TOC

INDEX

VIEW

B29 DB2 for Life Sciences

Lynn Everitt, Solutions Marketing, Data Management, IBM
Alick Law, Marketing Manager, IBM

IBM has made significant investments in this emerging industry. This presentation gives an overview of the life sciences industry and why this industry is receiving so much attention from the media and governments around the world. Current technology trends in this industry will be reviewed and how IBM's DB2 UDB and DiscoveryLink are used to help firms discover new drugs.

B29

DB2 for Life Sciences

DB2 Technical Conference - 2002

IBM Data Management Technical Conference

Anaheim, CA **Sept 9 - 13, 2002**

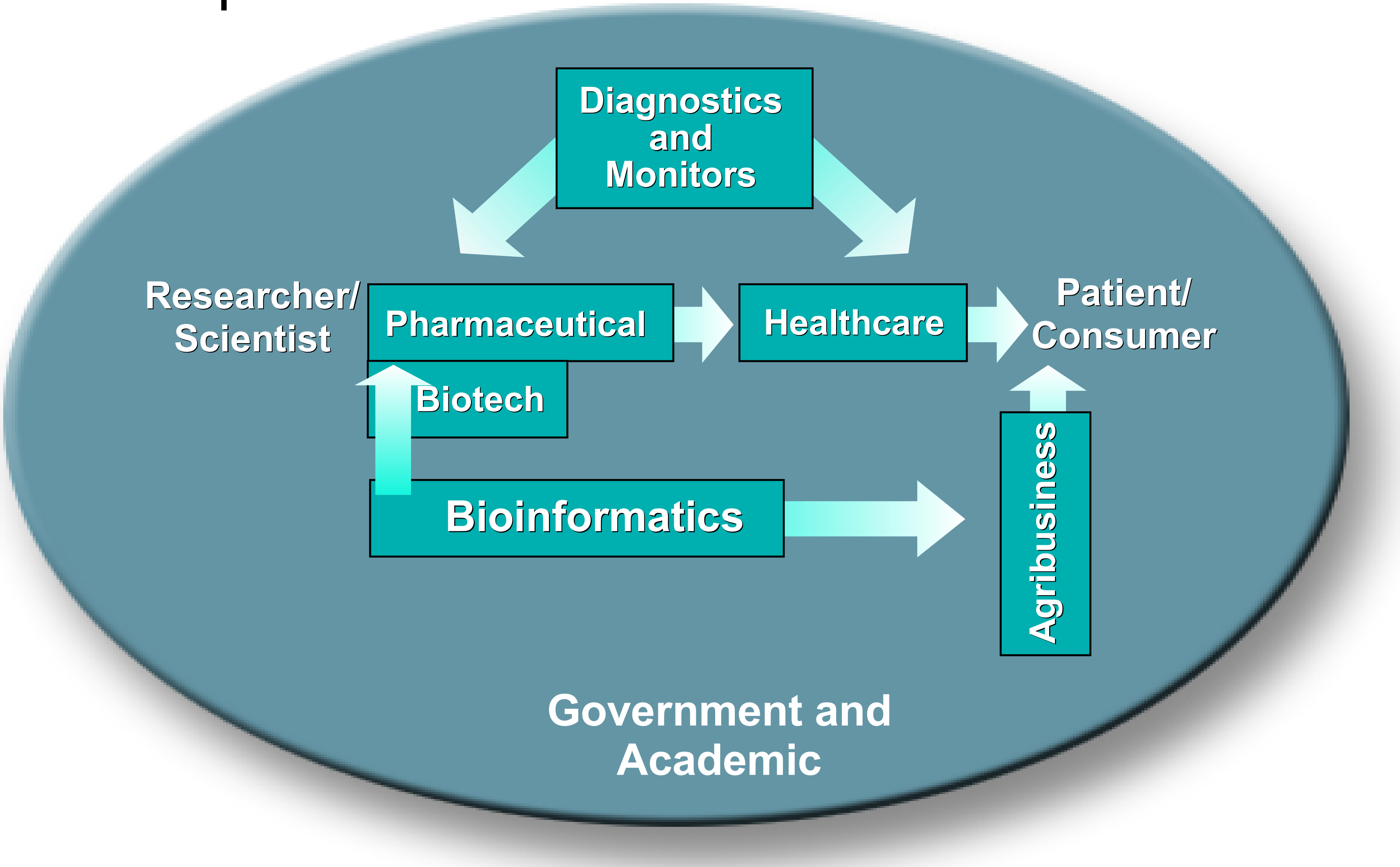
RETURN TO INDEX

Agenda

- What is Life Sciences
- Data challenges
- DM Solutions
  - DiscoveryLink
  - Office Connect
  - DB2
  - Data Mining
  - Content Manager

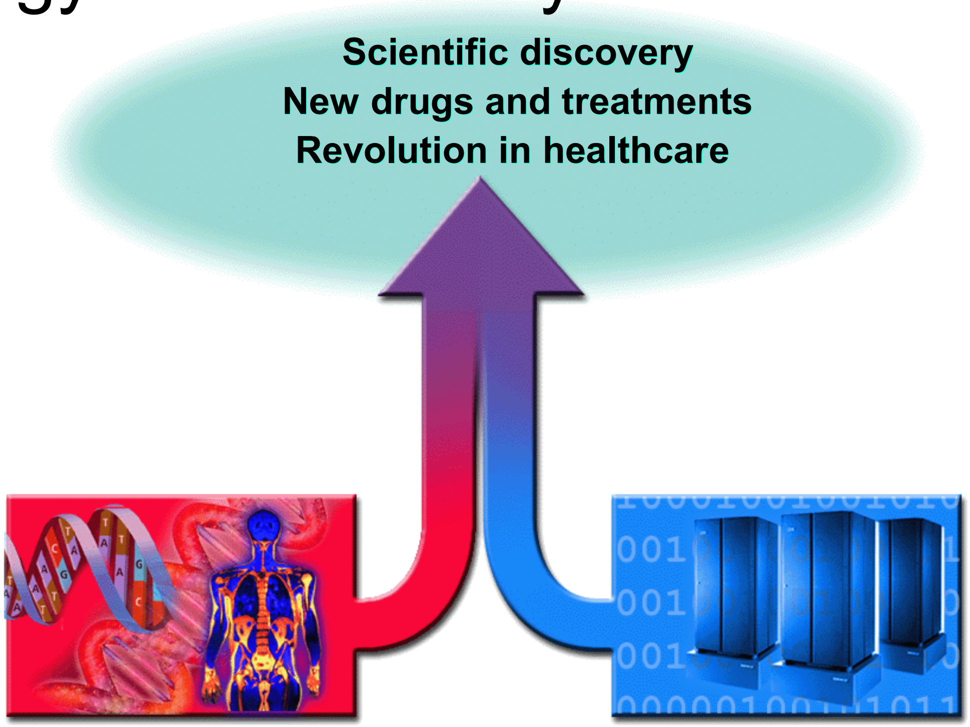
What is Life Sciences

The Life Sciences industry focuses on drug discovery and development.



What Has Changed

- Recent developments in laboratory automation combined with powerful computational and algorithmic capabilities have created a bioinformatics industry.
- Genomic and proteomic data will be readily available in massive amounts.
- This will revolutionize agriscience, drug discovery, biotechnology and ultimately human healthcare.

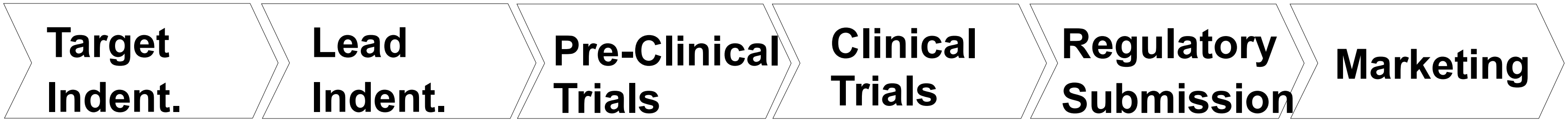


Scientific discovery
New drugs and treatments
Revolution in healthcare

Industry Challenges

- Data and Knowledge Management
- Collaboration
- Process transformation/improvement
- Regulatory compliance
- Clincial submission

Drug Discovery Process



Target Indent.

Lead Indent.

