Knowledge Management: Overview and Related Fields

Hsinchun Chen, Ph.D.

McClelland Professor,

Director, Artificial Intelligence Lab and Hoffman eCommerce Lab

The University of Arizona

CEO, Knowledge Computing Corporation

Acknowledgement: NSF DLI2, NIH/NLM, NCSA

美國亞歷桑那大學，陳炘鈞 博士

- My Background: (A Mixed Bag!)
  - BS NCTU Management Science, 1981
  - MBA SUNY Buffalo Finance, MS, MIS
  - Ph.D. NYU Information System, Minor: CS
  - Dissertation: “An AI Approach to the Design Of Online Information Retrieval Systems” (GEAC Online Cataloging System)
  - Associate/ Assistant/Full/Chair Professor, University of Arizona, MIS Department

- My Background: (A Mixed Bag!)
  - Founder, Artificial Intelligent Lab, 1990
  - Founder, Hoffman eCommerce Lab, 2000
  - PIs: NSF CISE DLI-1 DLI-2, DARPA, NIJ, NIH
  - Editors: JASIST, DSS
  - Conference Chairs: ADL1, ADL2, ADL3, ADL4
  - Industry Consulting: HP, IBM, AT&T, SGI, Microsoft, SAP
  - Founder, Knowledge Computing Corporation, 2000

Knowledge Management: Overview

Knowledge Management Overview

- What is Knowledge Management
- Data, Information, and Knowledge
- Why Knowledge Management?
- Knowledge Management Processes

Unit of Analysis

- **Data**: 1980s
  - Factual
  - Structured, numeric → Oracle, Sybase, DB2
- **Information**: 1990s
  - Factual
  - Unstructured, textual → Yahoo!, Excalibur, Verity, Documentum
- **Knowledge**: 2000s
  - Inferential, sensemaking, decision making
  - Multimedia → **???**

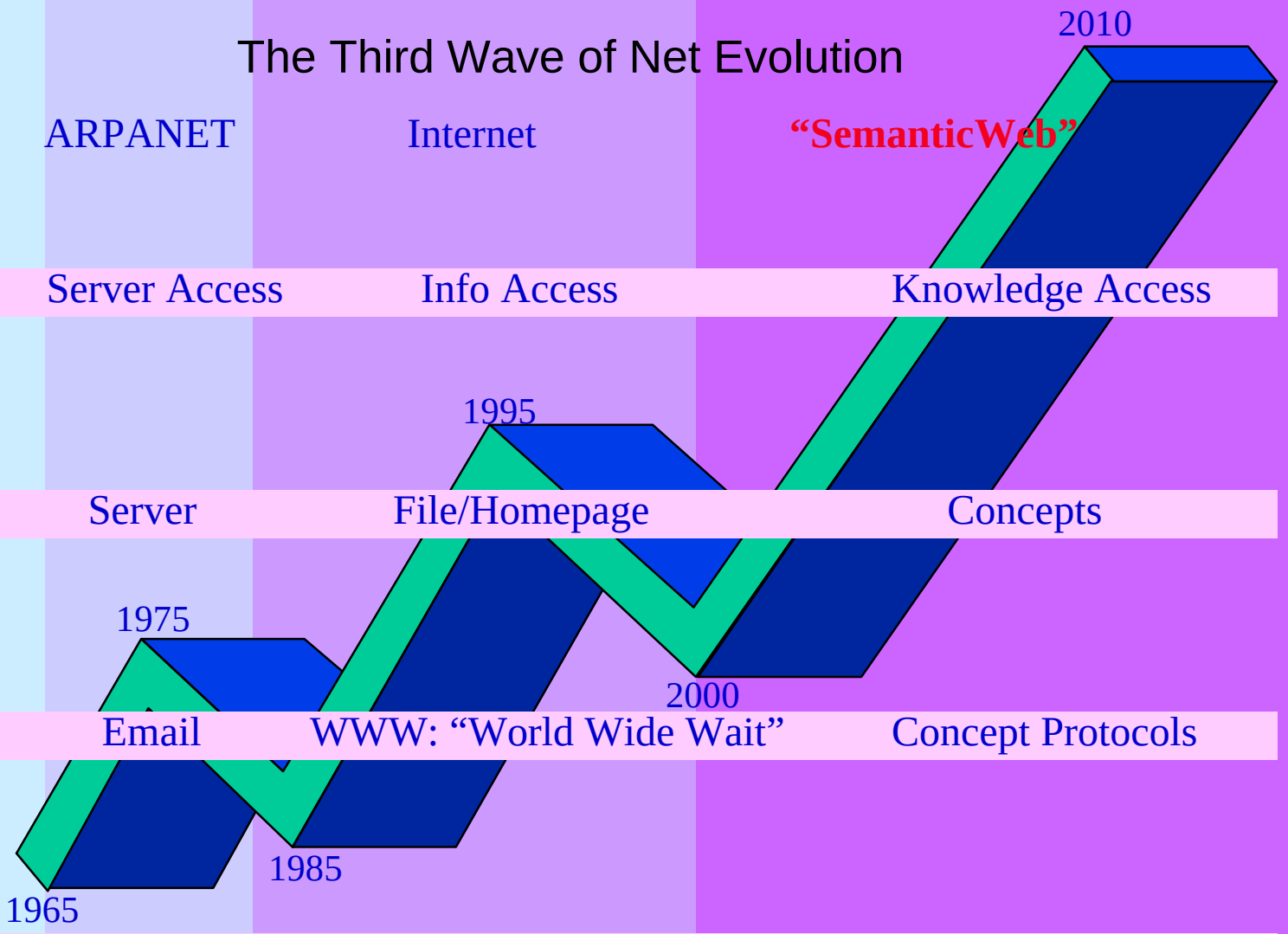
Data, Information and Knowledge:

