Issues in Information Integration for Drug Discovery

David Benton

Informatics and Knowledge Management

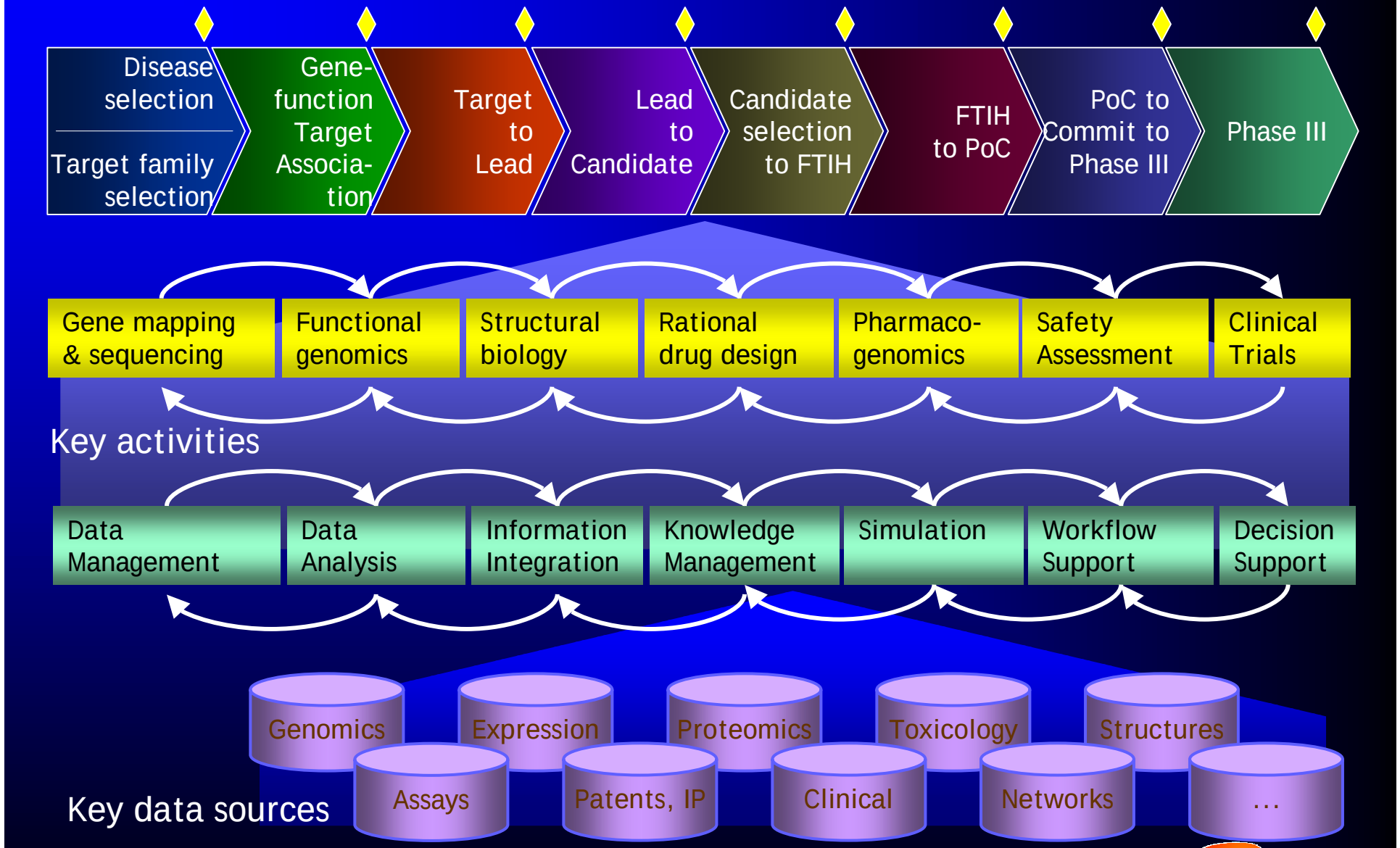
GlaxoSmithKline R&D

Overview

- Alternative and complementary architectures for information integration
- Challenges of multi-level heterogeneity to discovery information integration
- Illusion of technical solutions to the challenges of heterogeneity
- Importance of standards to information integration systems
- Sources of standards for life sciences research
- Model driven architecture & standards

