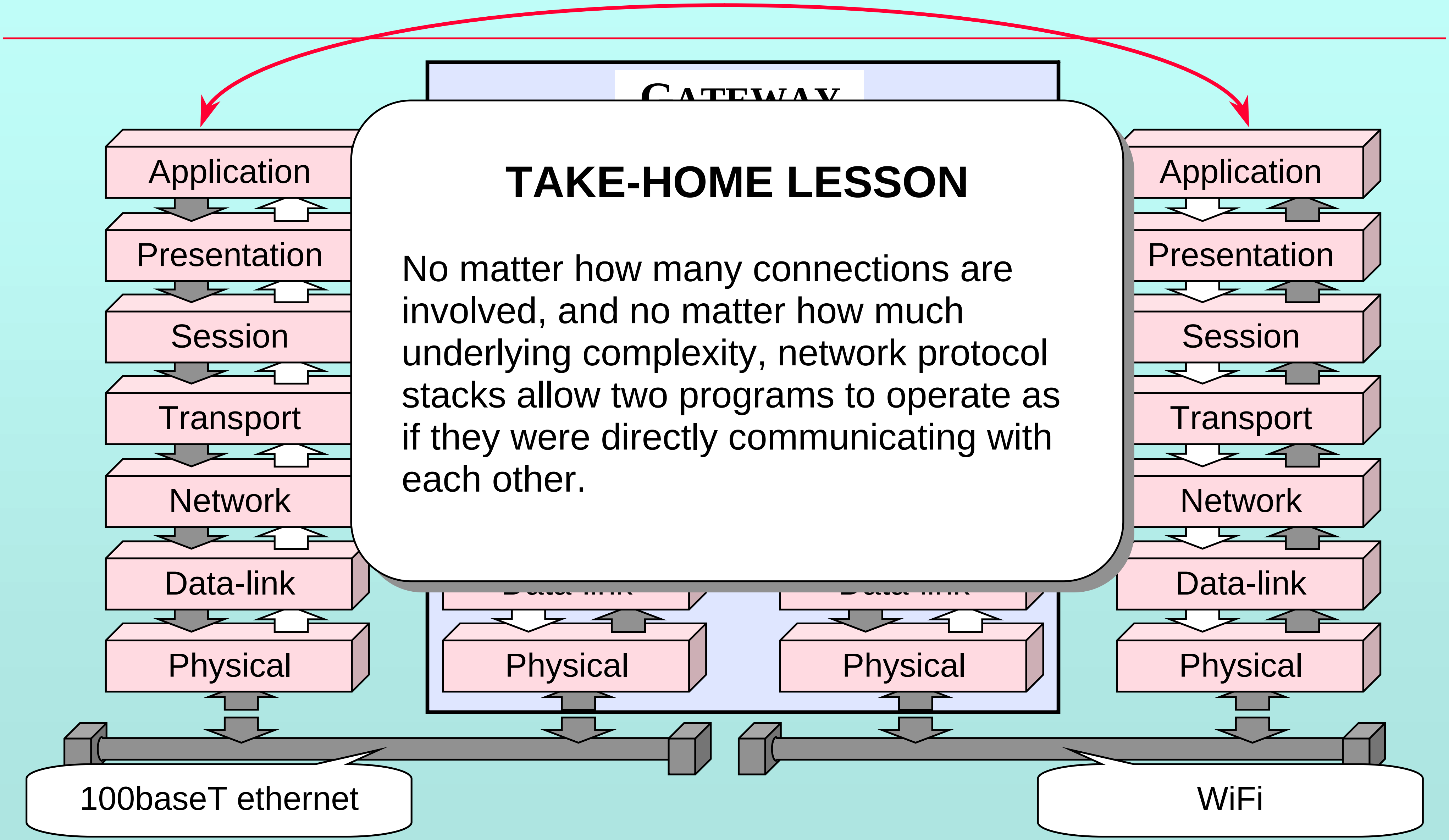
Physical Layer Protocols



ISO-OSI Network Model



© 2003, Robert J. Robbins



TAKE-HOME LESSON

No matter how many connections are involved, and no matter how much underlying complexity, network protocol stacks allow two programs to operate as if they were directly communicating with each other.

© 2003, Robert J. Robbins

Network Protocol Stacks

layer	ISO	TCP / IP	SNA	DECNET
7	Application	User	End User	Application
6	Presentation	ftp, telnet	NAU Services	
5	Session	(none)	Data-flow Control	(none)
4	Transport	Host-Host	Transmission Control	Network Services
3	Network	Source to destination IMP	Path Control	Transport
2	Data-Link	IMP-IMP	Data-Link Control	Data-Link Control
1	Physical	Physical	Physical	Physical

© 2003, Robert J. Robbins

Network Protocol Stacks

layer	ISO	TCP / IP	SNA	DECNET
7	Application	User	End User	Application
6	Presentation	ftp, telnet	NAU Services	
5	Session	(none)		
4	Transport	Host-Host Source to		es
		tion IMP	Path Control	Transport
		MP	Data-Link Control	Data-Link Control
		cal	Physical	Physical

TCP / IP protocols were developed to allow robust communication among distributed, heterogeneous computer systems, even under severely adverse conditions.

This automatically provides fine support for a rapidly changing environment.

Network Protocol Stacks

add HTTP, create the WWW

layer	ISO	TCP / IP		ET
7	Application	User		Application
6	Presentation	ftp, telnet, **http**	NAU Services	
5	Session	(none)	Data-flow Control Transmission Control	(none)
4	Transport	Host-Host Source to destination IMP		Network Services
3	Network		Path Control	Transport
2	Data-Link	IMP-IMP	Data-Link Control	Data-Link Control
1	Physical	Physical	Physical	Physical

© 2003, Robert J. Robbins

Network Protocol Stacks

layer	ISO	TCP / IP		ET
7	Application	User		Application
6	Presentation	ftp, telnet, **http**	NAU Services	
5	Session	(none)	Data-flow Control / Transmission Control	(none)
4	Transport	Host-Host		Network Services
3	Network	Source to destination IMP	Path Control	Transport
2	Data-Link	IMP-IMP	Data-Link Control	Data-Link Control
1	Physical	Physical	Physical	Physical

add HTTP, create the WWW

It cannot get any more extensible than this: add a protocol, create an industry.

© 2003, Robert J. Robbins

Declarative Interface

Declarative Interface

With TCP/IP networking, the commands to connect to HOSTNAME via PROTOCOL are:

telnet	snapple
ssh	foobar
ftp	shazbot

© 2003, Robert J. Robbins

Declarative Interface

With TCP/IP networking, the commands to connect to HOSTNAME via PROTOCOL are:

telnet snapple

ssh foobar

ftp shazbot

It cannot get any more declarative than this: there are two critical parameters and the command is just the concatenation of the parameters.

© 2003, Robert J. Robbins

Good Sociology

Sociological Attributes

- No definitive center
- Community participation
- Optional usage
- Avoid premature standards
- Evolving/extensible standards
- . . .

© 2003, Robert J. Robbins

RDBMS Technology

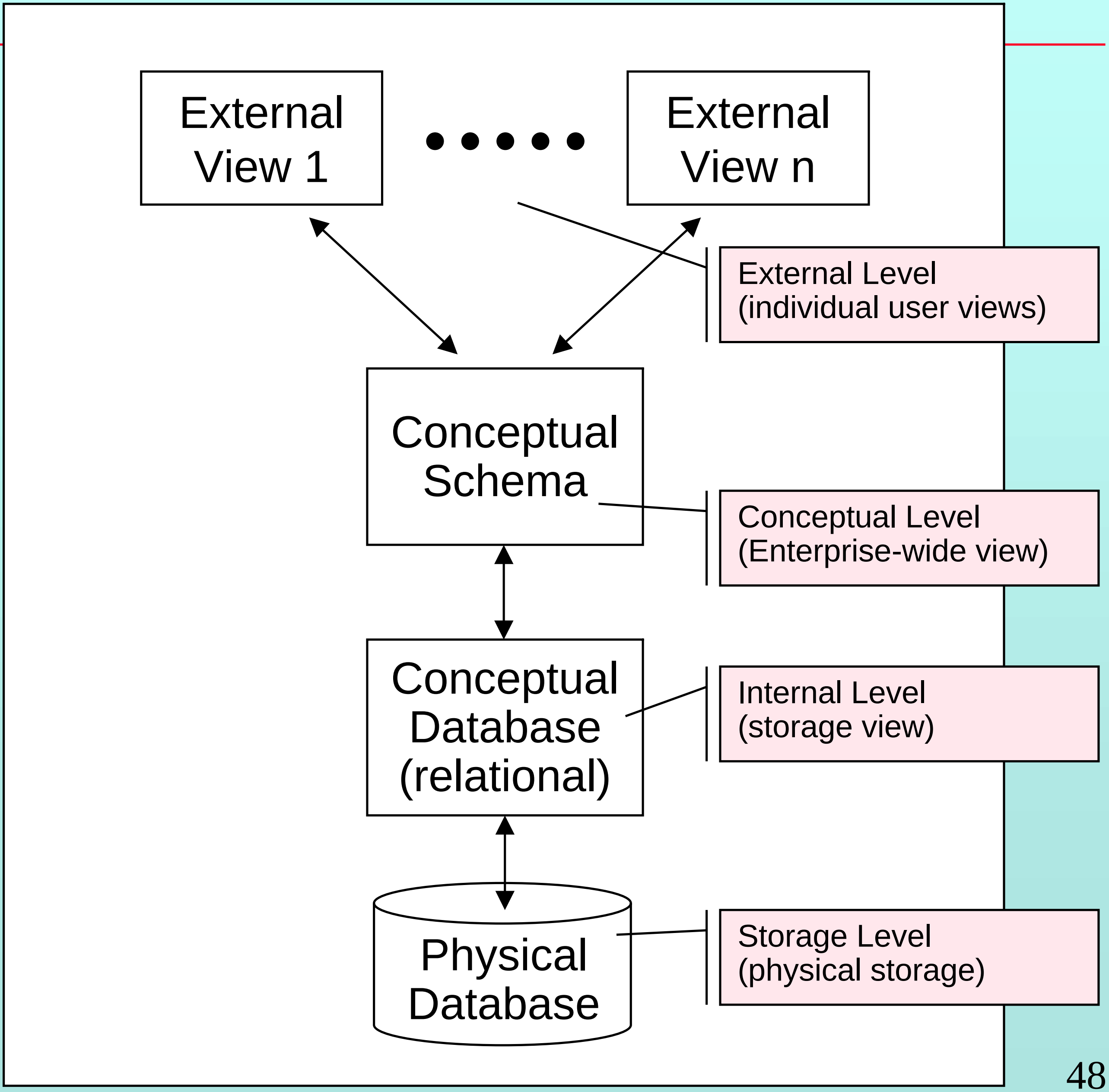
Technical Attributes

- Highly abstracted components
- Layered architecture
- Modular construction
- Clearly defined interfaces
- No interactions except through interfaces
- Declarative user interface

© 2003, Robert J. Robbins

RDBMS

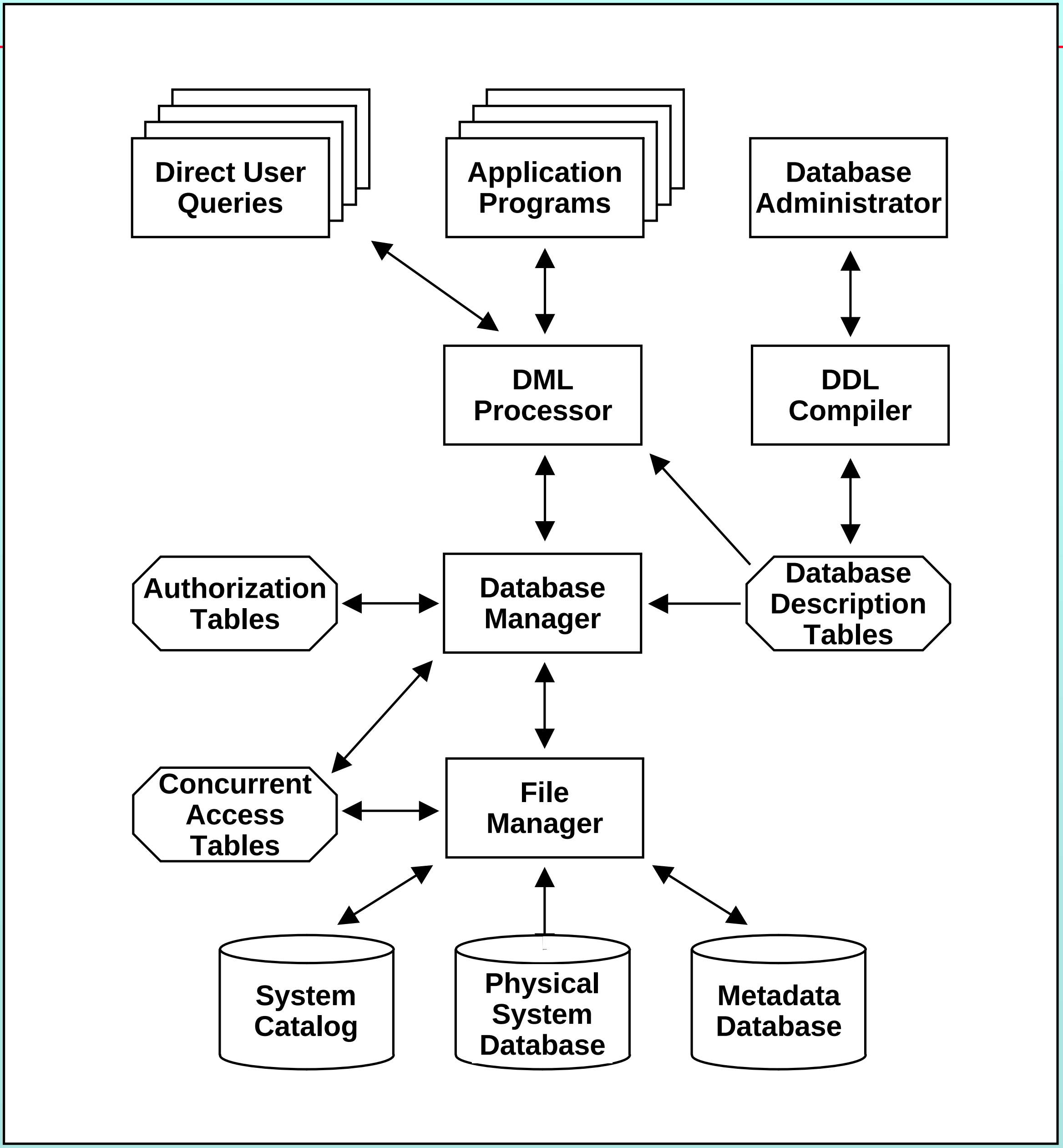
A database management system (DBMS) is a collection of programs that enables users to create and maintain a database. According to the ANSI/SPARC DBMS Report (1977), a DBMS should be envisioned as a multi-layered system:



© 2003, Robert J. Robbins

RDBMS

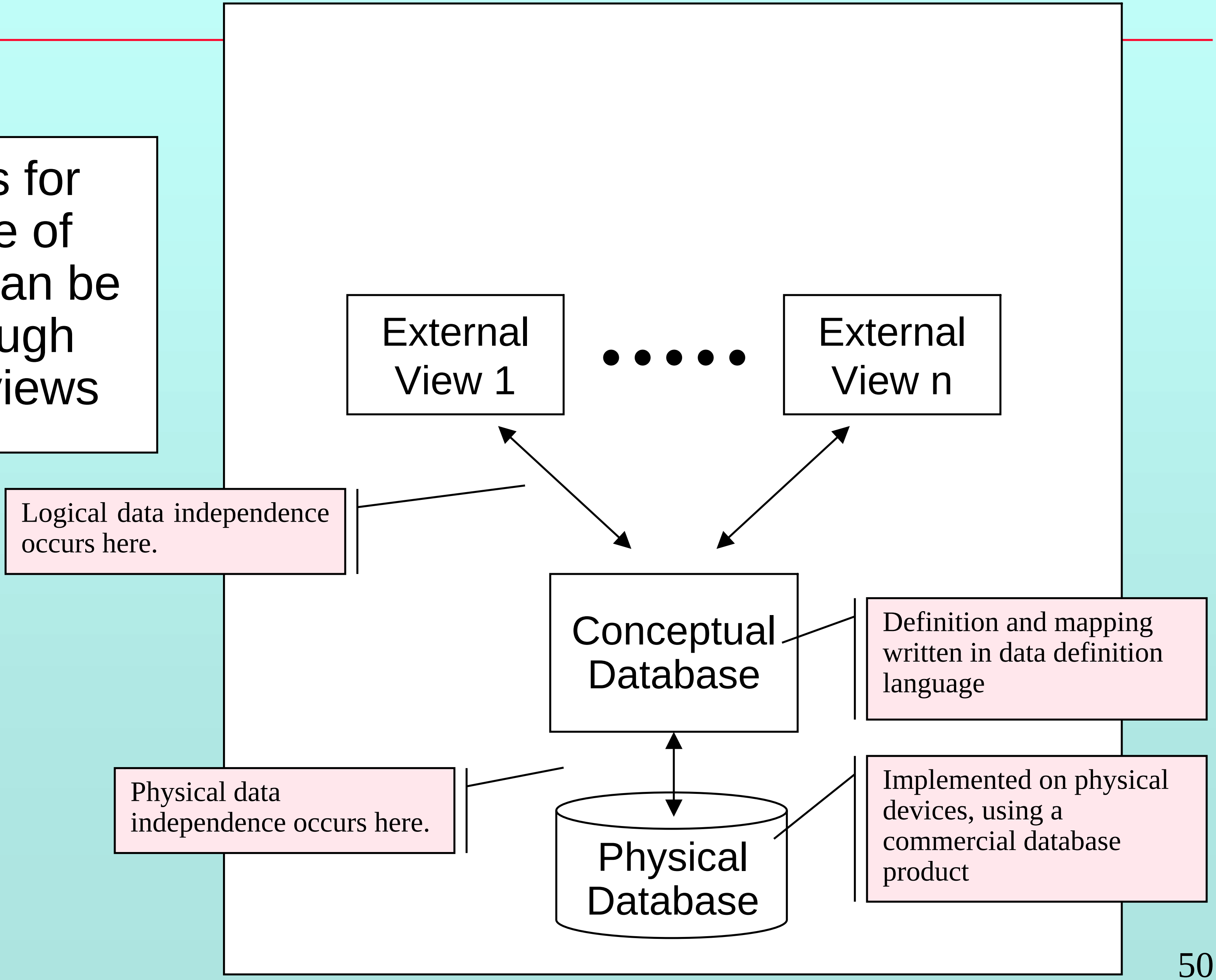
Many of the layers have identified, and separable sub-components…



© 2003, Robert J. Robbins

RDBMS

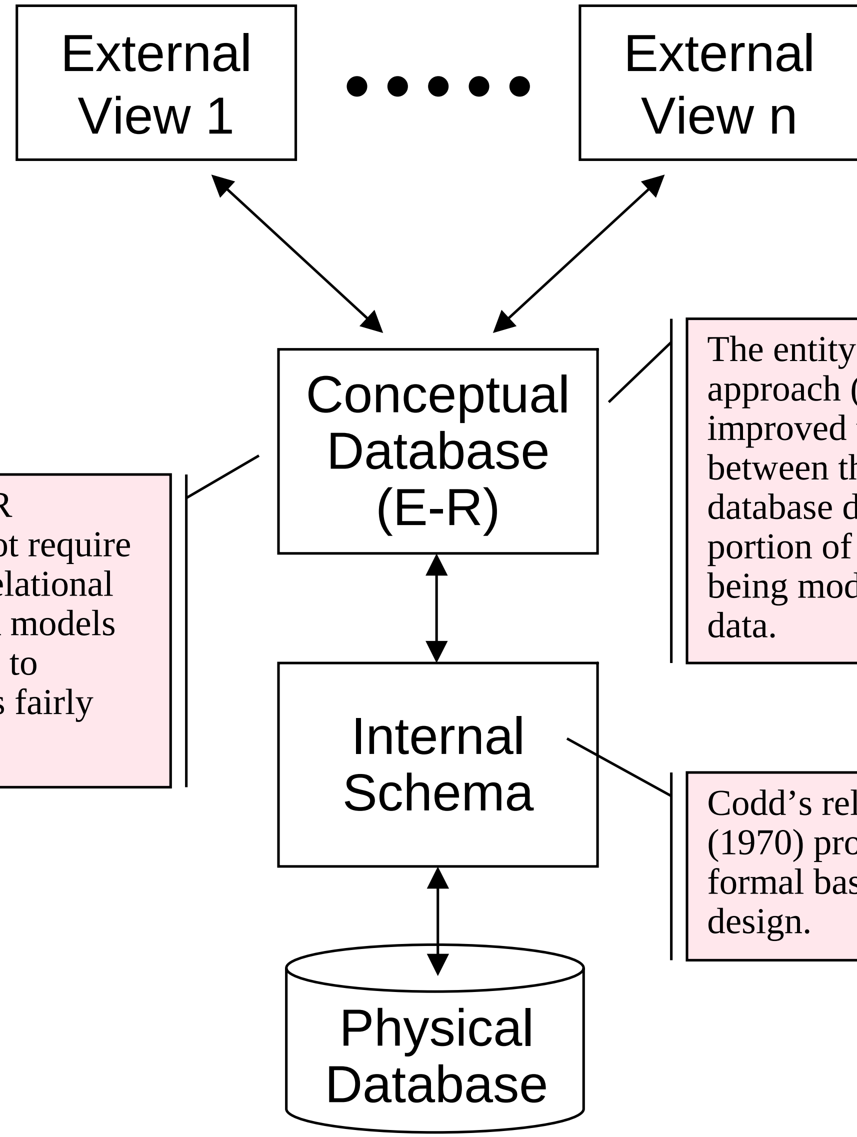
Different needs for access and use of the database can be supported through different user views



© 2003, Robert J. Robbins

RDBMS

Layers may be added to a conceptual design in order to increase the semantic richness available at the top design level.



