The Grand Challenge of Information Technology and The Illusion of Validity

Dr. Michael L. Brodie
Chief Scientist

Information Technology

The Global Computing Vision

- The elements and history of Global Computing
  - Computer: every object
  - Storage: every datum
  - Network: every place
  - Applications: every task
  - Processes: every endeavor

- But ... It's NEVER about technology
  - Value ?
  - Semantics: The Grand Challenge of IT
  - Impacts
    - Economic
    - Business
    - Technical
    - Cultural
    - Social
    - Religious

Outline

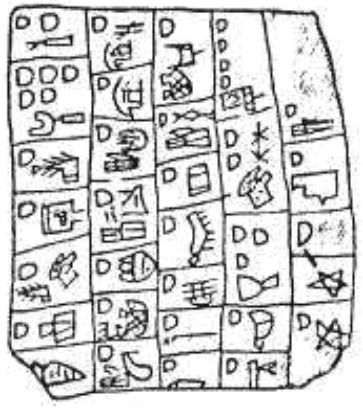
- Progress and Failure In Computer Science
- The Grand Challenge and The Illusion of Validity
- Roadblock to Current and Future Progress
- Why Attempt The Grand Challenge?
- Semantics: The Heart of The Grand Challenge
- Conclusions

Computing Has Changed The World

- Productivity
  - Business
    - Office work
    - ERP: finance, human resources
  - Government Services
    - Air Traffic Control
    - Taxes
- Science
  - Computing has replaced paper, pencil, and mathematics
  - Every domain depends on computing: Astrophysics
- Manufacturing / engineering
  - Boeing 777
- Communication / research
  - Web: *sine que non* for research - any topic in seconds

Profound Achievements

- Information Revolutions
  - 1st: c. 4,000 BCE - writing, Mesopotamia
  - 2nd: c.1300 BCE - book, China (Greece c.500 BC)
  - 3rd: 1450–1455 - printing press, Gutenberg
  - 4th: c. 2000 - information technology, Web



- 3rd Industrial Revolution
  - British 1750-1830 - steam
  - American 1880-1940 - mass production, electricity, …
  - Automation / Information 1946 - 2030?

Progress in Computer Science

- Frequent Paradigm-Shifting Leaps
  - Client/server
  - Objects
  - Intelligence
  - Knowledge
  - Understanding

- Steady Stream of Visions
  - E-Business
  - Semantic Web
  - Collaborative Design

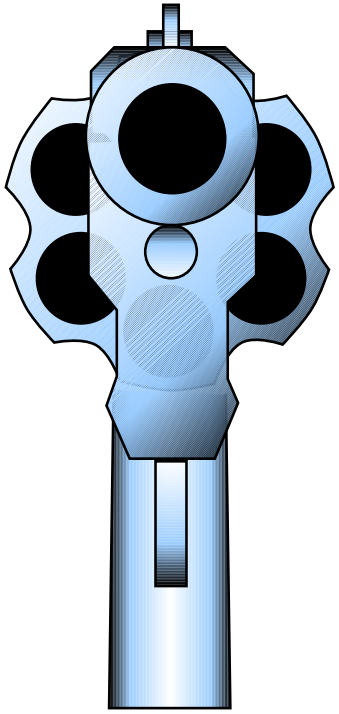
Amazing Failures

- Productivity Paradox
- International Conspiracy
  - Technology failure rate: 80% [Moore's Chasm]
  - Project failure rate on $250 B/year [Standish Group]
    - 30% fail
    - 52% "challenged"
    - 16% succeed
- Silver Bullets

Silver Bullets

Large Scale Industrial Trends

- Open Systems
- Distributed databases
- Legacy extension/optimization
- Legacy migration
- CASE
- Outsourcing
- Re-engineering
- Build integrated environments and applications
- Buy: best of breed, best practices
- Unified COTS / ERP
- Enterprise Integration
- Dot.Com
  - Internet Speed
  - 1st mover advantage



Technical Trends

- Client/server
- Expert systems
- Business process re-engineering
- Object-oriented products
- Workflow
- Enterprise modeling
- Conceptual modeling
- Domain orientation
- Business objects
- Business rules
- Re-use
- Class libraries
- Distributed object computing
- Agents
- Knowledge Management
- Business Intelligence

Future Silver Bullets?

- Post Dot.Bomb Hot Trends
  - Wireless Internet / anything “Mobilize or die”[1]
  - Instant messaging
  - Peer-to-peer (P2P)
- Business Intelligence
- Knowledge Management
- Adaptive Supply Chains
- Semantic Web
- Web Services
- Collaborative Commerce

[1]“Wireless E-Commerce Bombed”, Informationweek, June 17, 2002

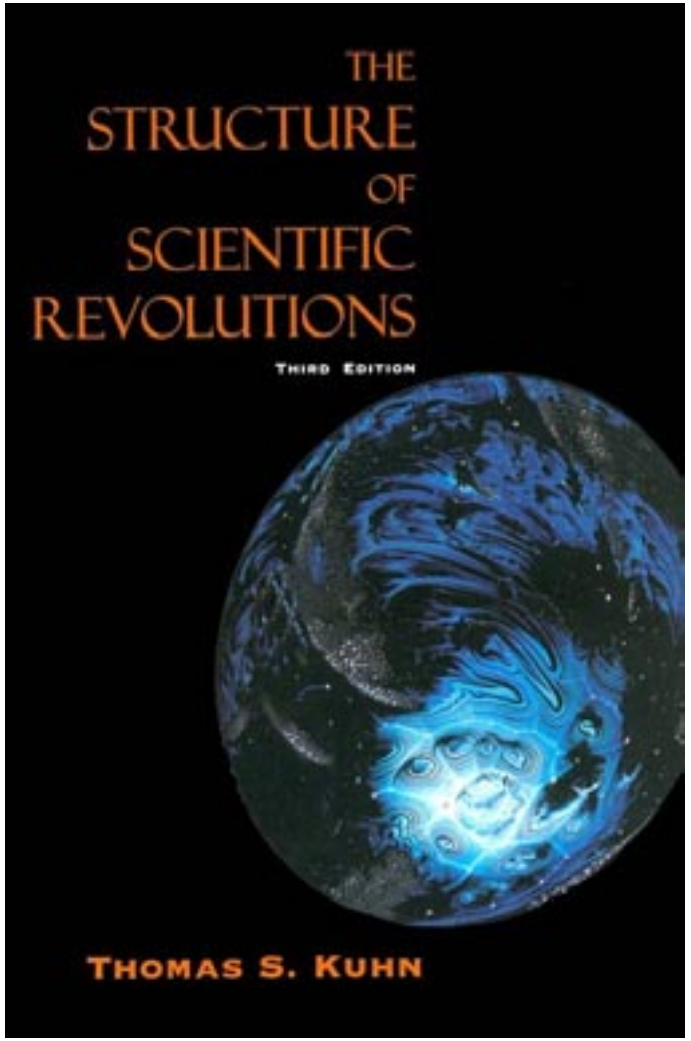
Silver Bullet Pattern

- Pattern
  - Big Vision1 (e.g., CORBA)
    - Dramatic claims / promises
    - Vision1 trouble
  - Big Vision2 (e.g., JAVA)
    - Dramatic claims / promises
    - Big Vision1 vanishes
    - Vision2 trouble
  - Big Vision3 (e.g., Web Services)

- Recurring Theme: Next-Generation Information Systems
    - Distributed
    - Service oriented
    - Scalable
    - Plug and play
    - Integrated
    - Re-use
    - Class libraries
    - Business objects
    - Process-oriented
    - Flexible

Normal Science[1], Not Revolutionary

- Visions
  - Great, inspiring, necessary
  - Rarely realized “as advertised”
    - Pull – Crisis or Necessity - Mother of Invention
    - Push - Unanticipated breakthroughs
- Perennial lack of progress
  - Integration: systems, process, data
  - Data hygiene: consistency, integrity, security
  - View construction and materialization
  - Data models, conceptual modeling
  - Technology evolution: systems, data, ...
  - Methodologies
- Why?
  - It’s not about technology
  - Adoption is a social (non-technical) issue
  - Research abstracts away critical issues: scale
  - Inherently hard - contains the Grand Challenge



Outline

- Progress and Failure In Computer Science
- The Grand Challenge and The Illusion of Validity
- Roadblock to Current and Future Progress
- Why Attempt The Grand Challenge?
- Semantics: The Heart of The Grand Challenge
- Conclusions

The Grand Challenge of IT

- Semantics: capturing real world "meaning"
  - Enhance Information Systems so that the automated actions and data more closely correspond to the real world actions and facts that they represent, with minimal human involvement
  - Stunning Example: "Books of Record" for all major corporations
- Reasoning
  - Enhance automated reasoning to assist human problem solving
  - Stunning example: "What if ...

Semantics Quiz

- Can machines think?
- What do Information Systems know?
  - Are answers to queries "The whole truth and nothing but the truth" ?
- Does you schema contain
  - Semantics?
  - More semantics than Fred's?
- Does your Information System deal with semantics?
- What role does semantics play in your problem / solution ?

Grand Challenge Properties

- Pervasive
  - Business requirements
  - Technologies
  - Visions
- Little progress in 30 years
- Cyclical re-appearance
  - From fascinating to mission critical
- Inadequate understanding

Progress in computer science and IT depends on a more principled and robust treatment of semantics

- Identify to role of semantics in your problem
- Model and analyze the solution for soundness, completeness, feasibility, ...
- Fix to avoid semantic problems (at least semantics preserving or lossless)

The Illusion of Validity

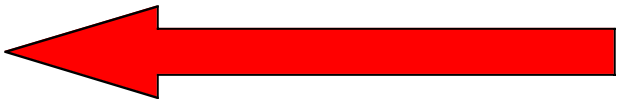
- Illusion of Validity[1]
  - Focus on evidence that would confirm your beliefs, creating and reinforcing your understanding of the world.

- Applications
  - Behavioral Finance (I.e., investing)
  - Silver Bullets: biases impede progress in computing
    - IT professionals and CEO accept strings of anecdotes as proof that IT spending boost productivity, instead of finding rigorous ways of assessing ITs contribution. The Squandered Computer, Paul Strassmann
    - "Despite the enormous investment in IT during recent years, demonstrating the effects of such investments on organizational performance has proven extremely difficult" The Journal of management Information
    - "... when pushed, decision makers, both individual and corporate, often describe their decisions as being based to a greater or less extent on instinct." Electronic Journal of Information System Evaluation

1 Einhorn HJ, Hogarth RM. Confidence in judgment: persistence of the illusion of validity. Psychol Rev. 1978;85:395-416

Outline

- Progress and Failure In Computer Science
- The Grand Challenge and The Illusion of Validity
- Roadblock to Current and Future Progress
- Why Attempt The Grand Challenge?
- Semantics: The Heart of The Grand Challenge
- Conclusions



Current Technologies and Research

Technologies heavily dependent on "semantics"

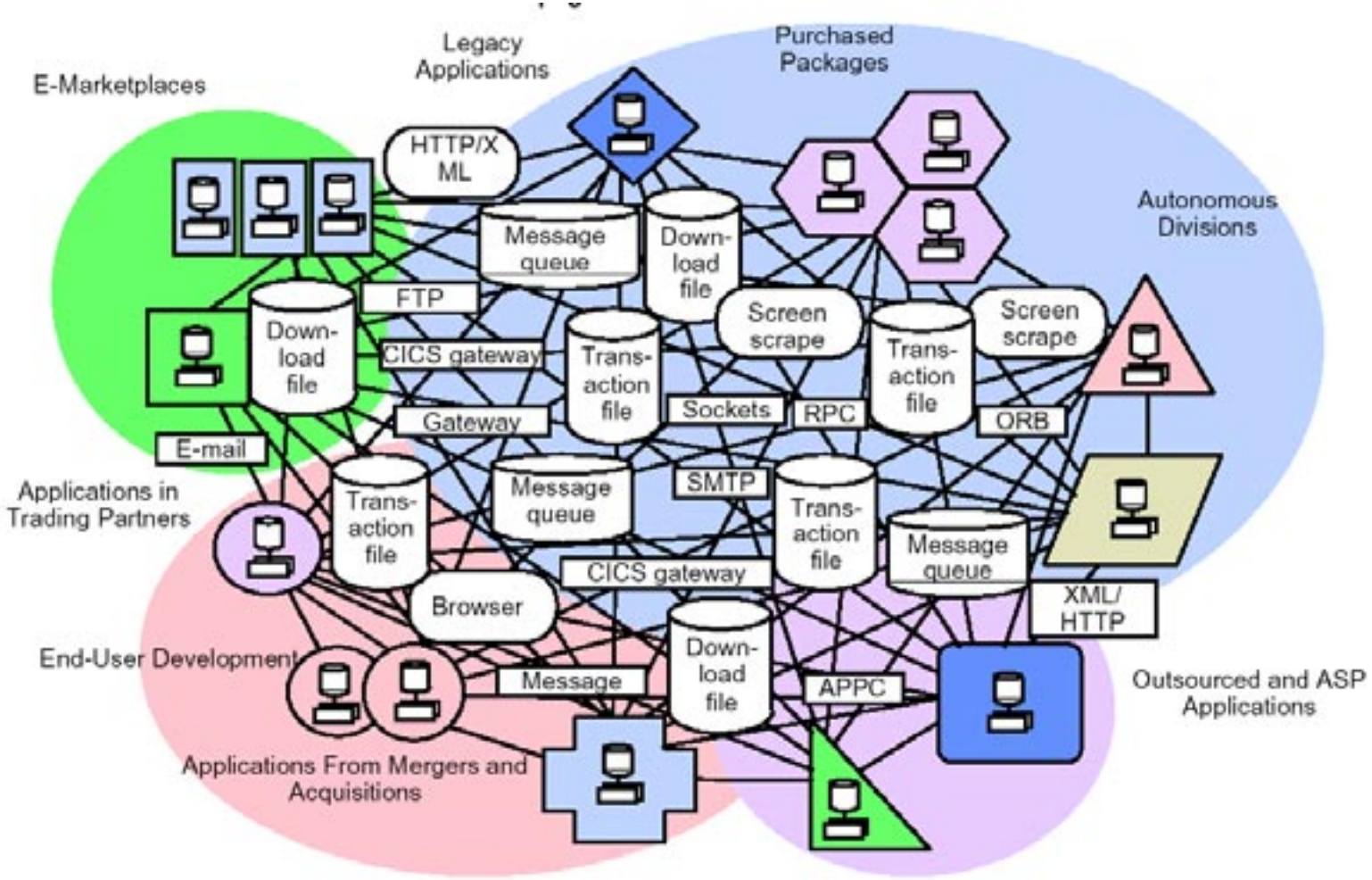
- Search
- Query processing
- Database design
- Database views
- Distributed databases
- Distributed computing
- Interoperability, Heterogeneous and Federated Databases, Mediators
- Data warehouse
- Data Mining and Knowledge Discovery
- Data Quality
- Data Transformation, Integration, Evolution, and Migration
- Data Warehousing
- Information Retrieval with Database Systems
- Meta-data management
- Personalized or Profile-Based Data Management
- Workflow Systems