GlaxoSmithKline

Drug discovery & development



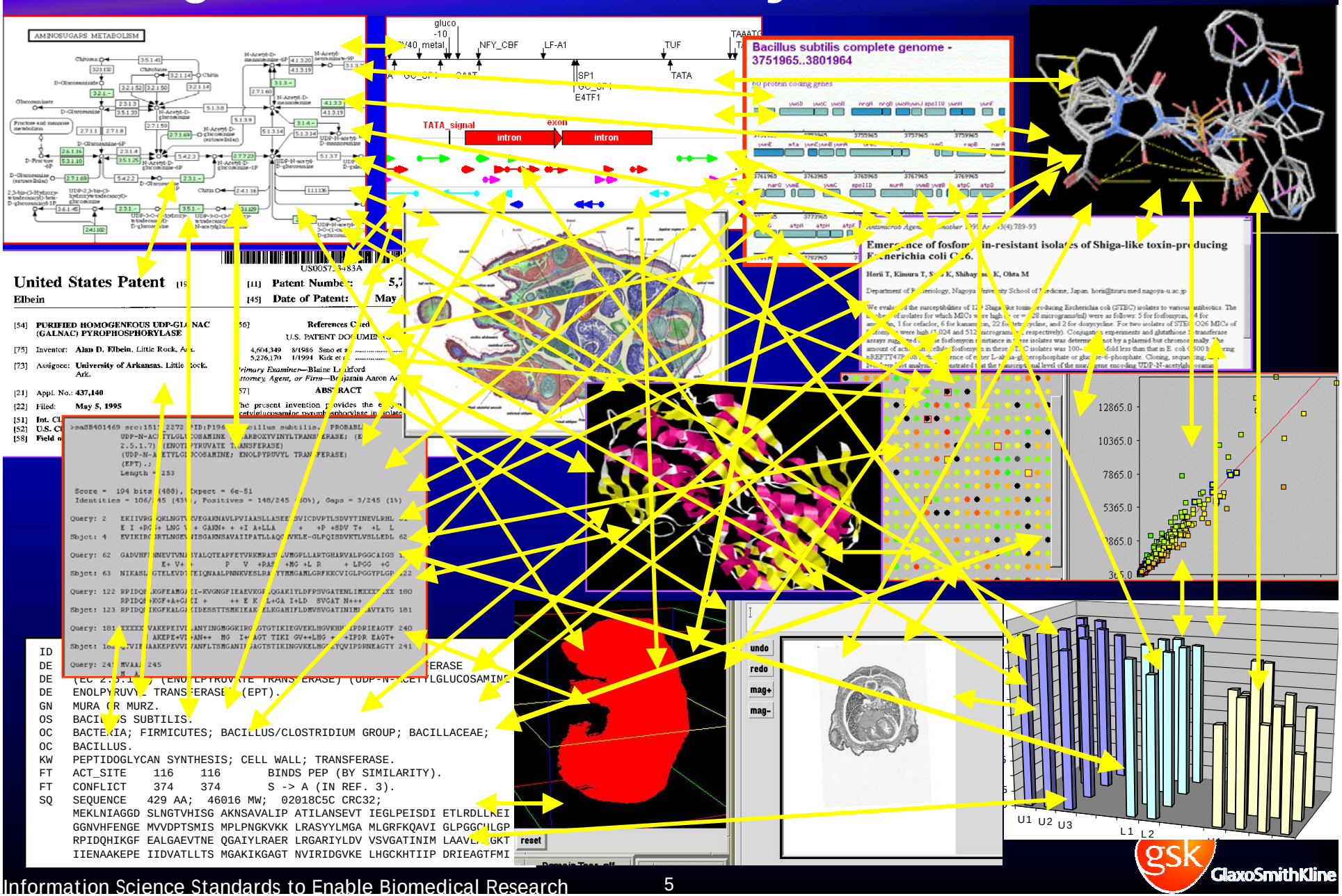
GlaxoSmithKline

Why integrate?

- Provide unified access to data regardless of source
  - Internal (proprietary)
  - Alliance
  - External (public or licensed)
- Connect related data (e.g., by target, pathway)
- Allows users to focus on questions, not access (schemata, query languages, source location)
- See platform data in context
- Provide all information required for decision support

GlaxoSmithKline

Integration of discovery information

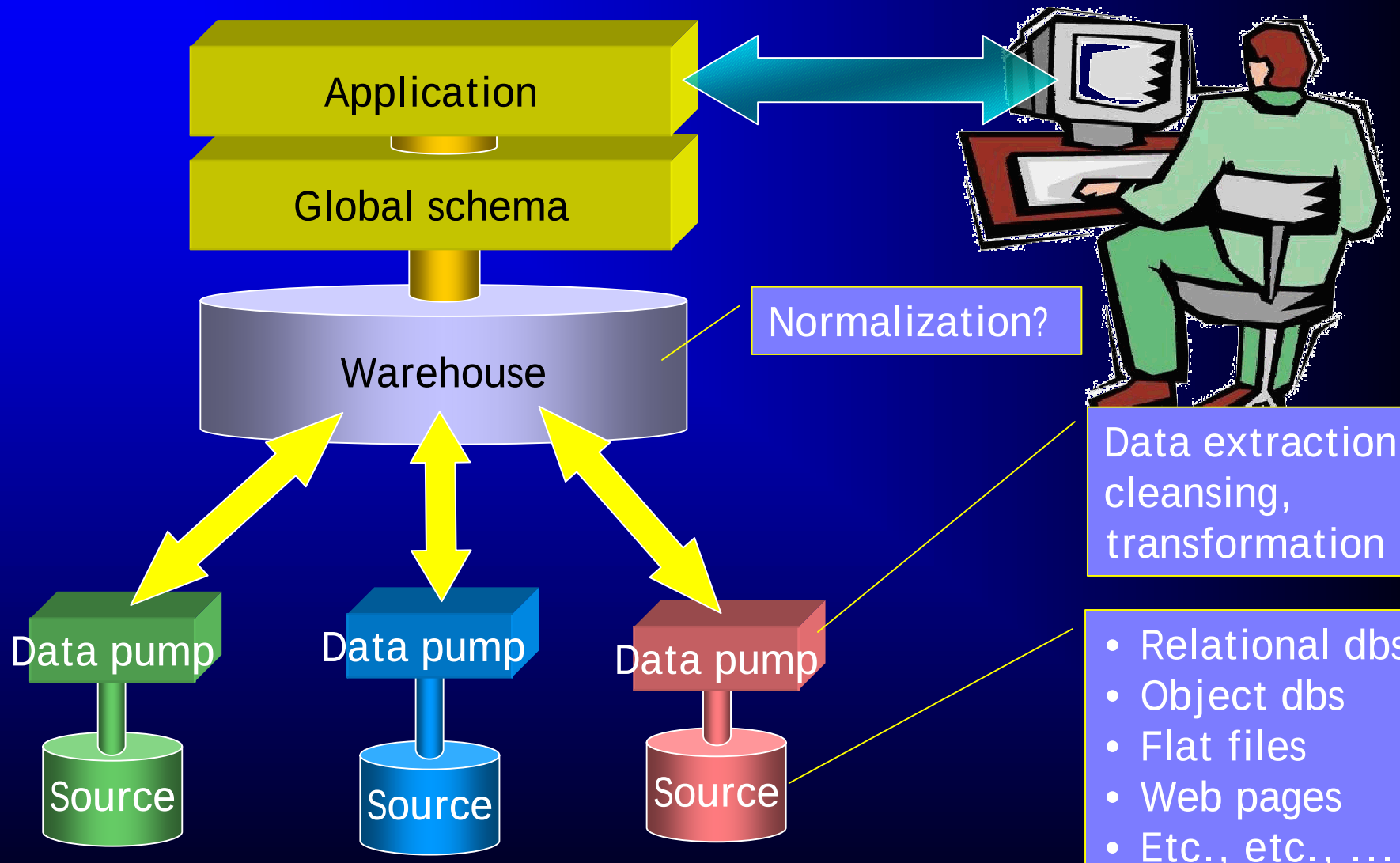


Approaches to Information Integration

- *Ab vitro*
  - Integration is left to the user
- Data warehouse
  - Instantiates data from multiple sources in a central database (materialized views)
  - Use if significant data cleansing is required
  - Best for database searching or queries that return large results sets
- Federated database
  - Uses mediated queries to access distributed data at the remote sources
  - Best for rapidly changing, high-volume data
  - Useful if data must be kept segregated