Although the E-R approach does not require an under-lying relational model, most E-R models can be converted to relational models fairly easily.

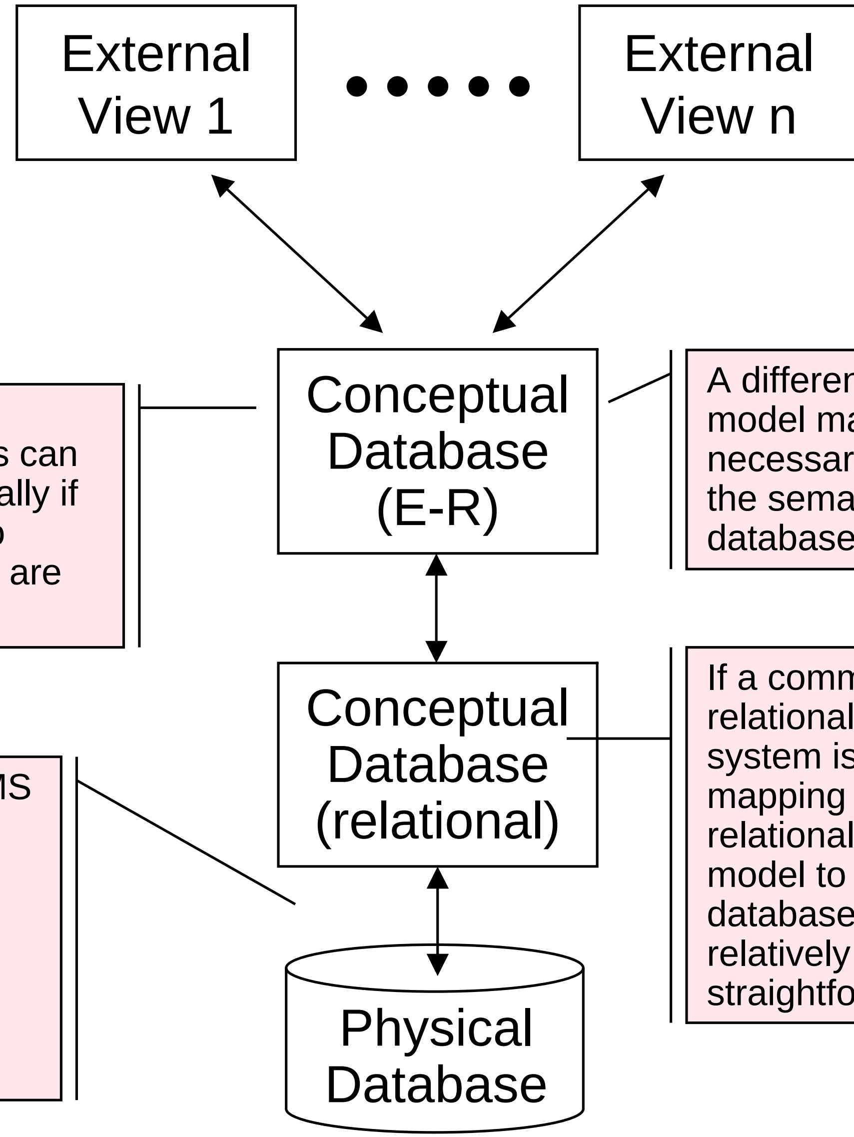
The entity-relationship approach (Chen, 1976) improved the mapping between the semantics of a database design and that portion of the real world being modeled with the data.

Codd's relational model (1970) provided the first formal basis for database design.

© 2003, Robert J. Robbins

RDBMS

If layered conceptual models are used, the layering may be perceived differently by the system's users and developers. Users often see the database only in terms of the views that they employ. System analysts and designers may think primarily about the E-R schema, whereas the database administrator is likely to deal primarily with the relational schema and the physical system.



Moving between conceptual models can be difficult, especially if automated tools to facilitate the move are not available.

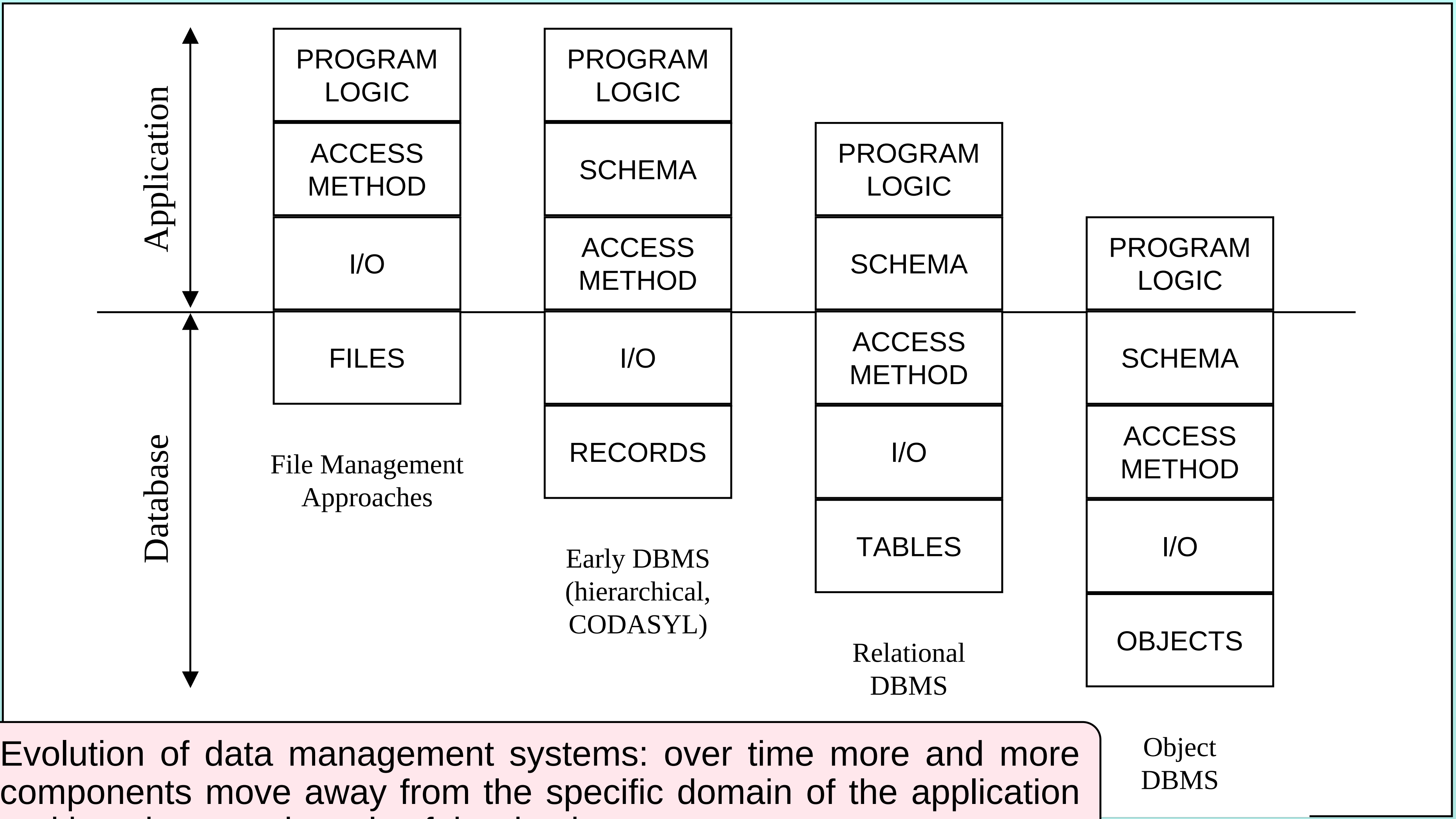
A different conceptual model may be necessary to capture the semantics of the database domain.

If a commercial relational database system is used, mapping from a relational conceptual model to the physical database should be relatively straightforward.

If a commercial RDBMS is used, a relational conceptual model provides a basis for designing and implementing an underlying physical database.

© 2003, Robert J. Robbins

RDBMS



Evolution of data management systems: over time more and more components move away from the specific domain of the application and into the generic tools of the database management system

© 2003, Robert J. Robbins

Declarative Interface

Declarative Interface

With RDBMS, the command to extract data from the system are generically:

SELECT data elements
FROM source tables
WHERE condition is true

It would be hard to get any more declarative than this: the syntax is pretty much limited to the minimum set of verbs, nouns, and logic.

© 2003, Robert J. Robbins

TCP/IP & RDBMS Patterns

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…

© 2003, Robert J. Robbins

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…

Expect lots of nay-sayers and skeptics along the way…

© 2003, Robert J. Robbins

Patience is a Virtue

Internet Time:

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

© 2003, Robert J. Robbins

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…

© 2003, Robert J. Robbins

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

© 2003, Robert J. Robbins

Working Group Assignments

For each module:

Background
- The Problem
- Available Solutions
- Remaining Challenges
  - *To be Solved in Other Modules*
  - *To be Solved in This Module*

An Ideal Solution
- Requirements
- Black-box Attributes
- Interoperability Interfaces
- Other Necessary Components
- Possible Implementation Details

Summary and Overview

© 2003, Robert J. Robbins

Possible Modules

Possible Modules

- Basic Infrastructure

© 2003, Robert J. Robbins

Possible Modules

- Basic Infrastructure
- Authorization, Authentication, Auditing

© 2003, Robert J. Robbins

Possible Modules

- Basic Infrastructure
- Authorization, Authentication, Auditing
- Digital Publishing Support

© 2003, Robert J. Robbins

Possible Modules

- Basic Infrastructure
- Authorization, Authentication, Auditing
- Digital Publishing Support
- Scientific Database I: Data Models & Design

© 2003, Robert J. Robbins

Possible Modules

- Basic Infrastructure
- Authorization, Authentication, Auditing
- Digital Publishing Support
- Scientific Database I: Data Models & Design
- Scientific Database II: Data Integration

© 2003, Robert J. Robbins

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

© 2003, Robert J. Robbins

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

© 2003, Robert J. Robbins

Possible Modules

- Clinical Research I: Research Access to Clinical Data

© 2003, Robert J. Robbins

Possible Modules

- Clinical Research I: Research Access to Clinical Data
- Clinical Research II: Research Trials

© 2003, Robert J. Robbins

Possible Modules

- Clinical Research I: Research Access to Clinical Data
- Clinical Research II: Research Trials
- Clinical Research III: Controlled Vocabularies

© 2003, Robert J. Robbins

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

© 2003, Robert J. Robbins

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

© 2003, Robert J. Robbins

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

© 2003, Robert J. Robbins

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

© 2003, Robert J. Robbins

Possible Modules

- More …

© 2003, Robert J. Robbins

Possible Methods

- Top down: ideal solutions

© 2003, Robert J. Robbins

Possible Methods

- Top down: ideal solutions
- Bottom up: current problems

© 2003, Robert J. Robbins

Possible Methods

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

© 2003, Robert J. Robbins

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…

© 2003, Robert J. Robbins

GLAAAS

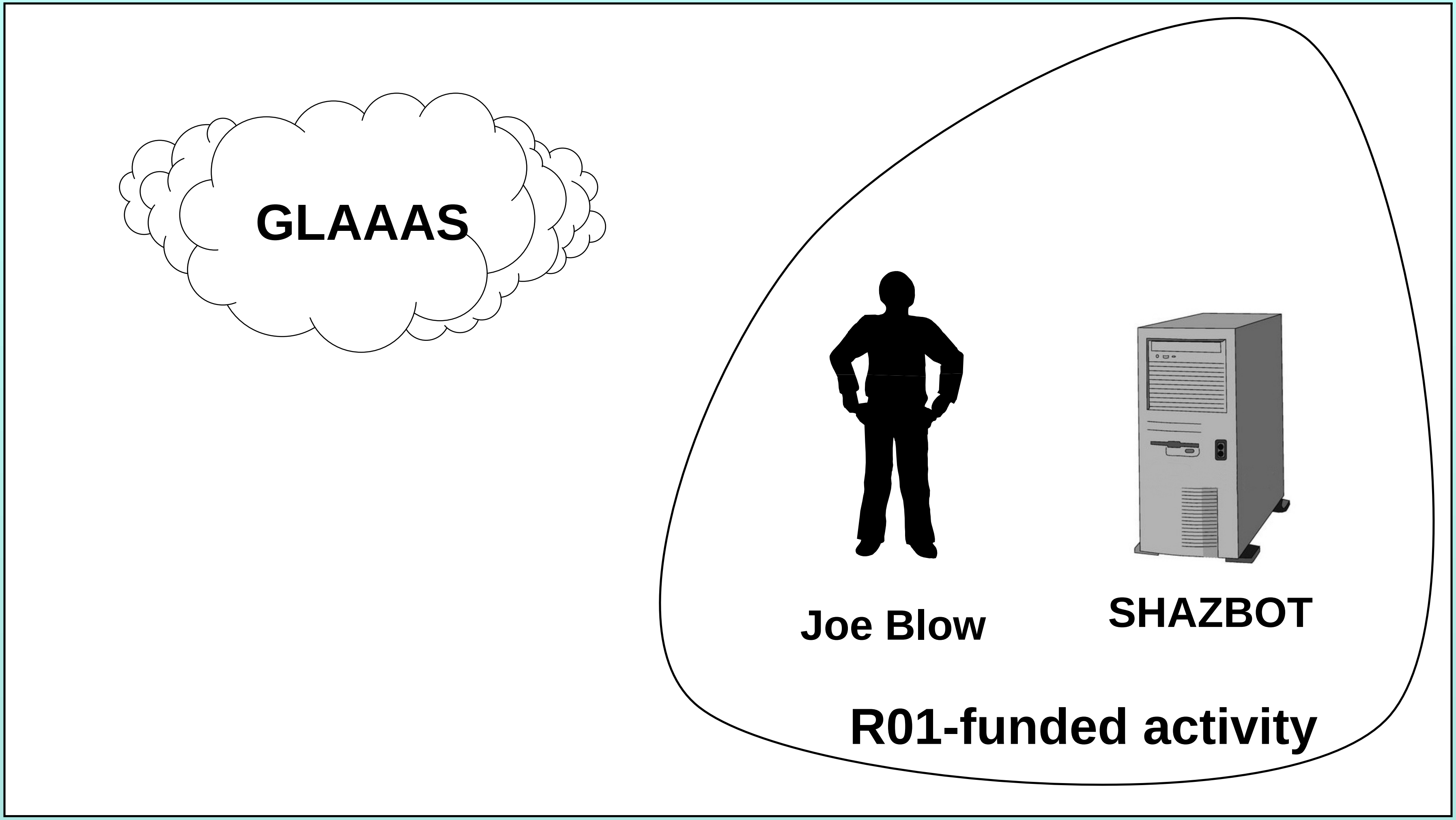
GLAAAS is a GLobal Authorization, Authentication, and Auditing System that can be used to assign, track, and audit permissions to use IT resources on any server that participates in GLAAAS.

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.

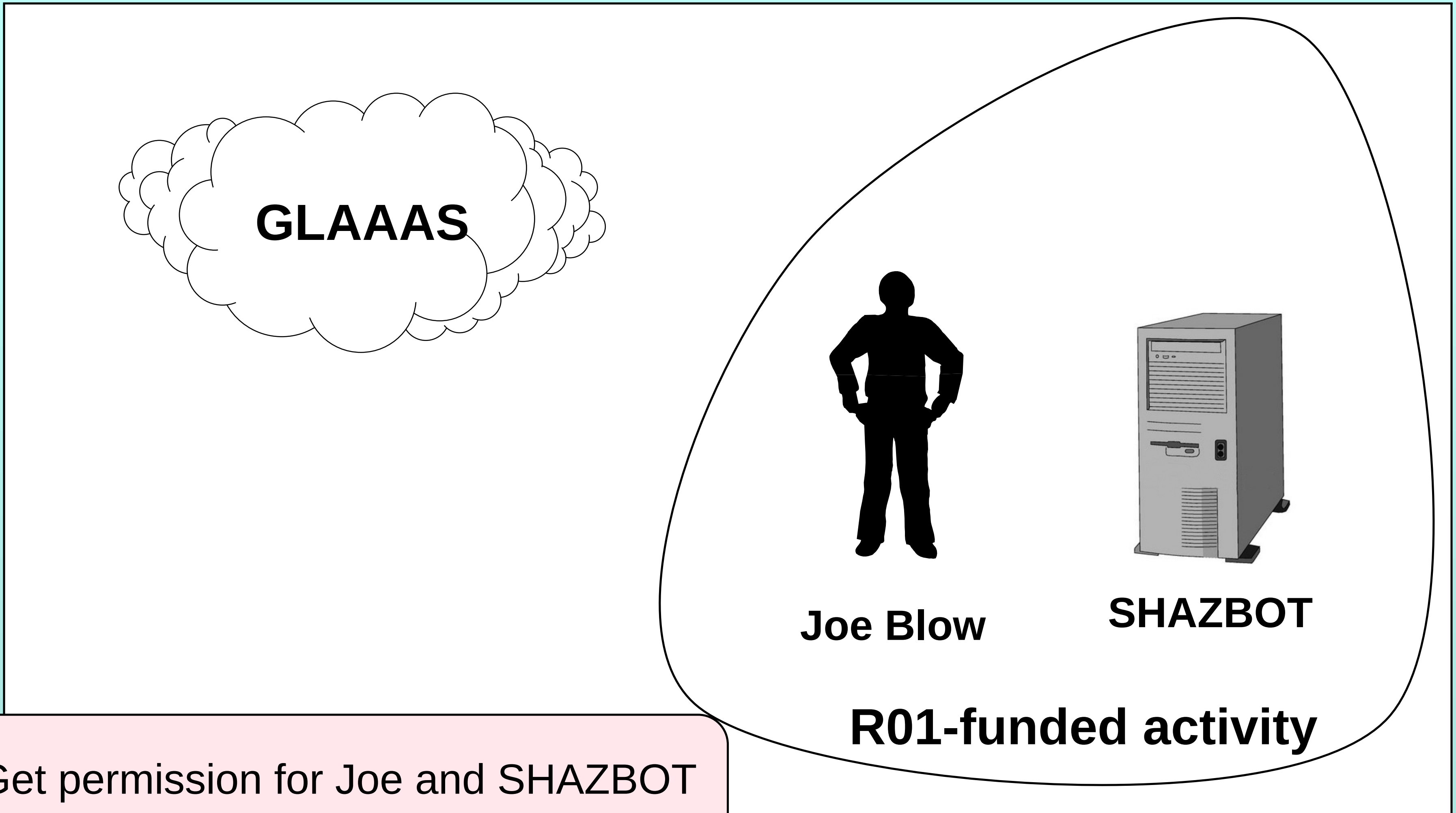
© 2003, Robert J. Robbins

GLAAAS



© 2003, Robert J. Robbins

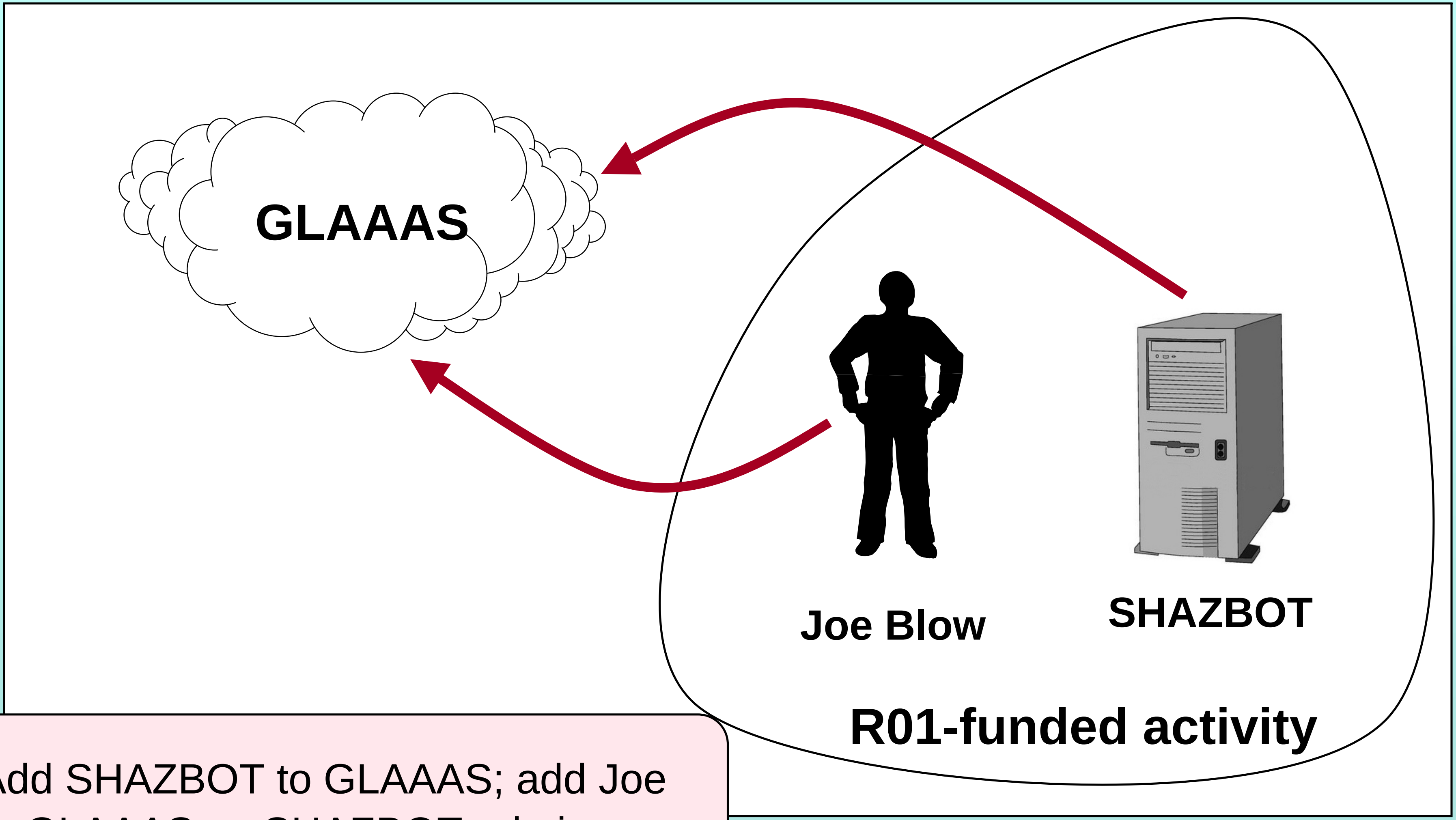
GLAAAS



Get permission for Joe and SHAZBOT to use the GLAAAS.