- **And lots more ..**

Lack of understanding in the research community

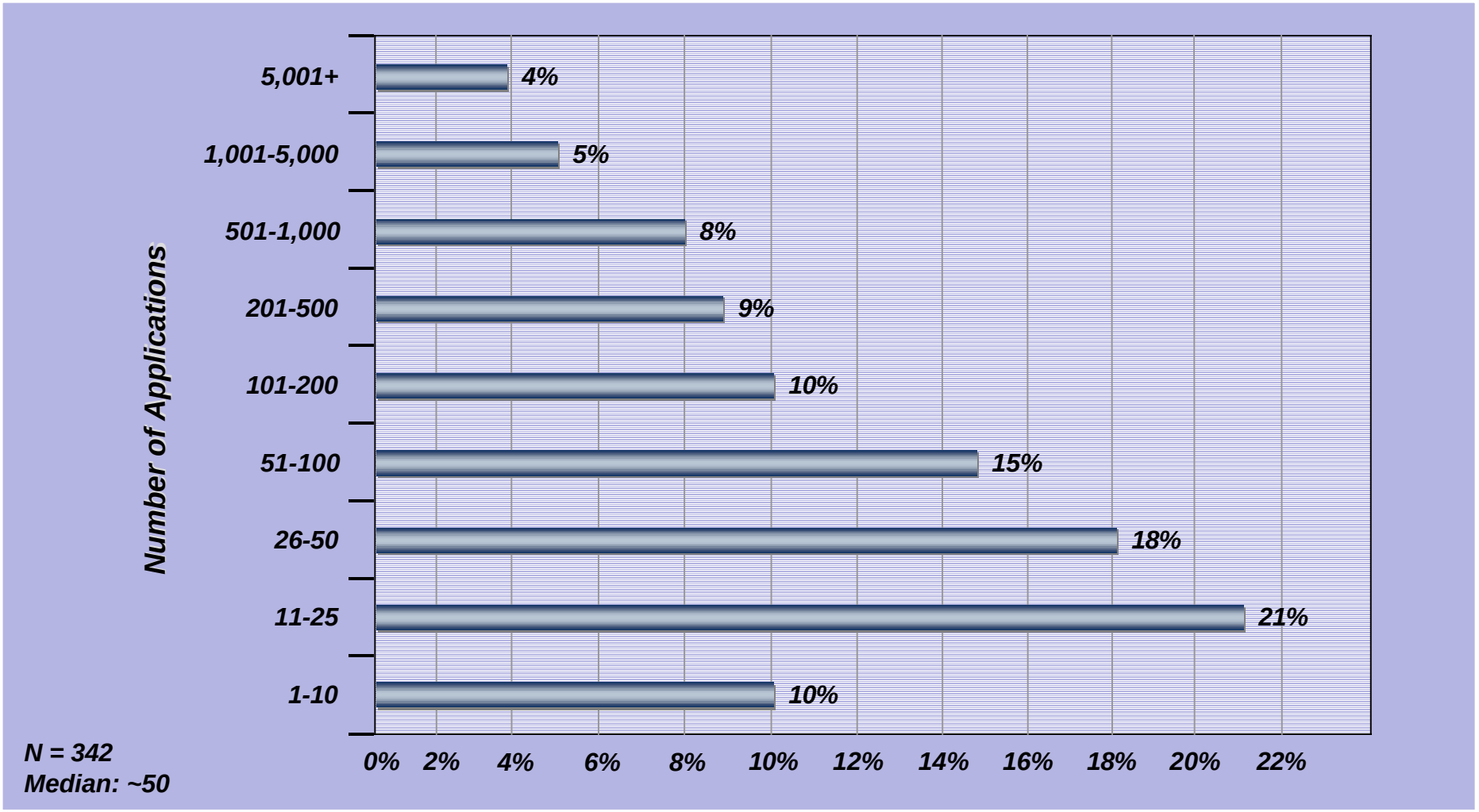
- Precision
  - Identify the role of semantics
  - Model and analyze the solution
  - Ensure feasibility, etc.
- Where
  - Papers
  - Presentations
  - Dot.com-like behavior
- Reality: scale

Reality: Constellations of System Clusters



Source: Gartner Research

How Many Applications?[1]



[1] 2002 Customer survey

Business Requirements: Integration

Characteristics

Integrate multiple (currently isolated)
- Processes
- Applications
- Data repositories

Processes Integration
- End-to-end business processes
  - Long-lived

Global Data Management
- Single logical data store (e.g., customer) over many sources
  - Heterogeneous
  - Structured and unstructured data w&w/o meta-data
  - Internal and external sources
  - Access restrictions
  - Varying “soundness”, cleanliness, content, ...
  - Vast number, vast growth (50+%)

Ensure
- Dynamic: Real-time access for accuracy
- Semantic equivalence of “equivalent” things - discount
- Seamlessness
- Flexible: systems enter and leave integration
- Performance

Areas

- Legacy evolution / migration
- Reverse engineering
- Integrated application suites
  - ERP: all finance and HR data
  - CRM: all customer data
  - Supply Chain / Logistics
  - Product Management
- Data warehouse
- Web
  - Search
  - Web-based Information Systems
  - Portals: enterprise, employee, customer
- Collaboration ($4.5 B sales in 2002, IDC)
  - Design
  - Ordering
  - Claims processing
- E-Business
  - E2E
    - Enterprise content management
    - Enterprise Portals
  - B2C
    - Multi-channel integration
  - B2B
    - E-Marketplaces

Business Requirements: Problem Classes

Legacy Modernization[1]

- Decompose: **EAI**- Enterprise Application Integration (real time access)
  - Break into “basic” functions
  - Expose via API
- Analyze
  - Identify common functions
- Re-engineer
  - Make common functions equivalent
- Publish: for enterprise use
- Combine: into new services
- Discover: dynamically
- Invoke: dynamically
- Debug: when errors detected

E-catalogue[2]

- E-Marketplace
  - Participants
    - Buyers (1,000s)
    - Supplier (1,000s) [Grainger 60,000]
  - Global catalogue (over supplier catalogues)
    - Description, price, availability, shipping, discounts, ...
    - High overlap, constant changes
- Customer query global catalogue
  - Find products and terms (fast)
  - Select products (eventually)
  - Commit to buy (legally)
  - Follow through
    - Logistics
    - Status
    - Payment
- Dynamic
    - Discovery
    - Partnering
    - Adaptation
    - Evolution

[1]Also: Distributed (Object) Computing, DCE, CORBA, COM+, CoopIS, Web Services, …
[2]Also: Distributed Queries, Web queries, product management, order status, manufacturing status, …

Other Business Requirements

- **Data Quality**
  - Industry average: 5-10% of data erroneous (Richard Wang, MIT - not validated)
  - Telecom
    - Finance databases: 0%
    - Network databases: 25-30%
  - **ETL**: Extract, translate, and load (for static integration)

- **E-mail**
  - BI, KM: Manage, search, understand
  - Filter: Spam, pornography

- **Document Management**
  - Content management

- Etc.

Summary: Industrial Challenges

Challenges

- Layers of integration
  - Humans
  - User interface
  - Business Processes
  - Applications
  - Data
  - Meta-data: tables / repositories / schemas / ontologies / ...
  - Platform
- Two+ resources - probably 1,000s
- Distributed
- Must communicate - agree
  - Query
  - Update
- Heterogeneous
  - Representations
  - Where “meaning” is represented

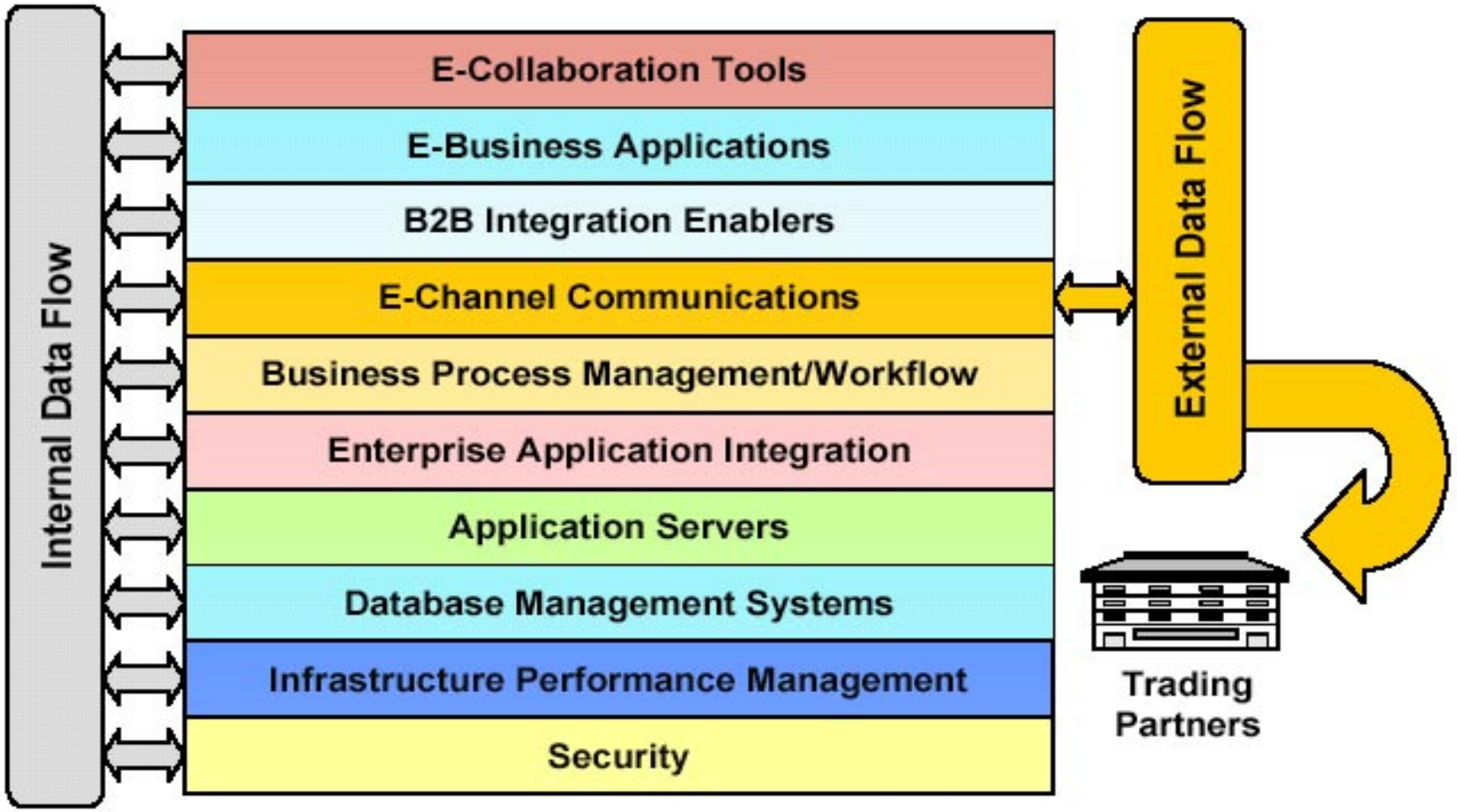
Perpetual IT Problem

- Or can you imagine a universal
  - Modelling language
  - Query language
  - Data model
  - Process model
  - Architecture
  - Computational model
- Forever ...?

Practical Solutions

- Spectrum of Solutions
  - Infrastructure / platform — Vendor Hard
  - Automation: language/ modeling / design — Turing Hard
    - How far can automation take you?
  - Semantics
    - Agreements — Politics Hard
    - Formal — Nobel Hard
- Community Agreement / Standards
  - Local
    - Enterprise
    - Powerful vendors, associations, ...
  - National / international
- Automation
  - Tools for specific problems
    - Automate
    - Reduce human error
  - Let's look at the current Chaos

The E-Business Integration Technology Stack



Source: Giga Information Group

Figure 1

Table 1: Security Vendors

Group	Group Description	Sample Vendors
Authentication	Who are you?	ActivCard, RSA Security, Computer Associates, Vasco, Entrust, Baltimore, VeriSign
Authorization	What may you do?	Check Point, VeriSign, WatchGuard, Computer Associates, Tivoli, Microsoft, ISS
Administration	How do I manage it all?	BMC, Access360, Microsoft
Audit	What happened?	Axent, ESM, PentaSafe, Counterpane
Enterprise Application Security	All of the above	Netegrity, Securant, Entrust, Entegrity, Oblix, Baltimore

Source: Giga Information Group

Table 2: Infrastructure Performance Management Vendors

Group	Sample Vendors
Infrastructure Performance Management	Computer Associates, Micromuse, Tivoli, HP

Source: Giga Information Group

Table 3: Database Management Systems

Group	Sample Vendors
Databases	IBM DB2/UDB, IBM Informix, Oracle 8i/9i, Microsoft SQL Server, NCR Teradata, Sybase ASE

Source: Giga Information Group

Table 4: Application Server Vendor Offerings

Group	Sample Vendors
Application Servers	BEA Weblogic, IBM WebSphere, iPlanet Application Server, Sybase/New Era of Networks EAS, HP/Bluestone Total e-Server, Oracle 9iAS

Source: Giga Information Group

Table 5: Enterprise Application Integration Vendor

Group	Sample Vendors
EAI Solutions	TIBCO, SeeBeyond, WebMethods/Active Software, Sybase/New Era of Networks, Vitria, Crossworlds, Viewlocity, Mercator

Source: Giga Information Group

Table 6: Business Process Management/Workflow Vendors

Group	Sample Vendors
Business Process Management/Workflow Solutions	Staffware, IBM, FileNet, Fujitsu, HP, icomXpress, Jetform, TIBCO, Peregrine, Savvion, Sun, Versata, Vitria, W4

Source: Giga Information Group

Table 7: E-Channel Communications Vendor Offerings

Group	Group Description	Sample Vendors
Direct Connections	Solutions that support direct, bilateral communications between trading partners over the Internet based on EDI/INT guidelines or Web Services protocols	IBM MQSeries, Microsoft MSMQ, Cyclone Commerce, IPNetSolutions, Syntrex
Electronic Trading Networks	Internet-based, managed network system designed to facilitate the exchange of B2B transactions between trading partners.	Internet Commerce Corp., eB2B Commerce, bTrade, CommerceQuest, Viacore, GE Global Exchange, Sterling Commerce, IBM, Peregrine
E-Marketplaces*	Solutions that provide many-to-many Internet-based connectivity in support of e-procurement and other more collaborative functions	Covisint, e2Open, Exostar, Omnexus, Transora, RetailersMarketXchange, GlobalNetworkExchange, WorldWideRetailExchange
Value-Added Networks	Traditional, managed network system designed to facilitate the exchange of EDI transactions	GE Global Services, EDS, Sterling Commerce, IBM, Peregrine

Source: Giga Information Group

* E-marketplaces are also included in the e-business applications category. They are included in this section due to their ability to support B2B communications that go beyond basic buying and selling transactions.