GlaxoSmithKline

Data Warehouse approach

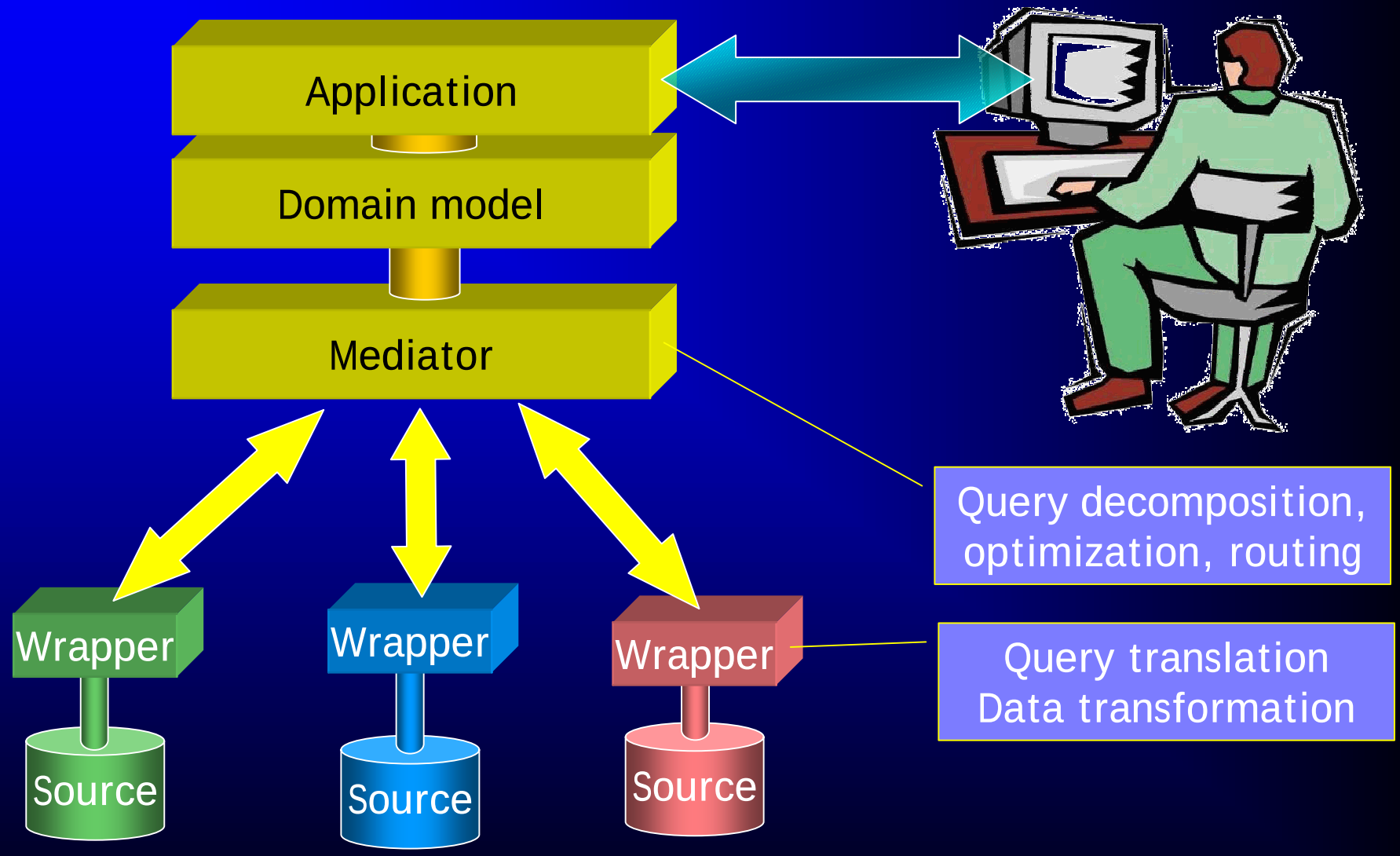


Data extraction, cleansing, transformation

- Relational dbs
- Object dbs
- Flat files
- Web pages
- Etc., etc., …

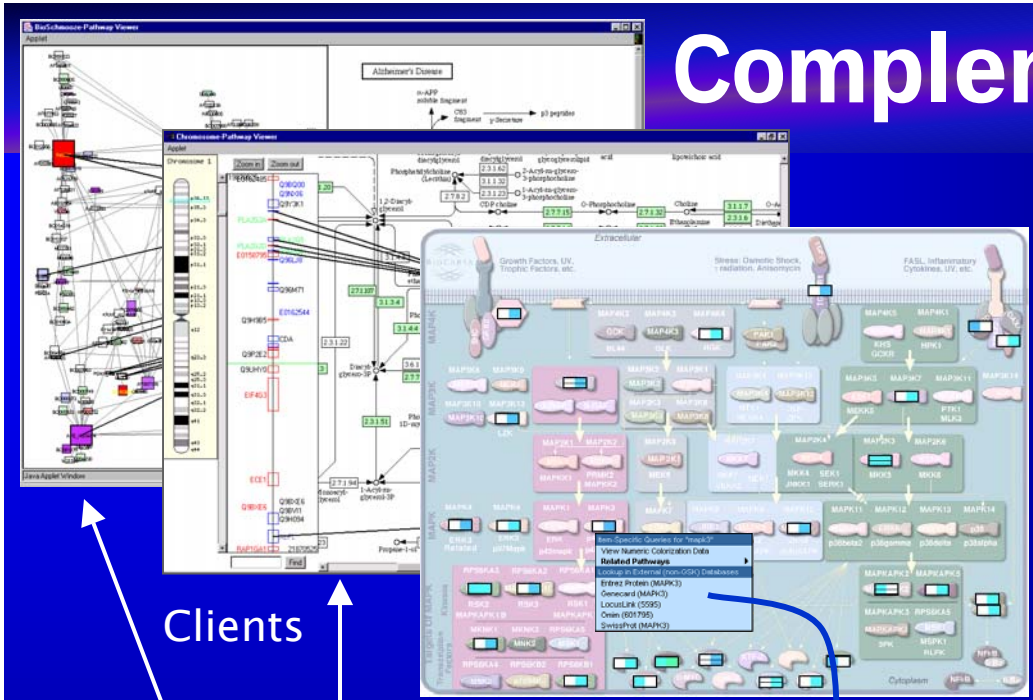
GlaxoSmithKline

Federated Database (mediated) approach



GlaxoSmithKline

Complementary approach



GSK tools & facilities

Clients

Mediator

Visualisation Tools

K2

Pathway Warehouse

KEGG

BioCarta

B-M

Pathway Maps

Exp. Data

Medline

Inhibitors

GATE

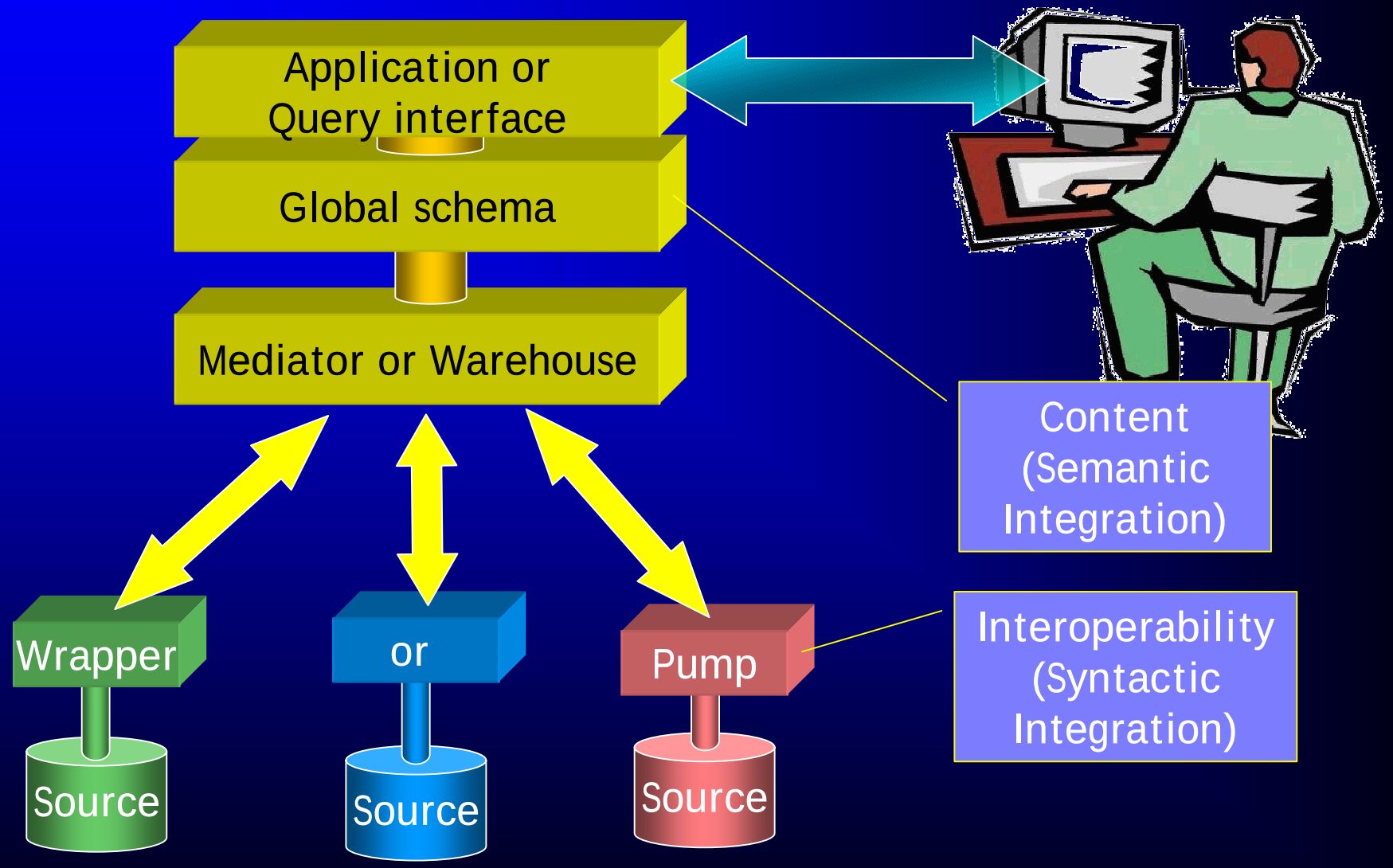
...

Pathway Sources

Real-time access to information sources

GlaxoSmithKline

Information integration architecture



Information integration requirements

- System interoperability
  - at the data pump or wrapper level
  - required to communicate query to source and data from source
- Semantic consistency
  - in information retrieved from varied sources
  - required to correctly map data into warehouse
  - required to correctly merge data from disparate sources

GlaxoSmithKline

Heterogeneity: the great obstacle

- Heterogeneous software systems
  - Hardware platforms, Operating systems, Network protocols
  - Programming languages, Application interfaces, Error handling
  - Data model, Query language & capabilities, Transaction protocol