© 2003, Robert J. Robbins

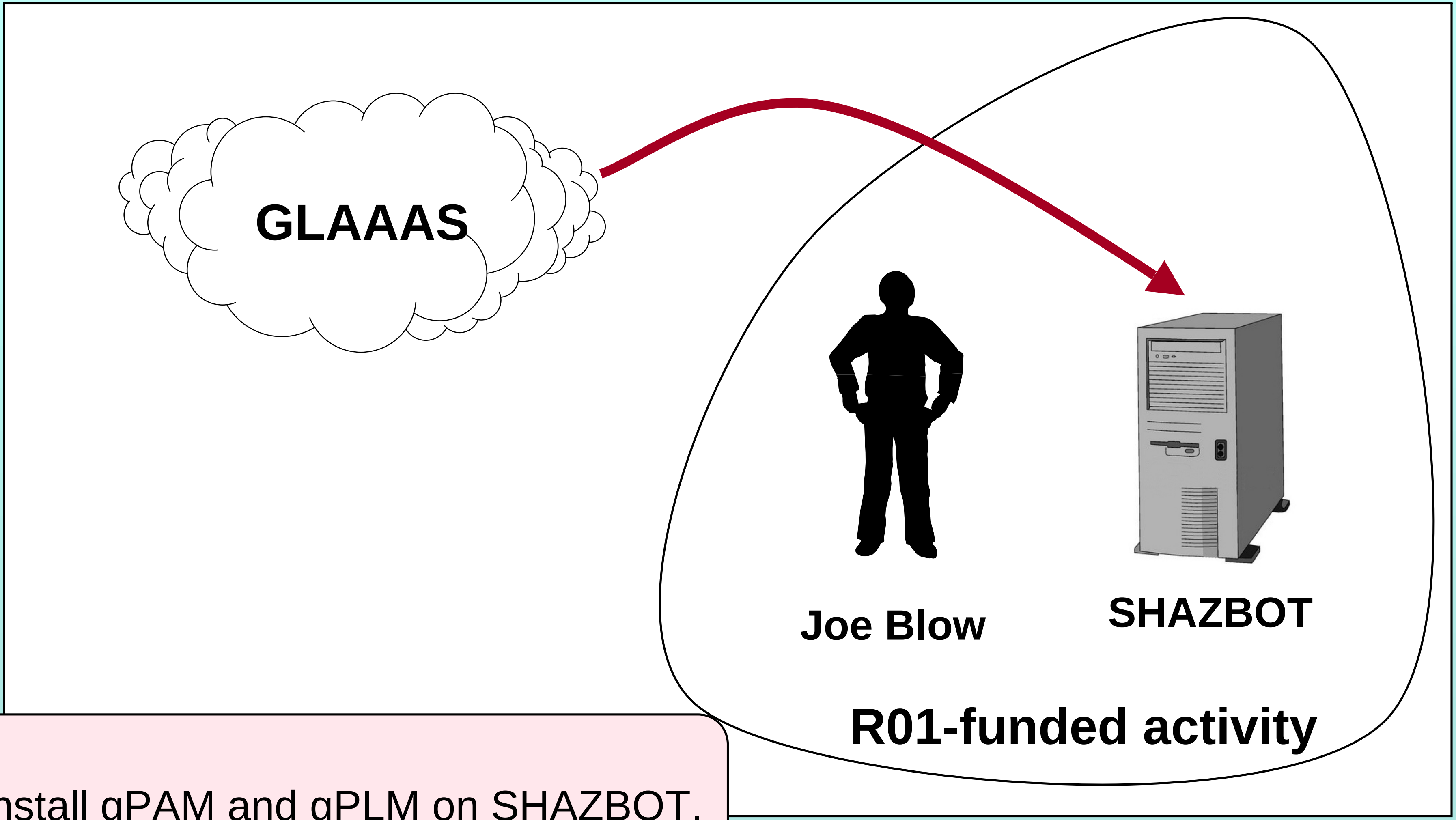
GLAAAS



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

© 2003, Robert J. Robbins

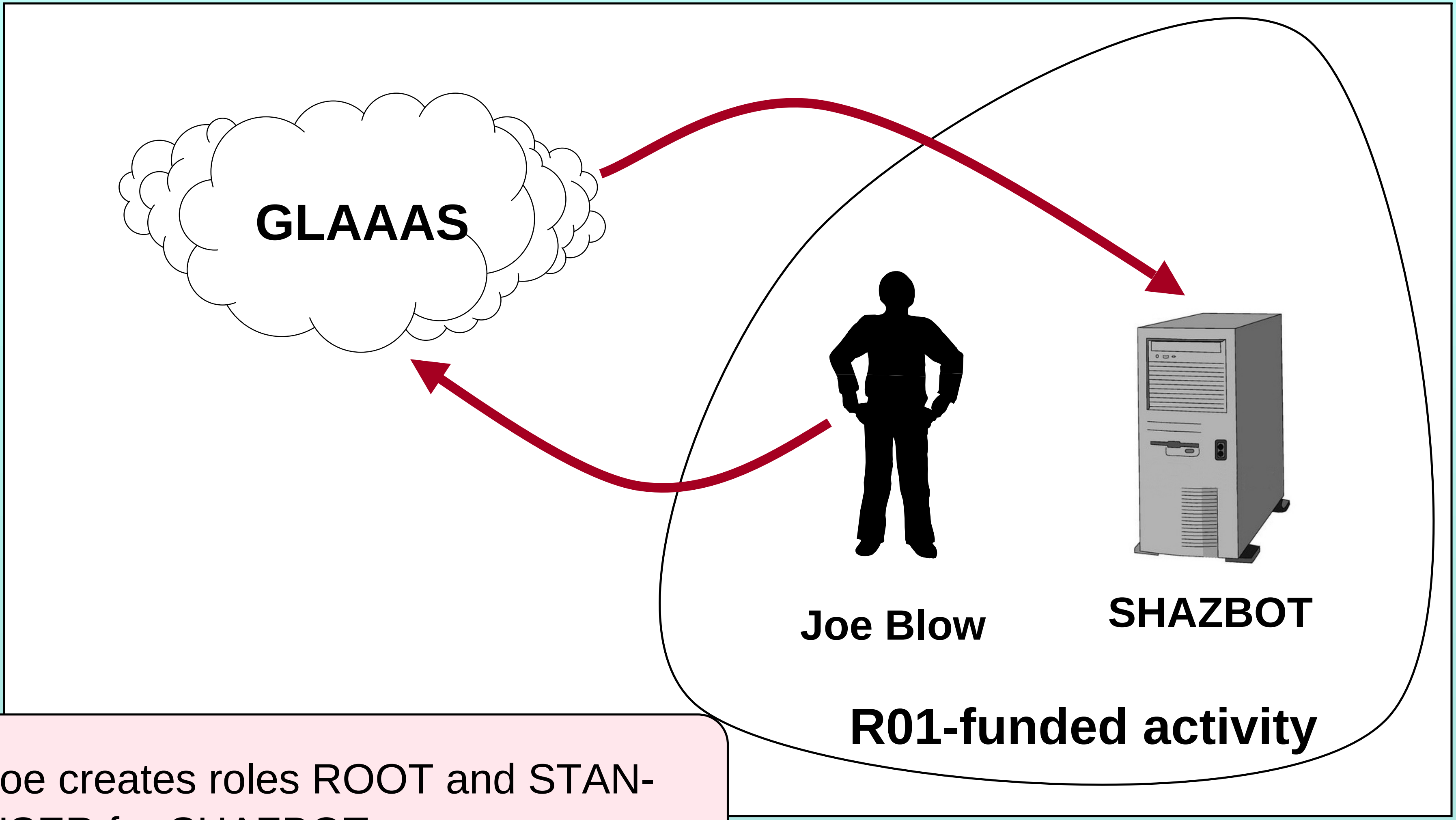
GLAAAS



Install gPAM and gPLM on SHAZBOT.

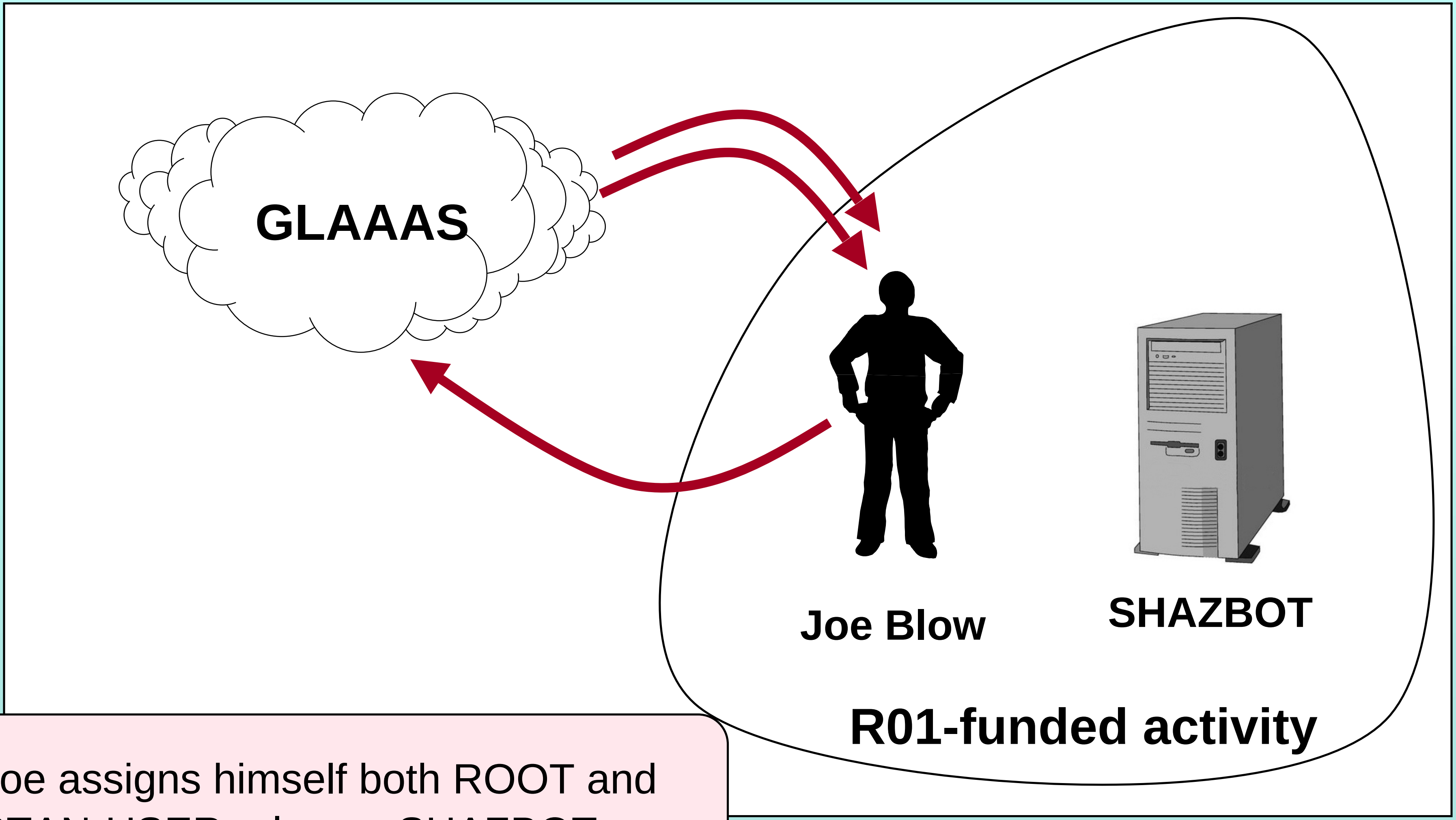
© 2003, Robert J. Robbins

GLAAAS



Joe creates roles ROOT and STAN-USER for SHAZBOT.