Table 8: B2B Integration Enabler Vendors

Group	Group Description	Sample Vendors
Electronic Data Interchange	Software that supports the transfer of internal application data to designated trading partners via standard business documents, such as purchase orders, invoices, electronic payments and vendor-managed inventory (VMI) transactions	GE Global Exchange, EDS, Sterling Commerce, IBM, Peregrine, SPS Commerce, Foresight, EC Outlook, ADX, QRS
Business Process Integration	Solutions to provide event-driven, real-time data exchanges between trading partners. Includes integrated EAI, workflow, trading partner management and e-channel communications functionality	IBM, webMethods, SeeBeyond, TIBCO, Vitria, Sybase/New Era of Networks, Healthcare.com, Iona, Microsoft, eXcelon, Metaserver, FileNet, Peregrine, Attunity, Fuegotech, Silverstream, BEA, NEON Systems, SAP, CommerceQuest

Source: Giga Information Group

Table 9: E-Business Applications

Group	Group Description	Sample Vendors
Enterprise Resource Planning	Integrated solutions to serve the needs of multiple departments	SAP, Oracle, PeopleSoft, Baan, J.D. Edwards
Customer Relationship Management (CRM)	Integrated solutions designed to support an organizations contacts with its base of customers	Siebel, Oracle, Nortel/Clarify, Remedy, PeopleSoft/Vantive, Trilogy, SAP, Epiphany
E-Procurement	Solutions to support online purchasing of indirect (and sometimes direct) materials.	Ariba, Commerce One/SAP, Oracle, i2, VerticalNet, FreeMarkets, Neuvis, Ventro
Supply Chain Management	Software that coordinates activities in the supply chain	i2, SAP, Oracle, PeopleSoft, Manugistics
Financial Management (FM)	Solutions for coordinating and integrating financial activities	SAP, PeopleSoft, Geac, Oracle, J.D. Edwards, Hyperion, Great Plains
Human Resources Management (HR)	Integrated solutions for coordinating various aspects of human resources activity	PeopleSoft, SAP, Oracle, Lawson

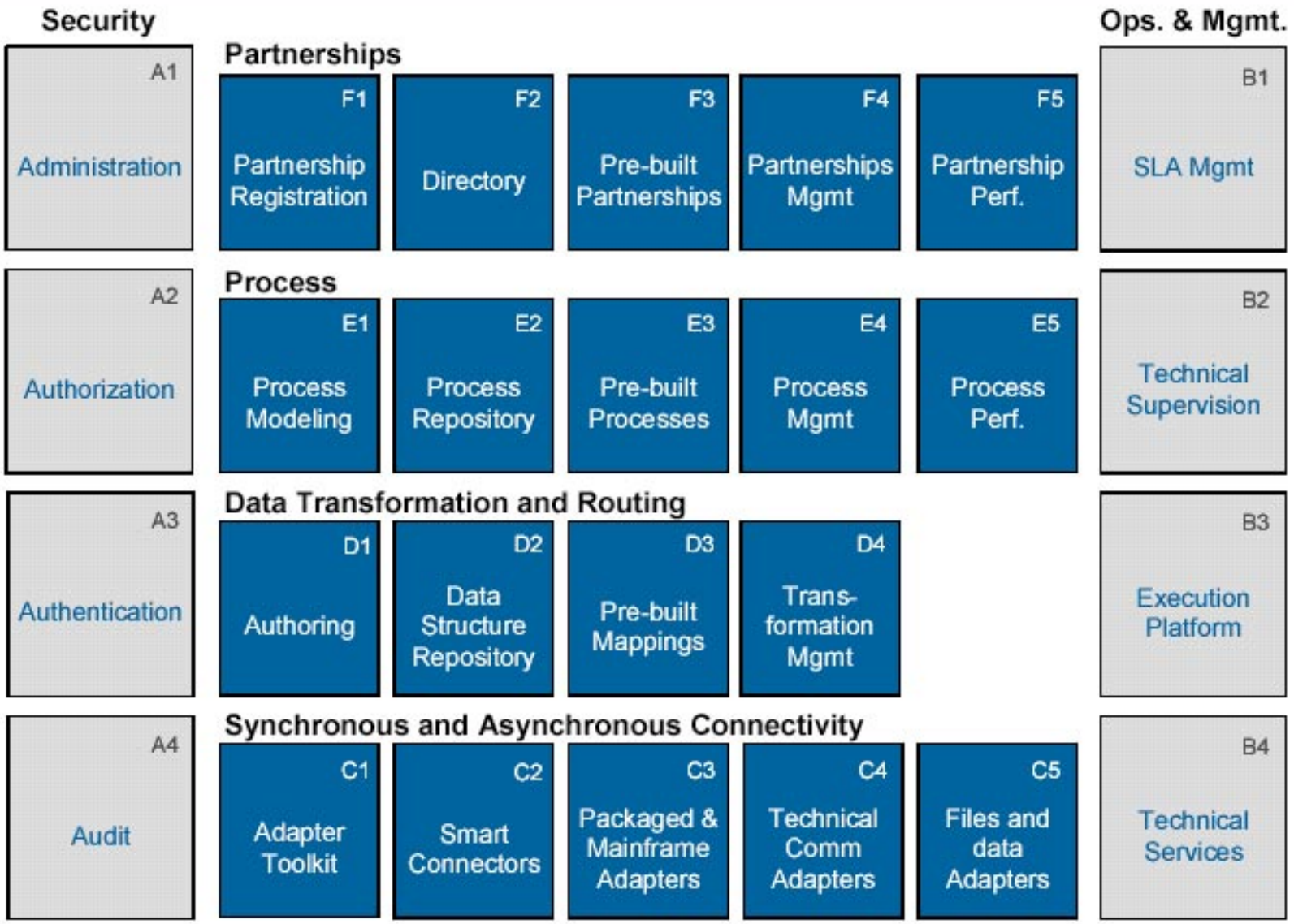
Source: Giga Information Group

Table 10: E-Collaboration Tools

Group	Sample Vendors
Groupware	IBM (Lotus Notes), Microsoft (Exchange)
Online Presentations	Placeware, WebX
Design Document Viewing and Development	Parametric Technology (PTC), UGS/SDRC, Dessault Systemes
Product Development	IDe
Parts and Service Manual Viewing	Enigma
Virtual Project Workspaces	NexPrise, e-Room
Collaborative Engineering	Agile Software
Electronic Design	Cadence, Synopsys, Menor Graphis, Avant!
Complex Project Management	Framework

Source: Giga Information Group

Application Integration Framework Model



Source: Giga Information Group

Figure 2

Integration Solution Trends

- Tool-Driven Solutions
  - Chaos: No community agreement
  - Non-integrated, point solutions
  - Semantics largely ignored
- Data-Driven Solutions
  - Mappers
    - With some semantics: IBM (Life Sciences), Vitria
    - Without: Microsoft, Oracle, ...
- Process-Driven Solutions (B2B)
  - Process integration: Microsoft, Oracle, Vitria
- Model-Driven Solutions
  - Vendor-centric Integrated Tools, Templates, and Architectures
  - Industry Templates and Architectures
  - Industry Ontologies: Vitria (e-Biz ontologies)
- Web Services

Siebel: Universal Application Network Architecture

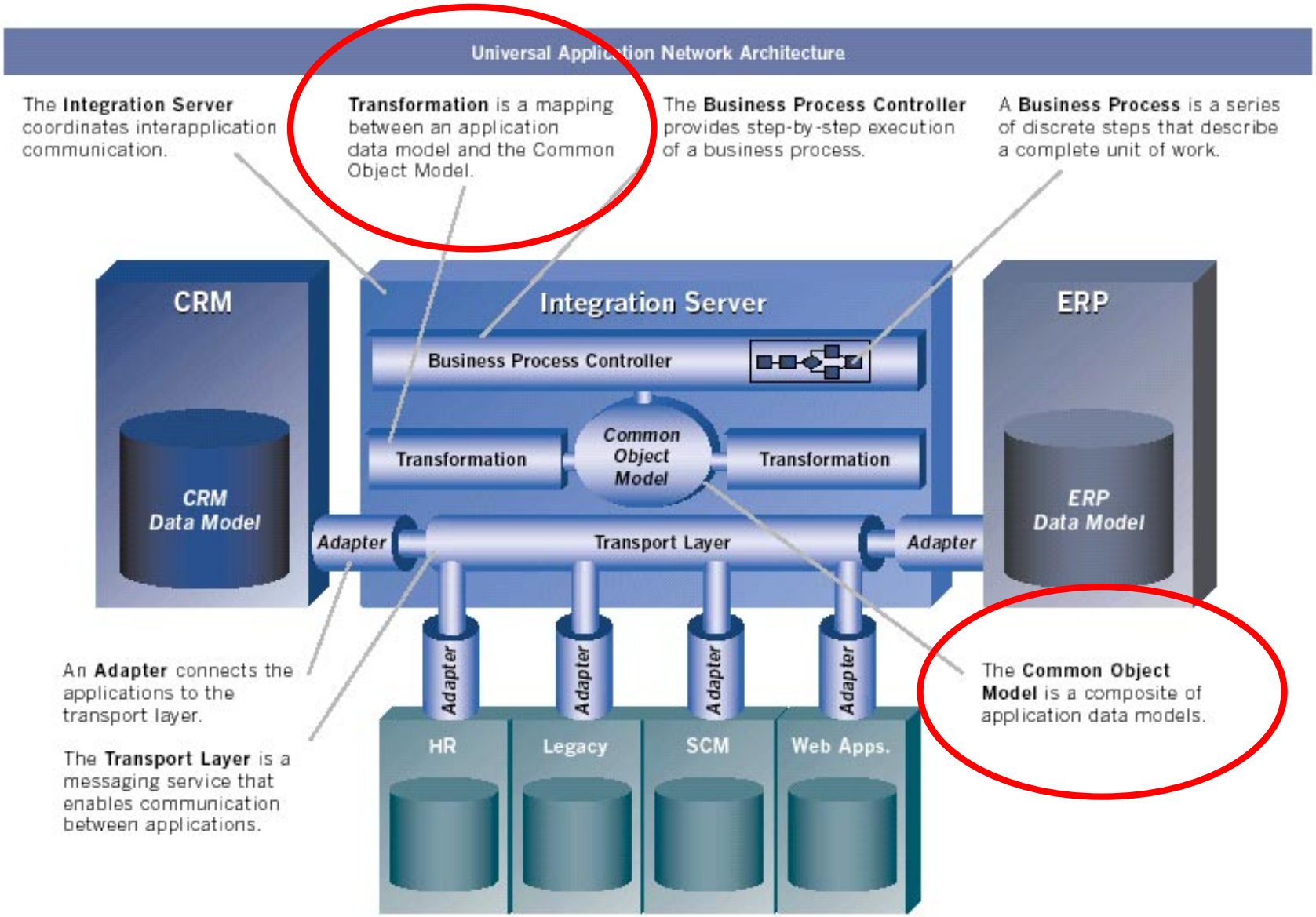
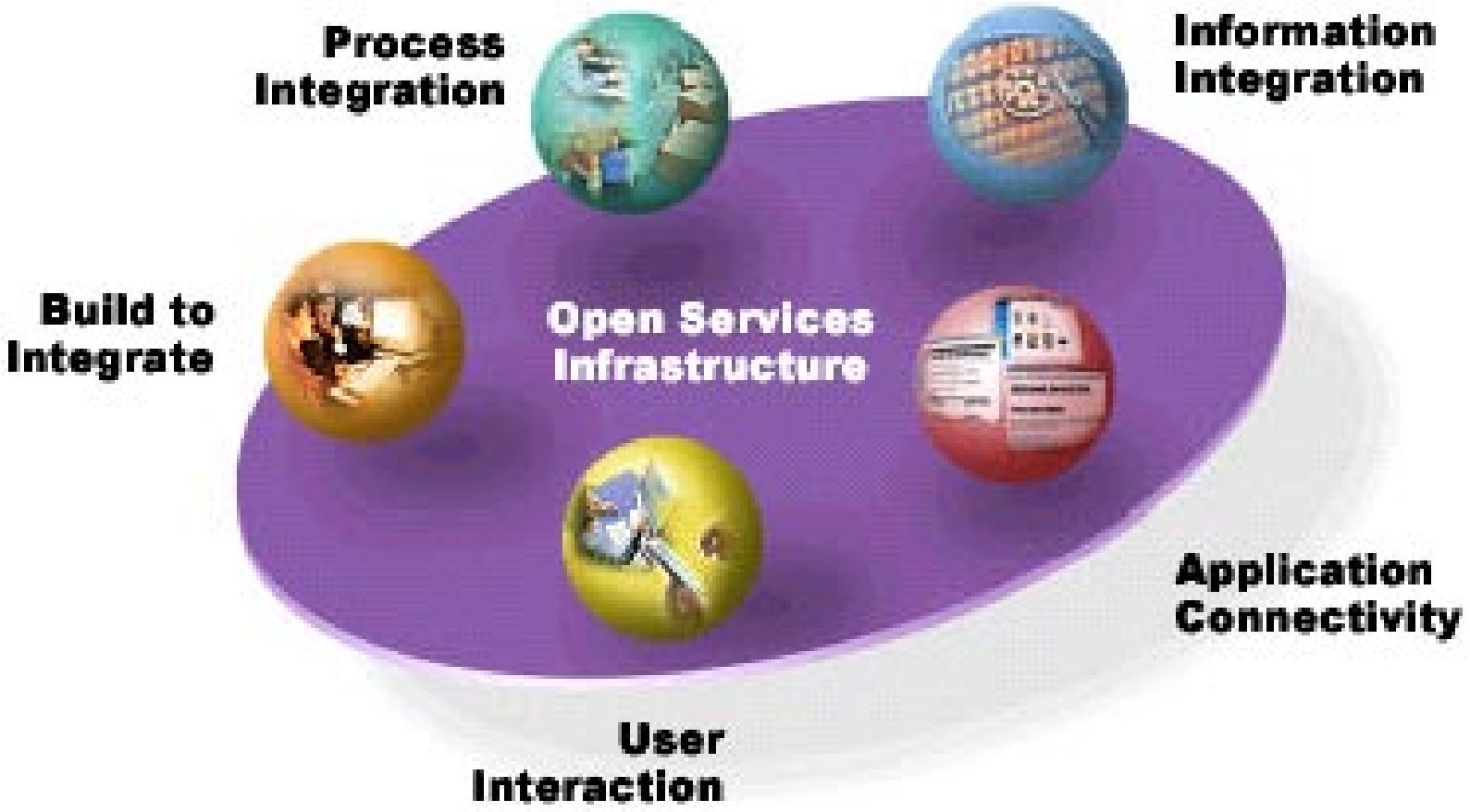


Figure 7: Universal Application Network enables prebuilt business processes to be deployed across a diverse set of applications.