- According to Alter (1996), Tobin (1996), and Beckman (1999):
  - Data: Facts, images, or sounds (+interpretation+meaning =)
  - Information: Formatted, filtered, and summarized data (+action+application =)
  - Knowledge: Instincts, ideas, rules, and procedures that guide actions and decisions

Application and Societal Relevance :

- Ontologies, hierarchies, and subject headings
- Knowledge management systems and practices: knowledge maps
- Digital libraries, search engines, web mining, text mining, data mining, CRM, eCommerce
- Semantic web, multilingual web, multimedia web, and wireless web

The Third Wave of Net Evolution



	ARPANET	Internet	"SemanticWeb"
Function	Server Access	Info Access	Knowledge Access
Unit	Server	File/Homepage	Concepts
Example	Email	WWW: "World Wide Wait"	Concept Protocols
Company	IBM	Microsoft/Netscape	**???**

Knowledge Management Definition

"The system and managerial approach to collecting, processing, and organizing enterprise-specific knowledge assets for business functions and decision making."

Knowledge Management Challenges

- "… making high-value corporate information and knowledge easily available to support decision making at the lowest, broadest possible levels …"
  - Personnel Turn-over
  - Organizational Resistance
  - Manual Top-down Knowledge Creation
  - Information Overload

Knowledge Management Landscape

- Research Community
  - NSF / DARPA / NASA, Digital Library Initiative I & II ($80M)
  - NSF, Digital Government Initiative ($60M)
  - NSF, Knowledge Networking Initiative ($50M)
  - NSF, Information Technology Research ($120M)
- Business Community
  - Intellectual Capital, Corporate Memory,
  - Knowledge Chain, Competitive Intelligence