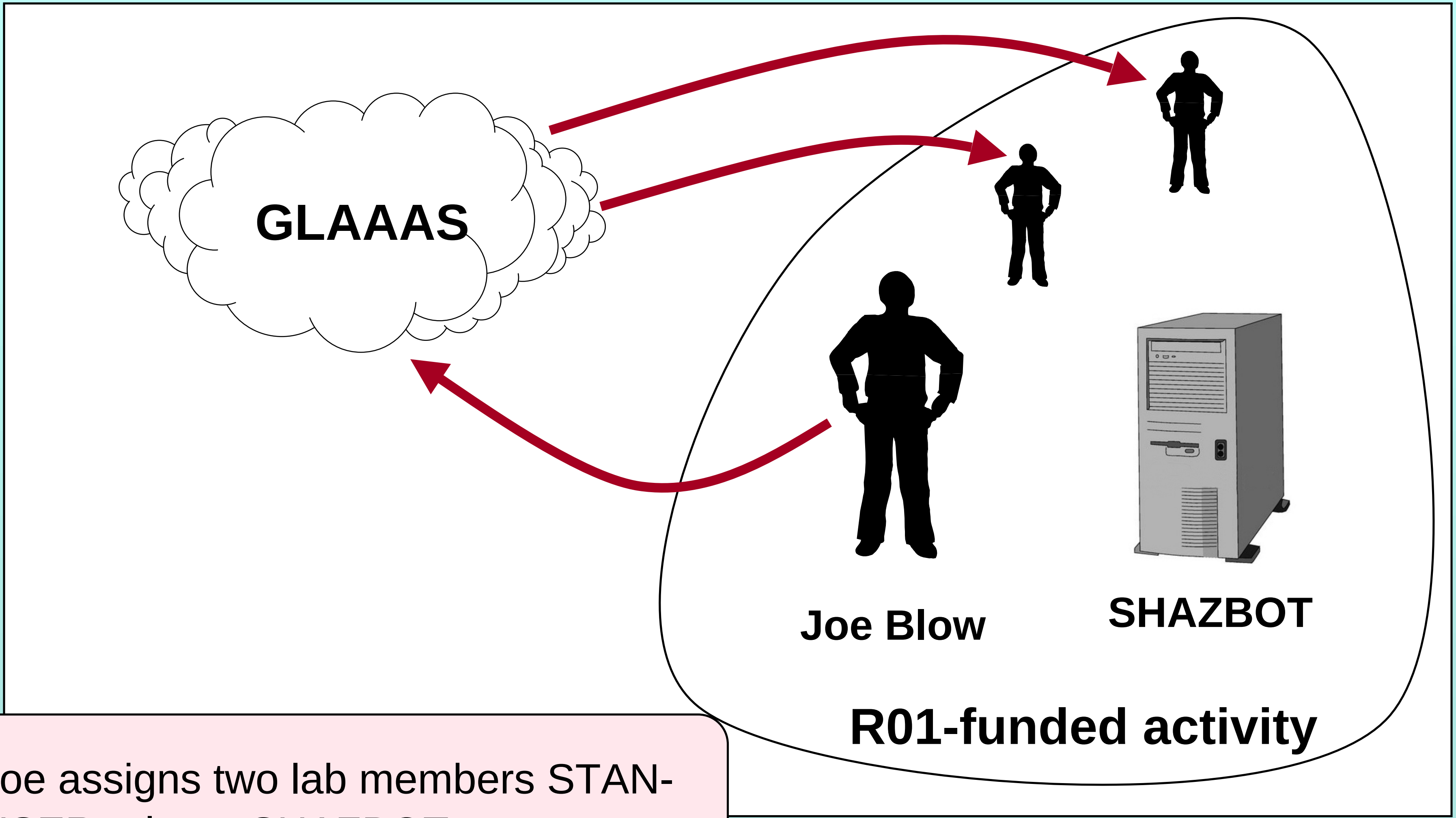
GLAAAS



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

© 2003, Robert J. Robbins

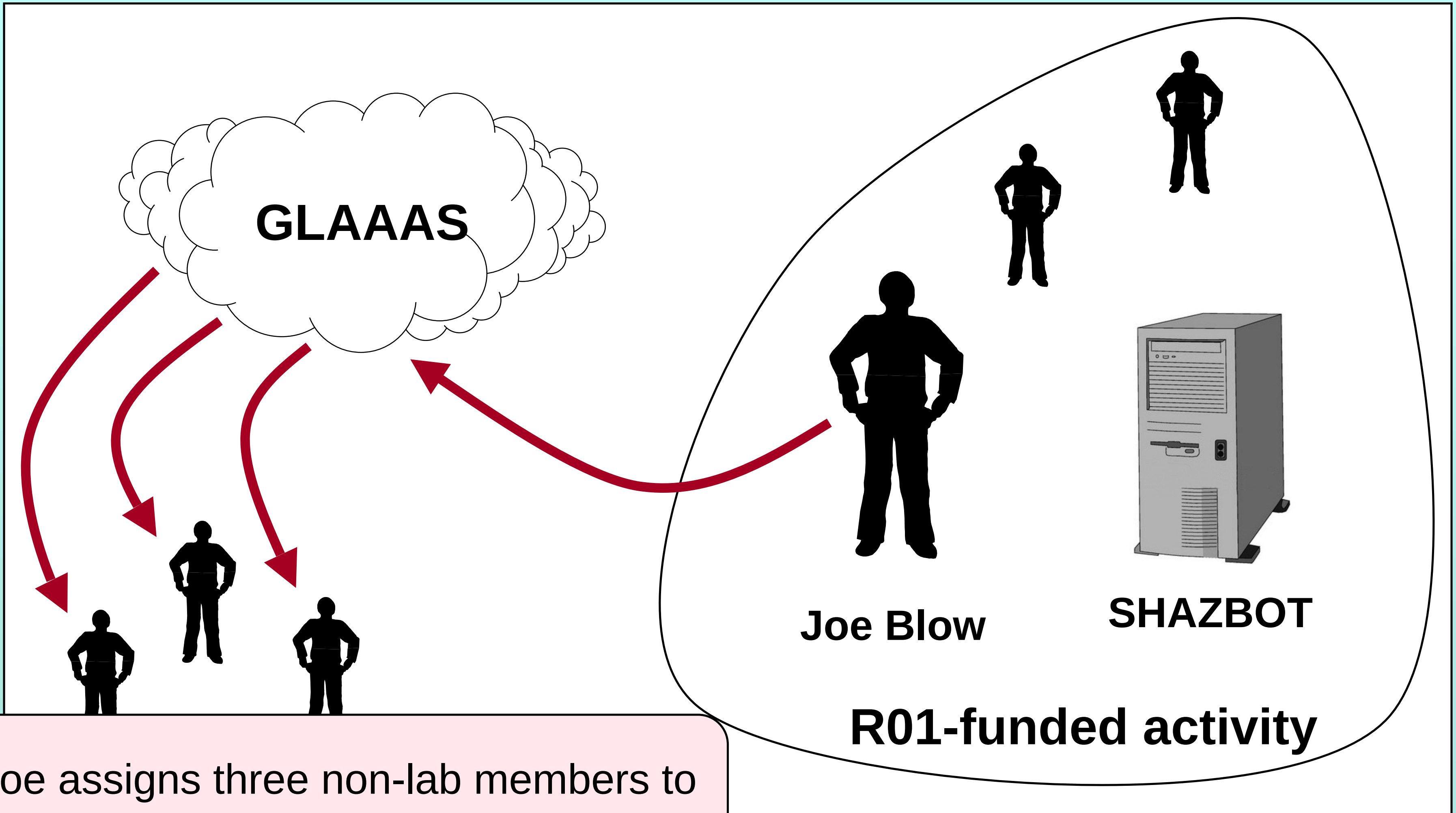
GLAAAS



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

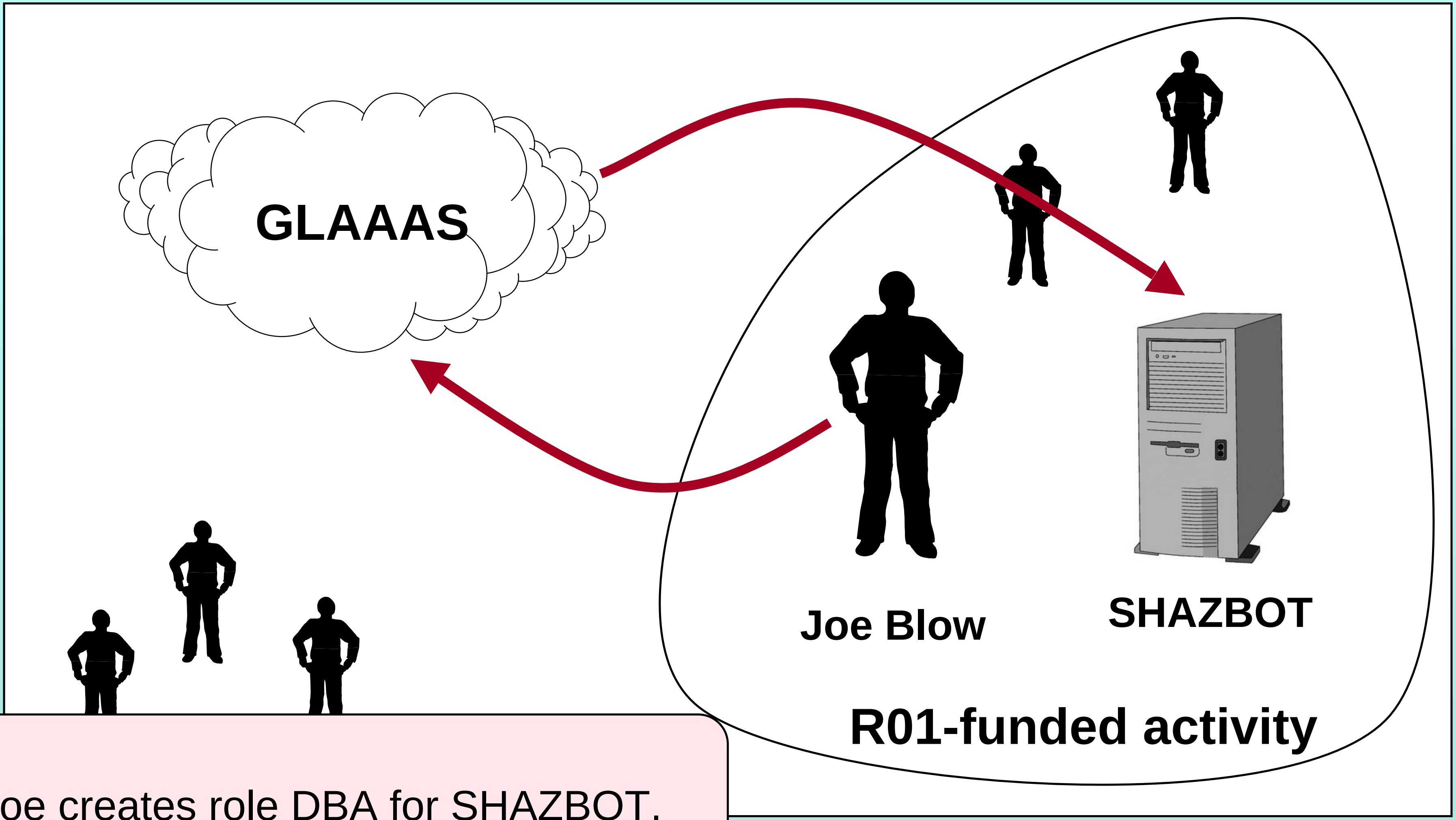
© 2003, Robert J. Robbins

GLAAAS



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

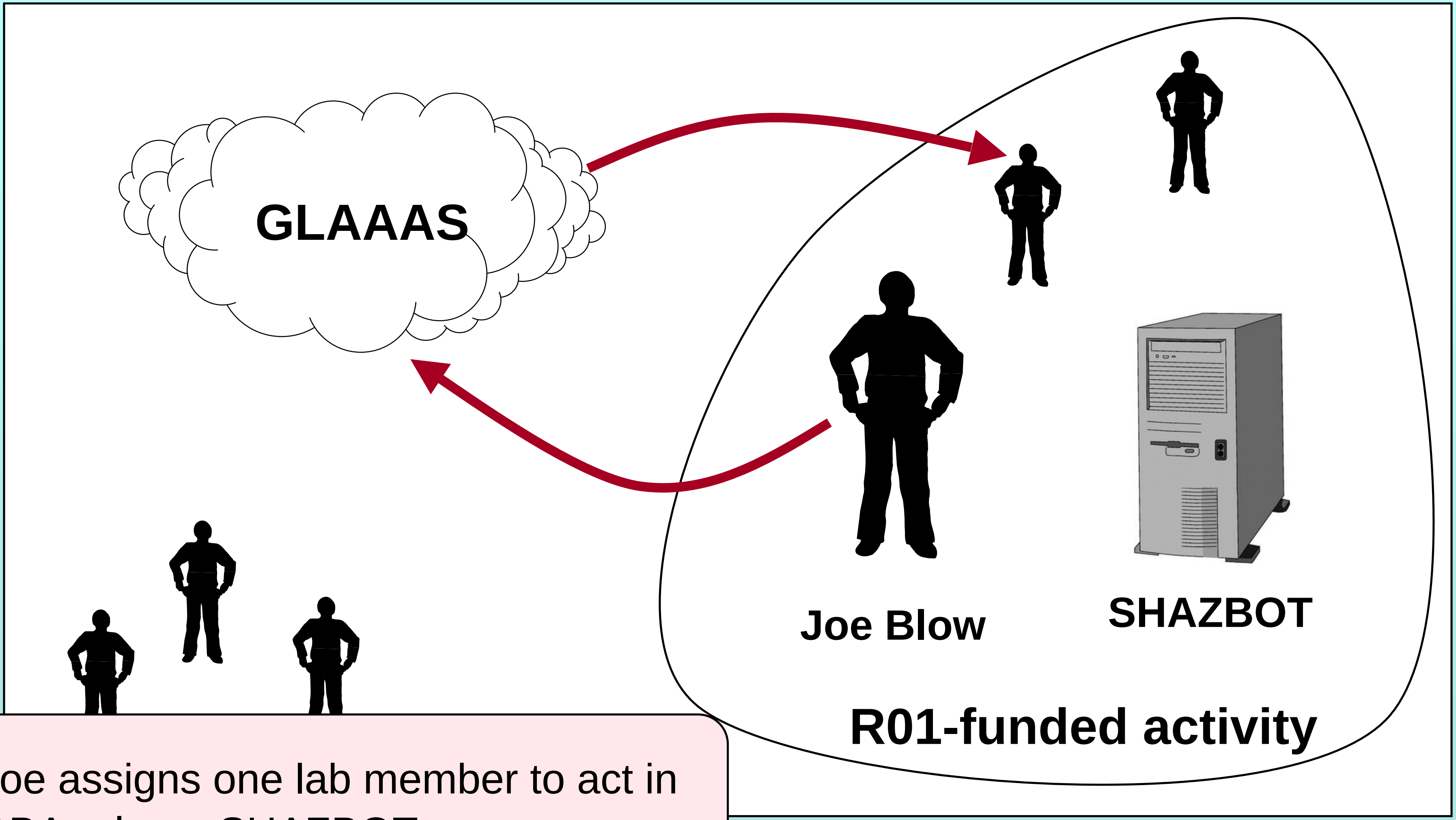
GLAAAS



Joe creates role DBA for SHAZBOT.

© 2003, Robert J. Robbins

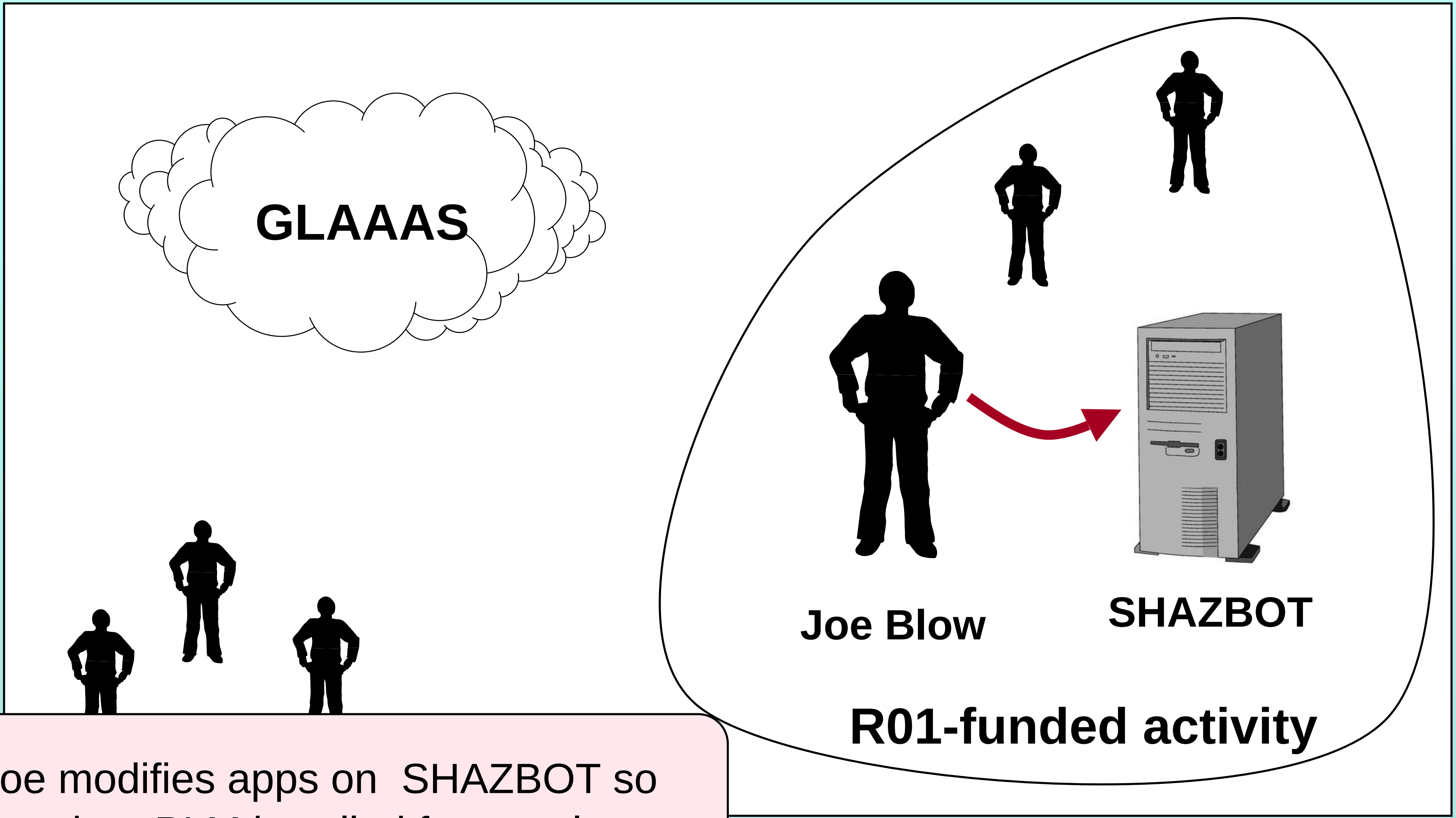
GLAAAS



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

© 2003, Robert J. Robbins

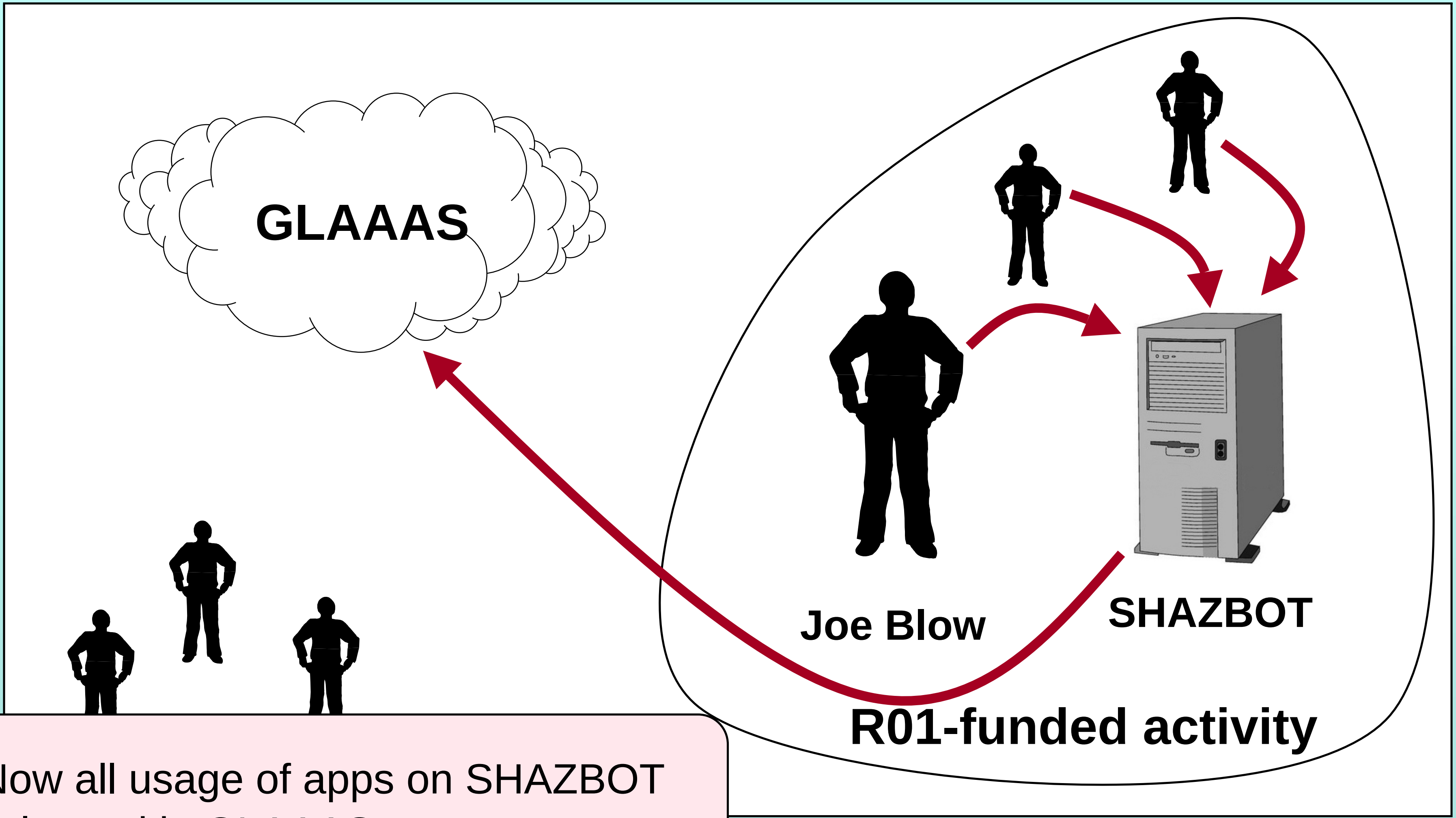
GLAAAS



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

© 2003, Robert J. Robbins

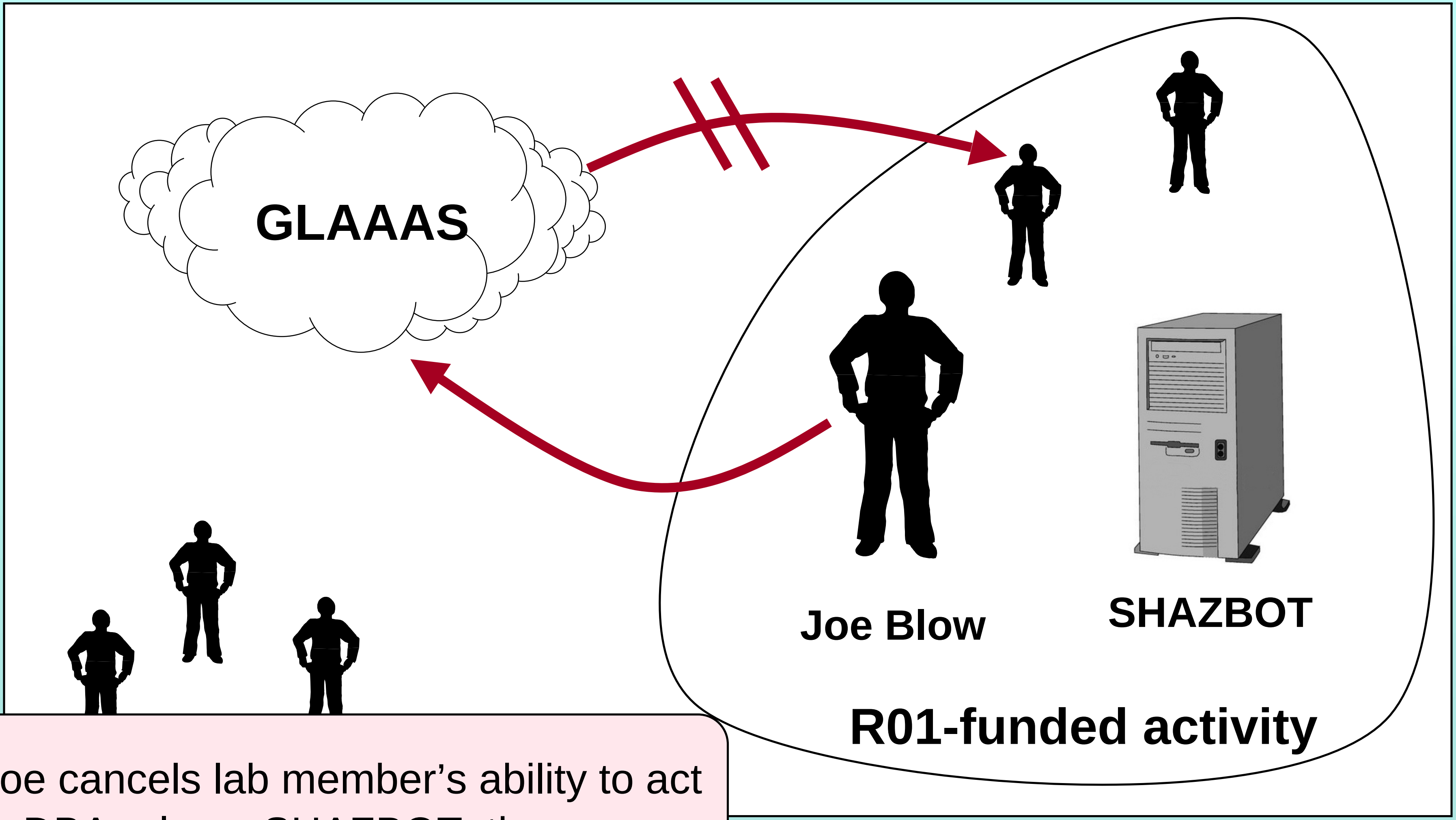
GLAAAS



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

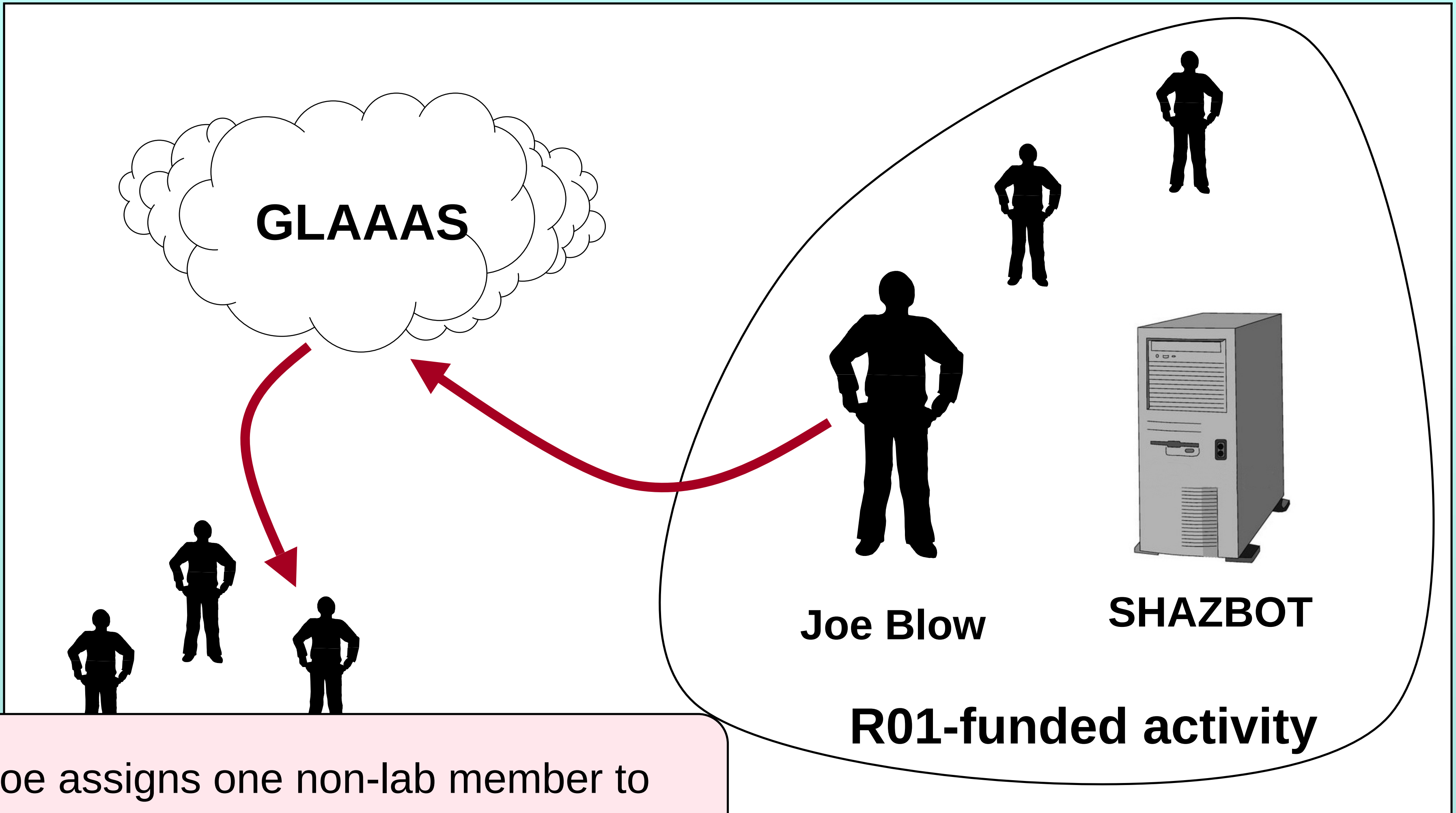
© 2003, Robert J. Robbins

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.