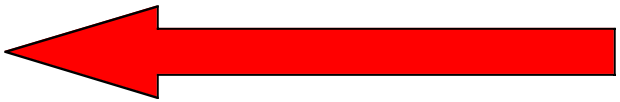
IBM Business Integration[1]



[1] Information Integration: At the Core of a Comprehensive Business Integration Infrastructure, IBM White Paper, May 2002

Outline

- Progress and Failure In Computer Science
- The Grand Challenge and The Illusion of Validity
- Roadblock to Current and Future Progress



- Why Attempt The Grand Challenge?
- Semantics: The Heart of The Grand Challenge
- Conclusions

Future Progress: X-Internet

X-Internet[1]

- Problem: the Web is
  - Dumb
  - Boring
  - Isolated

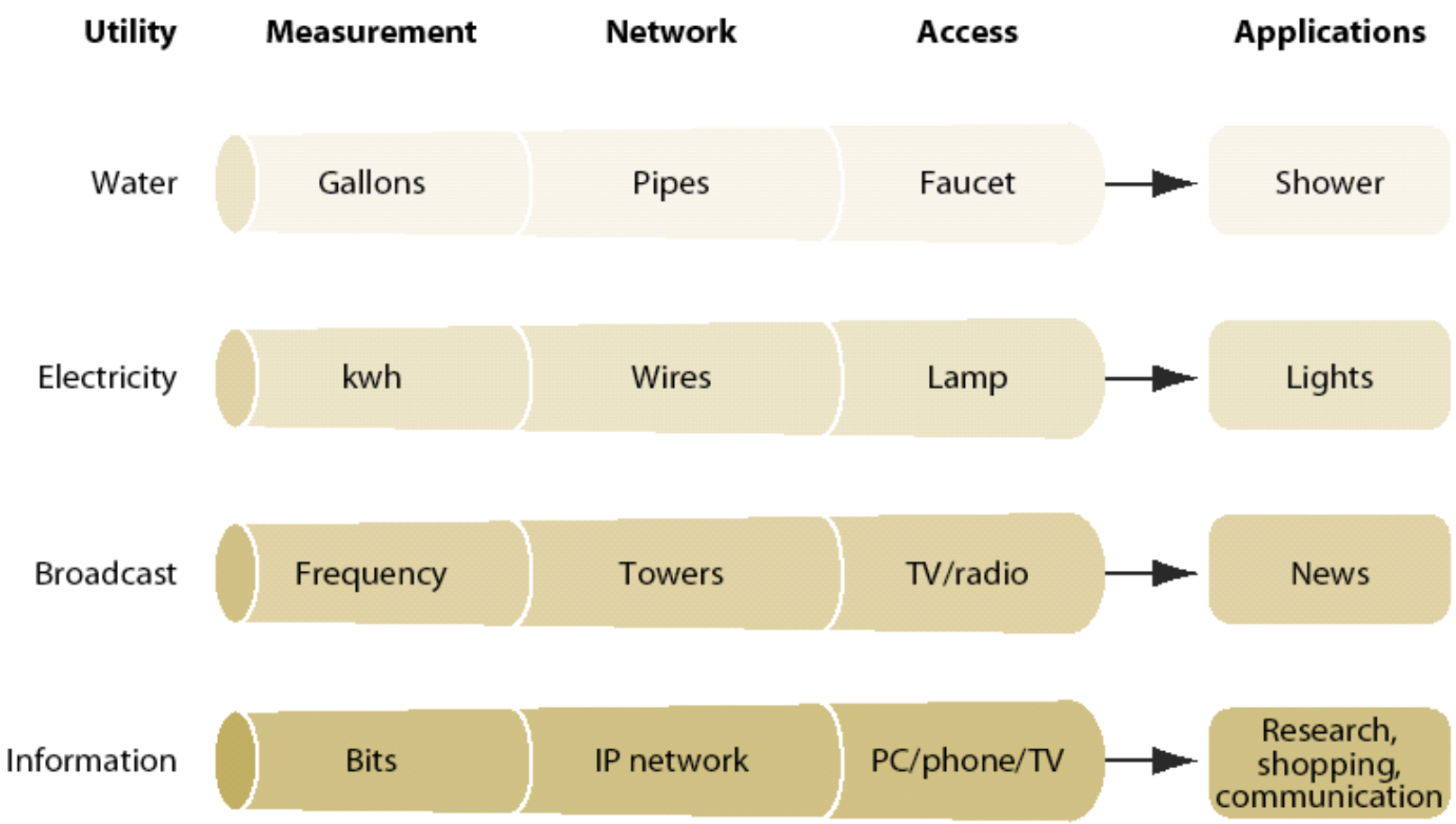
- Vision
  - Executable: *Intelligent applications that execute code near the user to create rich, engaging conversations via the Net*

  - Extended: *Internet devices and applications that sense, analyze, and control the real world*.

- Recognizing
  - XML does not address semantics
  - Industry agreement takes too long
  - 1-1 translation does not scale
  - Web Services help systems interact not understand
  - Centralized dictionaries have failed

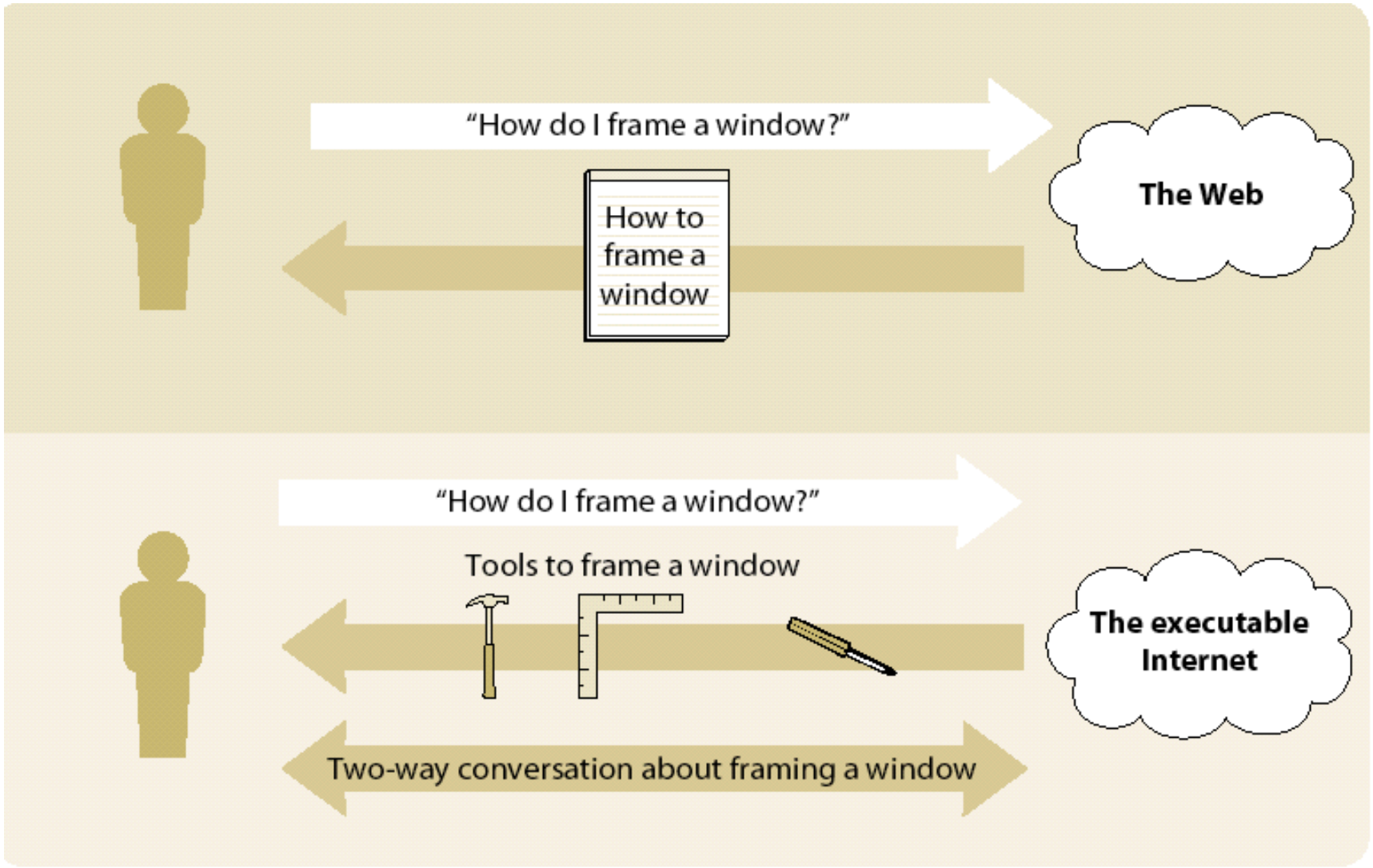
[1] The X-Internet, Forrester Research, May 2001

Information Exhibits the Characteristics of a Utility



Source: Forrester Research, Inc.

The Web Versus the Executable Internet



Source: Forrester Research, Inc.

Future Progress: Web Services[1]

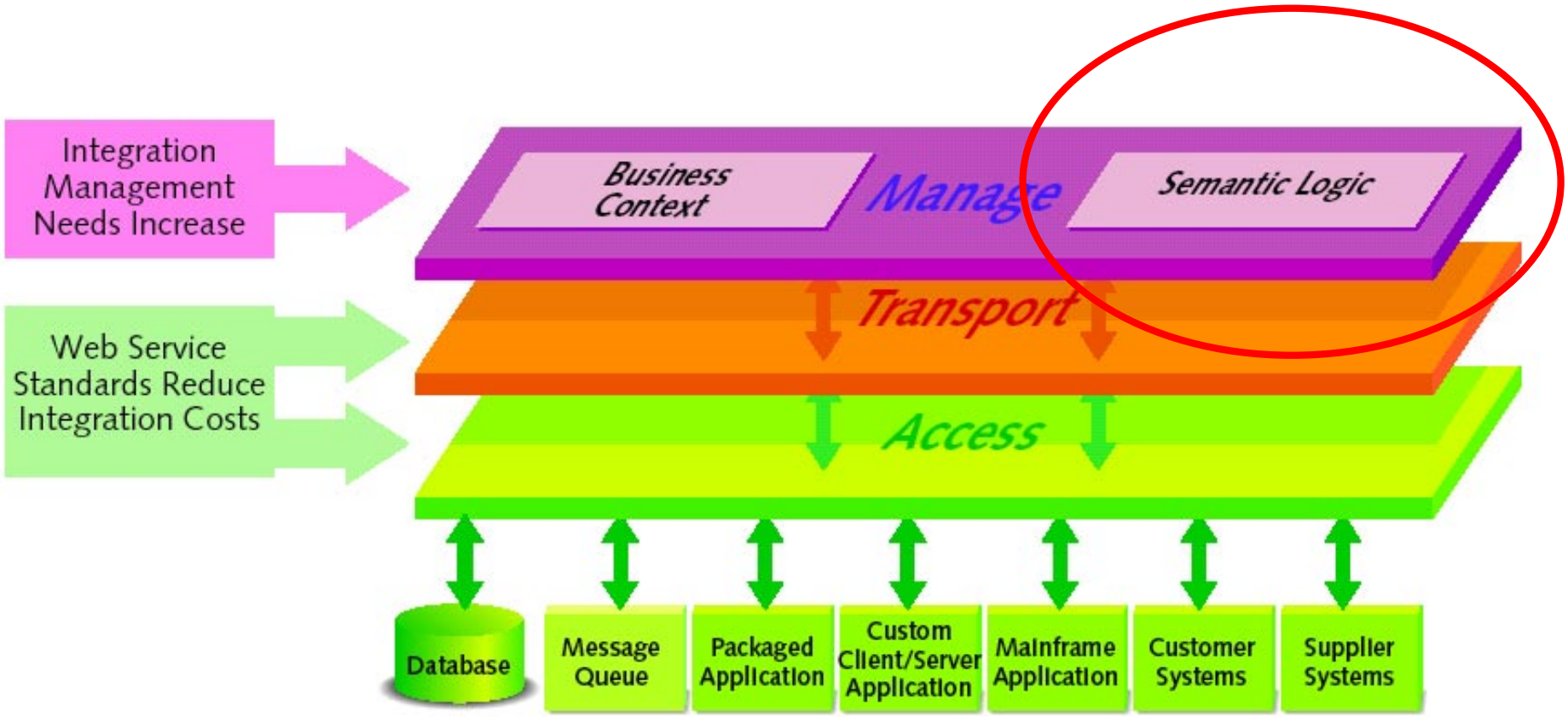
- Web Services[1]
  - Services: Any computing service you can build
  - Plumbing: Industry standard middleware for asynchronous, remote invocation of "services" distributed over the Web and everywhere else

- Using Web Services[2]
  - e-service: any code that you would like made visible to customers or applications
  - e-service description: attributes that characterize the service
  - e-service advertisement: publish service descriptions for discovery and access
  - e-service discovery and selection: discover and select a e-service (or combination of e-services) that fulfill specific requirements
  - e-service composition: combine basic e-services (possibly offered by different companies) can be combined to form value-added services.
  - e-service monitoring and analysis: to improve the service quality or efficiency

[1]2002 version of DCE, CORBA, COM+ PLUS asynchronous, loose coupling, vastly lower cost, industry agreement
[2]2002 version of: object-oriented programming - libraries, repositories, …