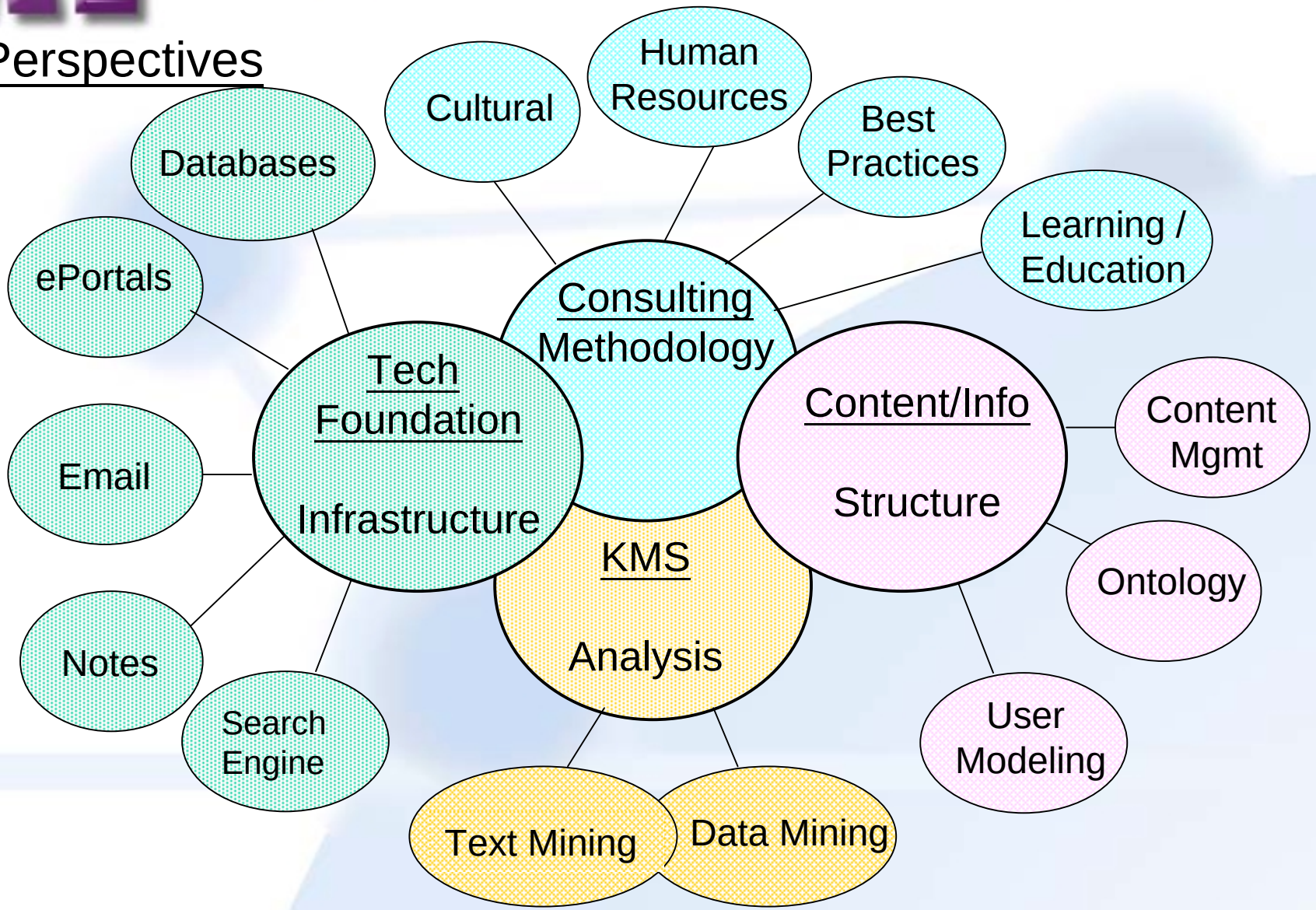
Knowledge Management Foundations

- Enabling Technologies:
  - Information Retrieval (Excalibur, Verity, Oracle Context)
  - Electronic Document Management (Documentum, PC DOCS)
  - Internet/Intranet (Yahoo!, Excite)
  - Groupware (Lotus Notes, MS Exchange, Ventana)
- Consulting and System Integration:
  - Best practices, human resources, organizational development, performance metrics, methodology, framework, ontology (Delphi, E&Y, Arthur Andersen, AMS, KPMG)

Knowledge Management Perspectives:

- Process perspective (management and behavior): consulting practices, methodology, best practices, e-learning, culture/reward, existing IT ➔ new information, old IT, new but manual process
- Information perspective (information and library sciences): content management, manual ontologies ➔ new information, manual process
- Knowledge Computing perspective (text mining, artificial intelligence): automated knowledge extraction, thesauri, knowledge maps ➔ new IT, new knowledge, automated process

KM Perspectives



- Dataware Technologies

  (1) Identify the Business Problem
  (2) Prepare for Change
  (3) Create a KM Team
  (4) Perform the Knowledge Audit and Analysis
  (5) Define the Key Features of the Solution
  (6) Implement the Building Blocks for KM
  (7) Link Knowledge to People

- Anderson Consulting

  (1) Acquire
  (2) Create
  (3) Synthesize
  (4) Share
  (5) Use to Achieve Organizational Goals
  (6) Environment Conducive to Knowledge Sharing