- Heterogeneous information semantics
  - naming conflicts
  - measurement conflicts
  - representation conflicts
  - computational conflicts
  - granularity conflicts

gsk GlaxoSmithKline

Sources of heterogeneity

- Irreducible complexity of biological systems
- Incomplete, evolving knowledge
- Extremely distributed nature of life sciences research activities permits (or encourages):
  - idiosyncratic experimental protocols, nomenclature, data representation, …
  - very heterogeneous distributed computational environment: systems, programs, languages, file formats, data types, concepts, organisations, …

GlaxoSmithKline

Technical solutions?

- COM / DCOM
- CORBA
- ICE
- .net
- REST
- SOAP
- XML

AGAVE:

```
<bio_sequence molecule_type="AA" length="8">
   <sequence>
     akffglkl
   </sequence>
</bio_sequence>
```

BSML:

```
<Sequence molecule="aa" length="8">
   <Seq-data>
     akffglkl
   </Seq-data>
</Sequence>
```


GlaxoSmithKline

Shared ontologies & domain models

For information integration, information sources and users must share:

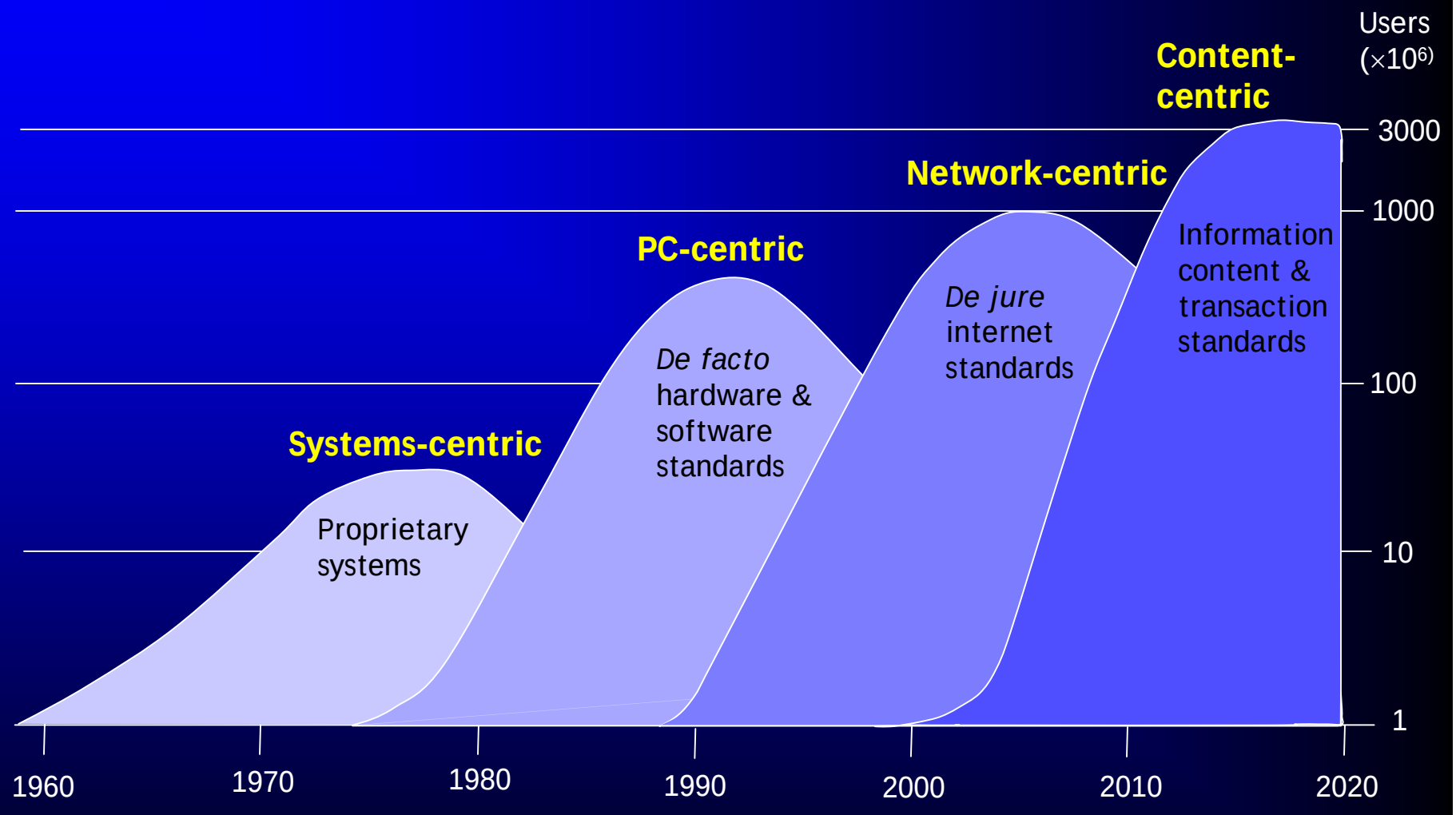
- Domain model of information space of interest
- Understanding of information source query interface
- Semantics of information representation