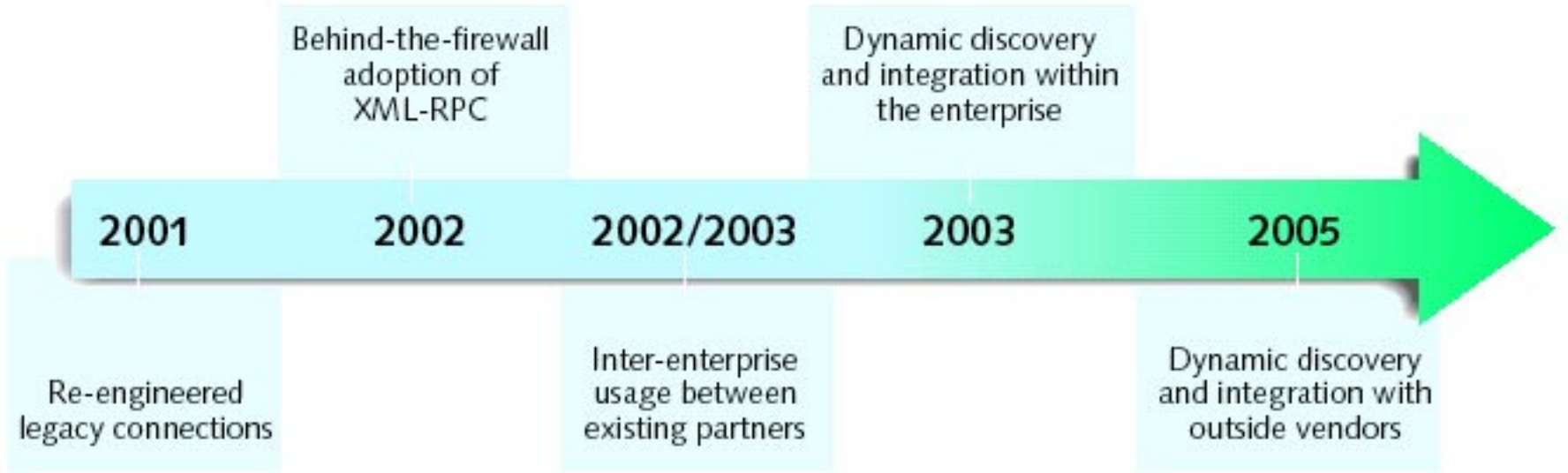
Web Services Integration

Semantics explicitly recognized



Source: the Yankee Group, 2002

Web Services Adoption Timeline



Source: the Yankee Group, 2002

Future Progress: Semantic Web

- Semantic Web = a machine-processable Web
  - Intelligent not Dumb
  - Engaging not Boring
  - Integrated and comprehensive not Isolated (no Deep Web)

- Intelligent applications collaborate to achieve goals with minimal human interaction

- Characteristics (partly based on[1])
  - Languages: express information in machine processable form
  - Search and discovery: to find the whole truth and nothing but the truth
  - Ontologically-integrated
    - Enhanced system and data interoperability: consistency based on semantics
    - Enhanced precision: queries and actions over the web
  - Supranet: billions of devices connected and integrated

- Compelling examples
  - Continuous tax preparation
  - Dynamically re-configuring, optimized supply chain

[1] The Semantic Web: Trying to Link the World, Gartner, August 2001

The Next Generation: Global Computing

- Every where - Ubiquitous
  - All devices (billions)
  - All locations (fixed, mobile)
  - Pervasive networks
- Every thing
  - All information resources (no Deep Web)
  - All services (applications)
- Everybody
  - Companies
  - Governments
  - Private citizens
  - Communities
- Examples: living your life on the “global computer”
  - Personal, continuous taxes
  - Optimized supply chain
- Integration
  - Transparent
  - Massive scale
  - It is just beginning

Outline

- Progress and Failure In Computer Science
- The Grand Challenge and The Illusion of Validity
- Roadblock to Current and Future Progress
- Why Attempt The Grand Challenge?
- Semantics: The Heart of The Grand Challenge
- Conclusions

Why Attempt the Grand Challenge?

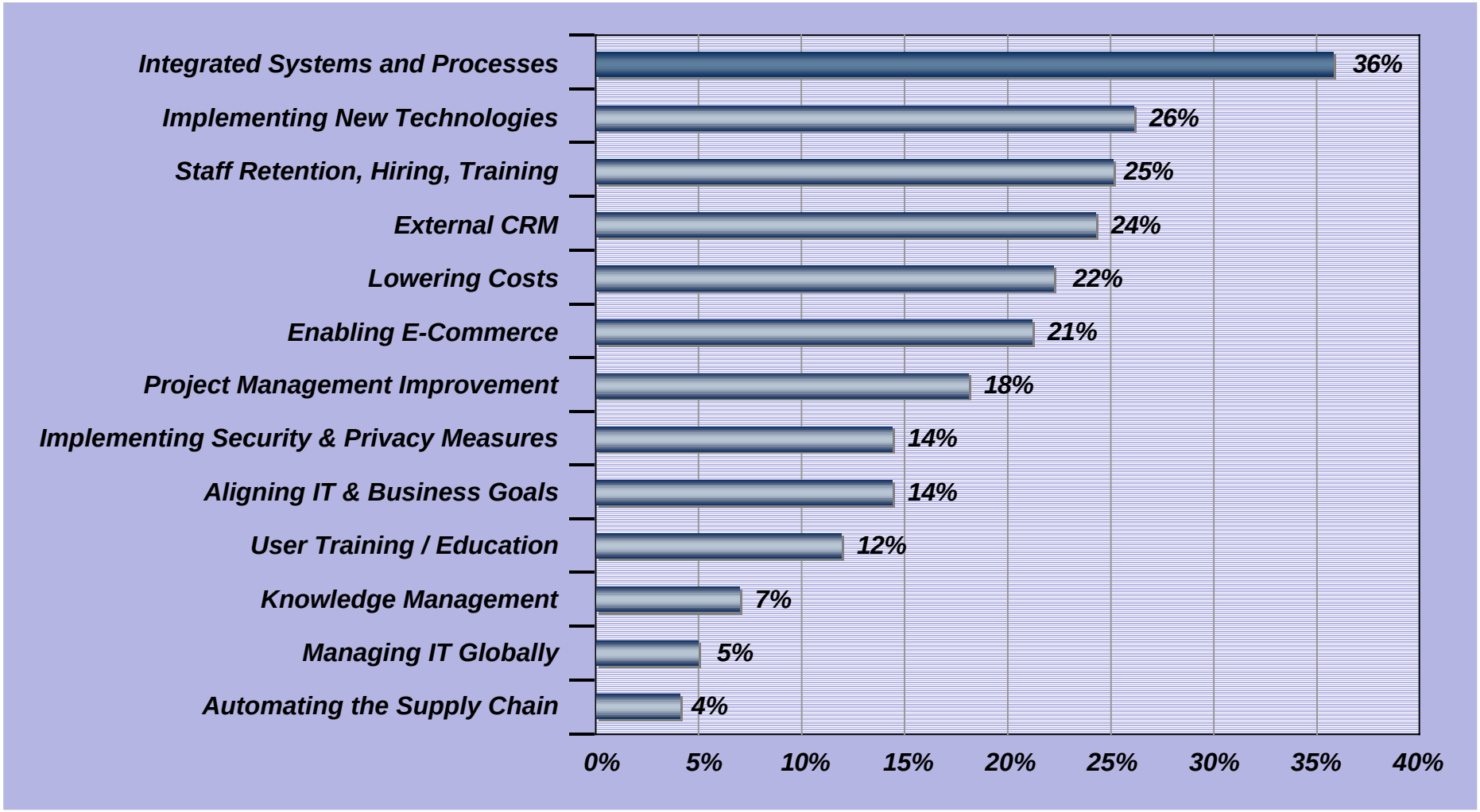
- Greater potential for
  - Precision
  - Automation
  - Optimization
  - Solutions - industrial problems
  - Visions

- Current solutions
  - May be imprecise or contain errors
  - Far too complex
  - Won't scale
    - Web-based integrated resources
    - More data to be generated in the next three years than in all of recorded history[1]

- Business need
  - CIO Priority
  - Economic Growth dependent on the Web working and scaling
  - Cost

[1] University of California, Berkeley P.Lyman, H.Varian, A. Dunn, A. Strygin, K. Swearingen, How Much Information? October 2000 [24 exabytes (2^{60} bytes)]

Top IT Spending Priorities[1]

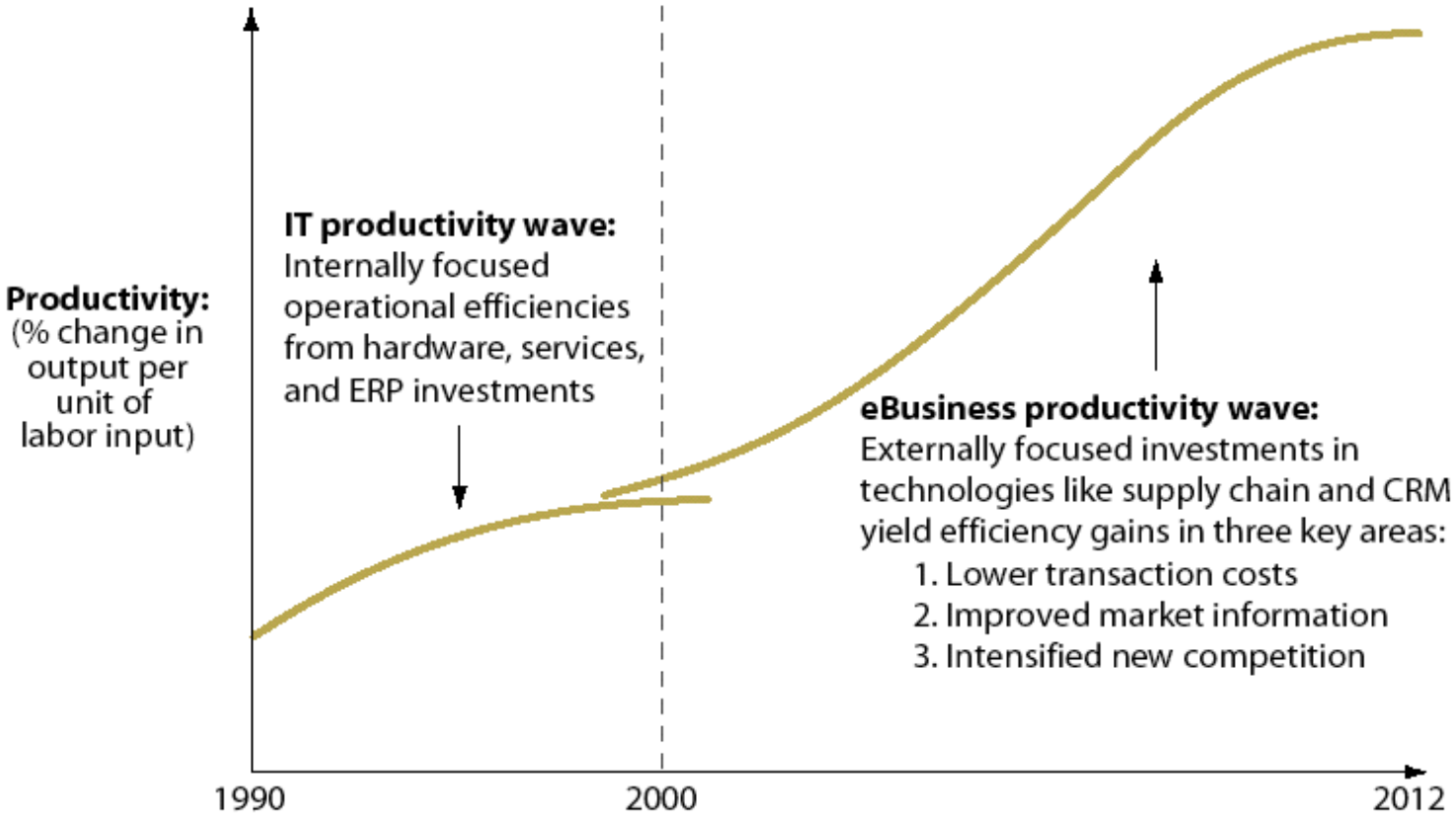


[1]CIO Magazine Survey, February 2002

33% of firms surveyed have EAI projects (Forrester, March 2002 Business Technographics benchmark)

Forrester eBusiness Productivity Model

"eBusiness will drive a new wave of productivity growth"



Source: Forrester Research, Inc.

Estimating The Grand Challenge Cost

- Integration's costs
  - 24% of IT budgets $180 B / year US (InfoWorld, January 2002 survey of 500 IT leaders)
  - 13% of IT spend $100 B of $752 B / year US (Giga estimate based on May 2002 report)
  - 25-40% of all IT projects (various)
  - 6% of US IT spending: $34 B of $610B / year US (IDC, May 2002)
  - 7% of IT spending: $90 B of $1.3T / year worldwide (IDC, May 2002)
  - 28+% of all consulting: $ 160 B / year worldwide (Gartner March 2002)
  - 43% of e-business consulting: $53 B / year worldwide (Gartner)
  - 1.75% to annual IT budget on EAI and B2Bi (Forrester,Dec 2001)
  - 10-30% of IT budgets (David Sink, IBM quoted in InformationWeek, May 27, 2002)

- Data Quality's costs
  - $600 B / year US (Data Warehouse Institute, 2002)

- Annual Integration + Data Quality Costs
  - Worldwide: order $1 Trillion / year

The Grand Challenge is now "mission critical".

Outline

- Progress and Failure In Computer Science
- The Grand Challenge and The Illusion of Validity
- Roadblock to Current and Future Progress
- Why Attempt The Grand Challenge?
- Semantics: The Heart of The Grand Challenge

- Conclusions

Recent Semantics Research

- Basic computer science
  - Theory
  - AI
  - Software Engineering
  - Programming Languages
- Database Area
  - 80's
    - Modelling, data models, query optimization, distributed databases, ...
    - Database theory, datalog, ...
  - Late 90's resurgence
    - Schema integration, mapping, equivalence
    - Query answering, equivalence, expressive power of query languages
    - View-based query answering
    - Web
      - Modelling
      - Querying
      - Information extraction and integration
      - Web site construction and restructuring
- Semantics largely avoided

Recent Semantics Research

- Information Systems + many communities
  - Mediators
  - Ontologies, terminologies, thesauri, vocabularies, ...