- The Delphi Group

(1) Key Concepts and Frameworks for Knowledge Management

(2) How to Use Knowledge Management as a Competitive Tool

(3) The Culture and Organization Aspects of Knowledge Management

(4) Best Practices in Knowledge Management

(5) The Technology of Knowledge Management

(6) Market Analysis

(7) Justifying Knowledge Management

(8) Implementing Knowledge Management

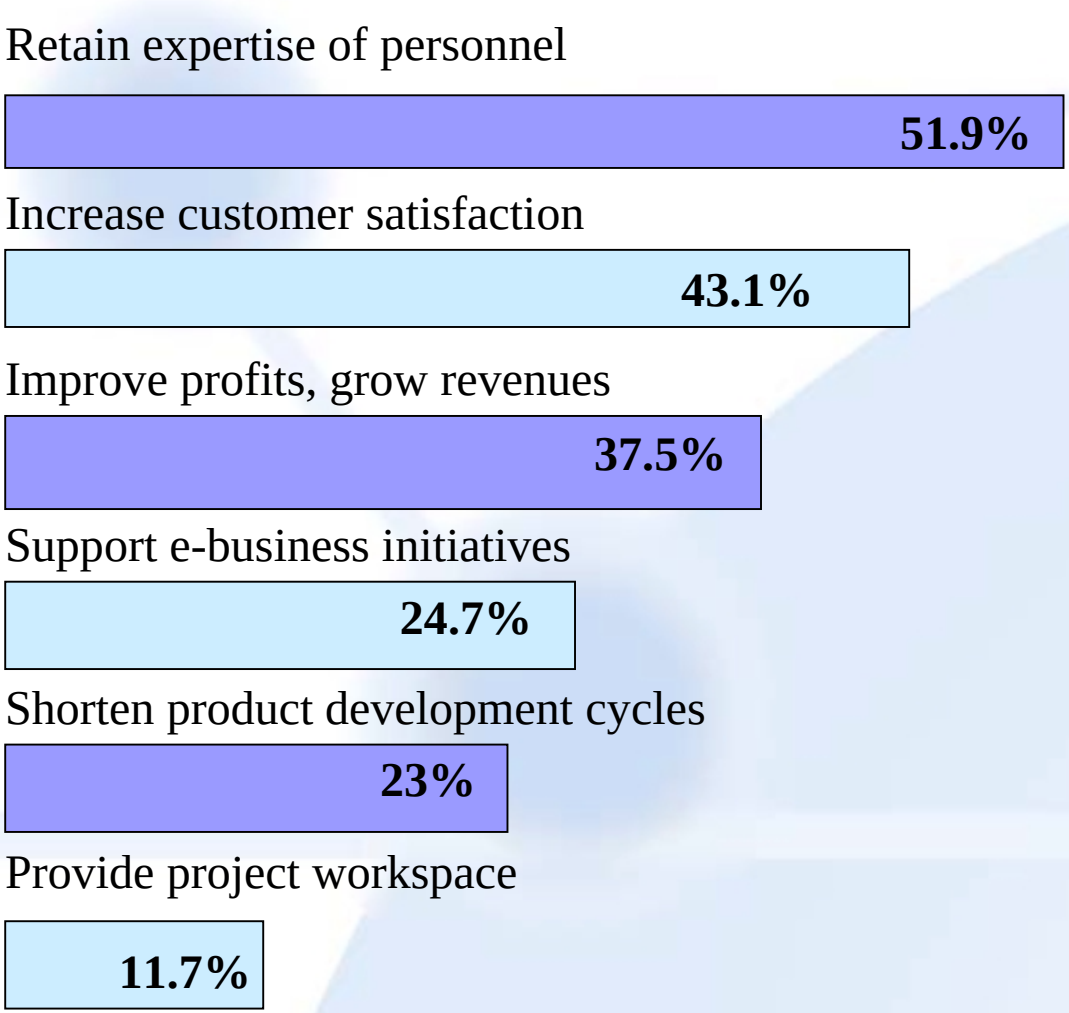
- Ernst & Young

  (1) Knowledge Generation
  (2) Knowledge Representation
  (3) Knowledge Codification
  (4) Knowledge Application