© 2003, Robert J. Robbins

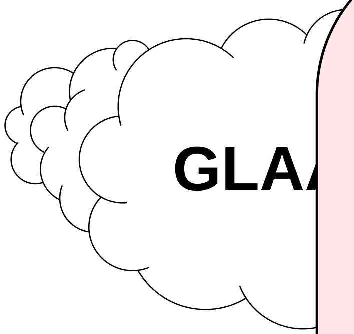
GLAAAS

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.

© 2003, Robert J. Robbins

GLAAAS



What else might GLAAAS do?

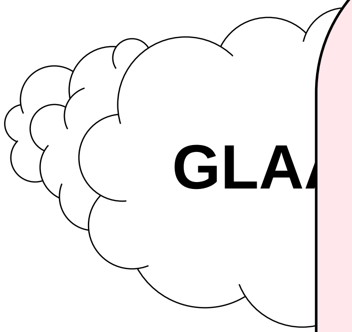
Provide truly GLOBAL support, by working with similar systems at other campuses?

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?

....?

© 2003, Robert J. Robbins

GLAAAS



Technically, how might GLAAAS actually work?

....?

Bottom-up Example

Database Issues

Relational Databases

Business Databases:

- FACTS
- REAL OBJECTS
- CLOSED UNIVERSE
- DEDUCTIVE REASONING

© 2003, Robert J. Robbins

Relational Databases

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

© 2003, Robert J. Robbins

Relational Databases

Facts:	**Observations:**
• SOLID	• SOFT
• STABLE	• CONSTANTLY CHANGING
• GLOBALLY CONSISTENT	• MUTUALLY INCONSISTENT

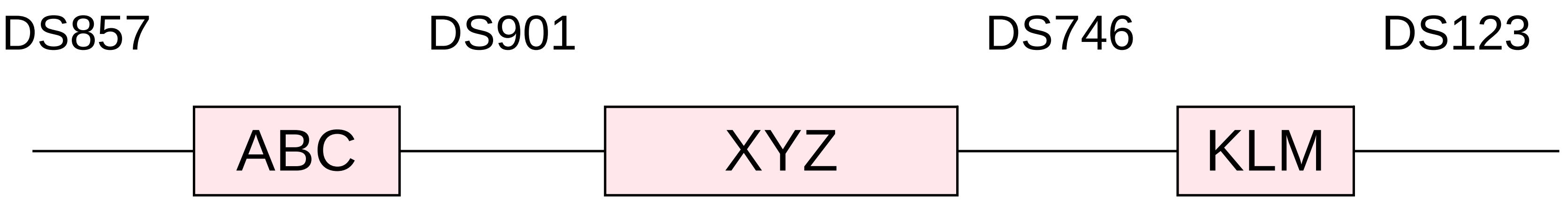
© 2003, Robert J. Robbins

Relational Databases

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

© 2003, Robert J. Robbins

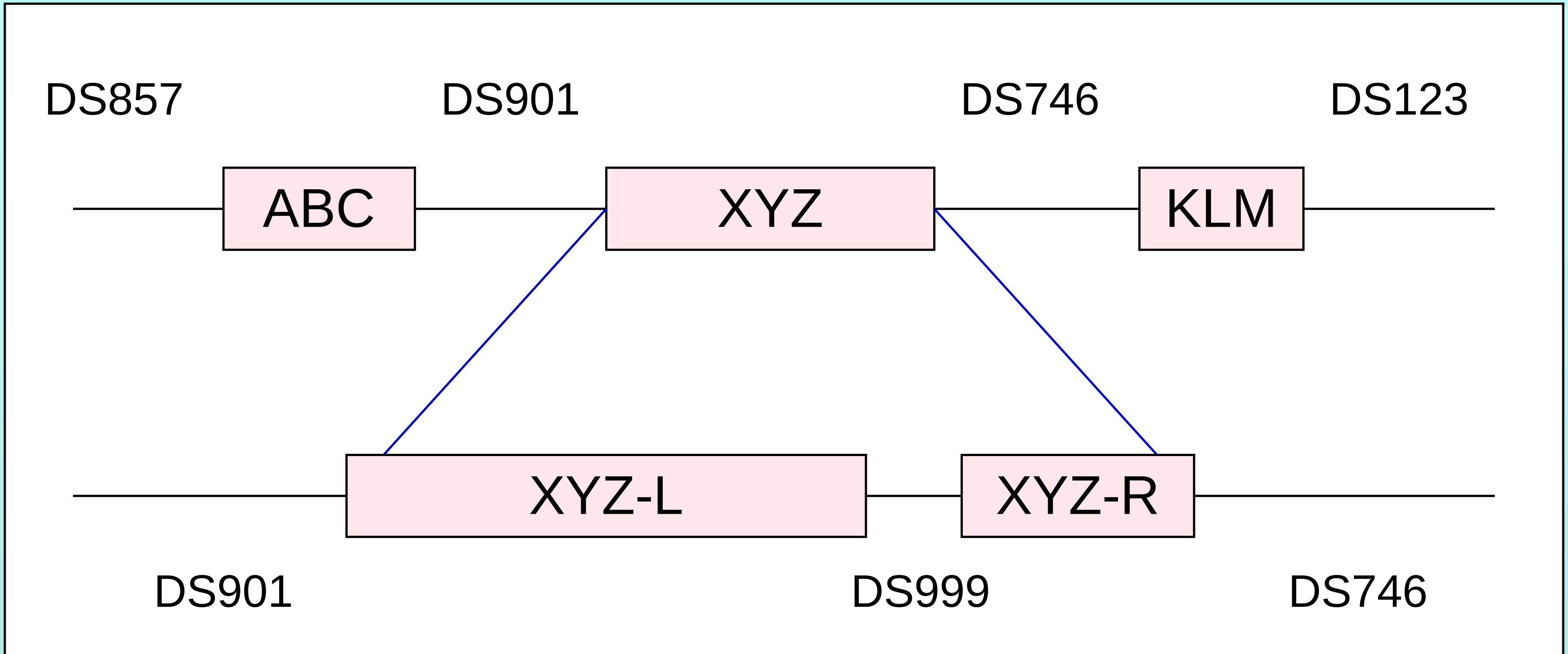
GDB Example:



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.

© 2003, Robert J. Robbins

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.

© 2003, Robert J. Robbins

Relational Databases

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

© 2003, Robert J. Robbins

Relational Databases

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

© 2003, Robert J. Robbins

Relational Databases

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

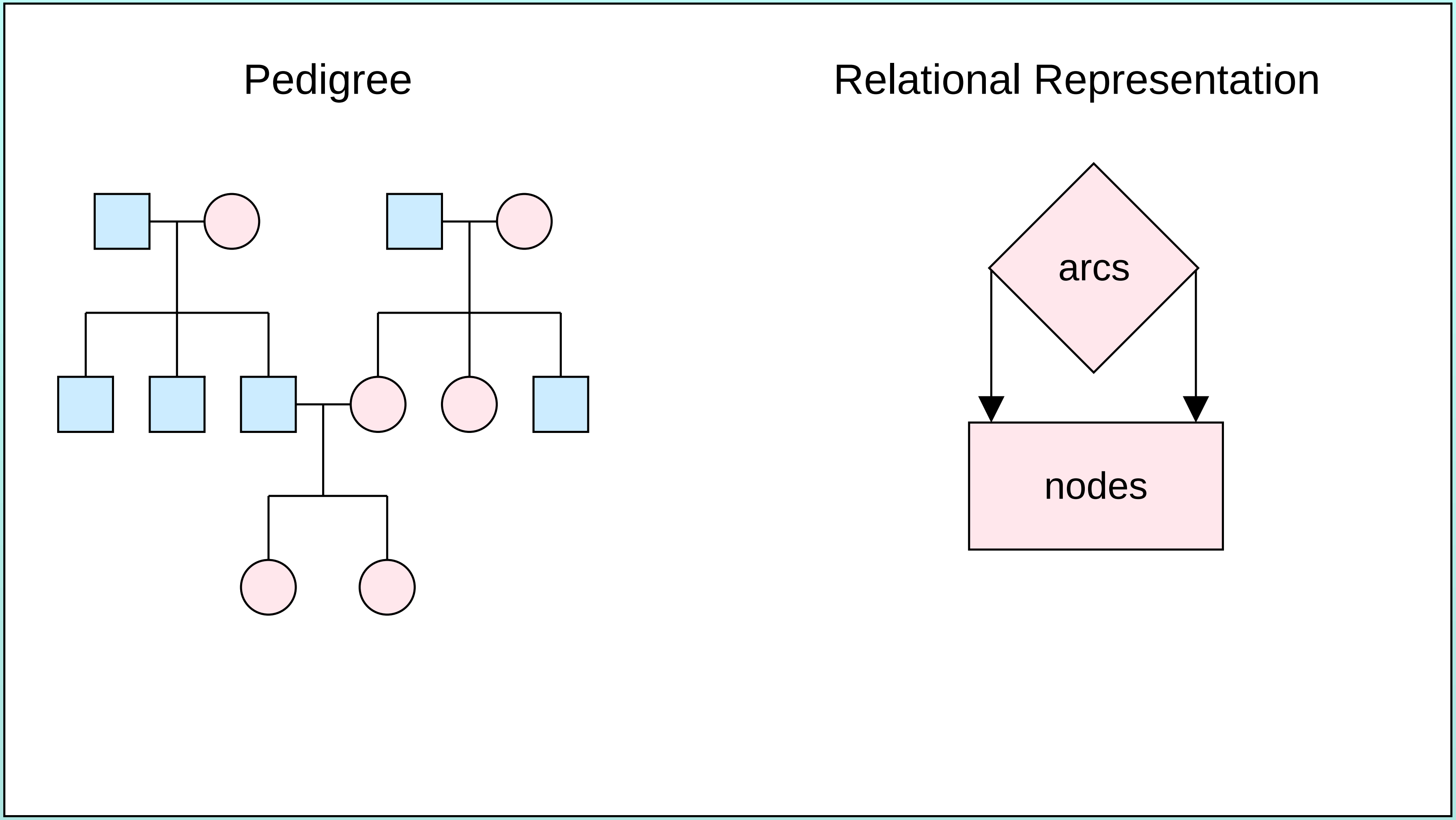
Relational Databases

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

© 2003, Robert J. Robbins

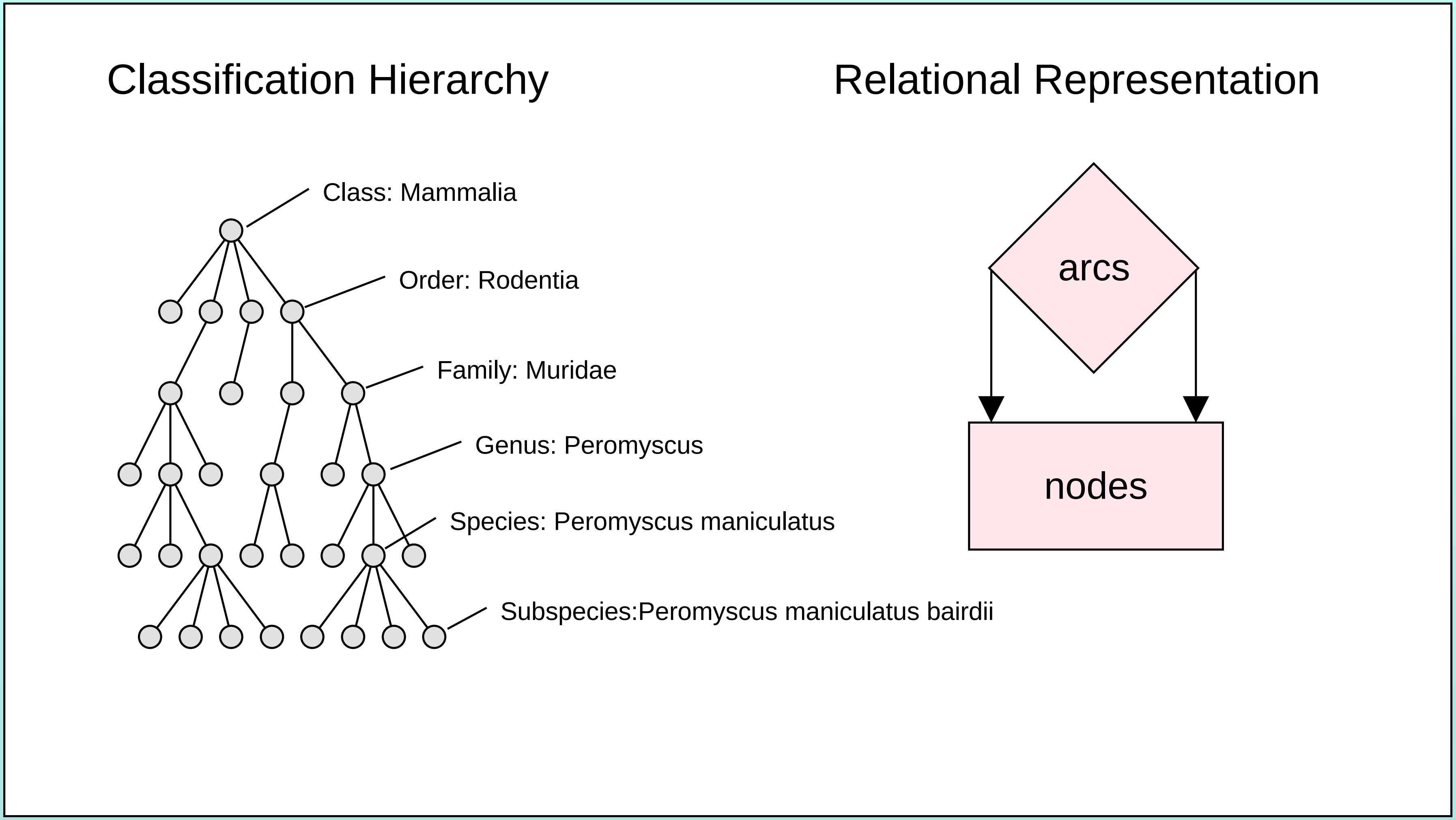
Data Model Problems

Graph Challenges



© 2003, Robert J. Robbins

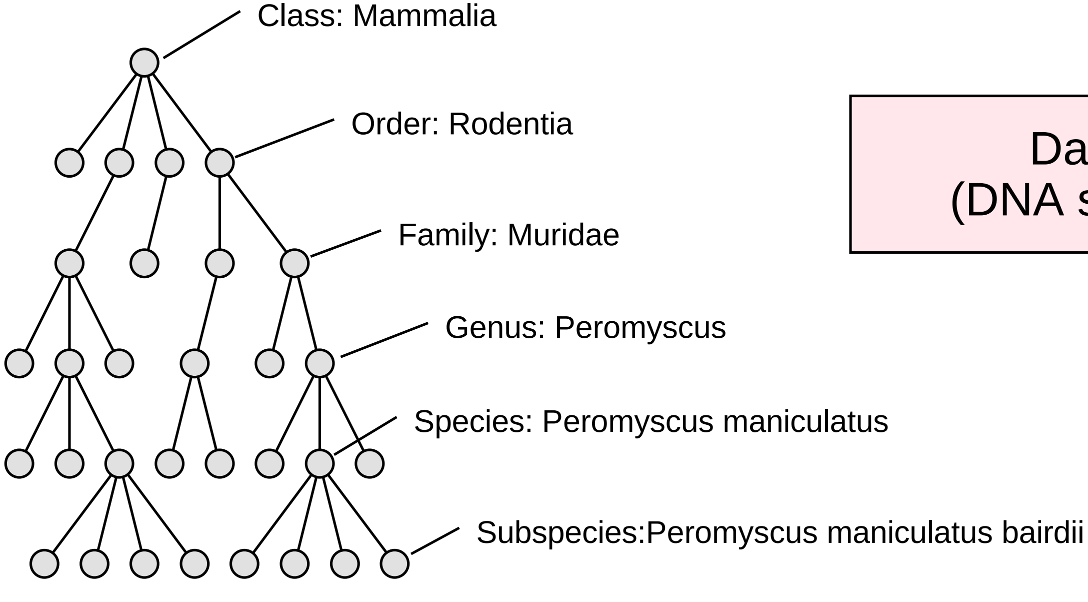
Graph Challenges



© 2003, Robert J. Robbins

Classification Challenges

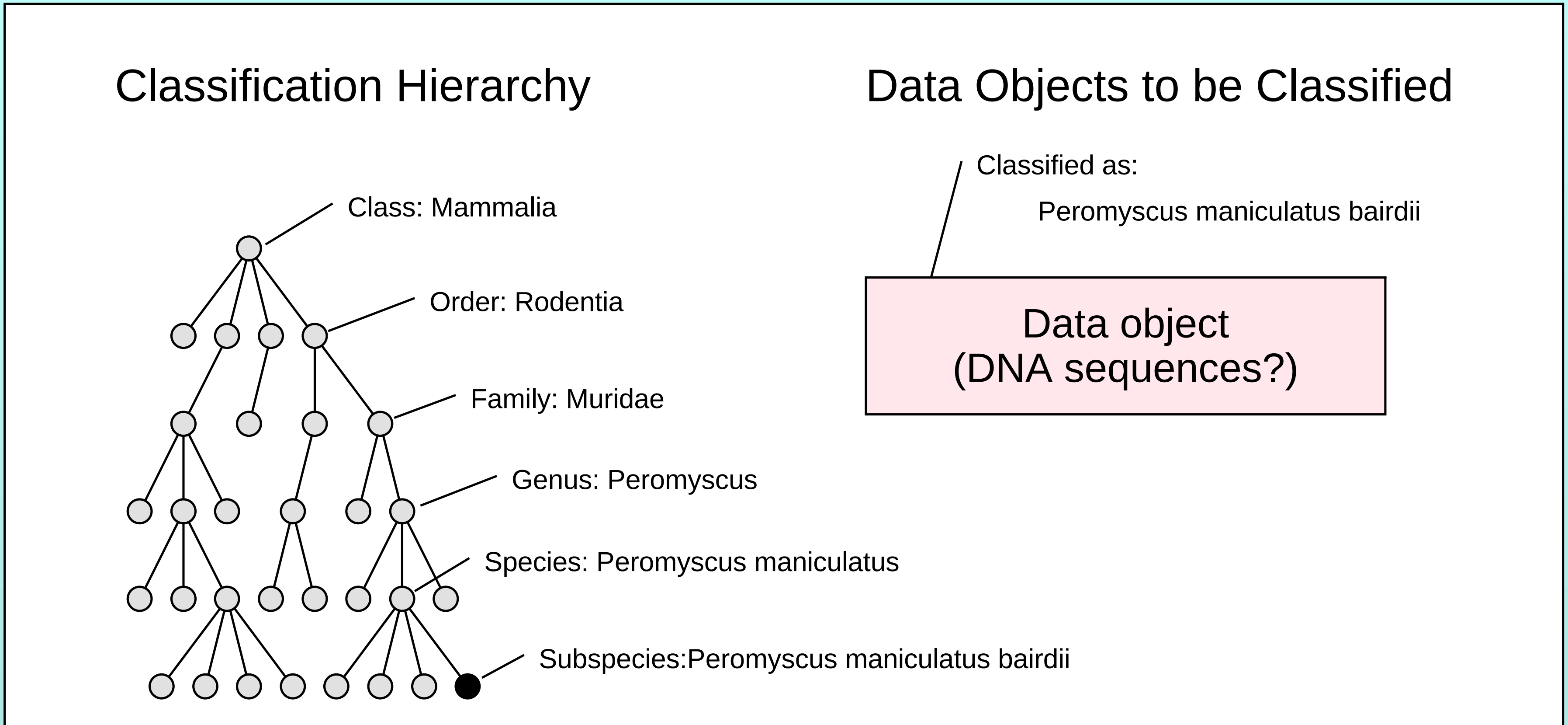
Classification Hierarchy



Data Objects to be Classified

Data object
(DNA sequences?)