Pre-Clinical Trials

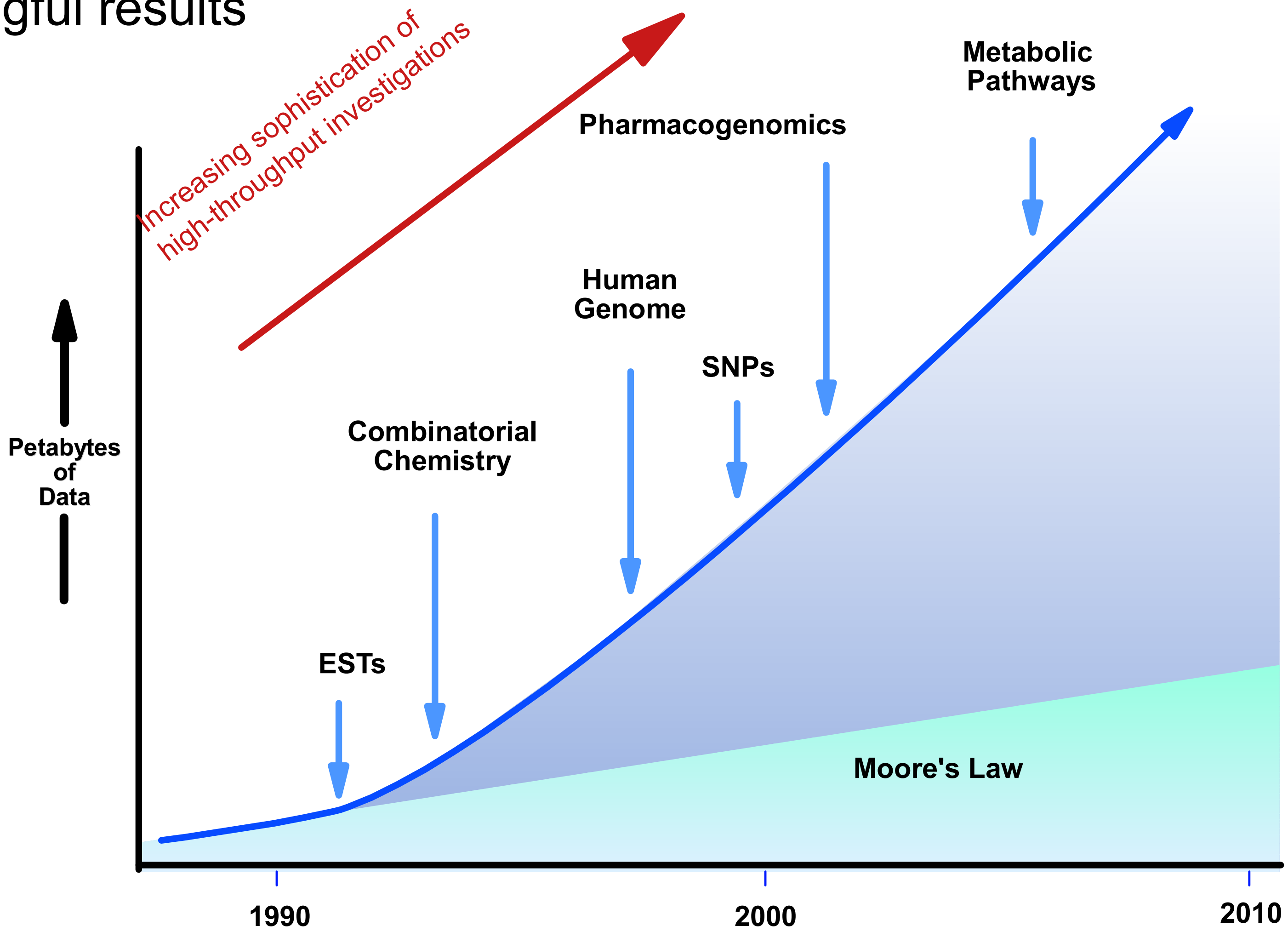
Clinical Trials

Regulatory Submission

Marketing

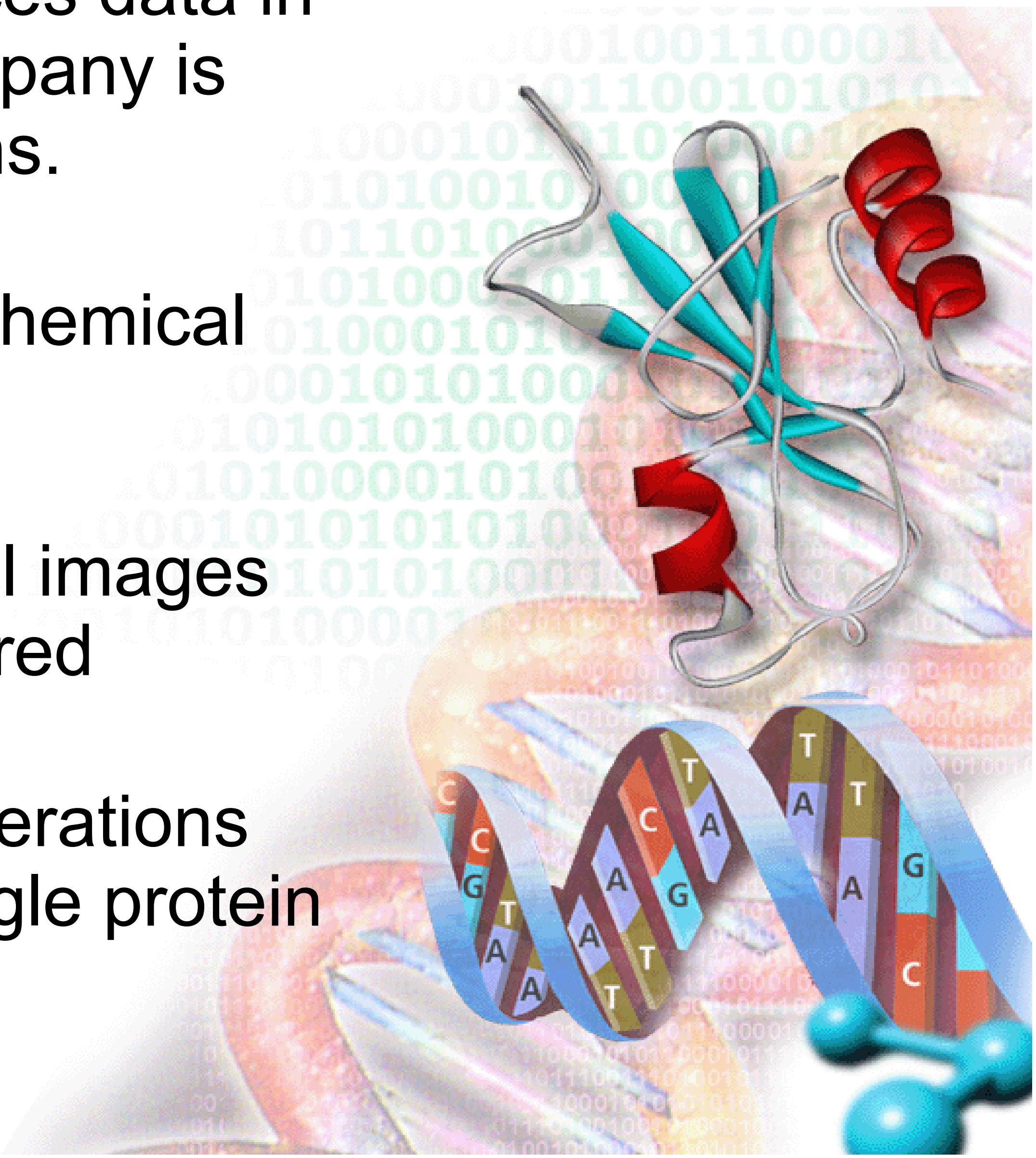
Challenge: Management and Access to Terabyte level of data

- Explosion of Biological Information Requires Effective Data Management and Computation
- Petabytes (10^{15}) of data are projected
- Data integration and data management are key to successfully deciphering meaningful results



Computational Intensity is Critical to Success

- The volume of life sciences data in the average biotech company is doubling every six months.
- 32,000 genes
  (and is at least 3 billion chemical letters)
  in the human genome
- 150 Petabytes of medical images annually worldwide if stored electronically
- $13x10^{21}$ Floating point operations are required to fold a single protein



IBM in the Life Sciences

- IBM Life Sciences Solutions launched in August 2000
  - ►Leader in supercomputing and high performance storage
  - ►Data integration and scalable database
  - ►Leader in KM solutions
  - ►Leader in e-business, security and privacy
- IBM Research
  - ►IBM Research is the largest research organization in the information technology industry
  - ►Basic research at the intersection of biology and computation since 1992

Partners Are Key to Our Success

[redacted]

turbogenomics
Accelerating the Bioinformatics Revolution™

Provides TurboBLAST, an accelerated, parallel implementation of BLAST and TurboBench, a comprehensive, enterprise-wide, automated high-performance bioinformatics software platform

proteome systems
Integrated proteome technology

Products range from protein sample preparation kits, through gel separation and sample excise instruments, to an integrated suite of Proteomic technologies

SB Structural Bioinformatics, Inc.

Provided technology and services for proteomics-driven drug discovery and optimization processes.

LabBook inc

Provided software for analyzing genetic data with LabBook desktop information retrieval, integration, mining, and visualization software.

documentum

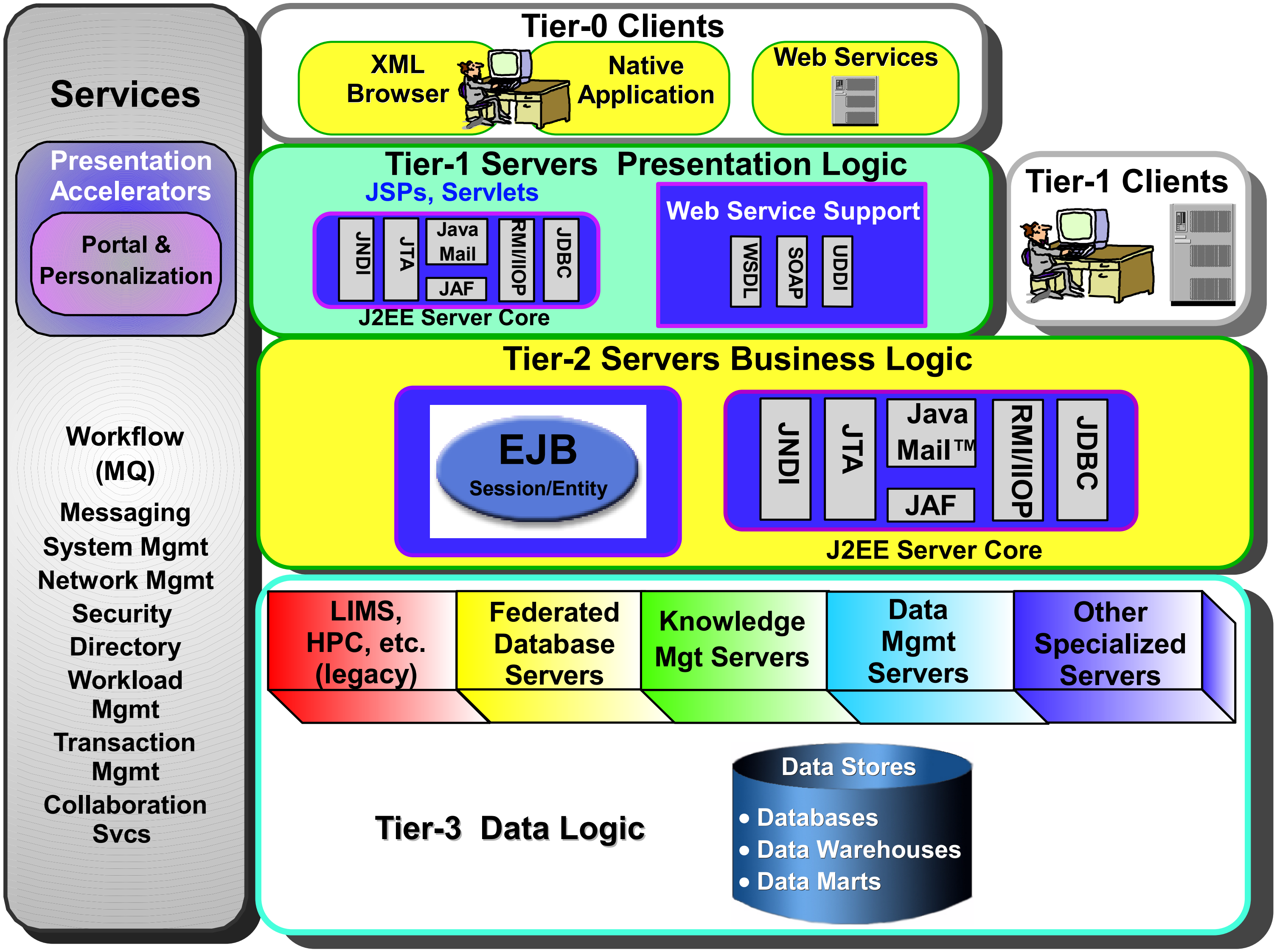
Documentum is the leader in content management in the Life Sciences industry

Life Sciences Framework

Creating a robust ecosystem to complement our offerings

- A collaborative research-centric environment
- Built on industry standards, proven technologies and methodologies
- Supporting openness
- Integrating domain-specific functions (legacy and new)
- Partnering with industry solution providers

Creating end-to-end solutions for Life Sciences



Services

Presentation Accelerators

Portal & Personalization

Workflow (MQ)

Messaging

System Mgmt

Network Mgmt

Security

Directory

Workload Mgmt

Transaction Mgmt

Collaboration Svcs

Tier-0 Clients

XML Browser

Native Application

Web Services

Tier-1 Servers Presentation Logic

JSPs, Servlets

JNDI, JTA, Java Mail, JAF, RMI/IIOP, JDBC

J2EE Server Core

Web Service Support

WSDL, SOAP, UDDI

Tier-1 Clients

Tier-2 Servers Business Logic

EJB

Session/Entity

JNDI, JTA, Java Mail™, JAF, RMI/IIOP, JDBC

J2EE Server Core

LIMS, HPC, etc. (legacy)