- PriceWaterhouseCoopers

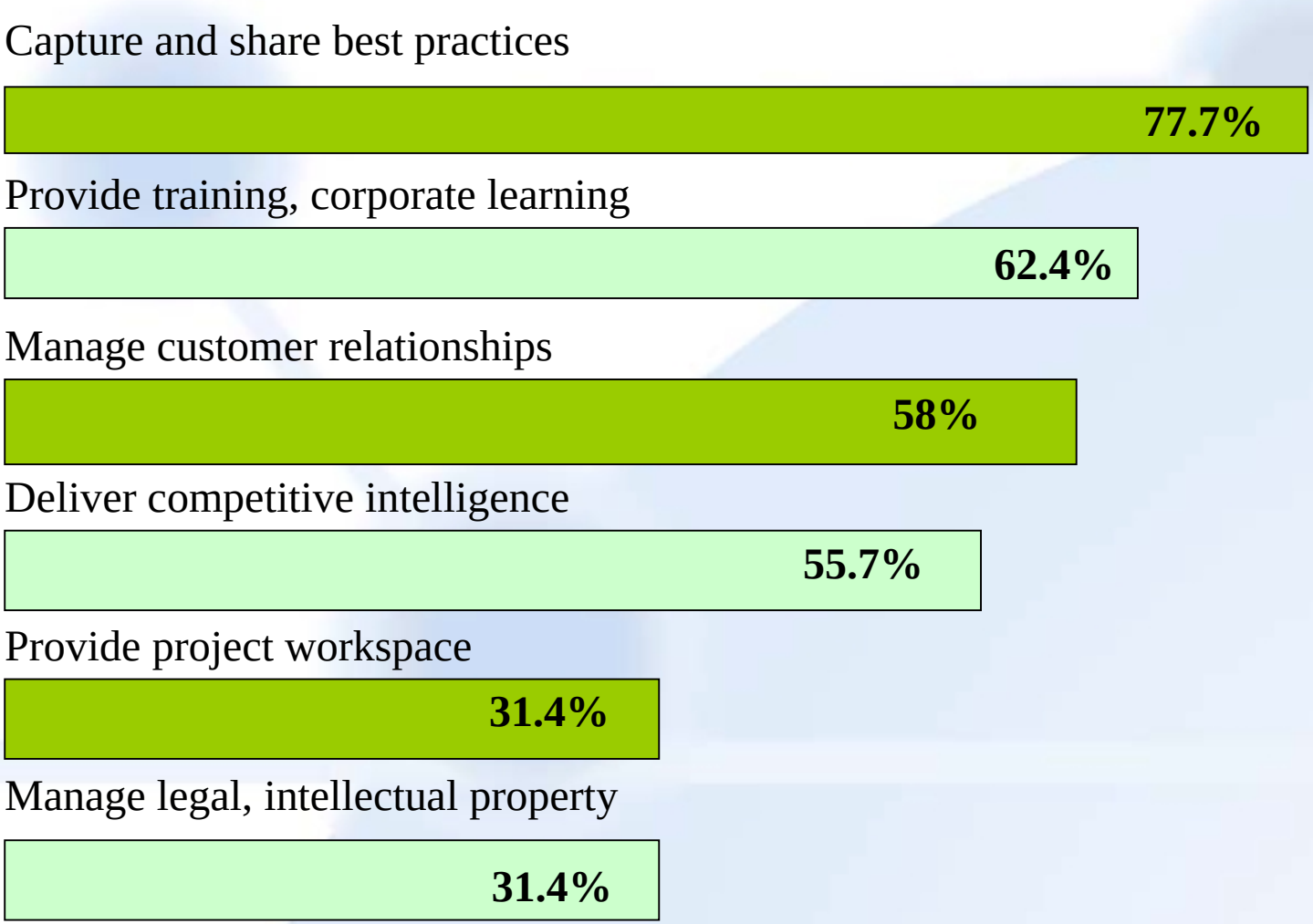
(1) Find
(2) Filter [for relevance]
(3) Format [to problem]
(4) Forward [to right people] and
(5) Feedback [from users]