- Semantic Web Community
  - Ontologies
  - Upper ontologies
  - XML variants: ebXML, ...
  - Mediators
  - Agents

Semantics is Harder Than You Thought

- Understand
  - Formal aspects of semantics
  - Where things might go wrong
  - The role that semantics plays in your problem
- Model the semantic problem
- Model a solution
- Analyze the solution
  - Soundness
  - Completeness
  - Complexity
  - Semantics preserving (lossless)
- Reduce intractable or costly solutions to tractable efficient solutions

So …

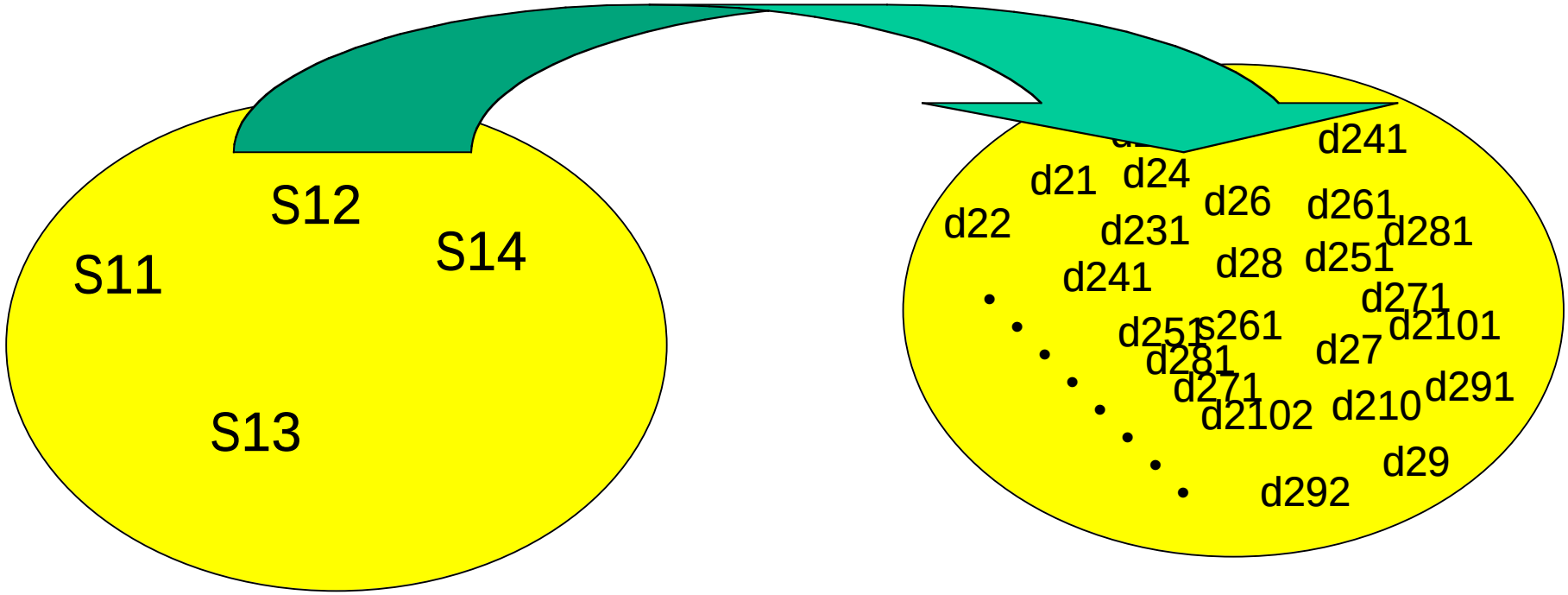
What are

- Semantics
- Ontologies

How do they help with

The Grand Challenge - Enhancing Information Systems to better represent real world facts and actions

Mathematical Semantics

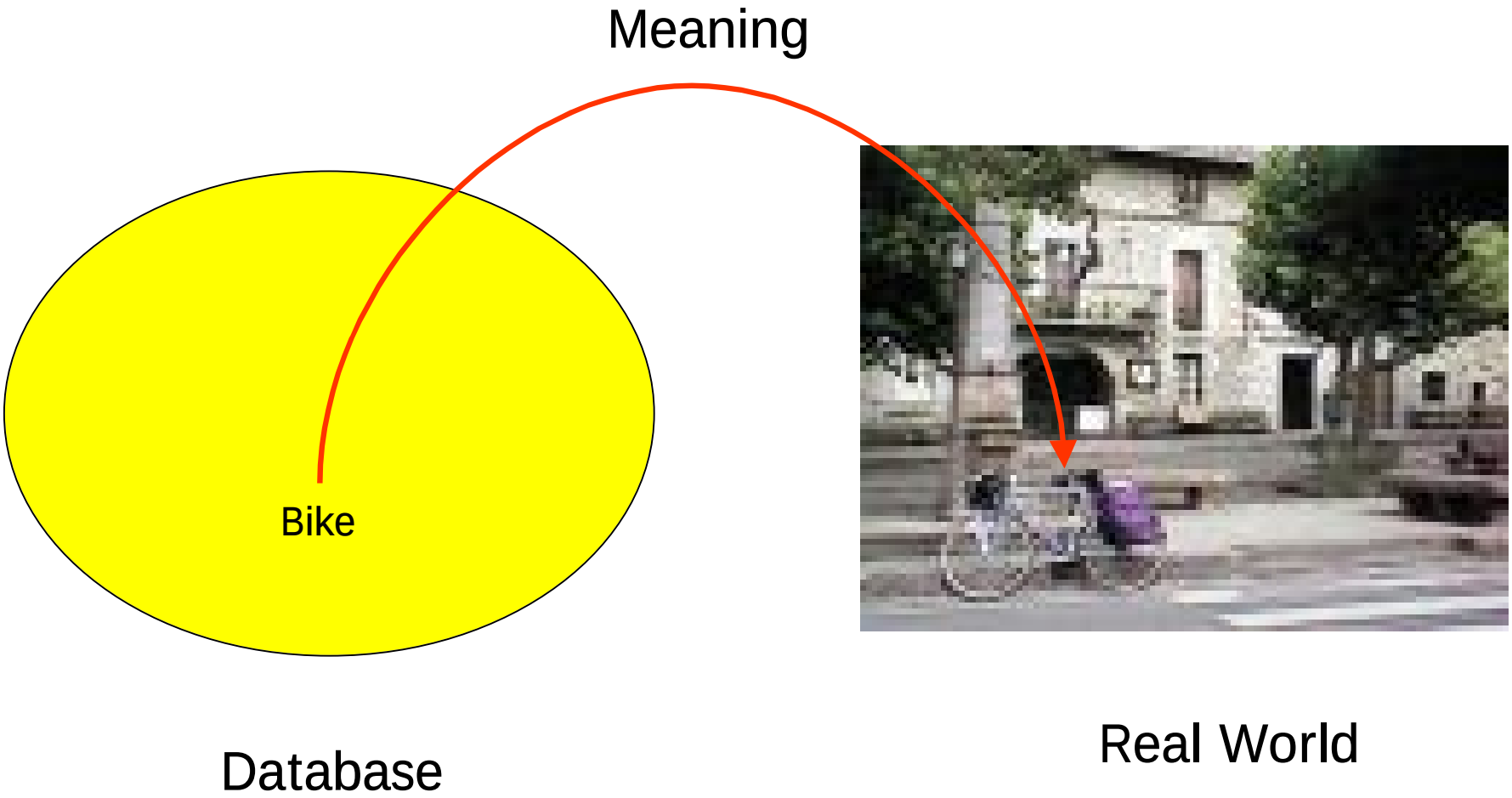


Mapping / interpretation / model for a (first order) language L

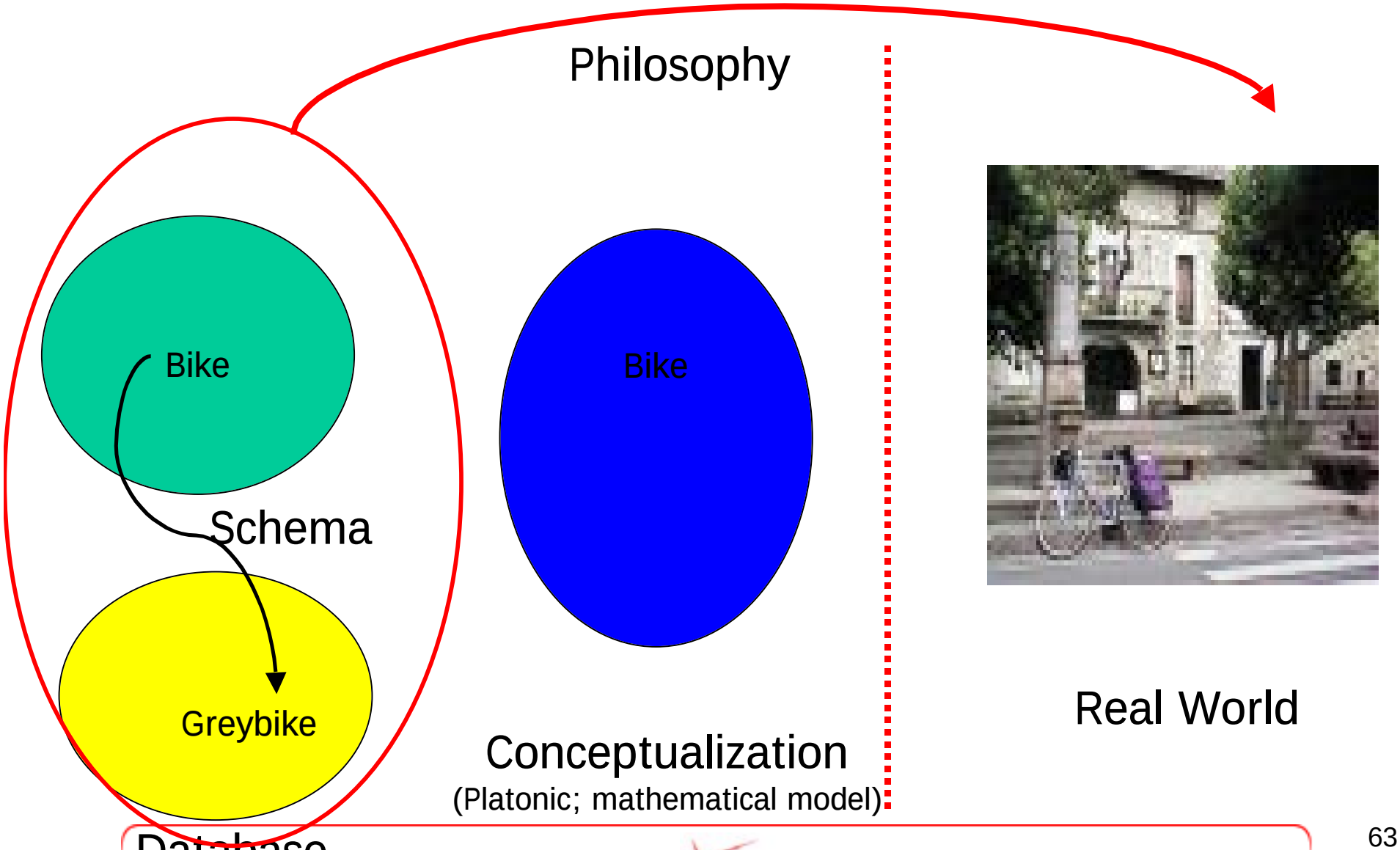
- Symbol set
- Relations, functions, and constants over Symbols
- Domain of interpretation
- Assignment of functions from symbols to domain elements

Ref: Elliot Mendelson, Introduction to Mathematical Logic, 4th Edition, Chapman & Hall, 1997

Database Semantics

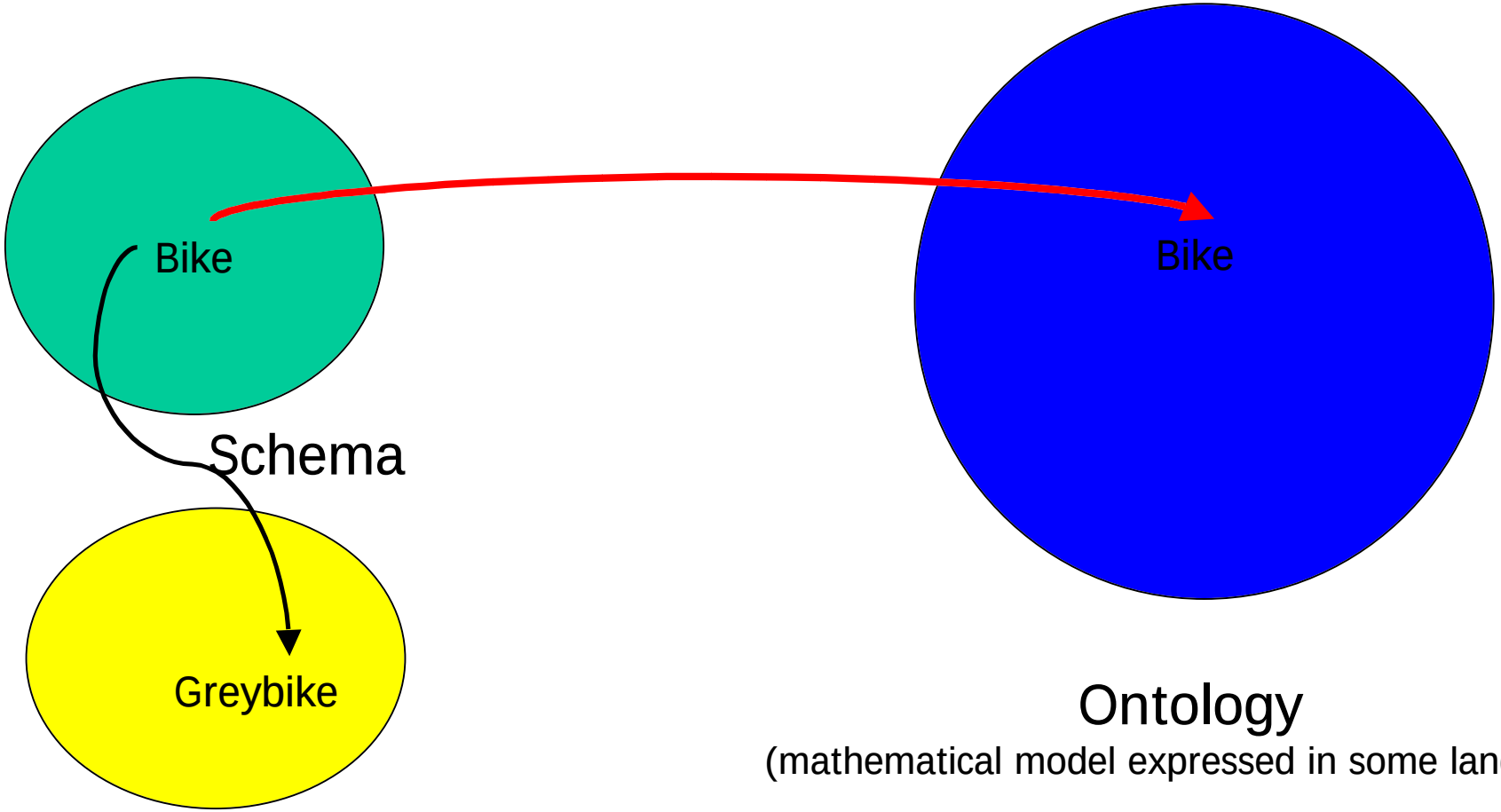


Formalizing Semantics (Tarski)