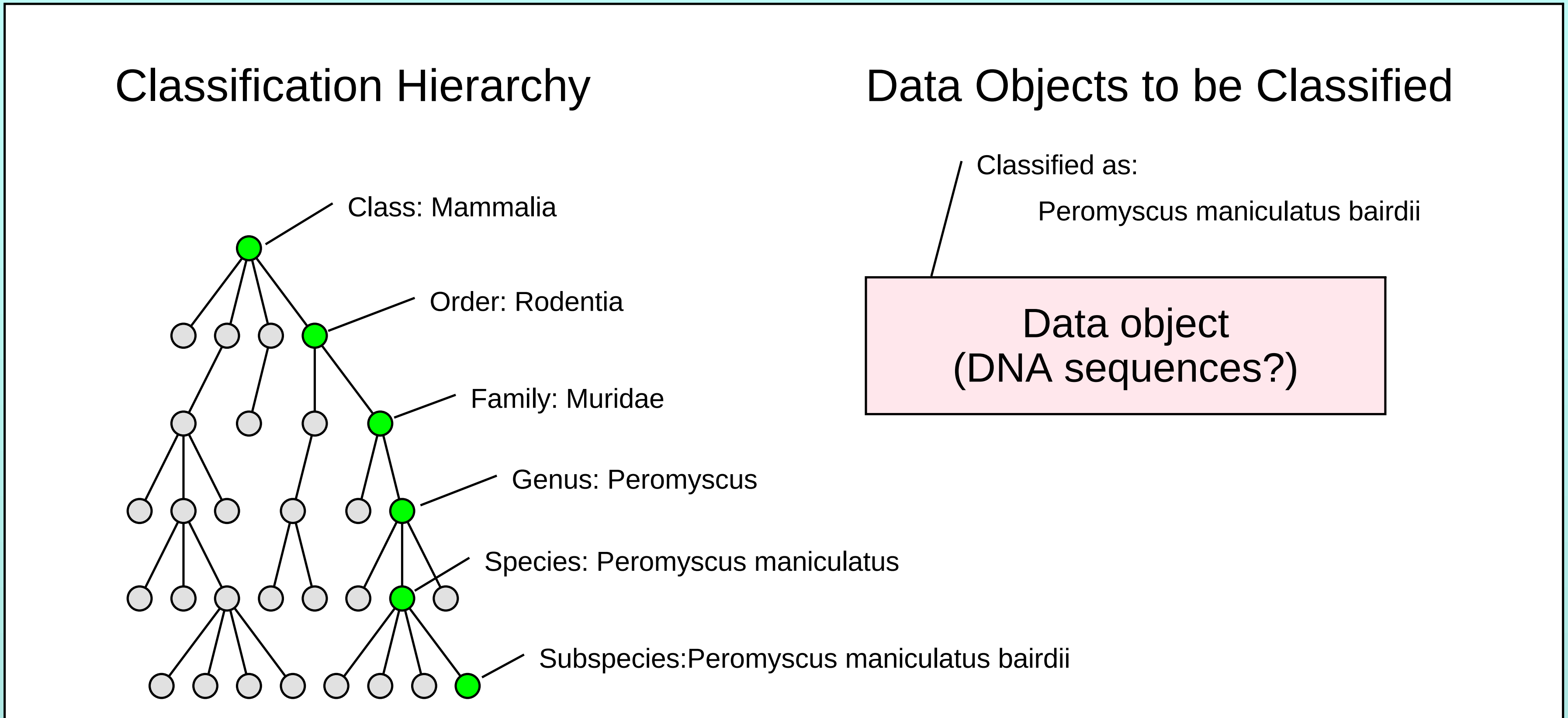
Classification Challenges



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

© 2003, Robert J. Robbins

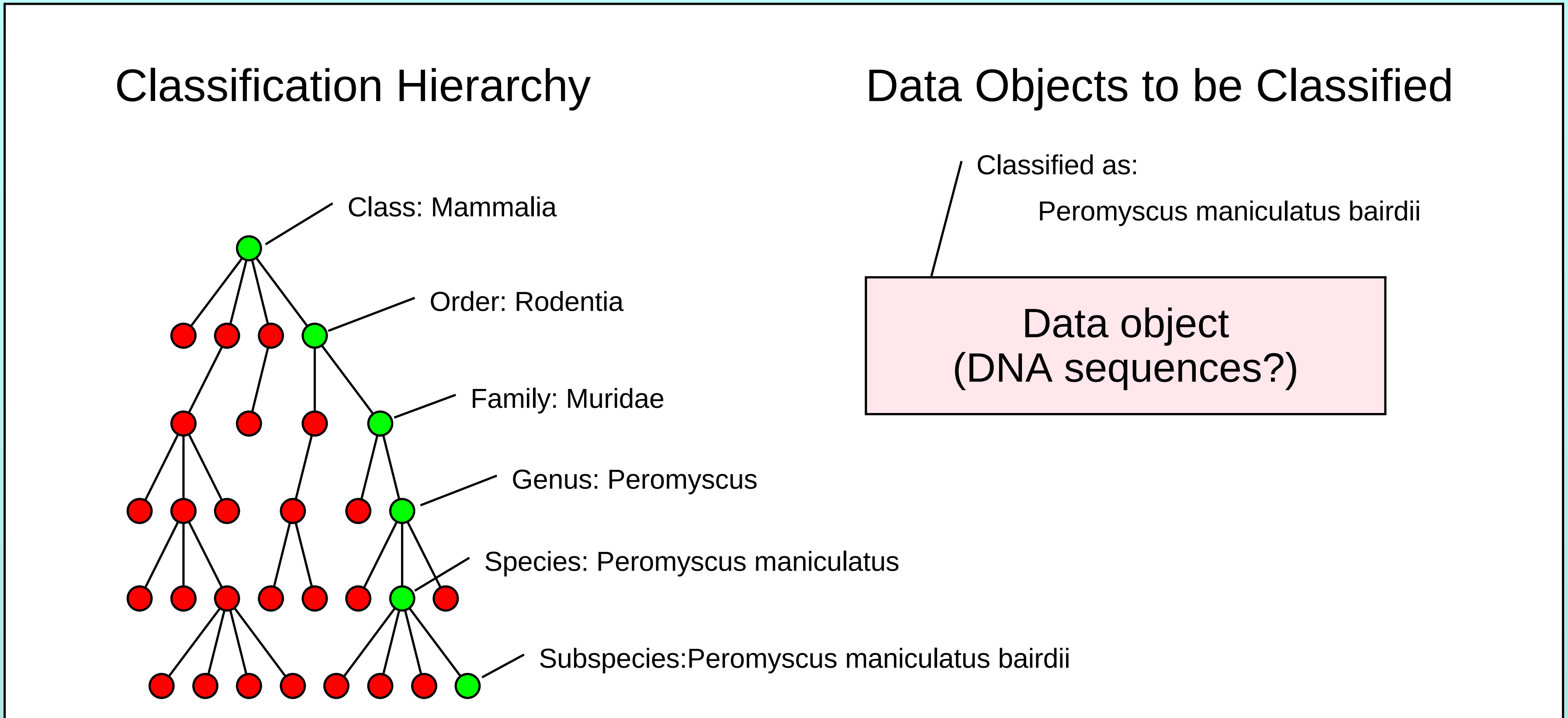
Classification Challenges



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**,

© 2003, Robert J. Robbins

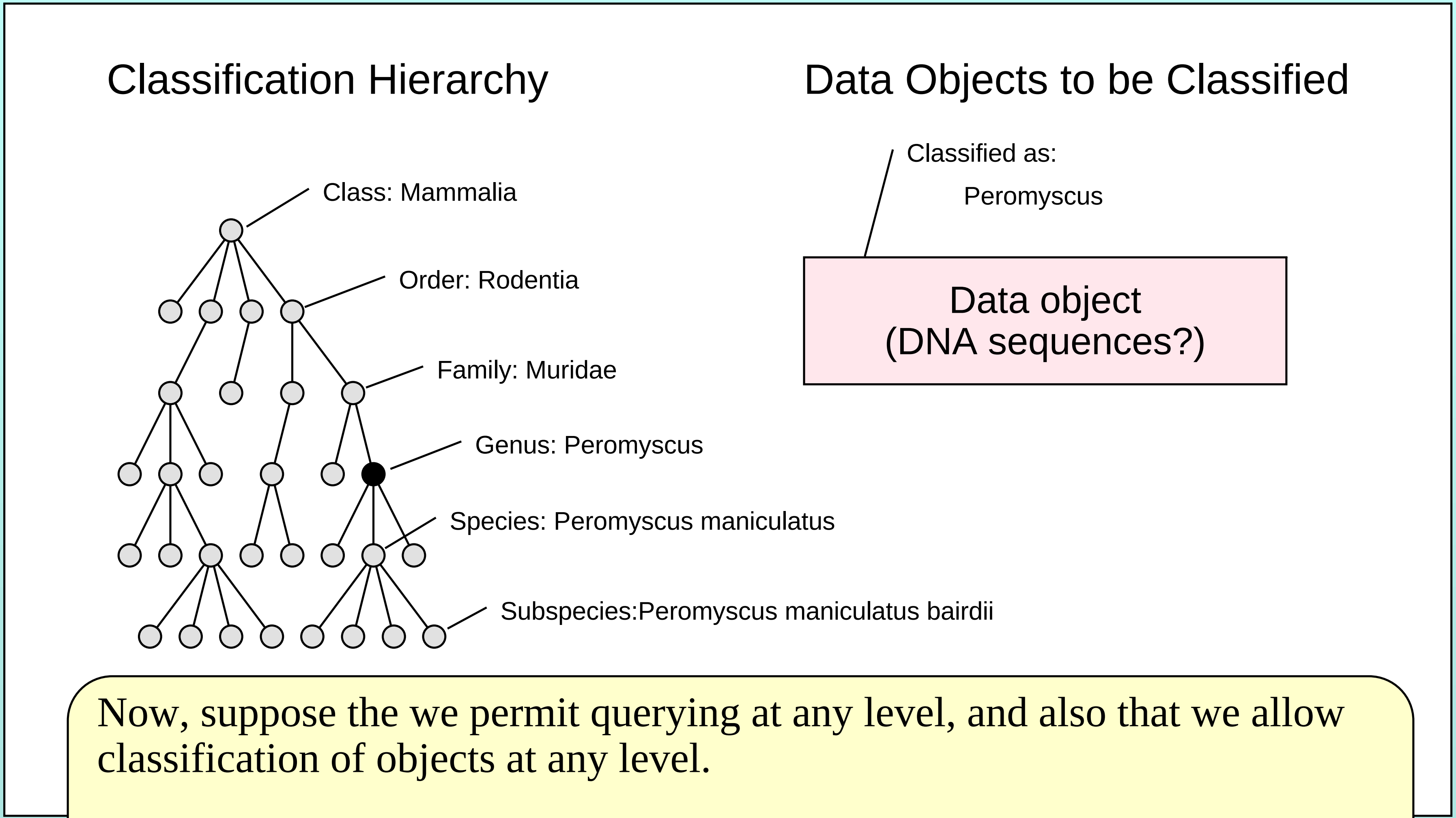
Classification Challenges



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**, all others **FALSE**.

© 2003, Robert J. Robbins

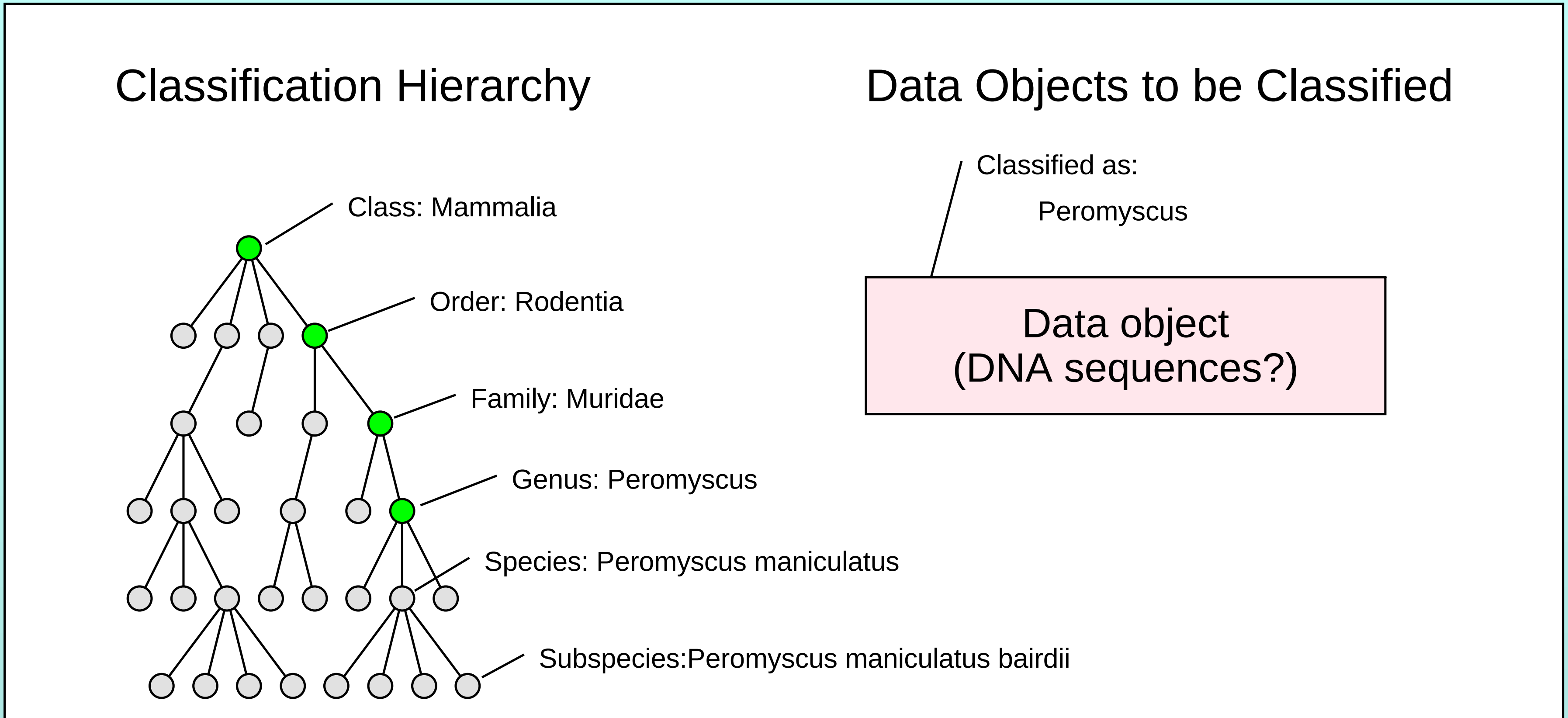
Classification Challenges



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

© 2003, Robert J. Robbins

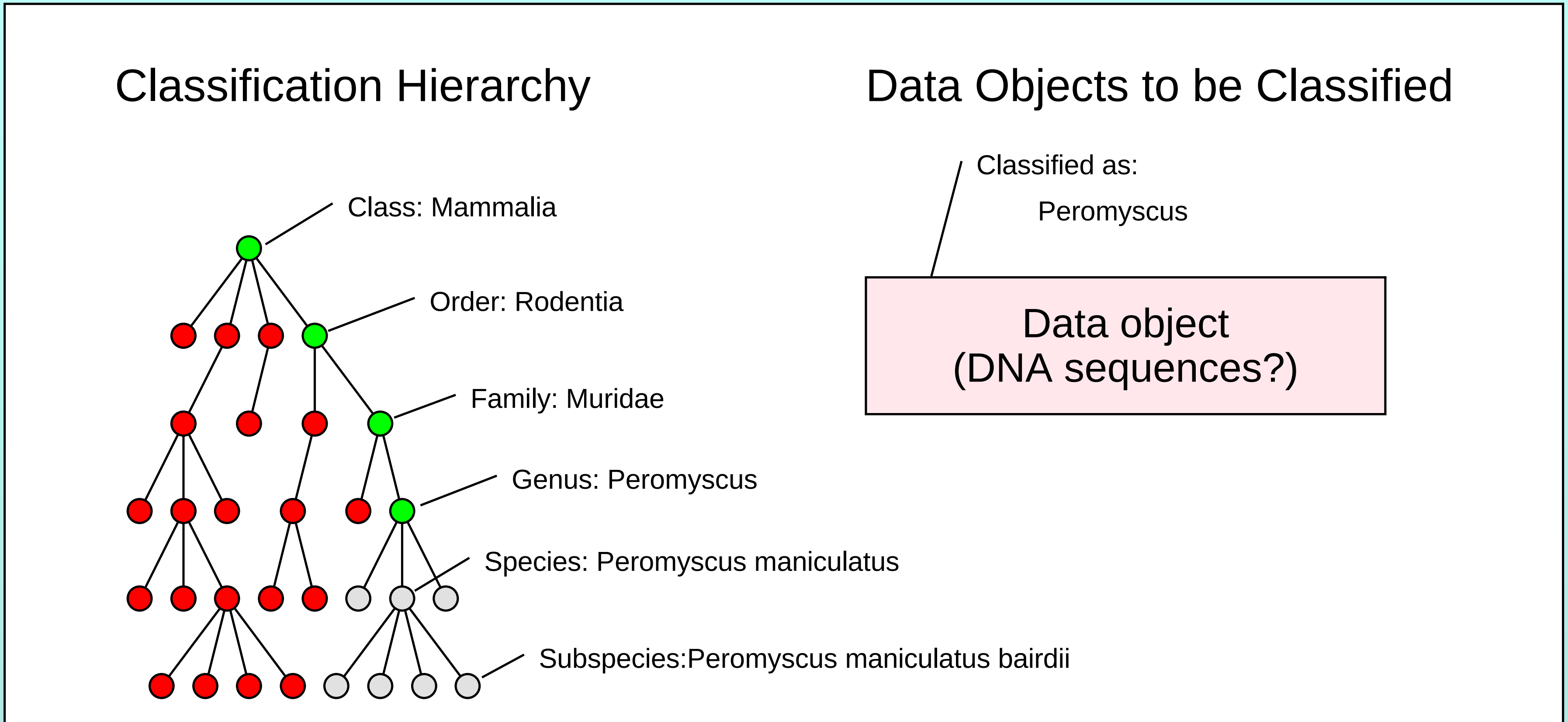
Classification Challenges



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**,

© 2003, Robert J. Robbins

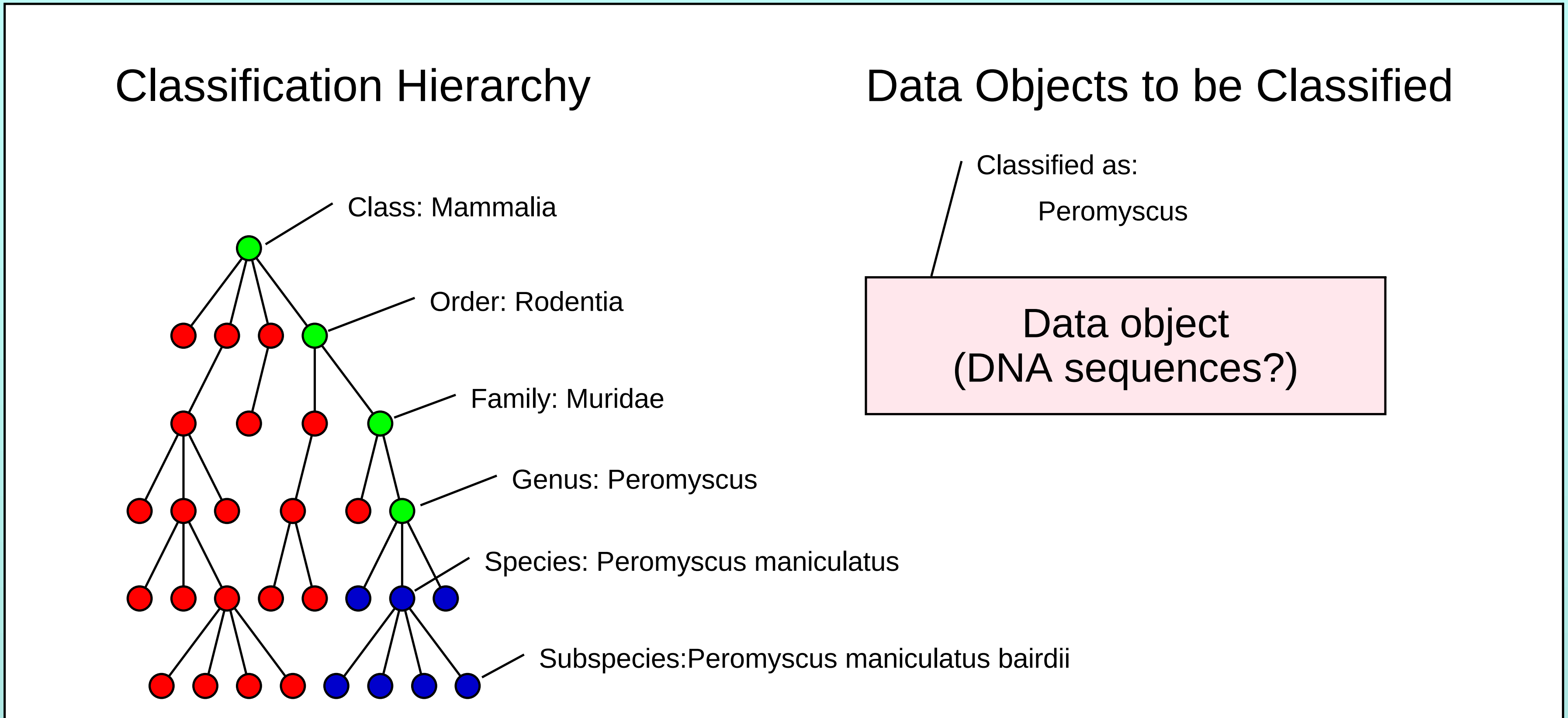
Classification Challenges



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**,

© 2003, Robert J. Robbins

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**.

© 2003, Robert J. Robbins

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.

© 2003, Robert J. Robbins

Data Integration Goals

Achieve conceptual integration of biomedical data.