Reason for Adopting KM



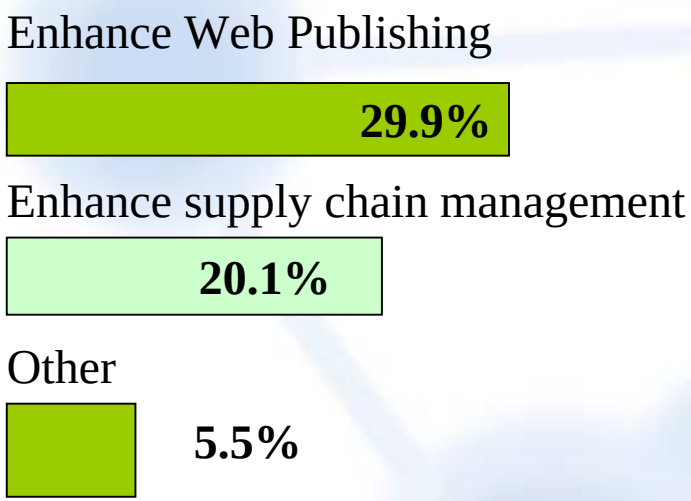
Knowledge Management and IDC May 2001

Business Uses Of KM Initiative



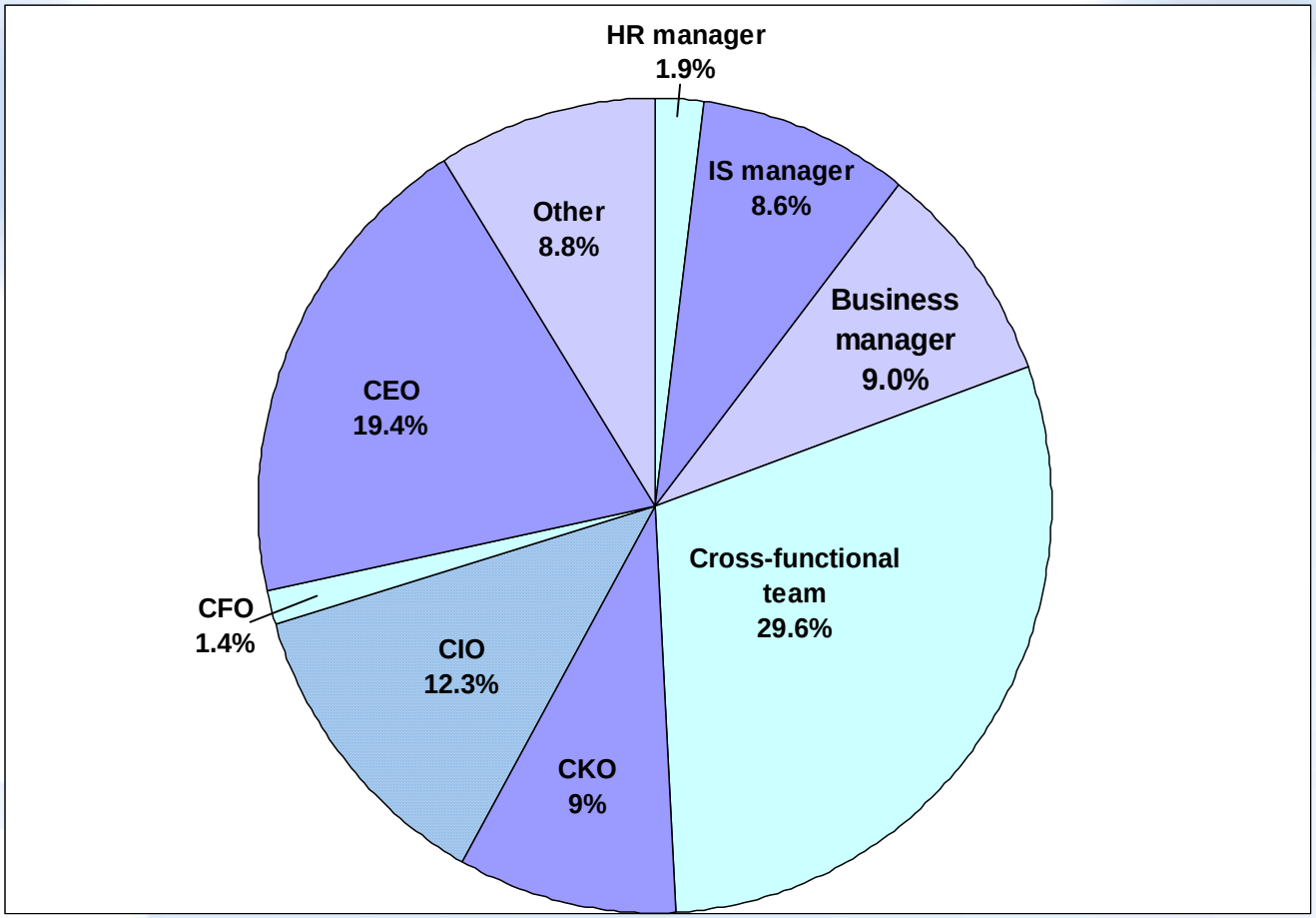