These understandings :

- are embedded in software components & services
- are expressed in information specifications
- can be achieved pair-wise or through standardization

GlaxoSmithKline

Stages of IT industry expansion



Source: David Moschella, *Waves of Power: Dynamics of Global Technology Leadership 1964-2010* and *The Economist*, May 10, 2003

Standardized components & services

Service / Component	Source	Scope
Biomolecular Sequence Analysis	OMG LSR (EBI, NetGenics)	Generalized analysis engine for sequences, alignments
Open Genomic Maps	OMG LSR (GSK/EBI)	Repository server for genomic maps
Macromolecular Structure	OMG LSR (SDSC)	Repository server for protein and nucleic acid structures
Bibliographic Query Service	OMG LSR (EBI)	Repository server for bibliographic information
BioJava, v. 1.3	BioJava Project	Open-source Java classes for manuipulating and analyzing sequences
BioPerl, v. 1.2.2	BioPerl Project	Open-source Perl classes for manuipulating and analyzing sequences
DAS (Distributed Annotation System), v. 1.53	BioDAS Project	Web-based system for exchanging genomic annotations

GlaxoSmithKline

Information standards

Standard	Source	Status
(CML) Chemical Markup Language , v. 2	P. Murray-Rust	Parsers and other tools available
SMILES	Daylight	Commercial software packages
MOL2	[illegible]	[illegible]mercial software packages
MIF (Molecular Inform[illegible] File)	[illegible]	[illegible]shed
Molfile, Rgfile, Rxnfile, [illegible], RDfiles	[illegible]	[illegible]shed. Commercial software packages
CXF-10	[illegible]	[illegible]d
mmCIF	[illegible]	[illegible] file manipulation [illegible], viewers
ASN.1 Bioseq	[illegible]	[illegible]d. Parsers and other tools available
AGAVE™	Life Code	Published
BSML	LabBook	Published. Viewers available
Microarray and Gene Expression (MAGE)-OM, -ML	OMG LSR	OMG adopted specification, implementations available
Gene Ontology	GO Consortium	Browsers, parsers available

How many standards is too many?

Is it time for some unification?

GlaxoSmithKline

Sources of ontologies & domain models

- Home-grown
- Vendors
- Information Providers
- Academic Research Projects
- Standards Organizations
- Industry Consortia

GlaxoSmithKline

Can consortia develop standards for pharma R&D?

- Benefits
  - Open process, full participation leads to broad acceptance & multiple implementations
  - Peer review leads to higher technical quality
- Costs
  - Time
  - Effort required to reach agreement
  - "Design by committee" doesn't usually produce the most direct answer to individual participants' immediate problems

GlaxoSmithKline

Standards development alliances & consortia

- ISO, ANSI, IEEE, IETF, OASIS, W3C
- Health Level Seven (HL7), CDISC
- OMG Life Sciences Research Domain Task Force
- Interoperable Informatics Infrastructure Consortium (I3C)
- Open Bioin
  Biopython, DAS, BioMOBY, BioSOAP, BioXML, BioRuby, BioGopher, BioPHP, …
- MGED
- Global Grid Forum Life Scien
- Open Science Alliance™
- Gene Ontology (GO) Consortium
- BioPathways Consortium

How many standards organizations is too many?

Is it time for some unification?

GlaxoSmithKline

Engagement with alliances

- Shopper / Victim

  No alliance engagement

- Watcher

  Semi-passive acceptance

- Navigator

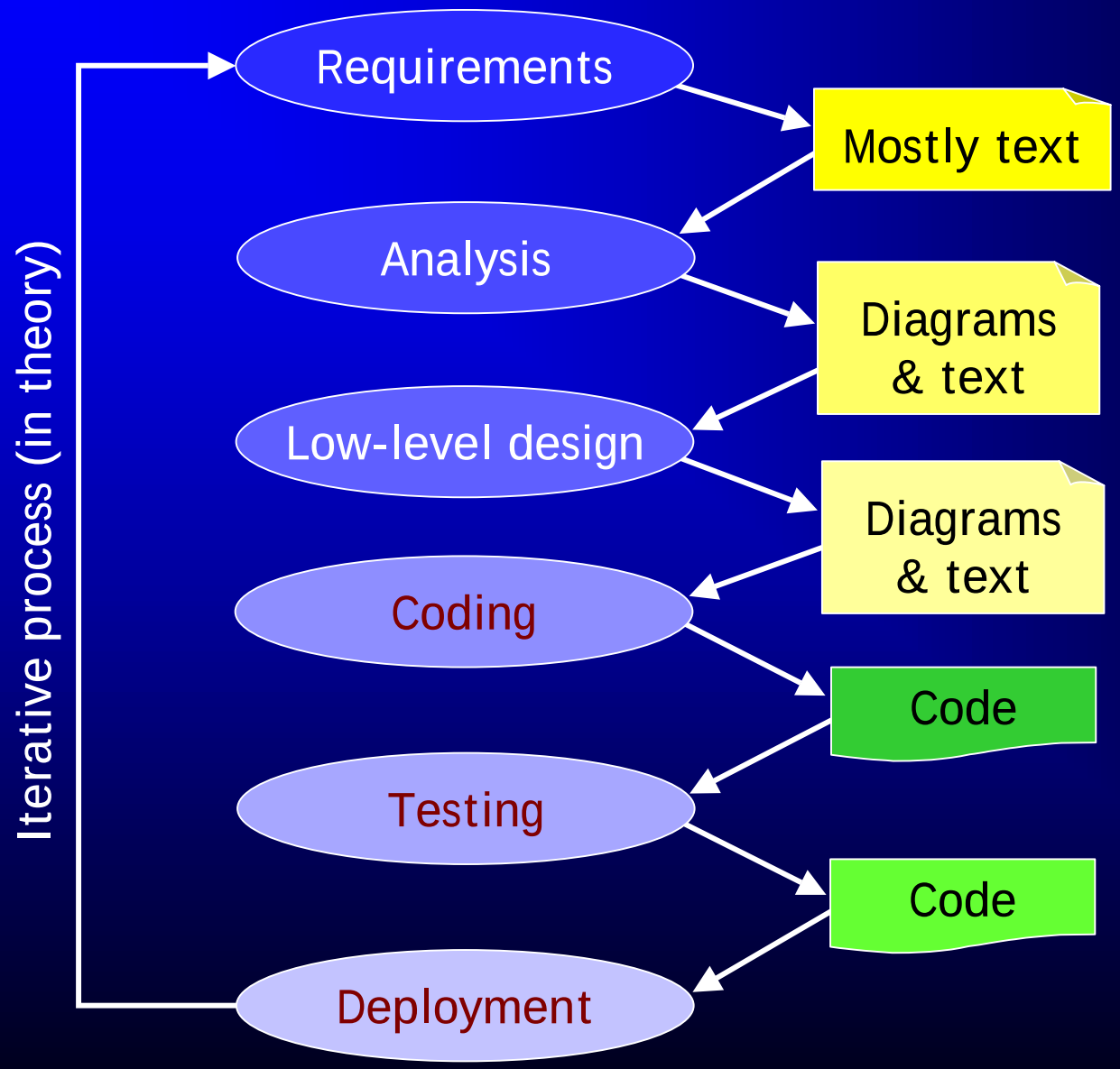
  Active participant: influences software & component development to suit enterprise strategic needs