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

© 2003, Robert J. Robbins

Data Integration Impediments

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

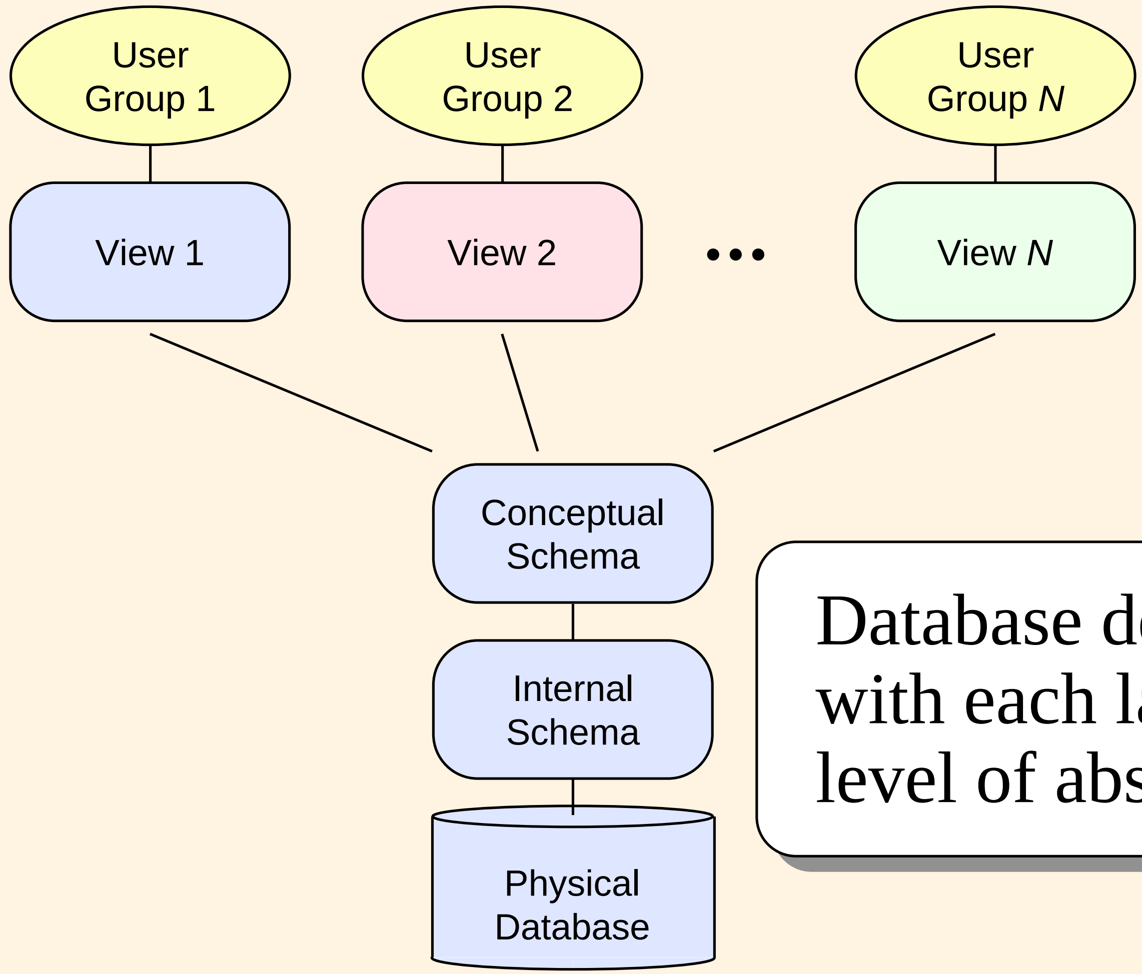
Sociological: Local incentives encourage competition, not cooperation.

Conceptual: Semantic mismatches exist among databases.

© 2003, Robert J. Robbins

Technical Impediments

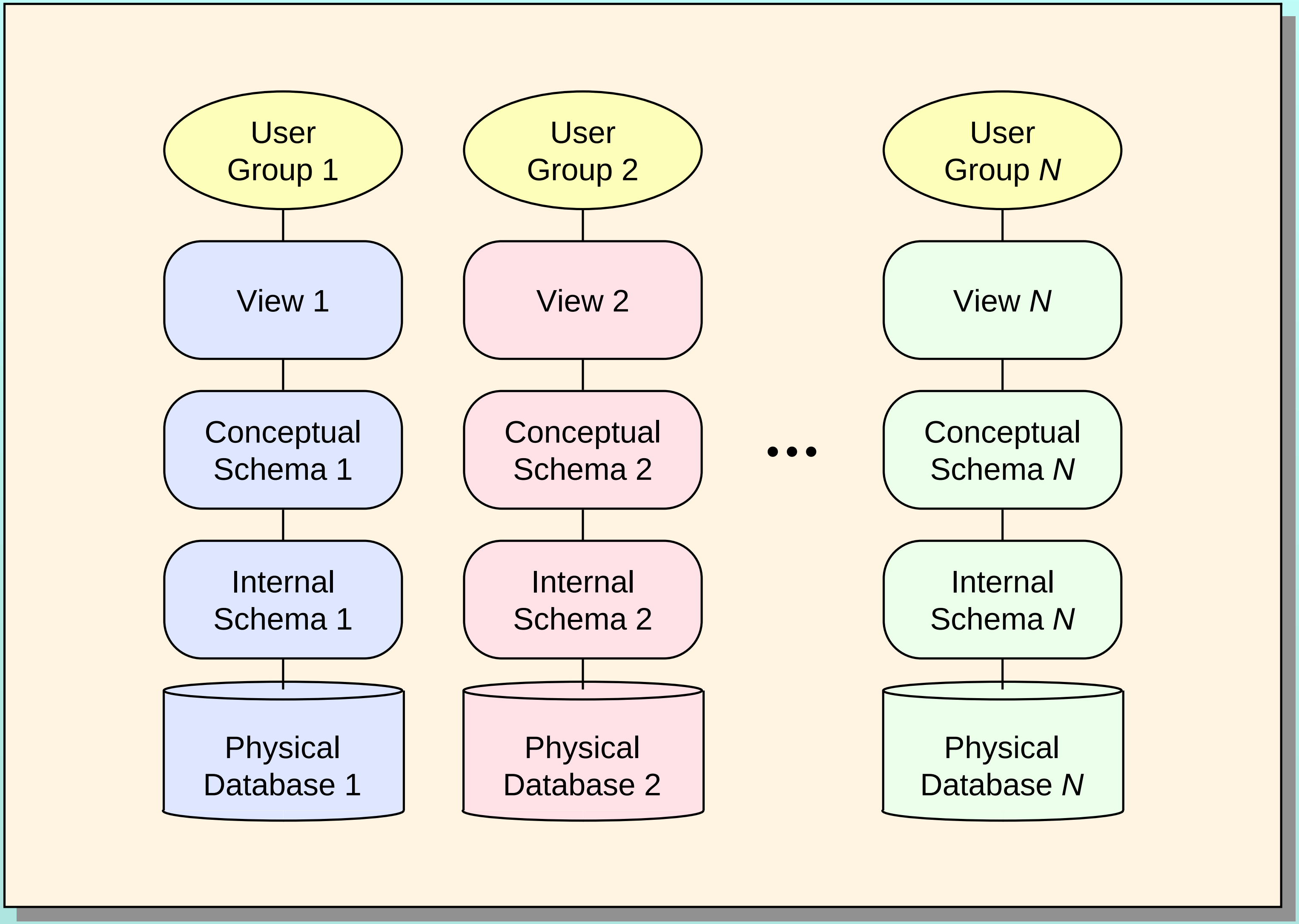
Multiple Views



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

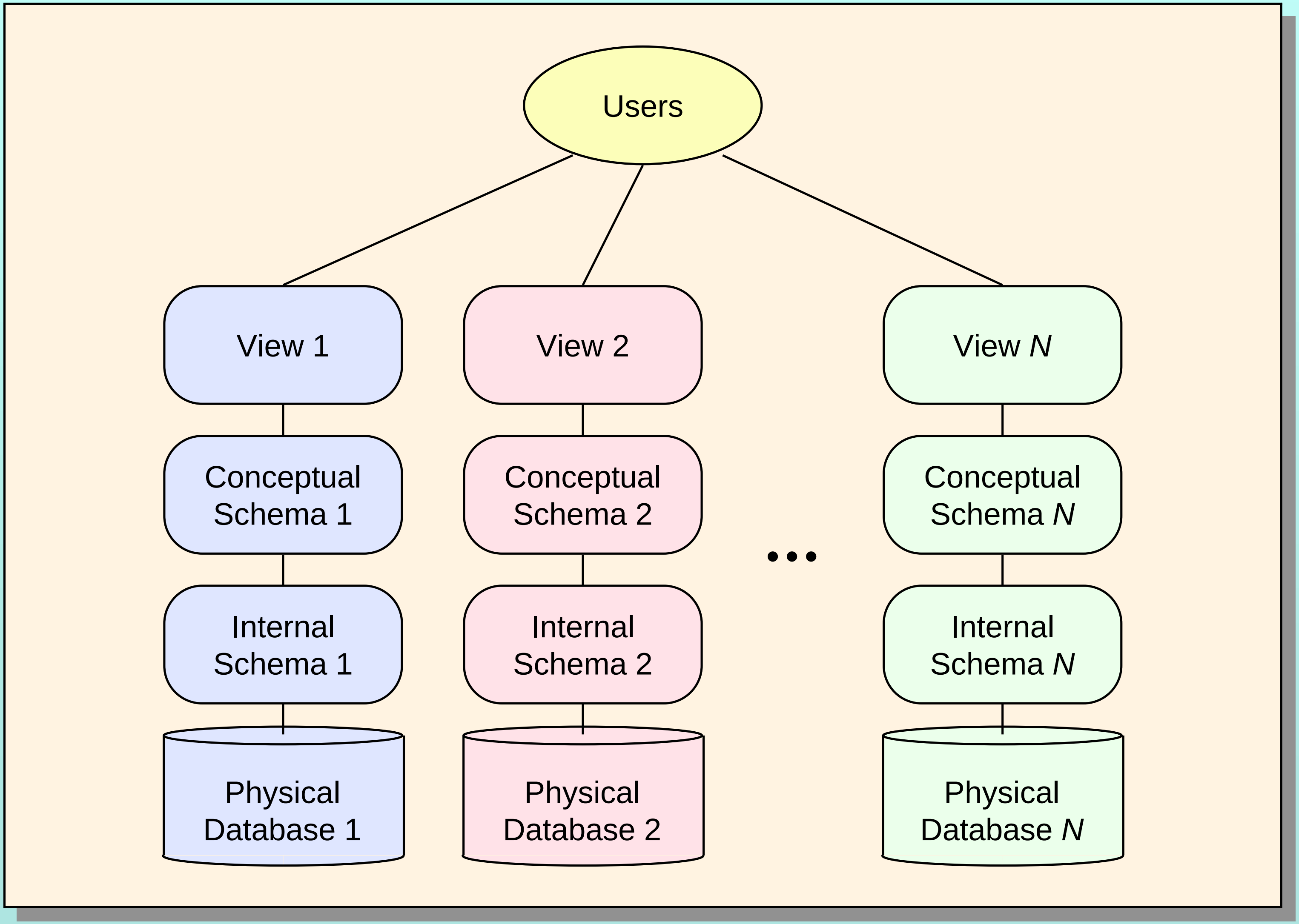
© 2003, Robert J. Robbins

Multiple Databases



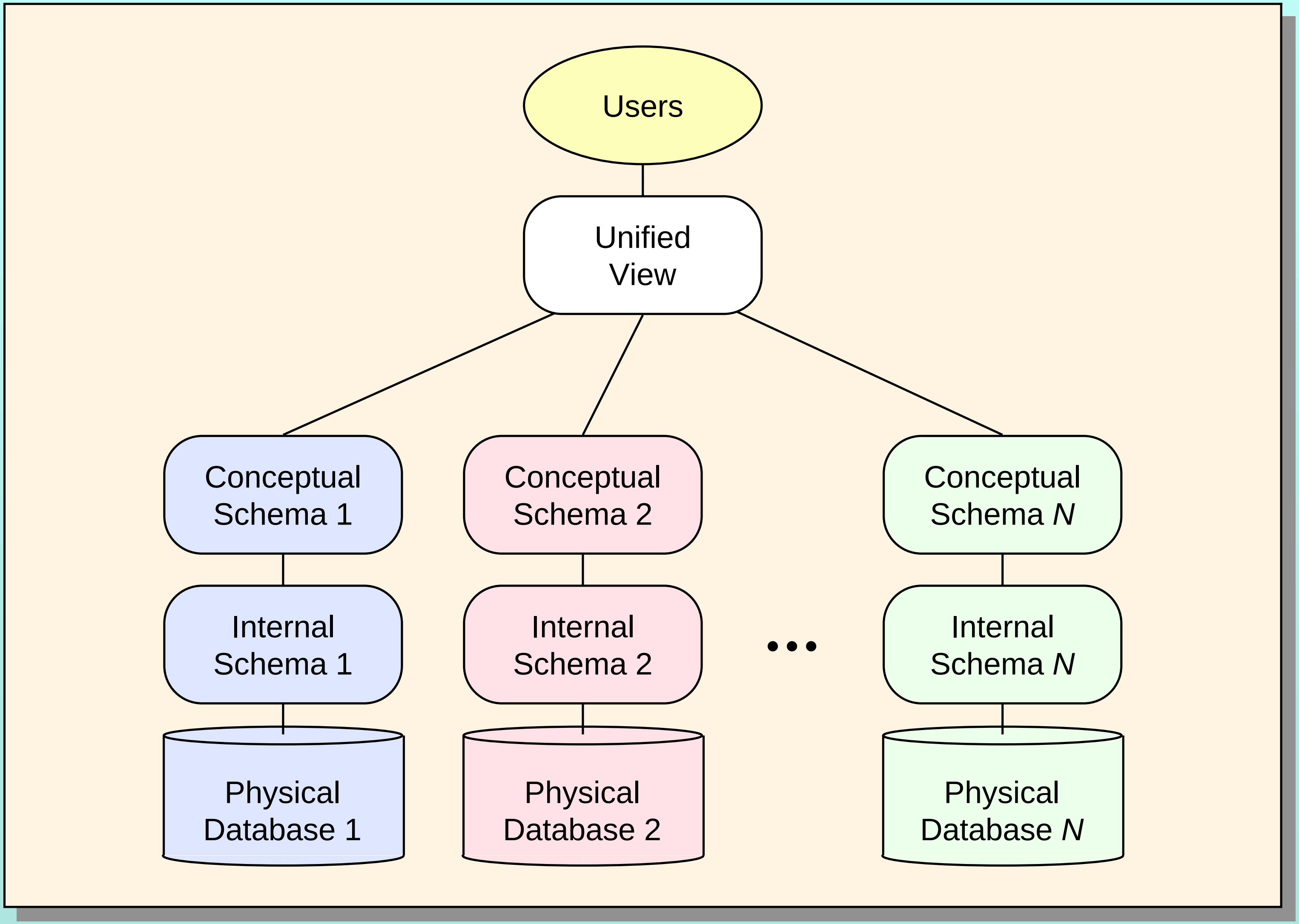
© 2003, Robert J. Robbins

Current Situation



© 2003, Robert J. Robbins

Desired Situation



© 2003, Robert J. Robbins

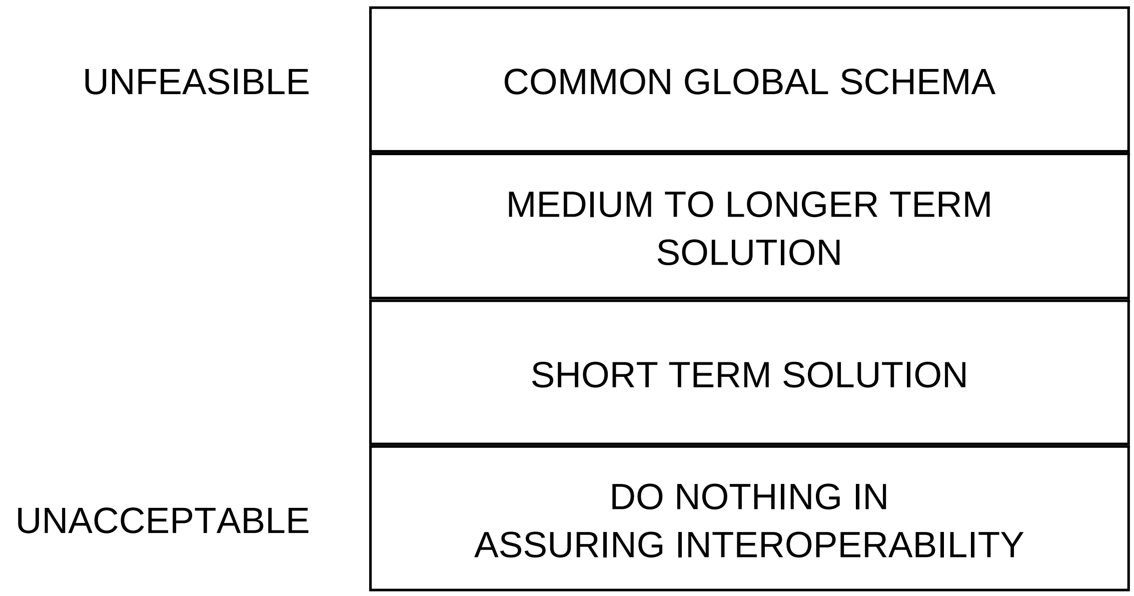
The Vision

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

© 2003, Robert J. Robbins

Taxonomy of Multidatabase Systems



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

© 2003, Robert J. Robbins

Taxonomy of Multidatabase Systems

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

•

•

•

adoption of common DBMSs at participating sites

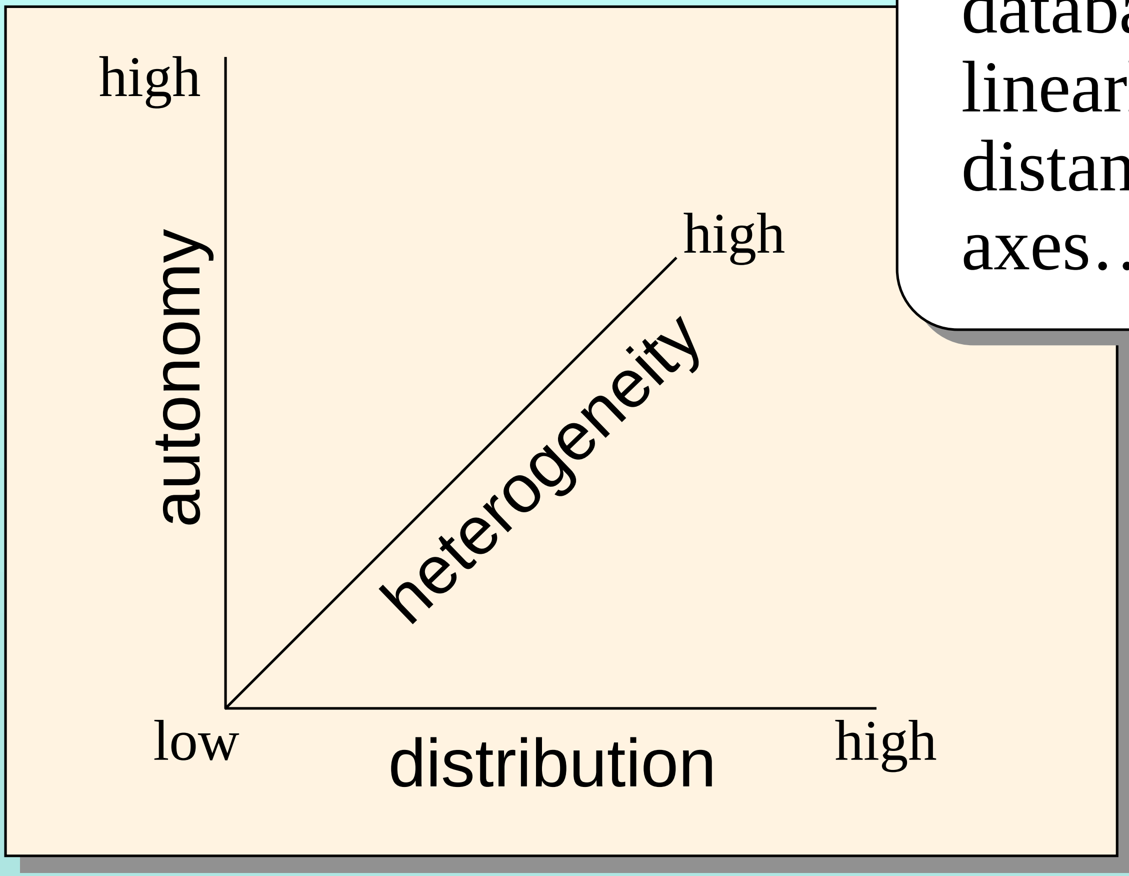
shared data model across participating sites

common semantics for data publishing

Loosely Coupled: common syntax for data publishing

© 2003, Robert J. Robbins

Difficulty Dimensions



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

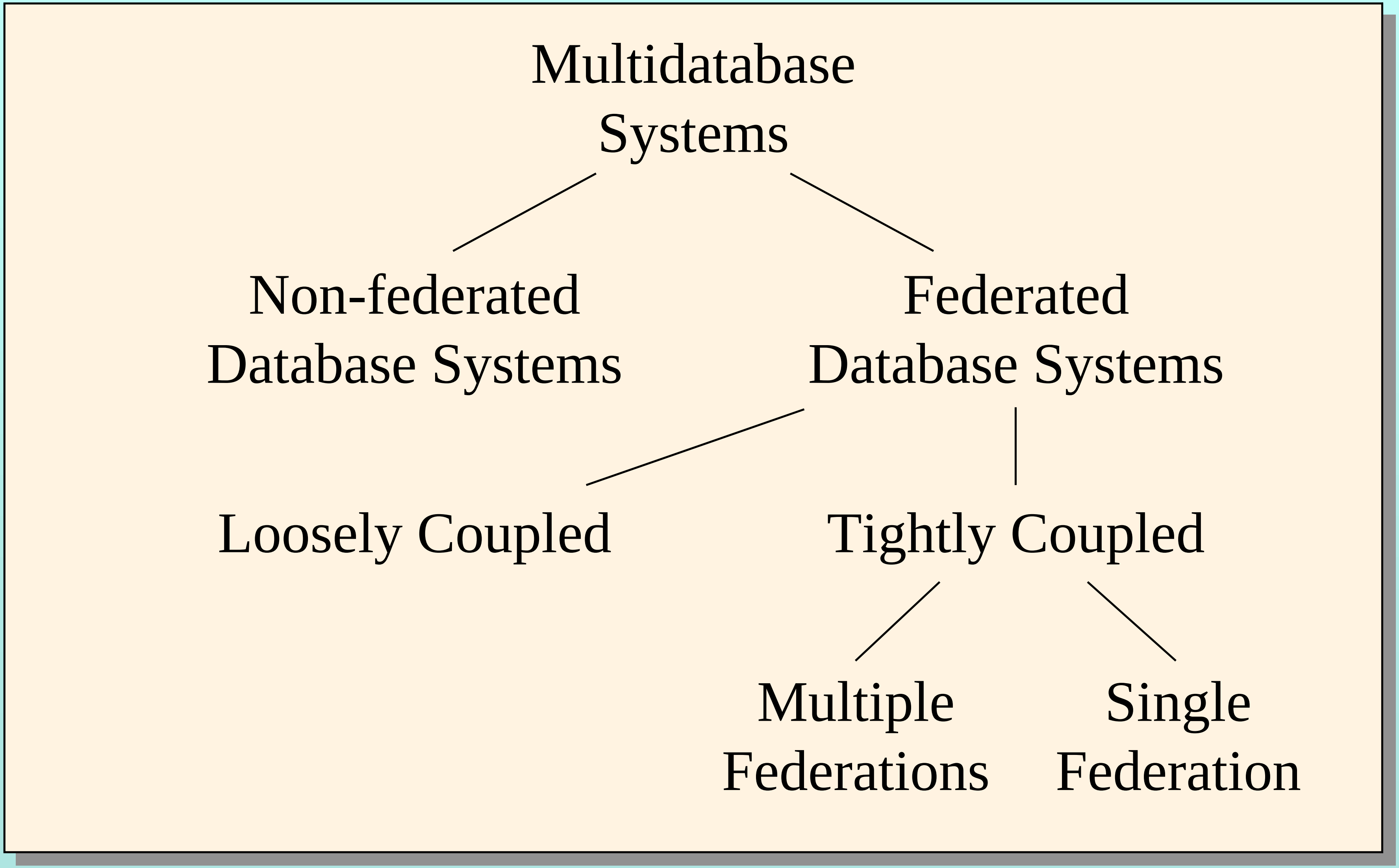
© 2003, Robert J. Robbins

Taxonomy of Multidatabase Systems

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.