Federated Database Servers

Knowledge Mgt Servers

Data Mgmt Servers

Other Specialized Servers

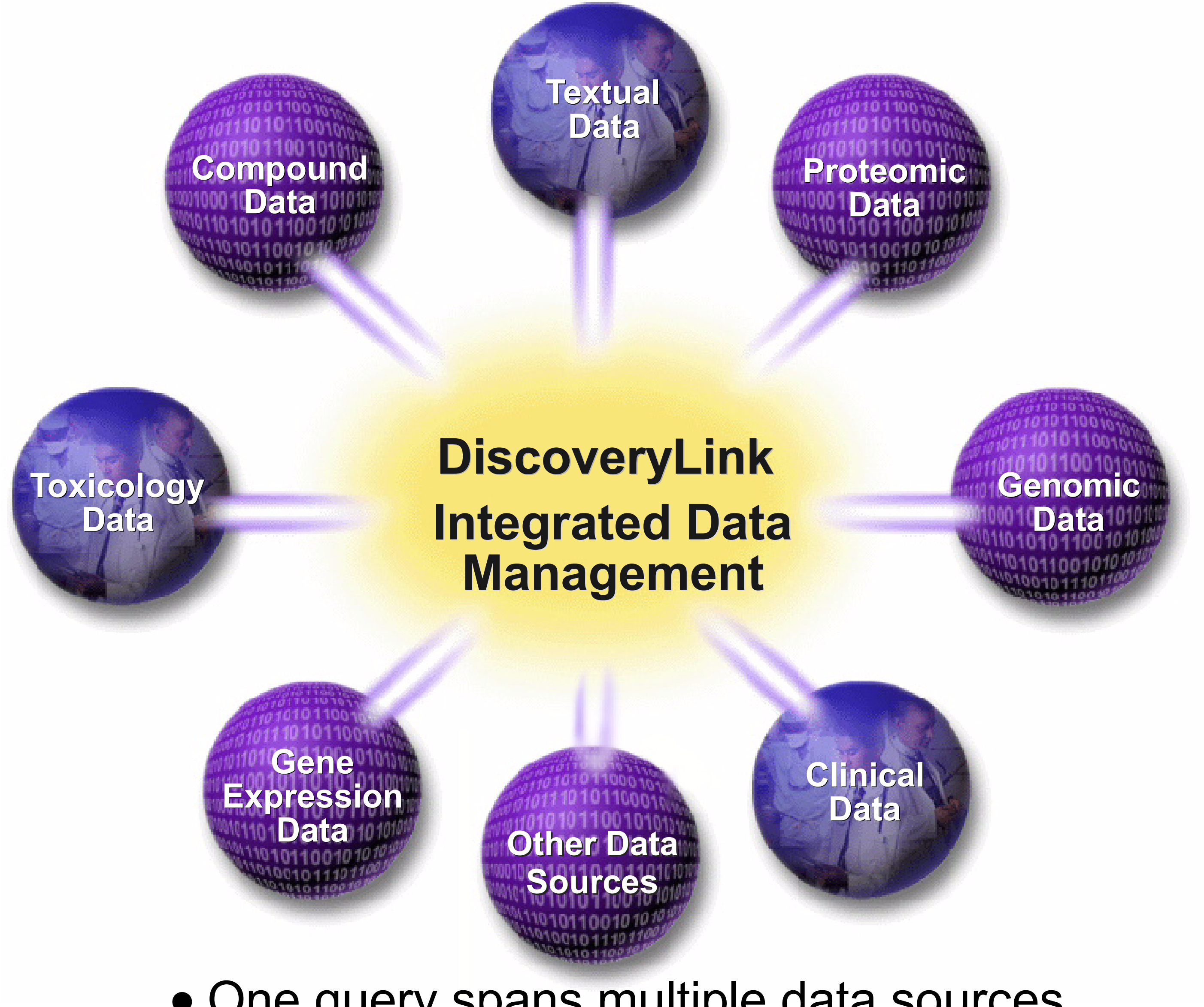
Tier-3 Data Logic

Data Stores

- Databases
- Data Warehouses
- Data Marts

Integrated Data Management

- **Link multiple heterogeneous data sources together**



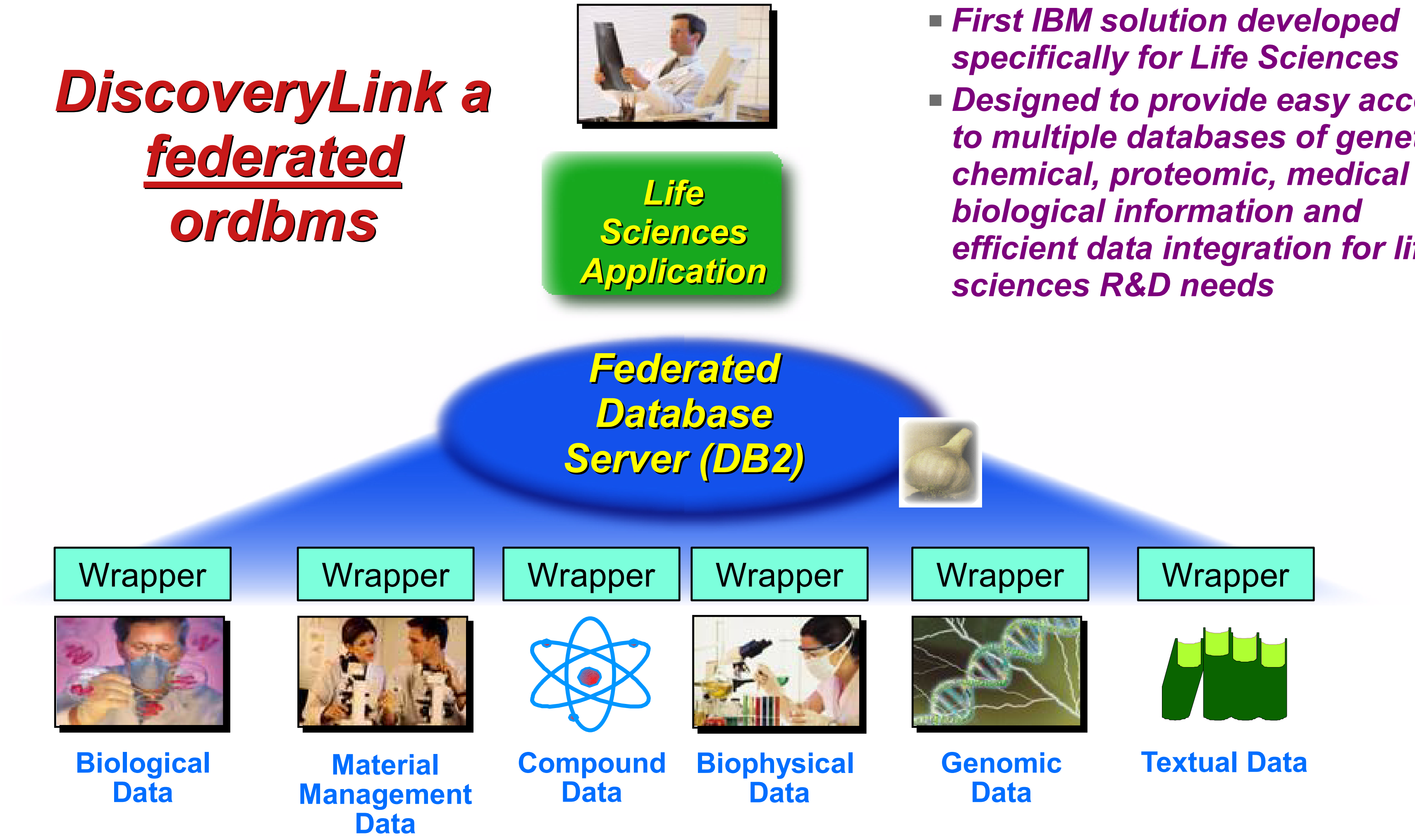
- One query spans multiple data sources

DiscoveryLink

Enabling researchers to find critical needles in a haystack of data and documents

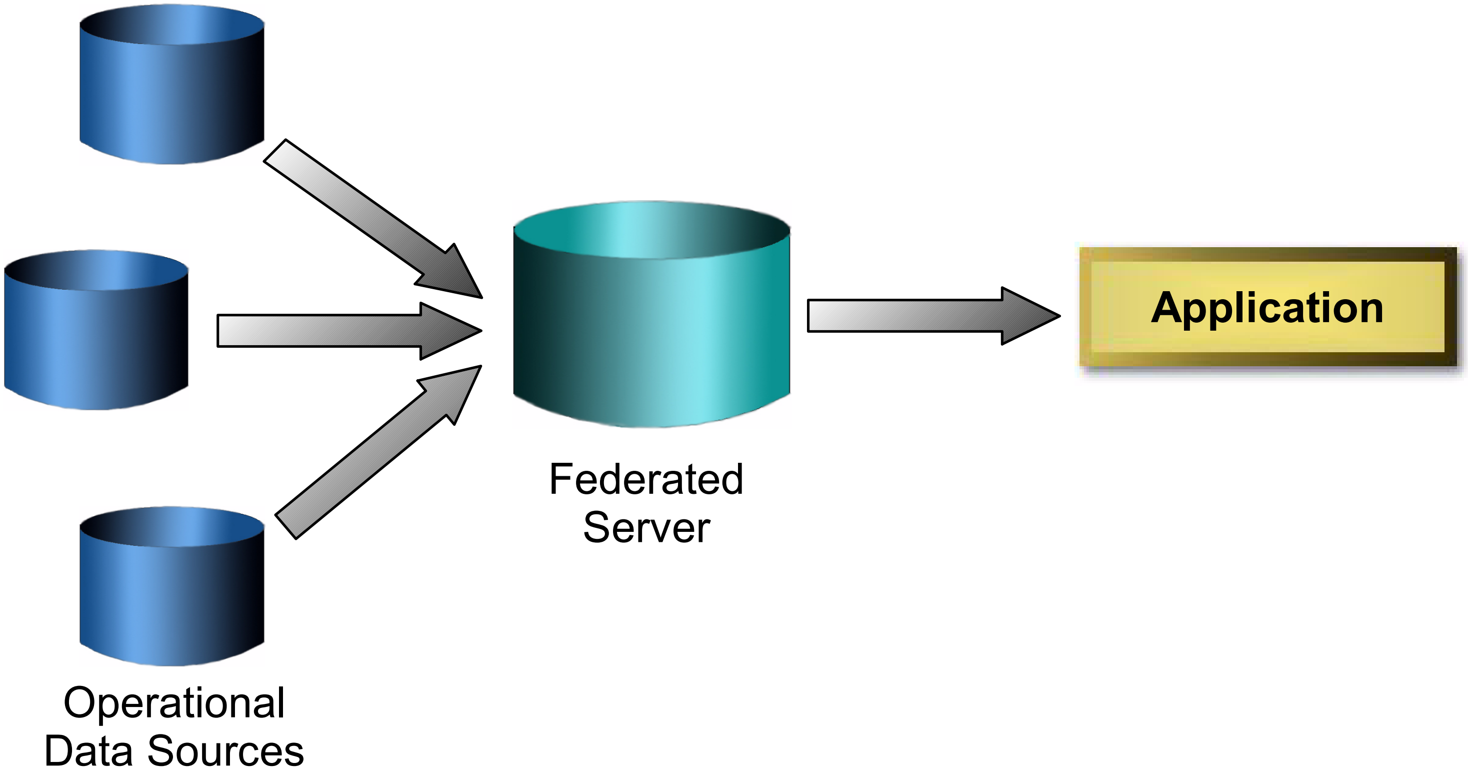
DiscoveryLink a <u>federated</u> ordbms

- *First IBM solution developed specifically for Life Sciences*
- *Designed to provide easy access to multiple databases of genetic, chemical, proteomic, medical and biological information and efficient data integration for life sciences R&D needs*