- Developer

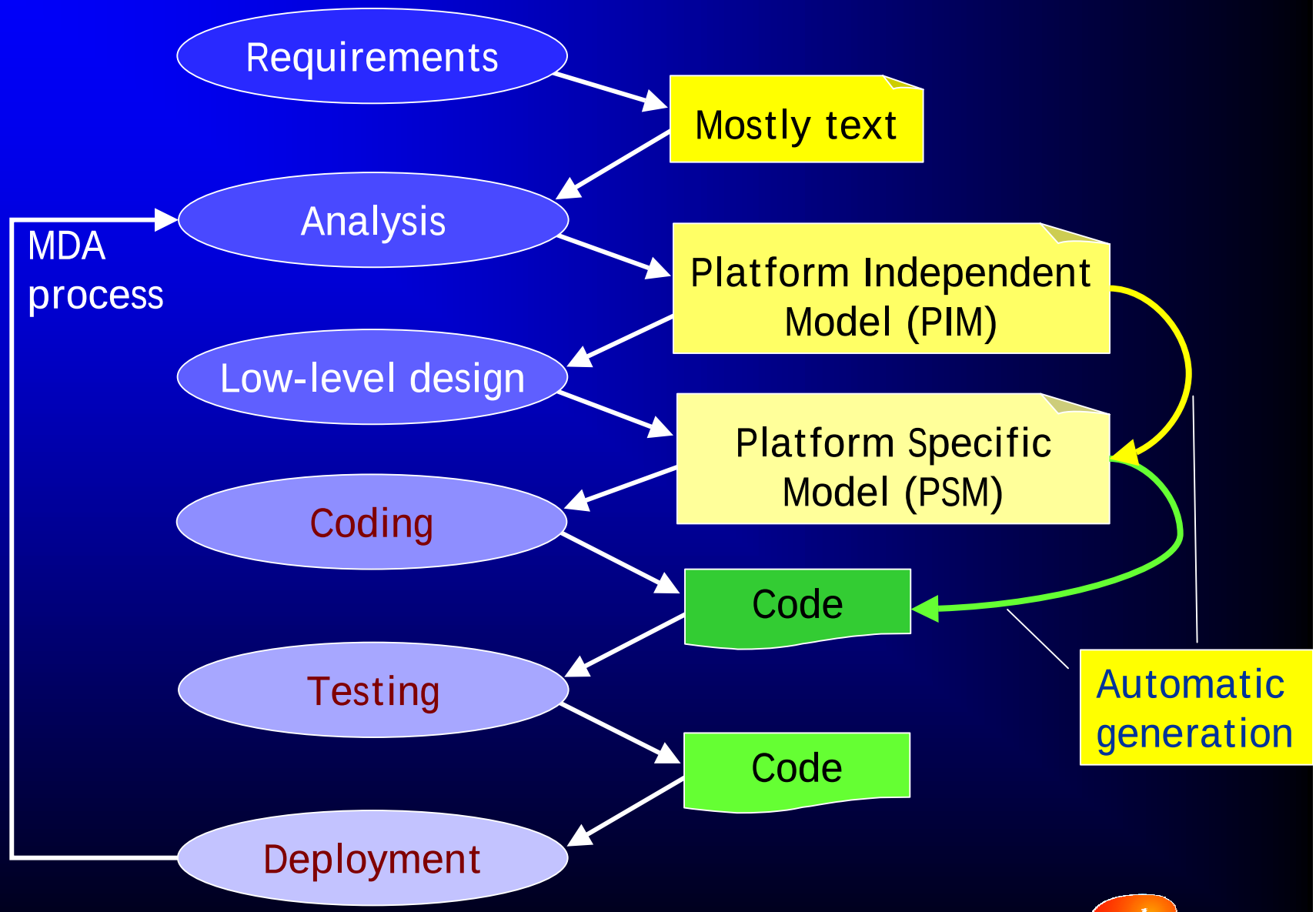
  Contributes technology / implementations