Denotational Semantics (Tarski)

Intensional mapping



Database

Ontology

(mathematical model expressed in some language)

Declarative Semantics (Tarski)

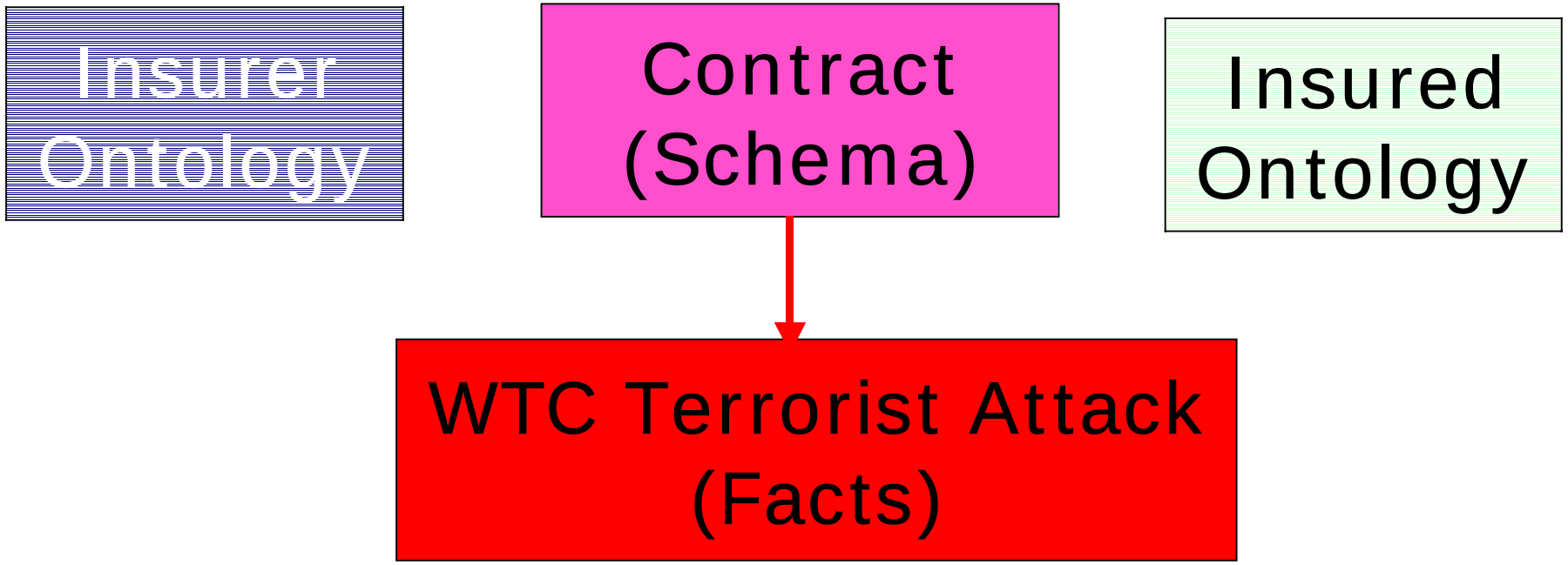
- Meaning = (mathematical) mapping of a representation (e.g. description in first order language) to an *agreed conceptualization* of the "real world"

- Meaning, in practice, cannot be absolute:
  - Requires agreement among all involved *cognitive agents*
  - About everything, in past, present and future for a particular application
    - on all *observations*, *facts, events, …*
    - on all *rules* in vigor
    - believed/enforced by large communities…

- Example
  - Was the World Trade Center terrorist attack one ($3.6B) or two ($7.2B) incidents?

Source: Prof Robert Meersman

Community Agreement Required



- Isolated schemas have
  - No semantics
  - Implicit ontology
- An ontology defines a semantic agreement

What is *an* Ontology?

- Poor definition:

  "Specification of a conceptualization" [Gruber, 1993]

- Better:

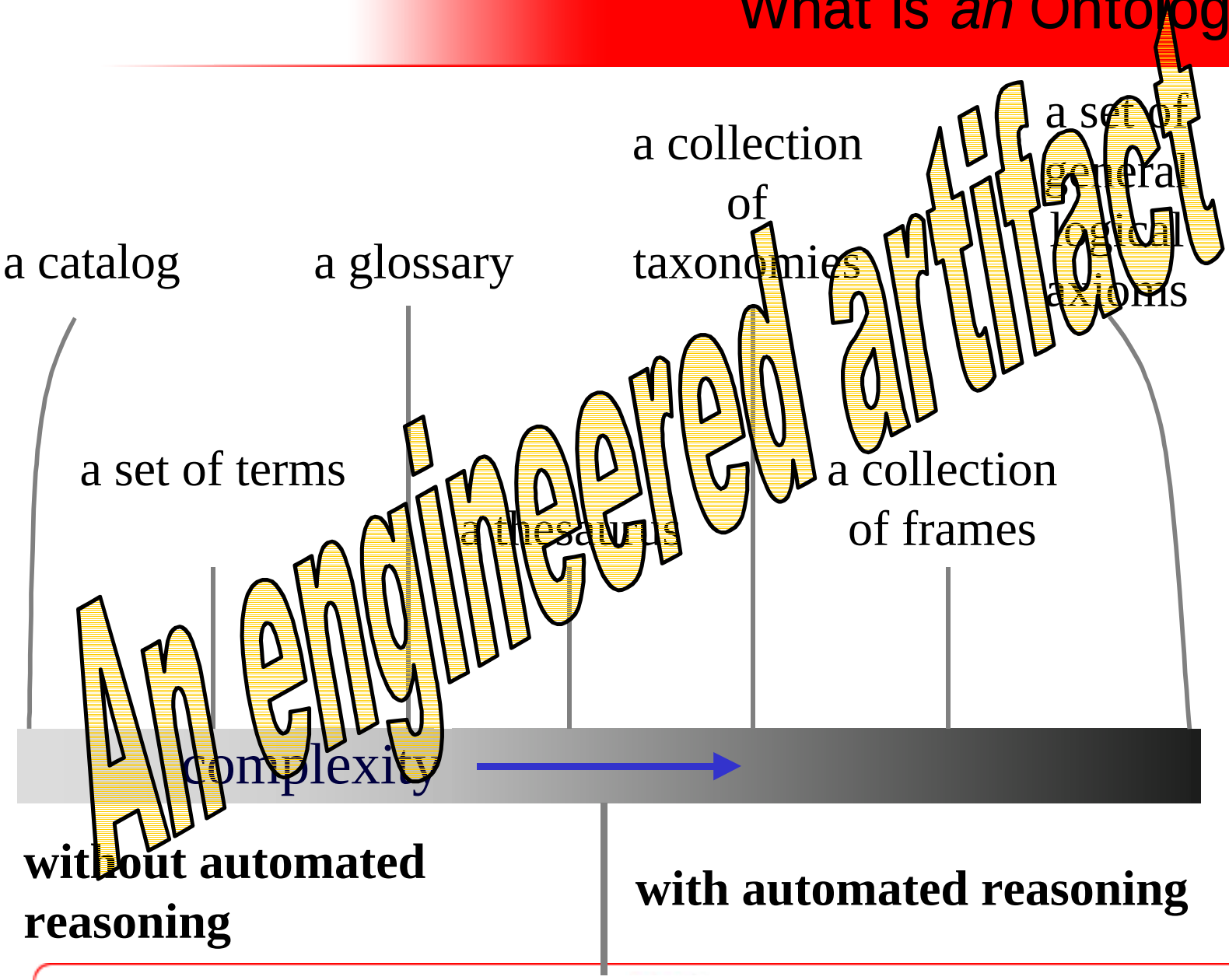
  "Description of the kinds of entities there are and how they are related."

- Good ontologies should provide:

  - **Meaning**
  - Organization
  - Taxonomy
  - Agreement
  - Common Understanding
  - Vocabulary
  - Connection to the "real world"

Source: Chris Welty, IBM Watson Research Center

What is *an* Ontology?



Source: Chris Welty, IBM Watson Research Center

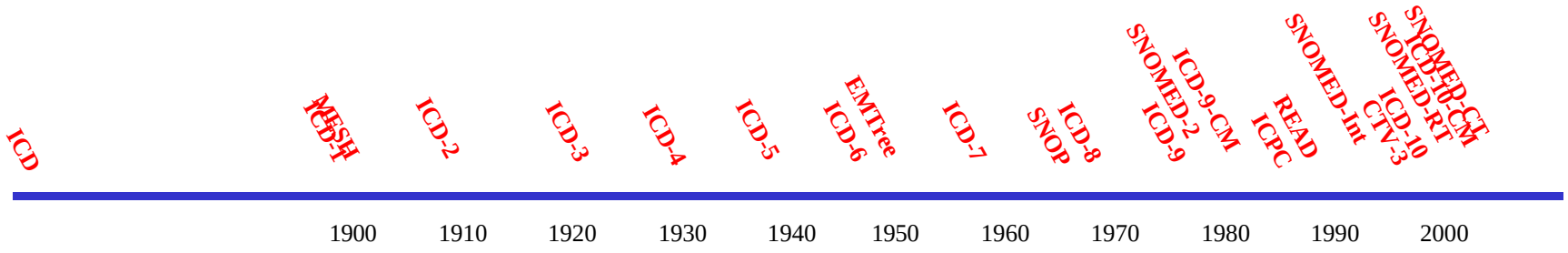
Ontologies go on ...

SNOMED-3
READ-2
MeSH
ICD-9
ICD-9-CM
ICD-O
NCSP
ICPM
OXMIS

CDAM
NGAP
ICPC
OPCS-4
CPT-4

NDC
NANDA
ICNP

ECRI-UMDNS
SNOP
HCFA
ACR-NEMA
IUPAC-NPU
LOINC
DICOM-SDM
MCTGE



THE UNIVERSITY of MANCHESTER

... and on ...

AIDSLINE
MED80
MED66
AIDSDRUGS
AIDSTRIALS
ChemID
CHEMLINE
GENE-TOX
HISTLINE
SDLINE
TOXLINE
TOXLINE65
TOXLIT
PDQ

AVLINE
BIOETHICS
CANCERLIT
CATLINE
DENTALPROJ
MEDLINE
POPLINE
SERLINE
DOCUSER
Dxplain
AI/RHEUM
Iliad
GenBank
OMS
PSY

TRIFACTS
NIOSH
NPIRS
NEDRES
MED85
MED75
HSTAT
HDA
MED90
HealthSTAR
ACR92
AIR93
BRMP96
NIC
ULT

BRMS96
COSTAR
CPM
CRISP
COSTART
DMD
DSM III & IV
DOR
HHC
INS
LCH
MCM
MIM
Neuronames
WHOART

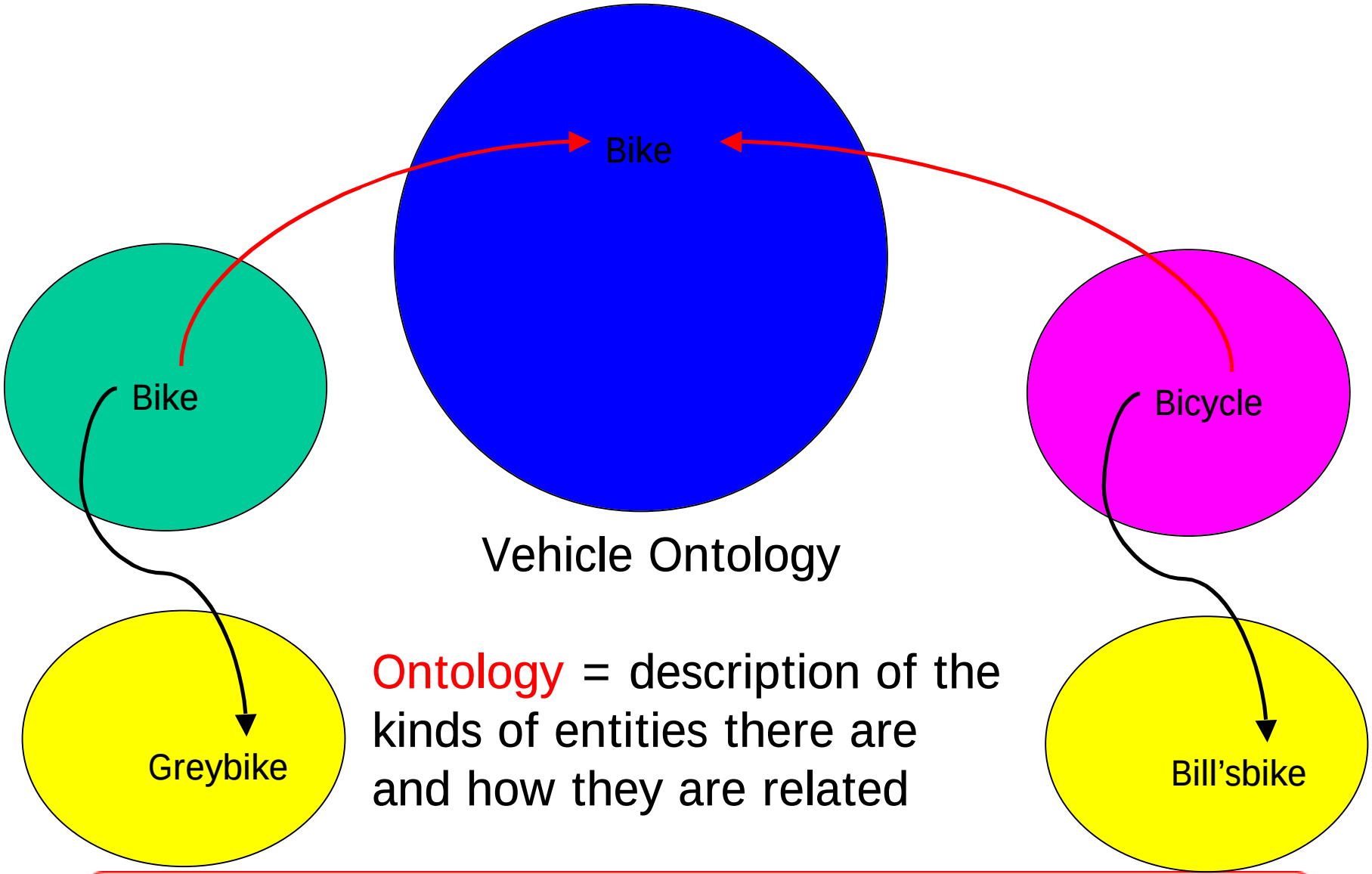
THE UNIVERSITY of MANCHESTER

...and on...