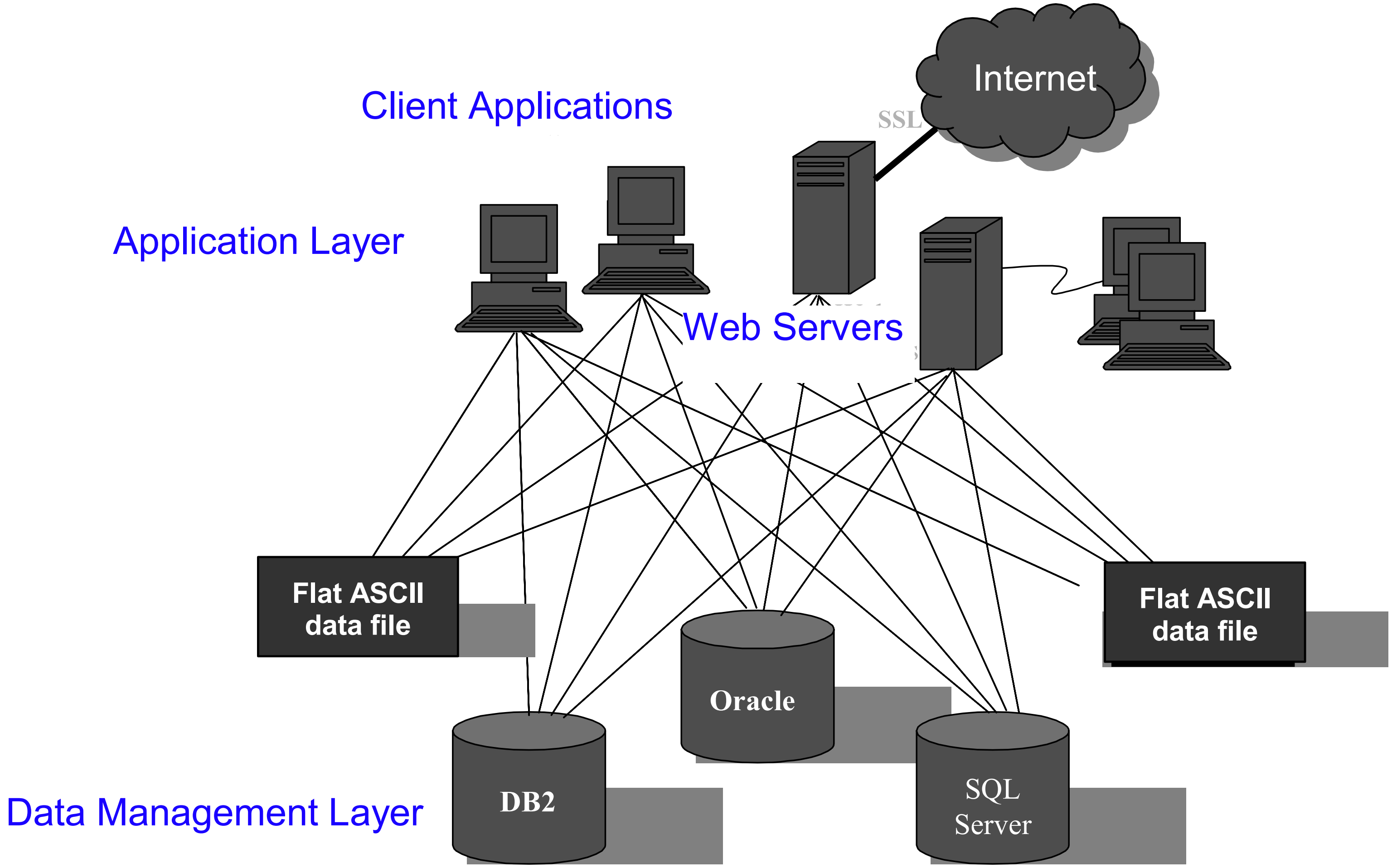


A Federated Database

- Data remains in the original separate sources
- All operational data sources accessible with a single query
- Query optimization on all data sources



Without integration layer



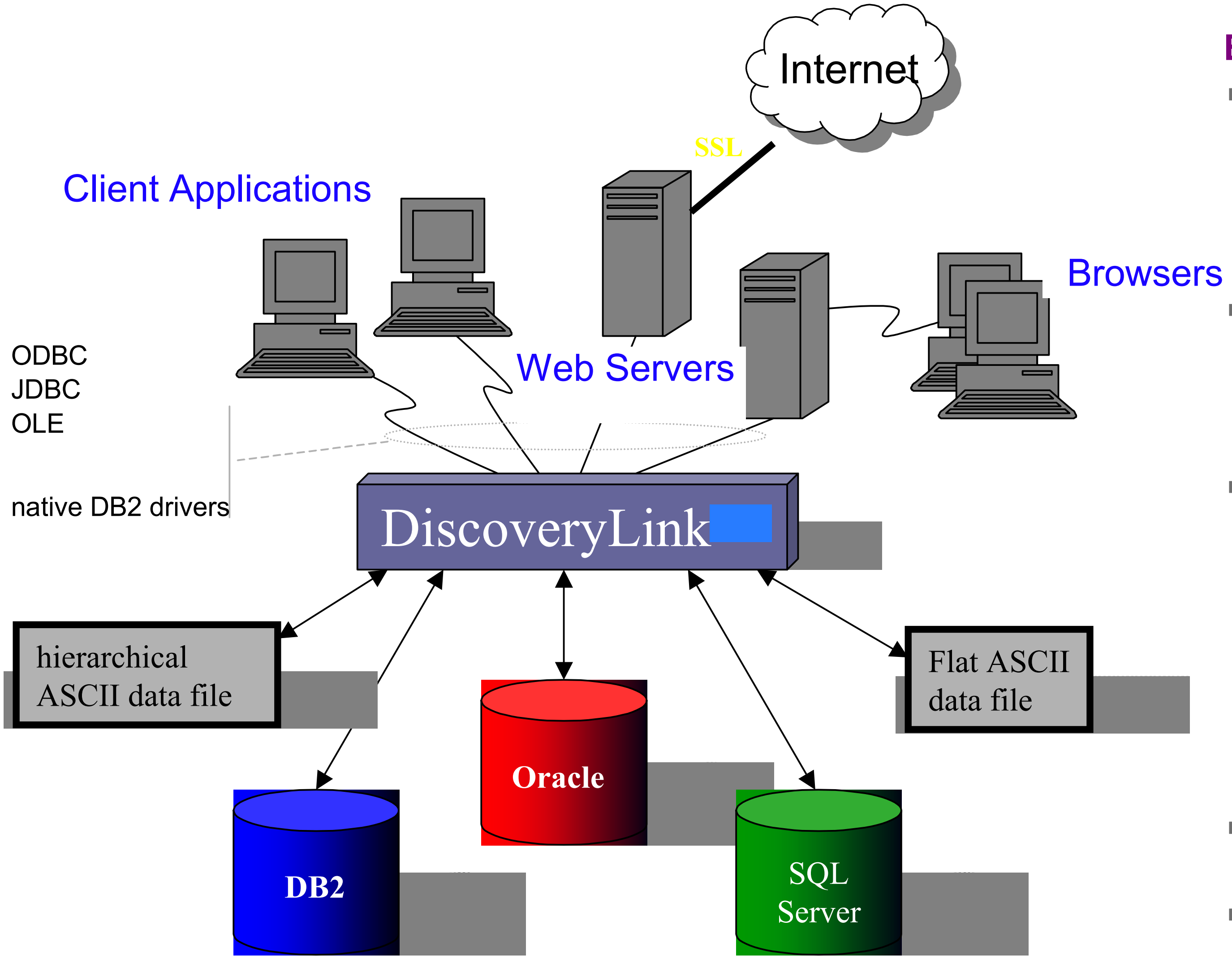
Client Applications

Application Layer

Web Servers

Data Management Layer

With integration layer



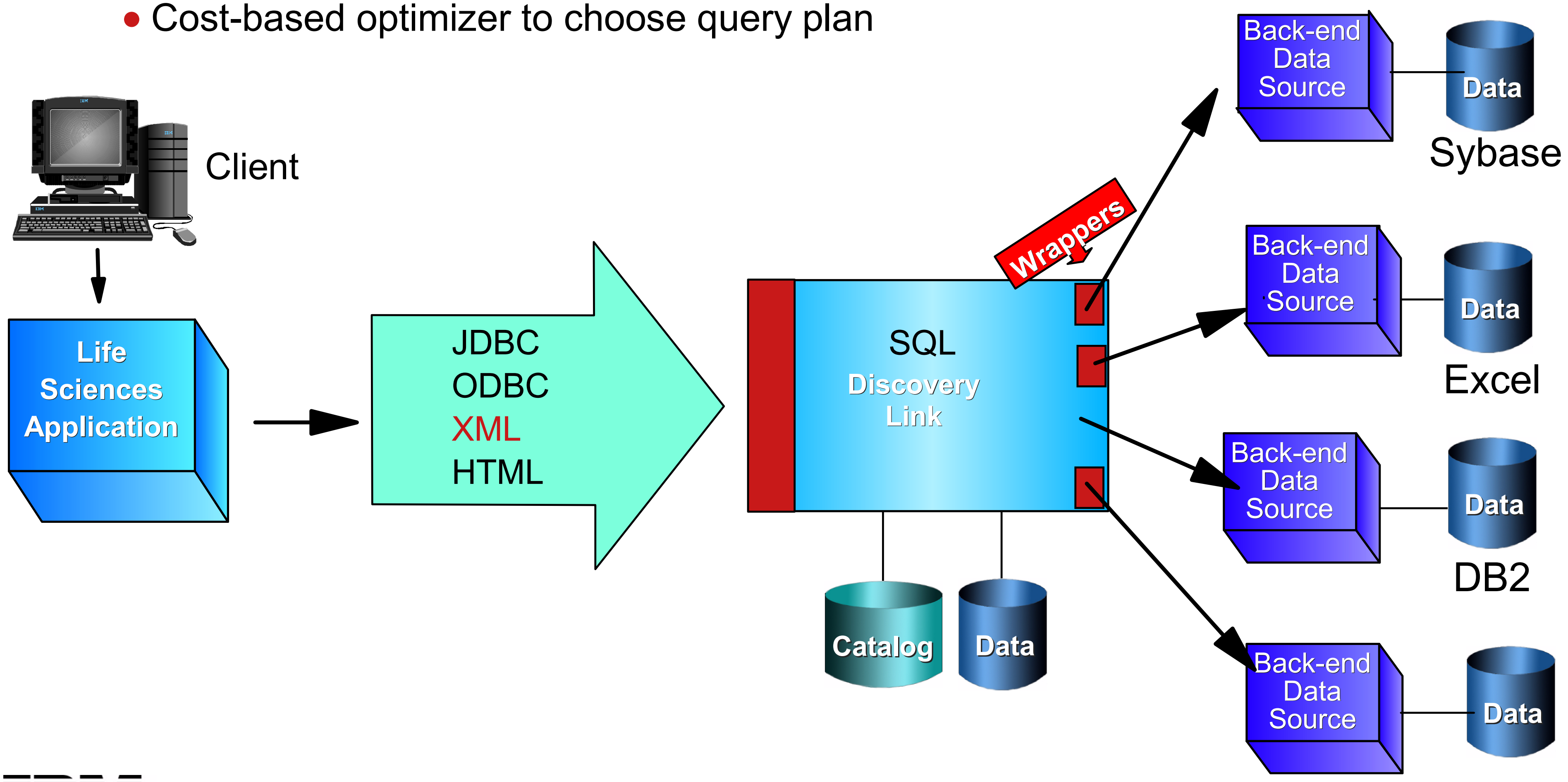
Benefits

- ***Transparency:*** Provides a single "virtual database" to applications
  - ► Appears to be one data source
  - ► Supports a high level query language
- ***Heterogeneity:*** Integrates data from different data sources
  - ► Diverse types of data
  - ► Diverse sources
- ***High Function:*** Capabilities of existing sources and of SQL
  - ► To search for and to manipulate data
  - ► Lose no functionality of source or of SQL language
  - ► One query can combine data from multiple sources
- ***Autonomy:*** No perturbation of existing data, sources
- ***Performance:*** Optimization of queries for good performance