GlaxoSmithKline

The software development process



Model-driven development process

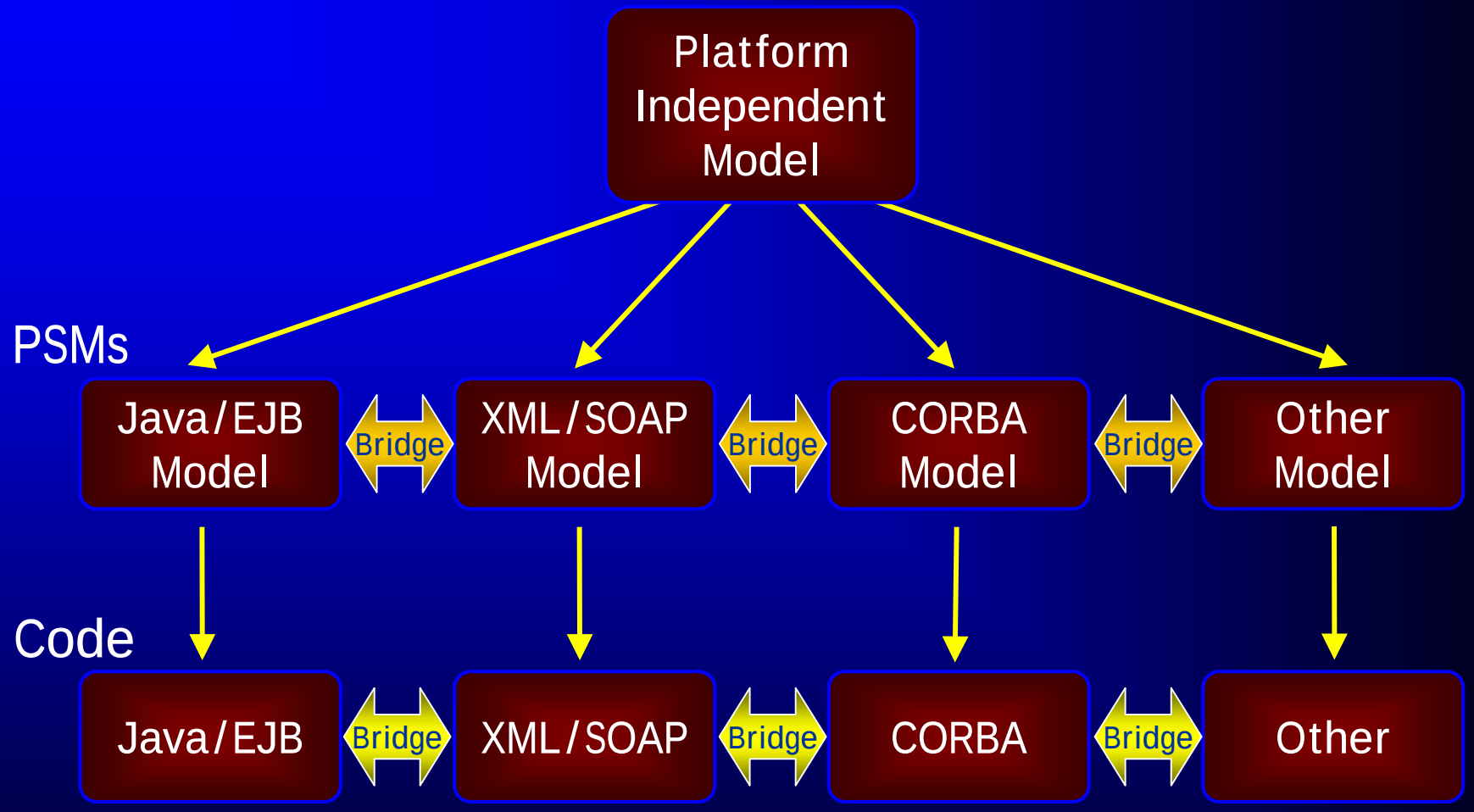


OMG's Model Driven Architecture (MDA)

- Solve the domain problem once: PIM in UML™
  - Formal specification of the structure and function of a system that abstracts away technical details
  - UML can be used to specify structures, behaviors, and constraints without committing to a specific platform
  - Distills the fundamental structure and meaning
- Add middleware details as needed: PSMs in UML
  - Introduces artifacts specific to a given platform, e.g., CORBA Services
  - By leveraging a PIM it's easier to produce implementations on different platforms while holding the essential structure and meaning of the system invariant
- Approach works with SOAP, XML, Java, EJB, CORBA, .NET, Web Services, "next best thing"

GlaxoSmithKline

Putting It All Together



How does MDA connect to integration?

- In Pharma R&D, it is common to have (need?) multiple middleware solutions for a given problem, e.g., Java and XML representations of a bio-sequence
- Interoperability and integration are much easier if these solutions have a PIM in common
- MDA provides a clean separation of domain and implementation issues
- Develop semantic standards at the PIM level
- Derive "platform-specific" standards from these PIM-level standards