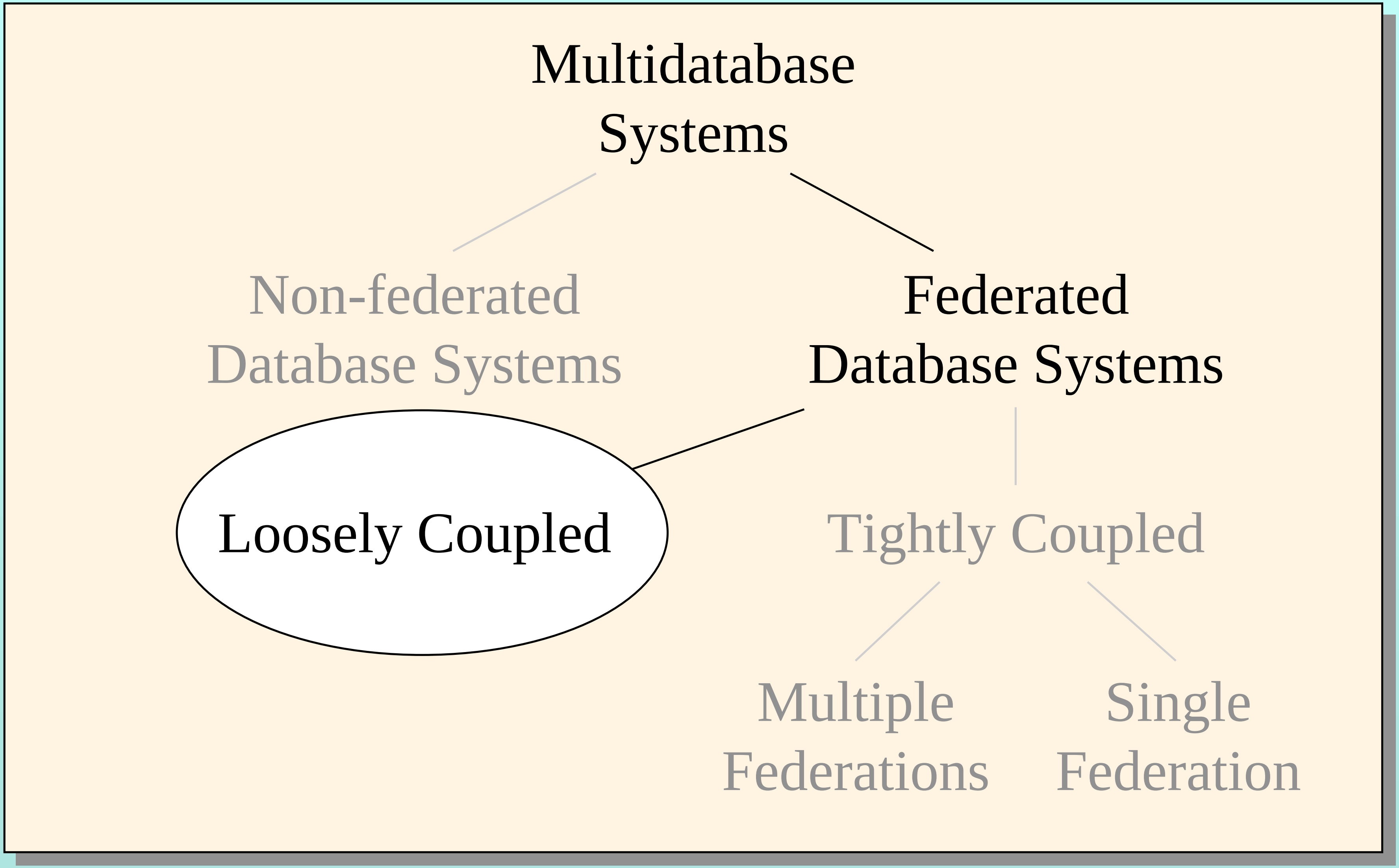
© 2003, Robert J. Robbins

Taxonomy of Multidatabase Systems



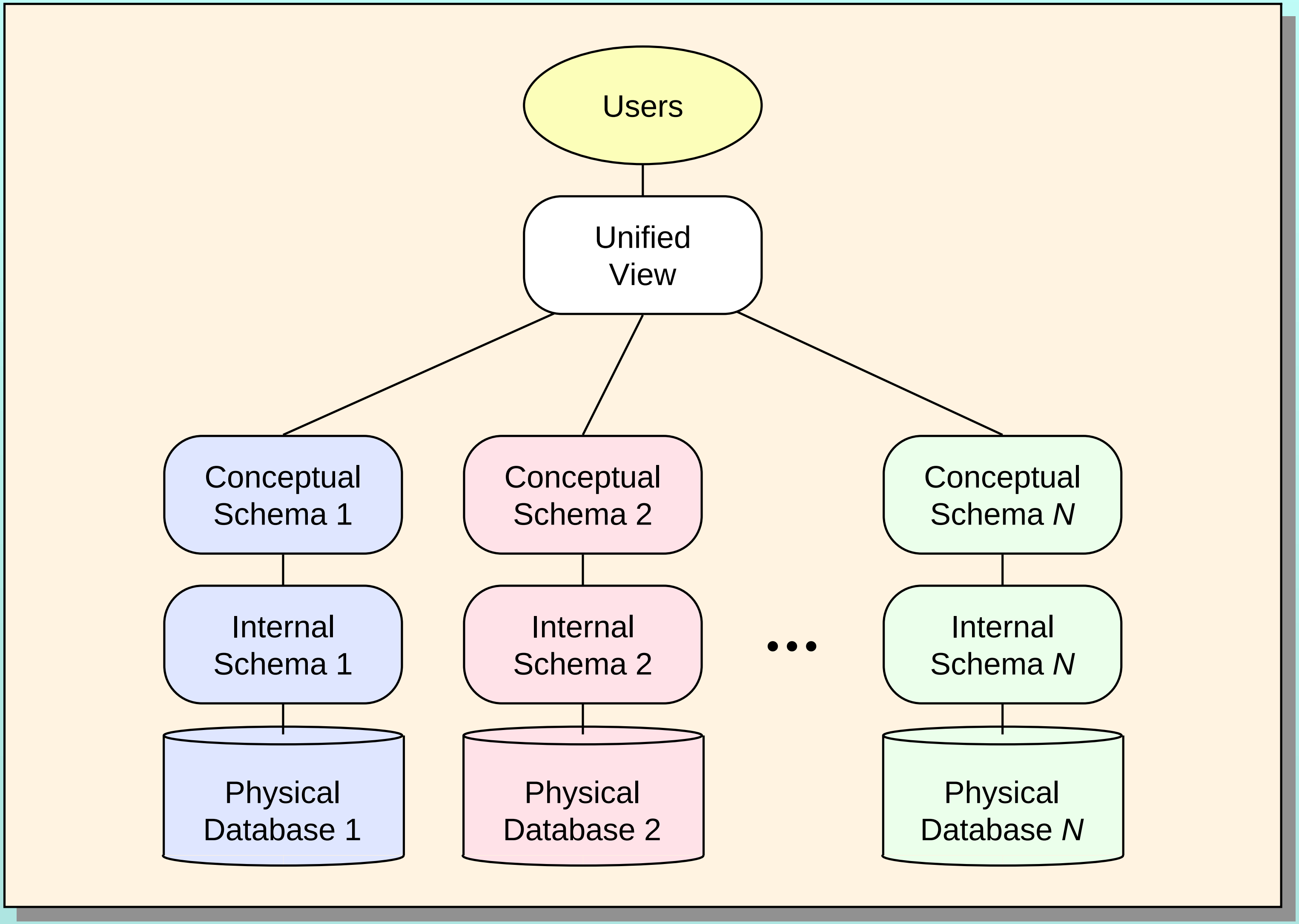
© 2003, Robert J. Robbins

Taxonomy of Multidatabase Systems



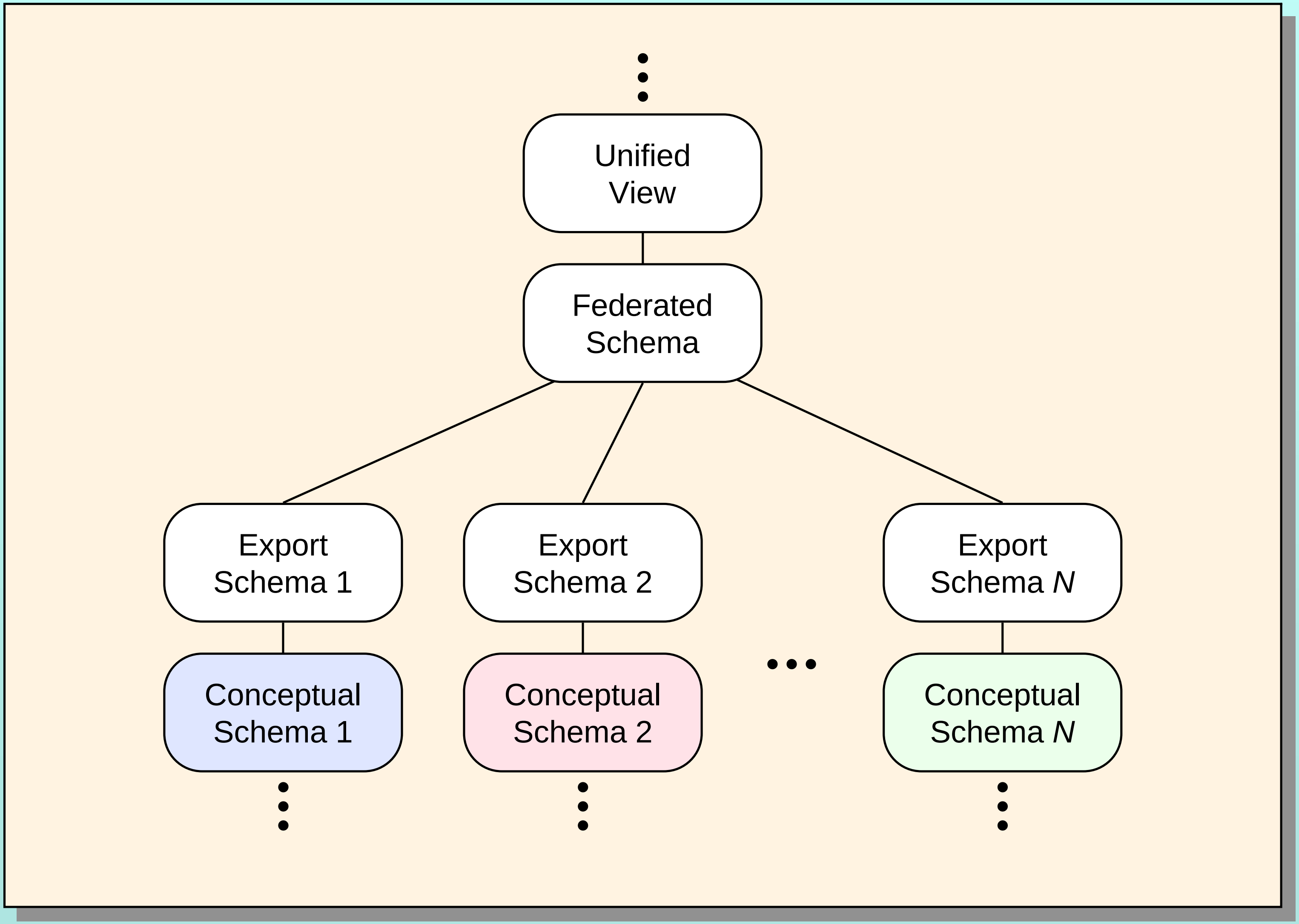
© 2003, Robert J. Robbins

Desired Situation



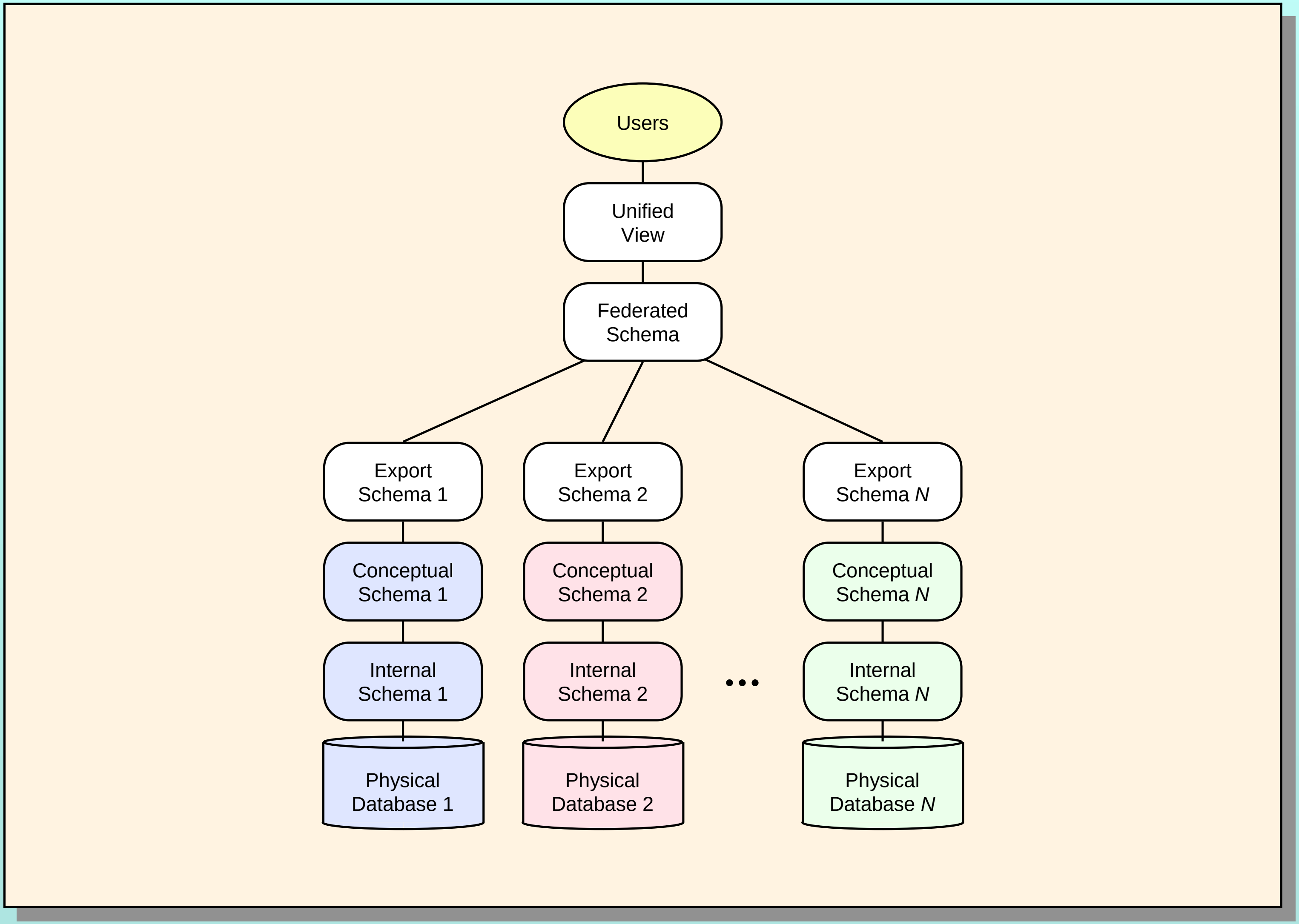
© 2003, Robert J. Robbins

More Layers



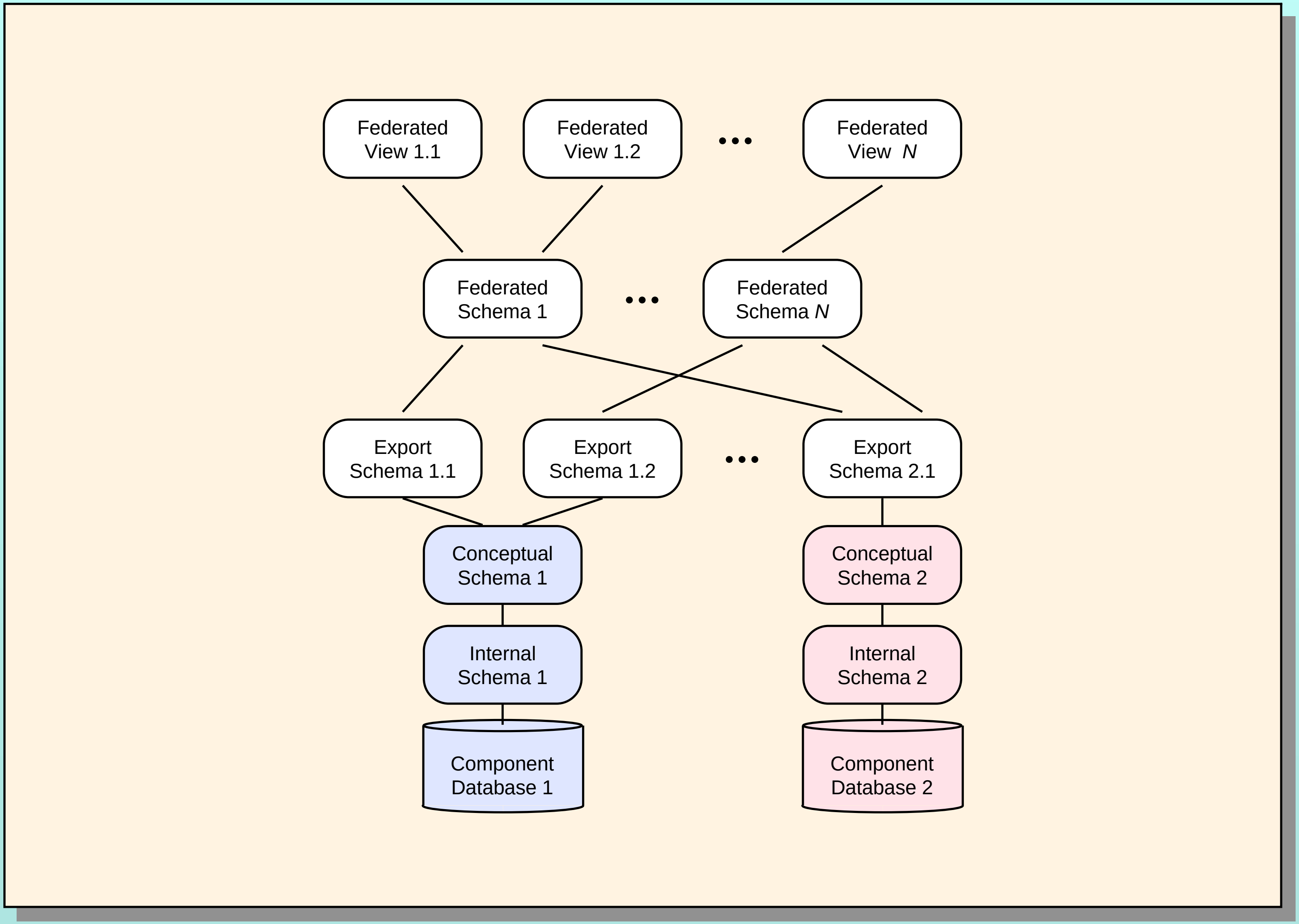
© 2003, Robert J. Robbins

Federated Schema



© 2003, Robert J. Robbins

Multiple Federations



© 2003, Robert J. Robbins

Federated Information Infrastructure

Public Funding of Databases

Stand-alone Criteria:

- Is there a need?
- Will this meet the need?
- Can they do it?
- Is it worth it?

© 2003, Robert J. Robbins

Public Funding of Databases

Global Criteria:

- Does it adhere to standards?
- Will it interoperate?
- Is there commitment to federation?
- Is it worth it?

© 2003, Robert J. Robbins

Information Resources and the GII

Guiding Principles:

- Global value explosion
- Componentry
- Anonymous interoperability
- Technical scalability
- Social scalability
- Value additivity

© 2003, Robert J. Robbins

Enough Examples!

Let’s Get to Work

Working Group Assignments

For each module:

Background
- The Problem
- Available Solutions
- Remaining Challenges
  - *To be Solved in Other Modules*
  - *To be Solved in This Module*

An Ideal Solution
- Requirements
- Black-box Attributes
- Interoperability Interfaces
- Other Necessary Components
- Possible Implementation Details

Summary and Overview

© 2003, Robert J. Robbins

Slides:

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

© 2003, Robert J. Robbins