Architecture

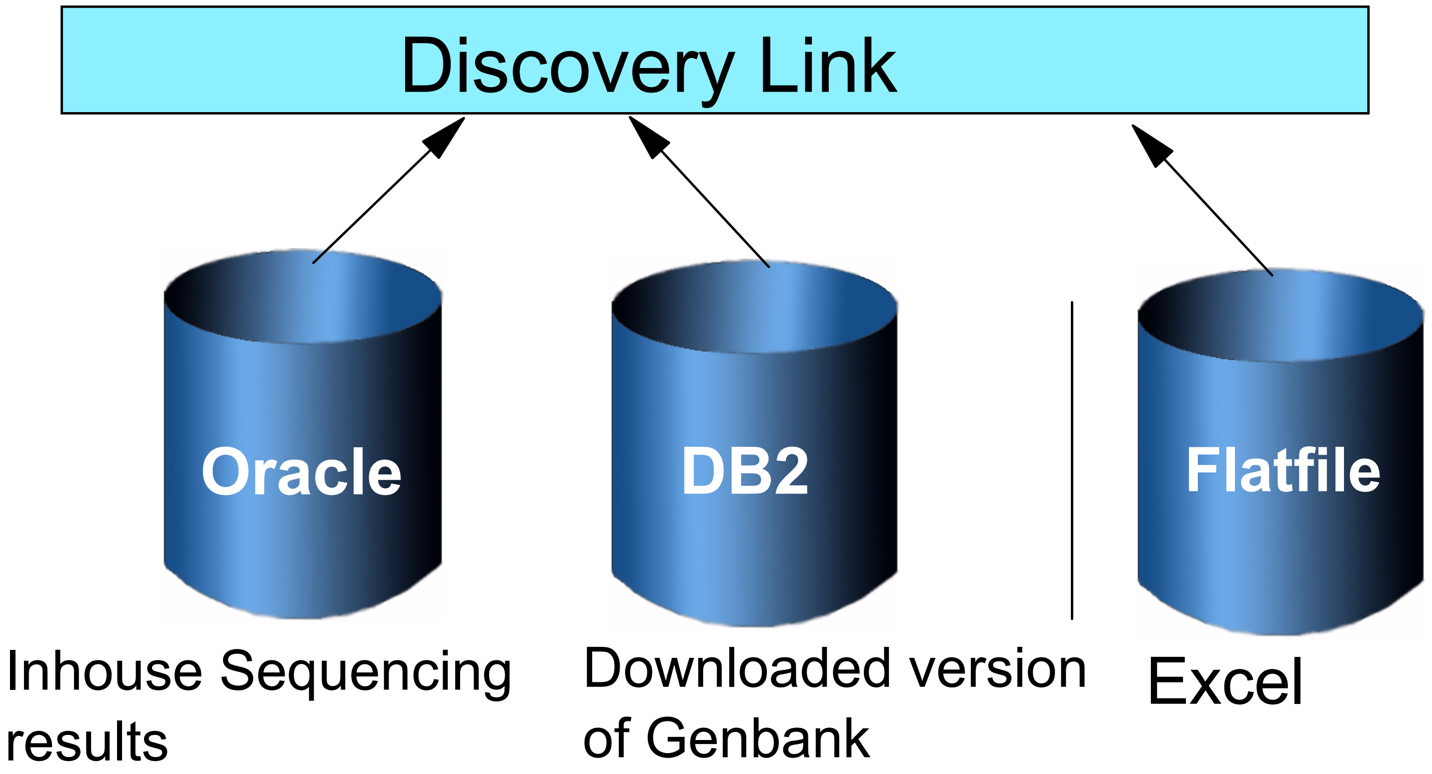
DiscoveryLink (DB2) Federated Database Engine

- DB2 drives DiscoveryLink, but it does **not replace** existing client databases!
- Powerful **query processing** engine in **federated server**
- Logical decomposition and distribution of queries
- Cost-based optimizer to choose query plan



Query 1

How similar is gene X to sequences within Genbank and within my inhouse proprietary genome and my research data on my spreadsheet?



Discovery Link

Oracle

DB2

Flatfile

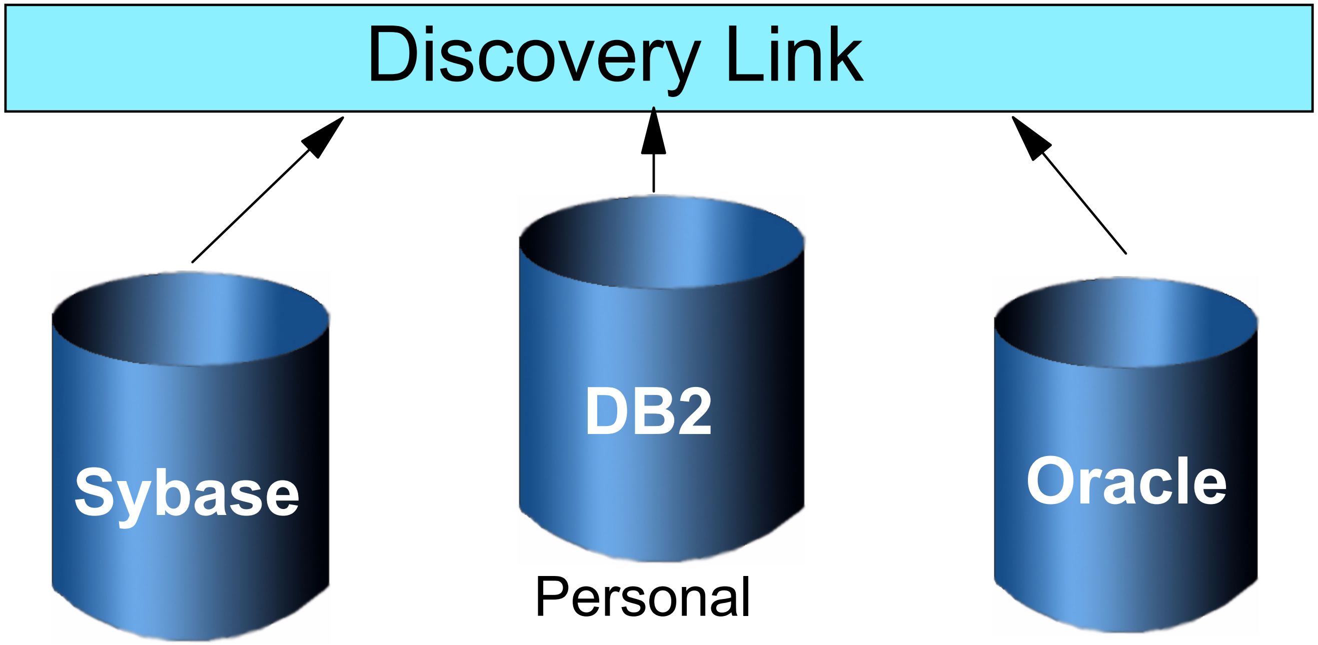
Inhouse Sequencing results

Downloaded version of Genbank

Excel

Query 2

What gene or genes affect the reaction of some people to antibiotic X?



Discovery Link

Sybase

DB2

Oracle

Demographic Information

Personal Genotypic Information

Annotated human genome/EST's

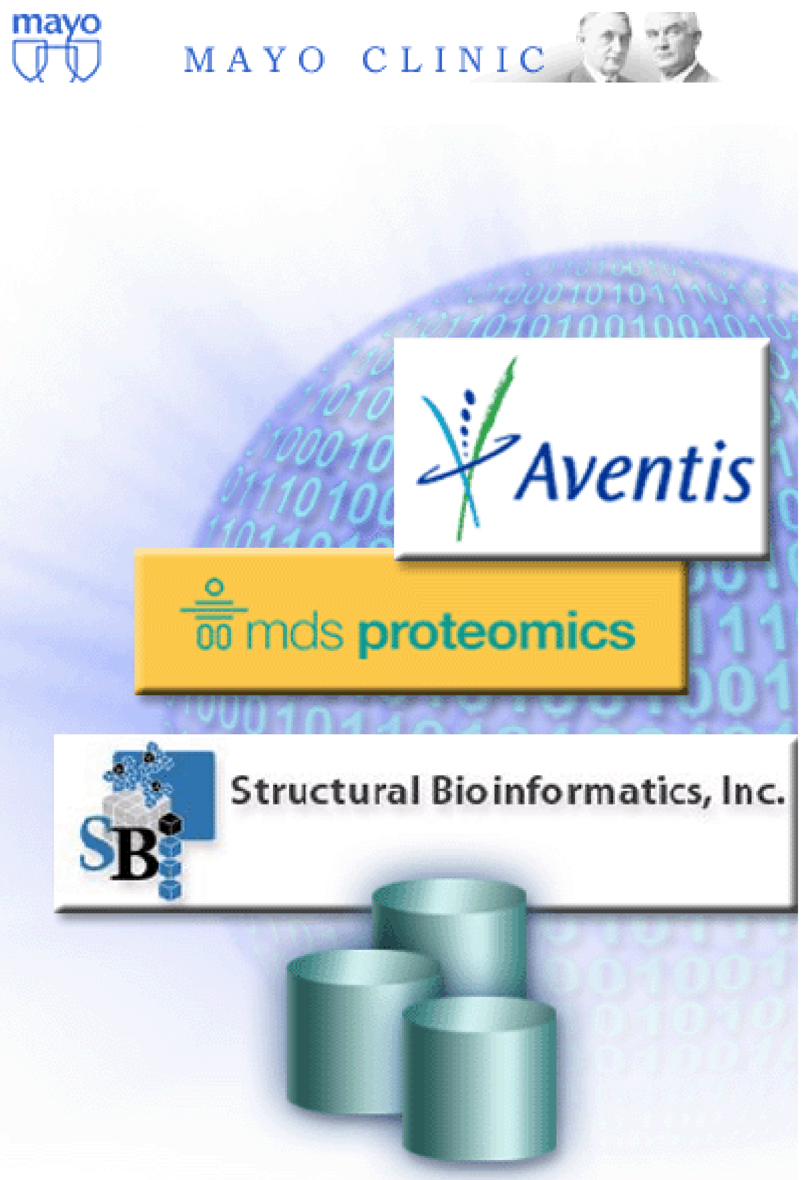
References

- **IBM, Mayo Clinic Develop Medical Database**

The new system could enable the clinic's medical staff to quickly draw meaning from data to support medical treatments, including genomic information from public and private databases and retrospective studies of millions of archived records from patients.