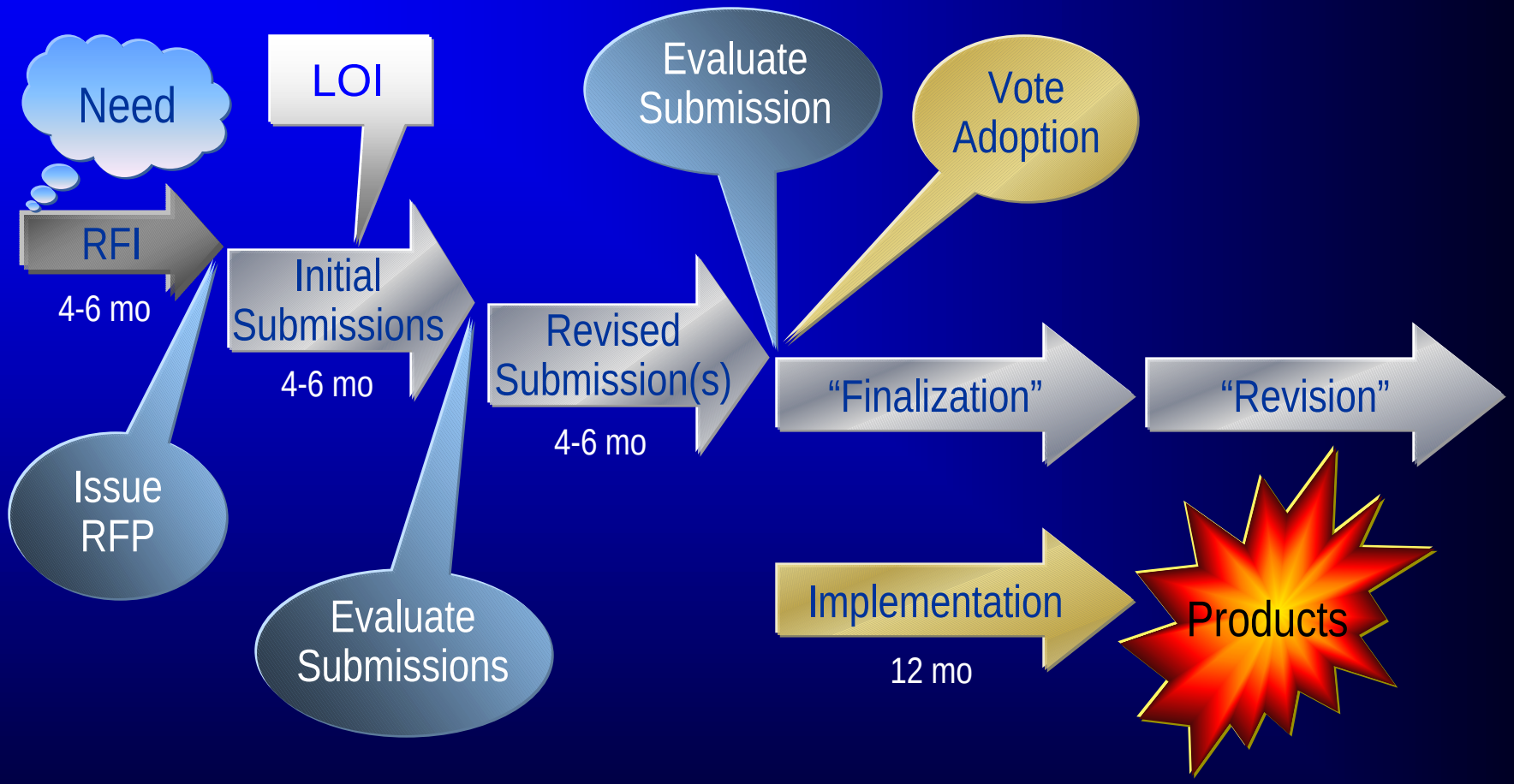
gsk GlaxoSmithKline

Essentials of Standards Development

- Participants
  - Need all significant stakeholders: vendors, users, practitioners, advocates
  - "Single-issue" & non-member organizations can play important rôles
  - Quality standards can only be developed on time if they are seen to be business critical, not "nice to have"
- Process
  - Well-defined "due process" is required for openness
  - Open process promotes (but cannot guarantee) broad acceptance
  - A defined process with deadlines prevents endless discussion
  - Open process allows essential leveraging & unifying of existing efforts
  - Education is key to ensuring standards are actually used
- Products
  - Standards should be as simple & as platform-independent as possible
  - Standards must be driven by good use cases (grounded in reality)
  - Begin by focussing on data models, add methods later
  - Process should require implementations to avoid "paper only" specifications

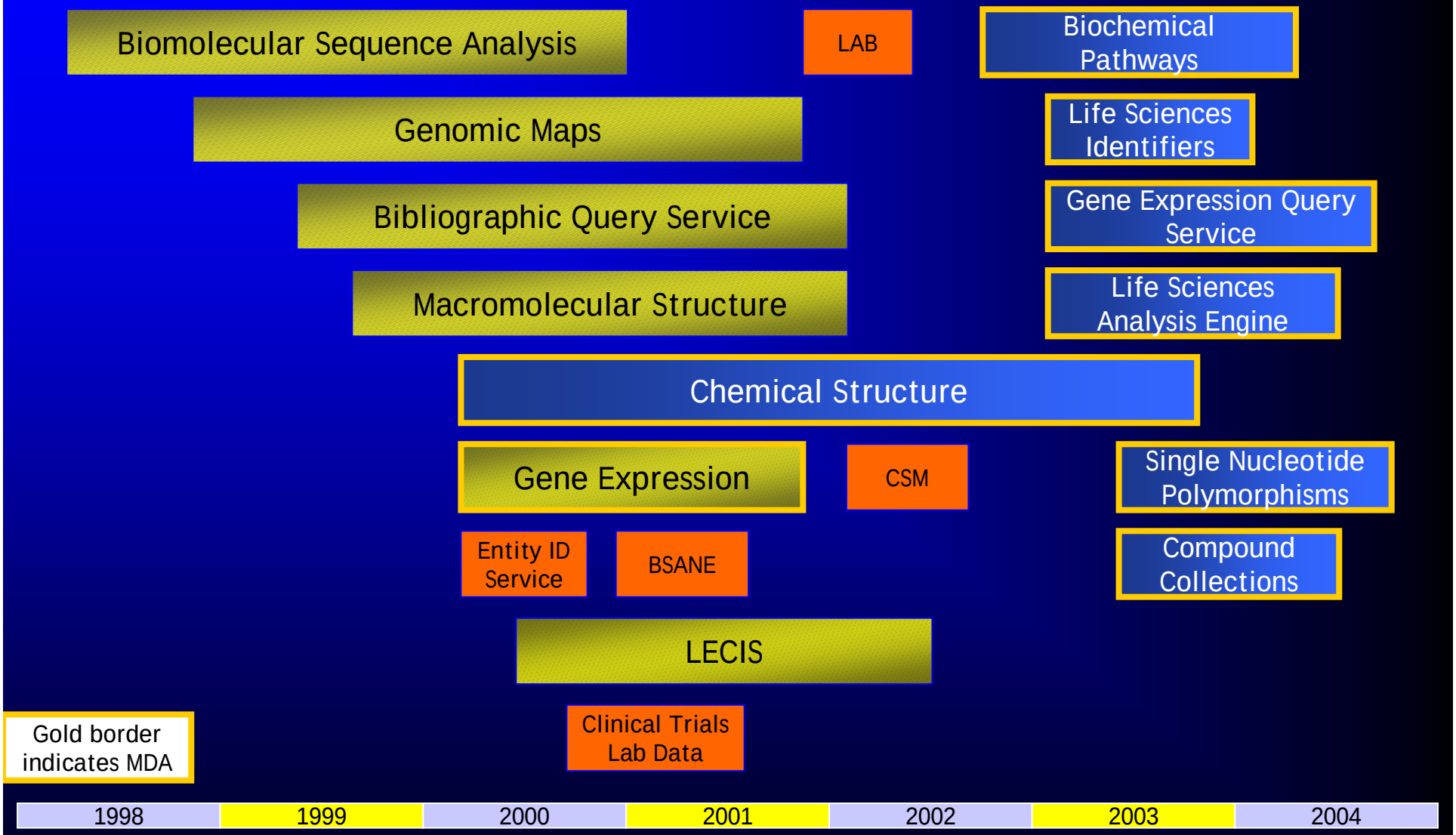
GlaxoSmithKline

OMG technology adoption process



GlaxoSmithKline

OMG LSR Technology Adoptions



GlaxoSmithKline

Availability

- OMG adopts & publishes model & interface specifications
- Specifications chosen from existing products in competitive selection process
- Specifications are freely available to both members and non-members
- Implementations must be available from OMG member (either commercially or open-source)

GlaxoSmithKline

In conclusion

- Information integration requires:
  - Complementary use of warehouse & mediation
  - Interoperable systems
  - Shared ontologies & domain models
- Syntactic & semantic heterogeneity challenges
- Standards can aid in meeting these challenges
- A well-defined, open standards development process offers distinct benefits
- Do we face a Tragedy or Triumph of the Commons?
- Will Pharma participate in the next IT wave?

GlaxoSmithKline

LSR Contact Information

- Web pages
  - OMG: www.omg.org
  - LSR: lsr.omg.org
  - Next meeting agenda: lsr.omg.org/latest/agenda.html
- Email addresses
  - lifesciences@omg.org
  - lsr-chair@omg.org

GlaxoSmithKline

Acknowledgements

- Jim Butler, Filip Fuma, Scott Harker, Paula Matuszek, Robin McEntire (GSK)
- Steve Chervitz (Affymetrix)
- Scott Markel (LION)
- Michael Miller (Rosetta)
- Tony Parsons (Pfizer)
- Richard Scott (De Novo)
- Martin Senger (EBI)
- LSR Members and Friends

GlaxoSmithKline