CCHI (Canada)
MBS-E (Australia)
ICD-10-PCS (USA)
WCC5 (Netherlands)
NCSP (Swedish Version)
NCSP (Finnish Version)
ICPM-DE (Germany)

CCAM (France)
READ 3.1 (UK)
SNOMED-RT (USA)
OPCS-5 (UK)
SKS (Denmark)
ICIDH (WHO)
Digital Anatomist (UW)
Nomina Anatomica

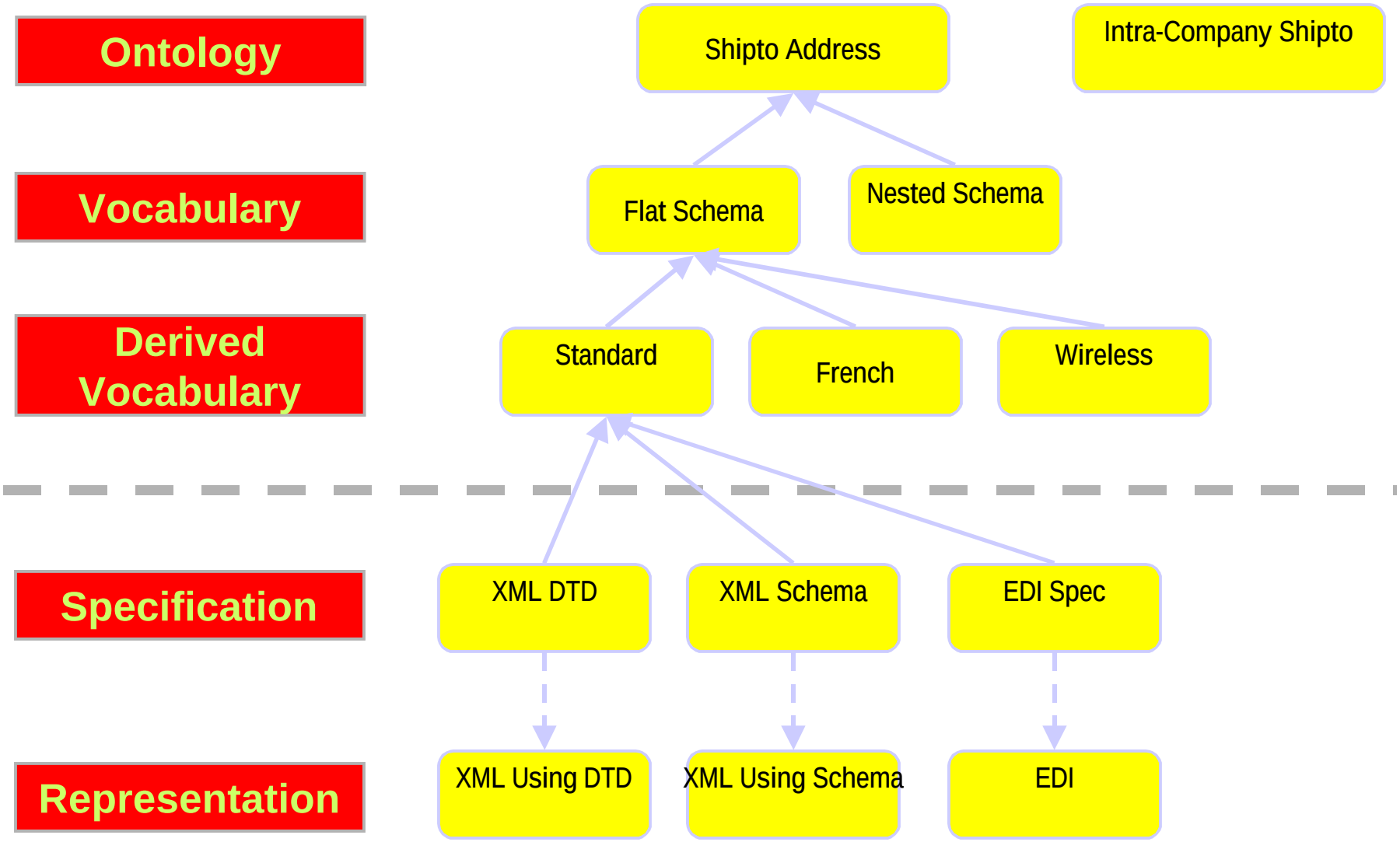
Schema Mapping



Ontology = description of the kinds of entities there are and how they are related

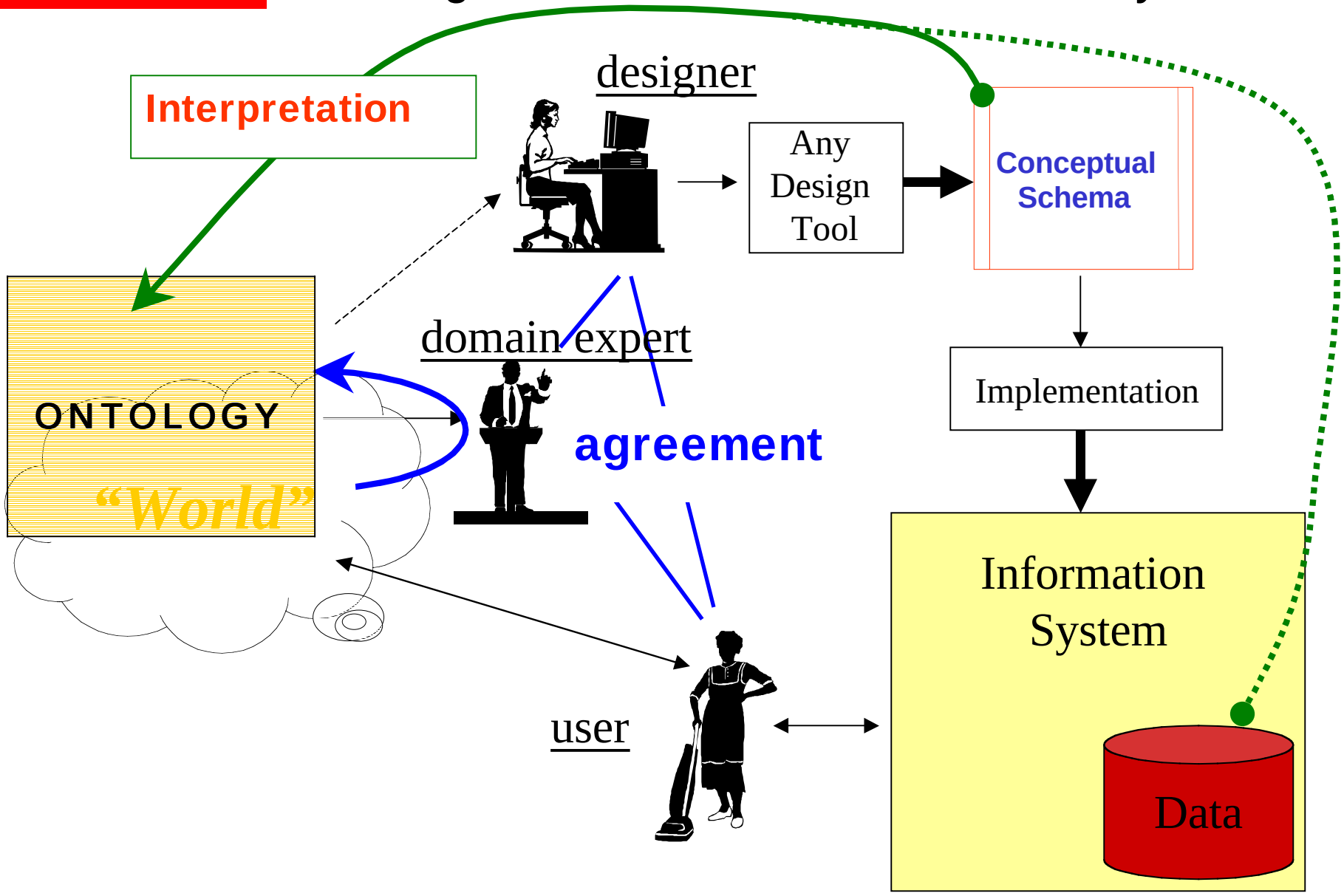
VITRIA

“Ship-to Address” Example



Source:
Prof Robert Meersman

Defining Semantics for an Information System



Analyzing A Problem

Determine the limitations and cost of solutions

- Identify the role of semantics in the your problem, e.g., database integration
- Model the semantic problem / solution, e.g., schema mapping, database mapping
- Model the solution formally, e.g. mapping equations
  - If undecidable: stop or **restrict or reduce** the solution until you find a decidable solution
  - If decidable, analyze for various properties
    - Soundness: under what condition will you get “nothing but the truth”?
    - Completeness: under what conditions will you get the whole truth?
    - Complexity
      - Exponential
      - Non-deterministic polynomial (NP) / coNP: the complement of the algorithm is NP
      - Polynomial (P)
      - NP-complete
    - Look for
      - Worst case
      - Average case
      - Best case

Schema/Database Equivalence[1]

- Problem: Determine if the two schemas are equivalent and under what conditions could data in one schema be restructured in a lossless way and represented in another schema?

- Analysis
  - **Solution**: equivalence modeled as equations is undecidable if you look at all combinations of maps between schemas
  - **Reduced solution**: a transformation language and operators reduced the number of matches but still NP-complete, in general
  - **Restricted**: for schemas with bounded out-degree it is quadratic

- Result
  - Efficient, lossless algorithm for schema equivalence and database mapping

[1]Renée J. Miller, Yannis E. Ioannidis, Raghu Ramakrishnan: Schema Equivalence in Heterogeneous Systems: Bridging Theory and Practice

View-Based Query Answering[1]

- **Problem**: Answer a query Q (expressed in query language QL1) over a global schema give a mapping between the global and local schema with
  - Sources characterized in terms of a query over the global schema expressed in a query language QL2
  - Sources are either
    - SOUND: all the data in the sources satisfy the corresponding query in the mapping, but there may be other data that satisfy such query
    - EXACT: all and only the data in the sources satisfy the corresponding query in the mapping

- **Result: Query answering**
  - In traditional databases is very efficient (PTIME complexity)
  - In data integration it is fundamentally different
    - Few cases are PTIME
    - Most are coNP and undecidable
  - Conditions under which you get the "whole truth" and "nothing but the truth"

[1] Serge Abiteboul, Oliver M. Duschka: Complexity of Answering Queries Using Materialized Views. PODS 1998
Data Integration: A Theoretical Perspective, Maurizio Lenzerini, PODS 2002

Complexity of View-Based Query Answering

Sound	CQ	$CQ^{\neq}$	PQ	Datalog	FOL
CQ	*PTIME*	*coNP*	*PTIME*	*PTIME*	*undec.*
$CQ^{\neq}$	*PTIME*	*coNP*	*PTIME*	*PTIME*	*undec.*
PQ	*coNP*	*coNP*	*coNP*	*coNP*	*undec.*
Datalog	*coNP*	*undec.*	*coNP*	*undec.*	*undec.*
FOL	*undec.*	*undec.*	*undec.*	*undec.*	*undec.*
Exact	**CQ**	**$CQ^{\neq}$**	**PQ**	**Datalog**	**FOL**
CQ	*coNP*	*coNP*	*coNP*	*coNP*	*undec.*
$CQ^{\neq}$	*coNP*	*coNP*	*coNP*	*coNP*	*undec.*
PQ	*coNP*	*coNP*	*coNP*	*coNP*	*undec.*
Datalog	*undec.*	*undec.*	*undec.*	*undec.*	*undec.*
FOL	*undec.*	*undec.*	*undec.*	*undec.*	*undec.*

Data Integration: A Theoretical Perspective, Maurizio Lenzerini, PODS 2002

Web Queries[1]

- **Problem**: Complexity analysis of certain types of web-based queries
- **Result:** hierarchy of web-based query types from
  - Efficient: since they access a bounded set of web pages
  - Infeasible: since they have to access every page on the Web
- **Utility**
  - Some queries that have efficient solutions on a bounded set of Web pages can become undecidable when those bounds are changed

[1] Serge Abiteboul, Alberto Mendelson, Tova Milo, and others

Outline

- Progress and Failure In Computer Science
- The Grand Challenge and The Illusion of Validity
- Roadblock to Current and Future Progress
- Why Attempt The Grand Challenge?
- Semantics: The Heart of The Grand Challenge
- Conclusions

Conclusions

- The Grand Challenge of IT
  - Enhancing Information Systems to better represent real world facts and actions
- The Illusion of Validity
  - Visions and solutions are frequently offered and believed without a principled, robust treatment of semantics, and consequently fail
- The Grand Challenge has become mission critical
  - Current solutions
    - May be imprecise or contain errors
    - Far too complex
    - Won't scale
  - Business need
    - Economic Growth dependent on the Web working and scaling
    - Cost: $ 1 Trillion / year

Industrial Community Recommendation

- Understand the Grand Challenge
  - Role in business requirements
  - Role in tools, technologies, ...
  - Limitations and risks
    - Current systems and technologies
    - Visions and plans
    - Methods / algorithms and their properties
    - Limitations of automation
- Work to address the Grand Challenge
  - Community agreements
  - Ensure soundness of systems, products, tools, and methods
  - Working with the research community

Research Community Recommendation

Focus on the Grand Challenge

- Theory of semantics
  - Methods / algorithms and their properties
  - Limitations of automation
- Systems: develop semantically-aware
  - Tools and techniques
  - Languages
  - Architectures
  - DBMSs
  - Search
- Semantics: Domain-specific
- Make your work
  - Relevant - realistic
    - Perfect match for academics
    - Watch out for old KR guys
  - Understandable

Final Thought

... Imagine how we could change the world with the power of computing, considering what Mother Teresa did with her resources