- **Structural Bioinformatics supports drug discovery with IBM and Linux**

SBI needed a life Sciences infrastructure solution for a complex protein modeling simulation program. They now have a faster, more robust, scalable IT environments, able to generate more data in shorter times with faster and more secure database access.



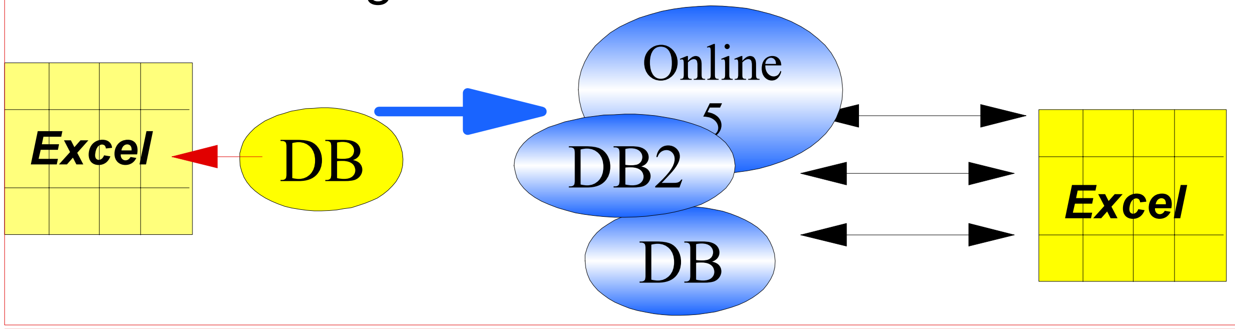
Aventis

- Business Need
  - Access to chemical and biological data stored in both local and remote databases
  - Increasing drug research efficiency
- Solution
  - Software: Brio, DB2 for AIX, DB2 UDB Developer's Edition, DB2 UDB Enterprise Edition, Query Patroller
  - Life Sciences Solutions: Data Integration, IBM Global Services Life Sciences Consulting and Solutions, DiscoveryLink
  - Business Partner Information: IBM's business partner, Computers & Communications (C&C) configured and implemented the new system.
- Benefits
  - Increased productivity
  - Chemical and biological research linked
  - Ability to search across both chemical and biological databases provided

"We have research organizations in four countries that need to collaborate and share chemical compound and biomedical data, from sources within Aventis and many public databases. DiscoveryLink allows us to access and mine the physical data in a way never before possible, significantly speeding up the drug discovery and development process." --Peter Loupos, Global Head of Drug Innovation and Approval Information Systems

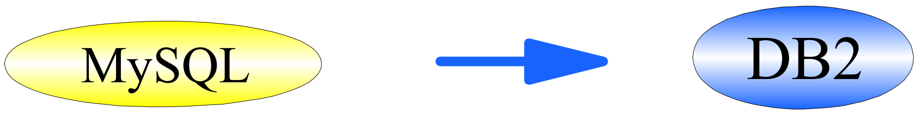
Customer Scenarios

Need to manage data and collaborate



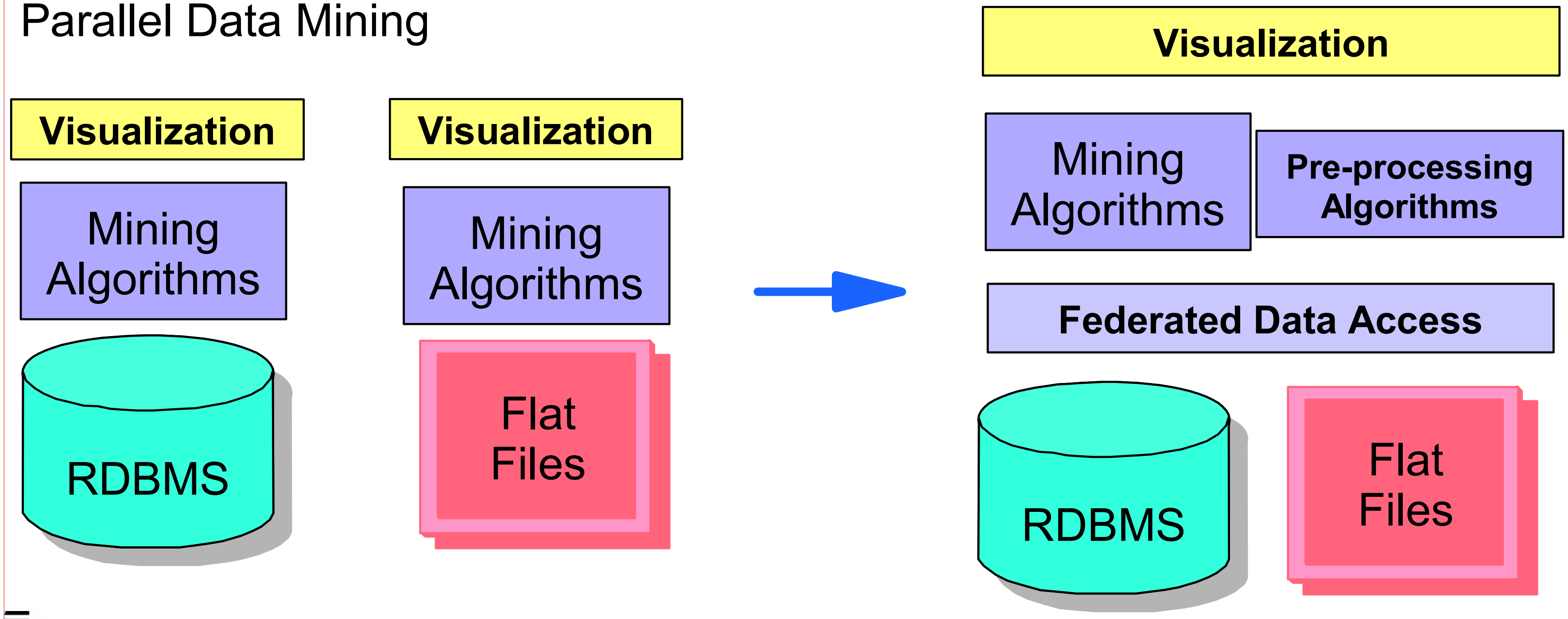
- Sophisticated database query and update
- Dynamically updating spreadsheets

Need scalability and performance

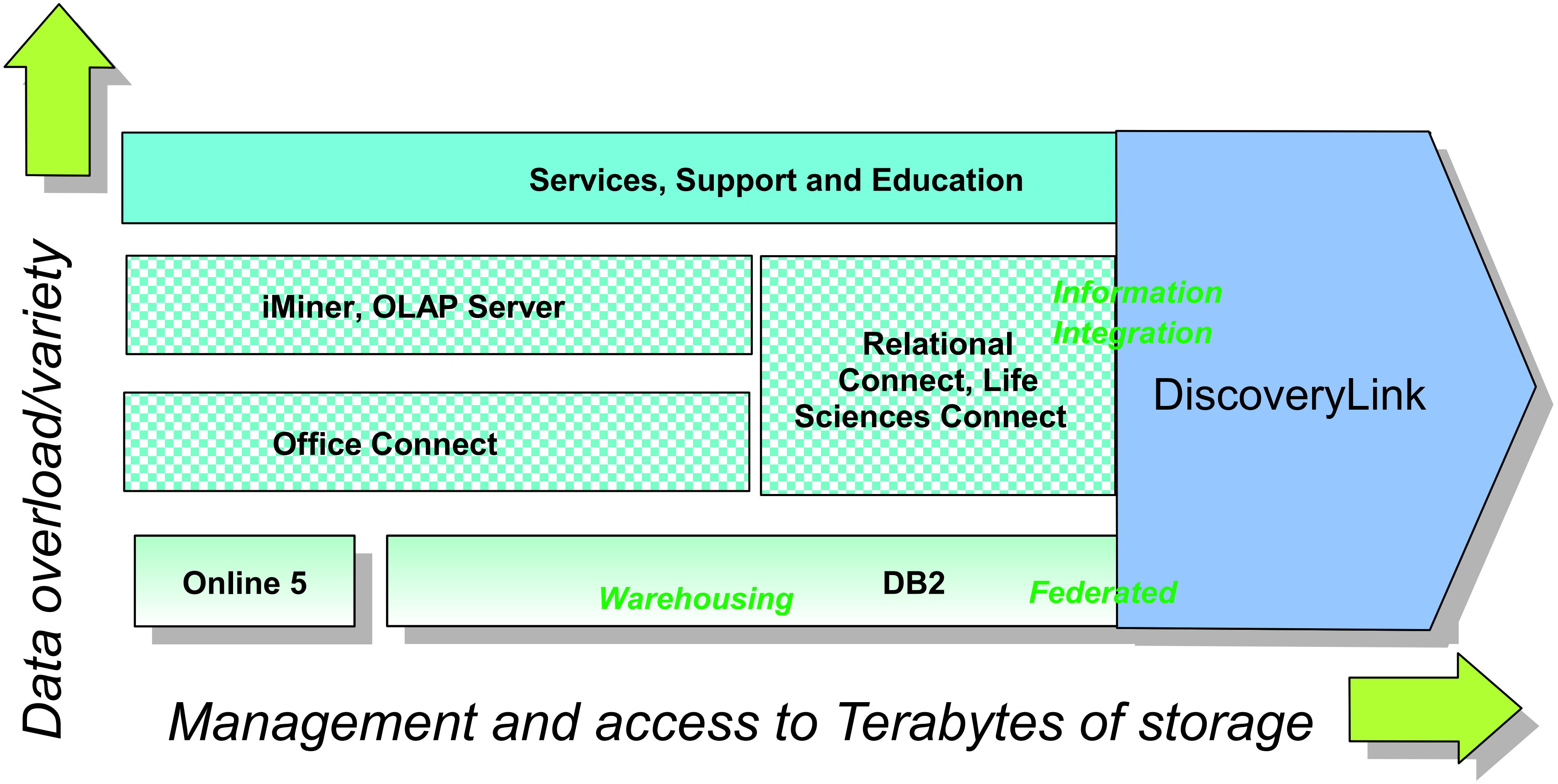


- Scale to Terabytes
- Supported, Functionality
- Platforms

Parallel Data Mining



How DM is addressing the Biotech Market Needs



Data overload/variety

Services, Support and Education

iMiner, OLAP Server

Office Connect

Relational Connect, Life Sciences Connect

Information Integration

DiscoveryLink

Online 5

DB2

Warehousing

Federated

Management and access to Terabytes of storage

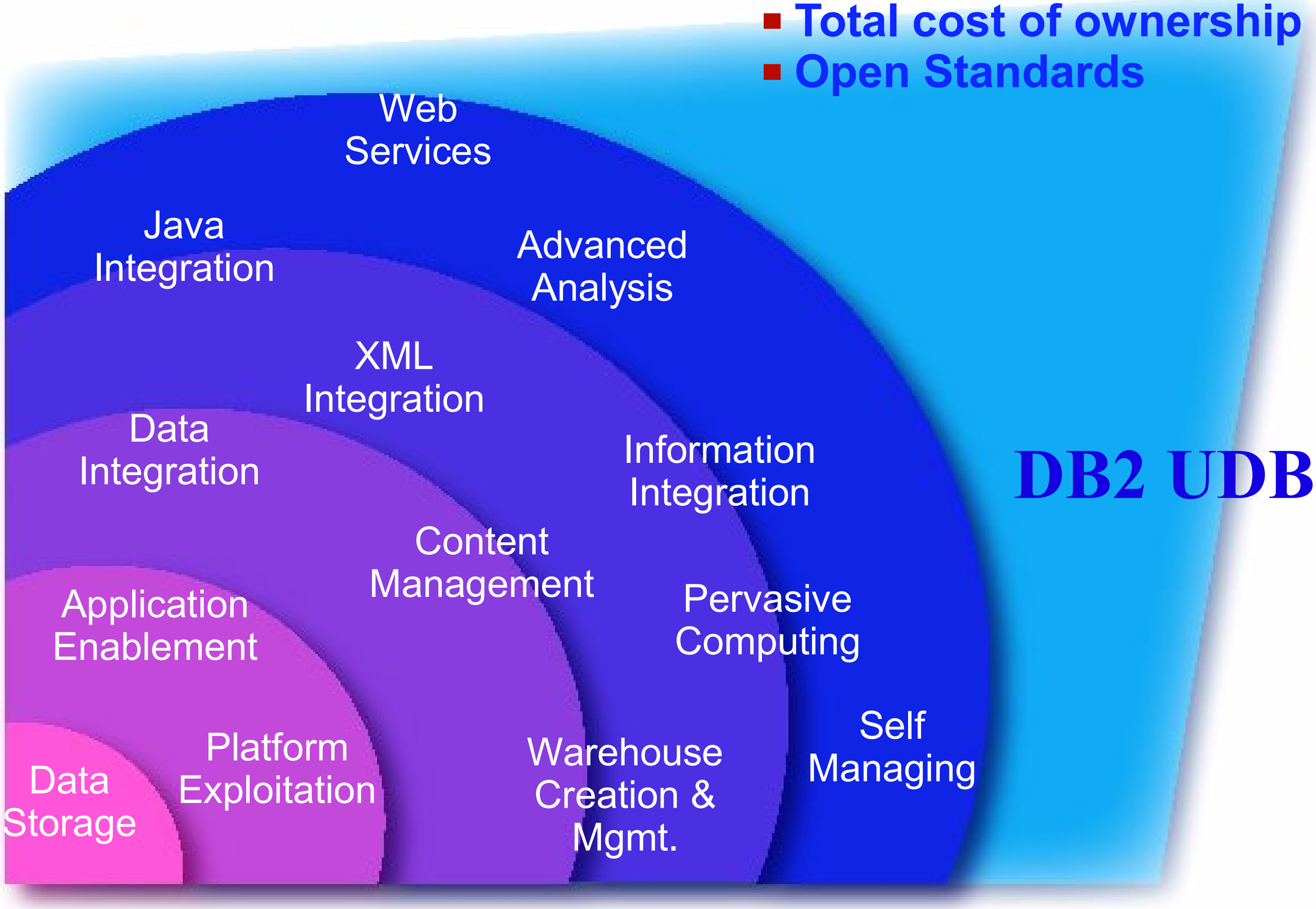
The perfect database for Biotech companies large, small and growing.

What is Office Connect?

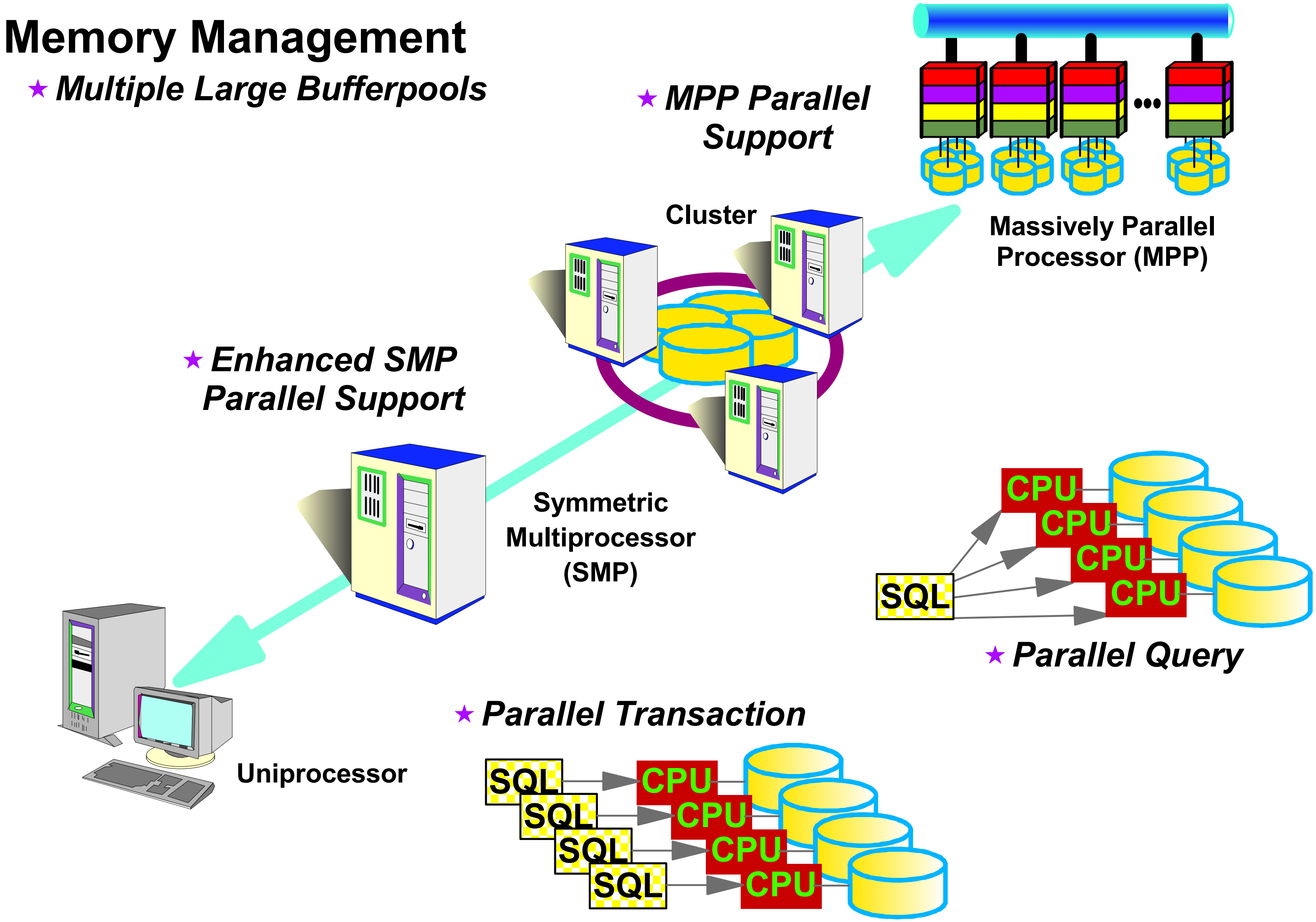
- Microsoft Excel Add-In
  - Binds spreadsheet columns to database columns or stored procedures
  - GUI interface with model viewer and query builder
- Web Edition supports deployment of pre-built spreadsheets over the web
  - Spreadsheets stored in a repository database
  - Access authorization via groups, users and roles

DB2 Value Propositions for the Life Sciences

- DB2 scalability, performance and availability
- Federated Search Optimization
- Total cost of ownership
- Open Standards



Parallel Performance and Scalability



Memory Management

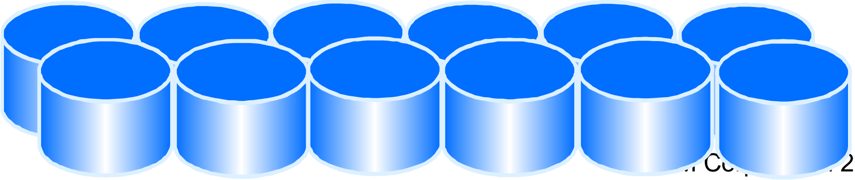
- ★ *Multiple Large Bufferpools*
- ★ *MPP Parallel Support*
- ★ *Enhanced SMP Parallel Support*
- ★ *Parallel Query*
- ★ *Parallel Transaction*

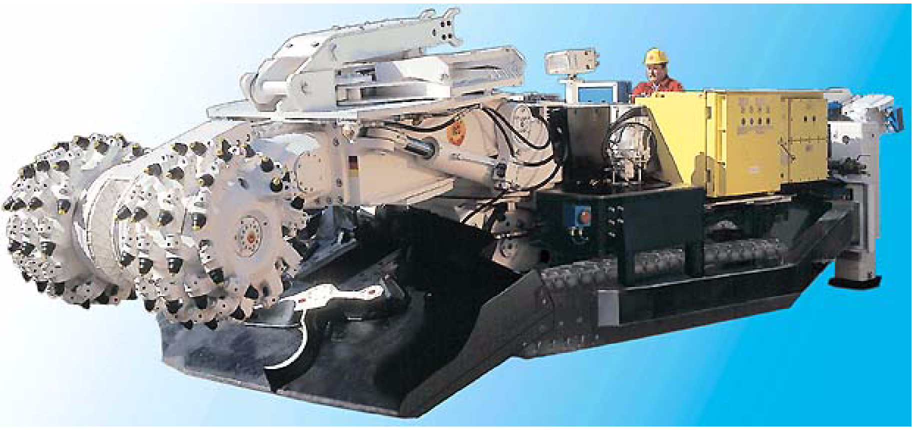
Cluster
Massively Parallel Processor (MPP)
Symmetric Multiprocessor (SMP)
Uniprocessor
SQL
CPU
CPU
CPU
CPU
SQL
SQL
SQL
SQL
CPU
CPU
CPU
CPU

When Data Mining is Valuable in Life Sciences

- **Large data volume (chemical compounds/genes)**
  - Difficult to understand the underlying relationships
  - Hundreds of Variables, Terabyte Level Storage
  - Need for automated analysis
- **Database Performance and Scalability**
  - Complex queries across distributed data sources
  - Multiple heteregenous sources
- **Drug Discovery**
  - Shorter cycle times, cost efficiencies, increased number of products to market

Analyzes the entire database





What is Datamining?

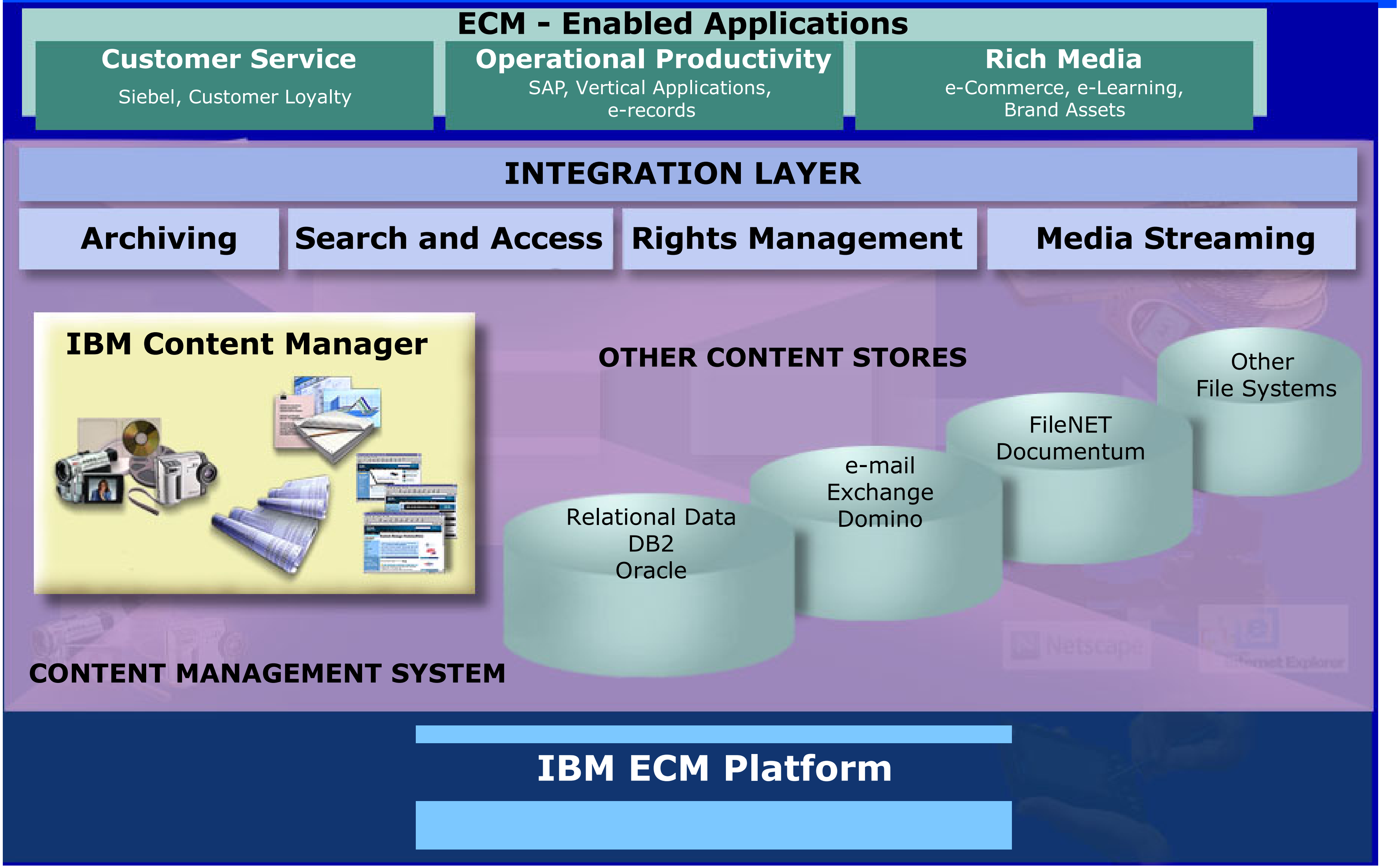
The process of automatically extracting previously unknown, comprehensible and important information from data

- Supervised
  - Classification
  - Prediction
- Unsupervised
  - Clustering
  - Associations
- Complementary analysis
  - Iterative & ad-hoc query
  - Multi-dimensional analysis (OLAP)

Data Analysis needs of Scientists

- What questions are scientists asking?
  - For genome datasets:
    - Are gene expression levels in these samples indicative of cell proliferation? **(classification in IMiner)**
    - How does the complex interaction over time between genes control **(Teiresias,Similar sequences, association rules in IMiner,OLAP Miner)**
    - cellular differentiation during development aging and disease?
    - Are there genes of similar function? **(clustering in IMiner)**
    - How does gene expression vary across tissue types, protein class ...? **(DB2 OLAP Server)**

IBM ECM Total Solution



ECM - Enabled Applications

Customer Service
Siebel, Customer Loyalty

Operational Productivity
SAP, Vertical Applications,
e-records

Rich Media
e-Commerce, e-Learning,
Brand Assets

INTEGRATION LAYER

Archiving | **Search and Access** | **Rights Management** | **Media Streaming**

IBM Content Manager

OTHER CONTENT STORES

Relational Data
DB2
Oracle

e-mail
Exchange
Domino

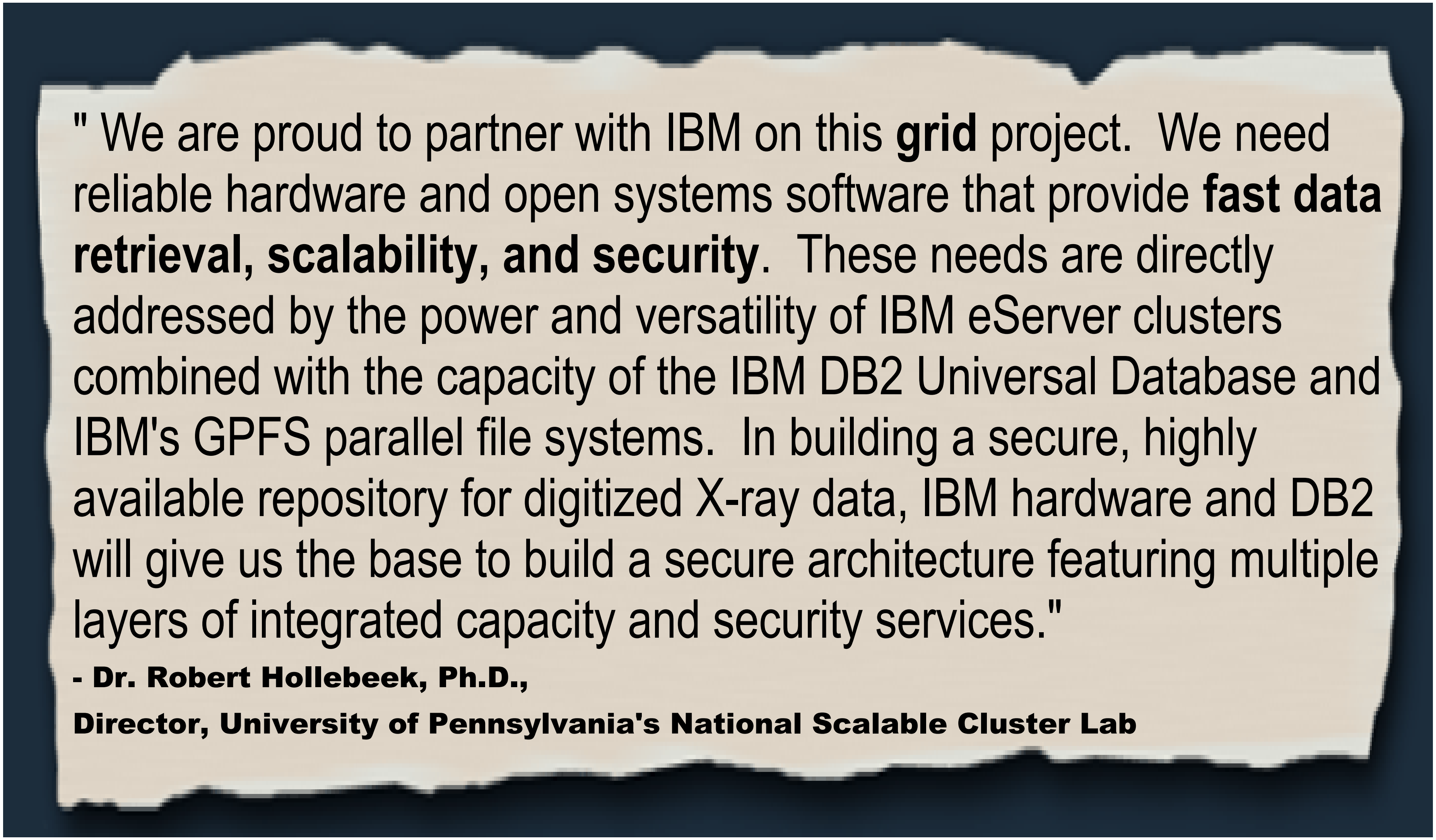
FileNET
Documentum

Other
File Systems

CONTENT MANAGEMENT SYSTEM

IBM ECM Platform

What Customers Are Saying...



" We are proud to partner with IBM on this **grid** project. We need reliable hardware and open systems software that provide **fast data retrieval, scalability, and security**. These needs are directly addressed by the power and versatility of IBM eServer clusters combined with the capacity of the IBM DB2 Universal Database and IBM's GPFS parallel file systems. In building a secure, highly available repository for digitized X-ray data, IBM hardware and DB2 will give us the base to build a secure architecture featuring multiple layers of integrated capacity and security services."

- Dr. Robert Hollebeek, Ph.D.,

Director, University of Pennsylvania's National Scalable Cluster Lab

Helping the demand for DB2 DBA skills

- **Technology**
  - **Built-in productivity**
    - Optimization
  - **Productivity tools**
    - Wizards
    - Control Center
    - DB Tools...
  - **SMART Databases**
    - Self-managing
    - Self-tuning
    - Self- administering

- **People**
  - Certifications
  - DBA cross training
  - DB2 Skills Plus Network
  - DB2 Scholars Programs
    - >4,000 Universities
  - Training Institutions

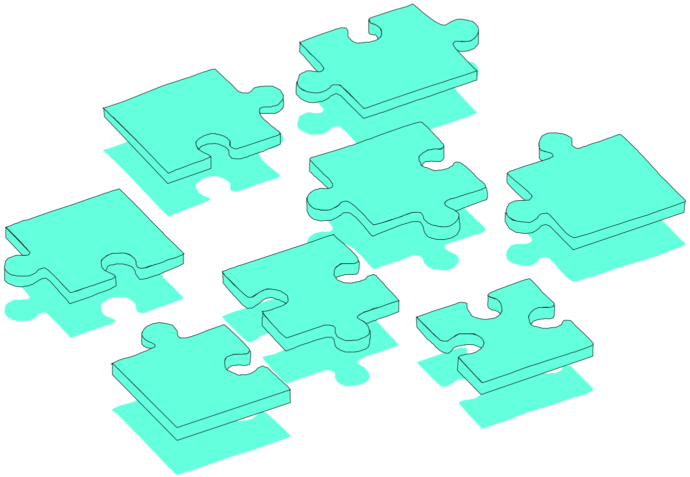
> **"…We have one database administrator. We would have needed three times that many [DBAs], at least, to run Oracle…"**
>
> Customer Quote, BusinessWeek Online, Nov. 2001

> **"…DB2 efficiencies yield an overall reduction in the work effort of 6% for OLTP systems, 15% for large OLTP systems, 20% for Internet-enabled databases, and 18% for data warehousing.**
>
> **DB2 vs. Oracle8i: D.H. Brown, Total Cost of Ownership, December, 2000**

Plus utilities designed to migrate a database from any concurrent RDMS to IBM DB2/UDB

DB2 and Life Sciences Summary

- IBM Partnership
  - Leverage Industry Leadership
- Competitive Advantage
  - Significant Savings
- Technology Leadership
  - Performance
  - Scalability
  - Accessibility
  - Openness
- Leverage Investments
  - Applications
  - Data
- Best of Breed Solutions
- Investment Protection
- Service and Support



Bringing it all Together