Continue

Business Uses Of KM Initiative



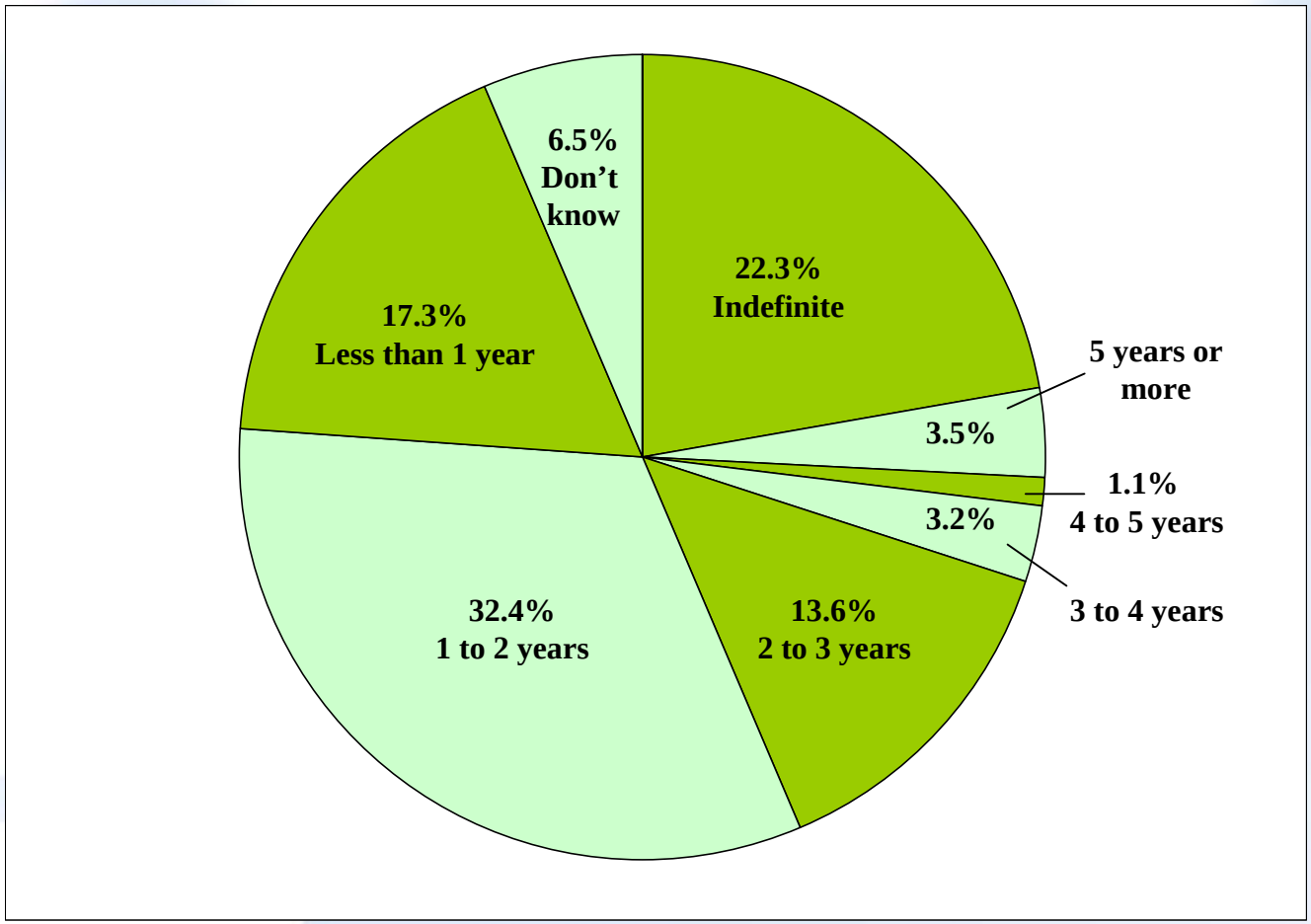
Knowledge Management and IDC May 2001

Leader Of KM Initiative



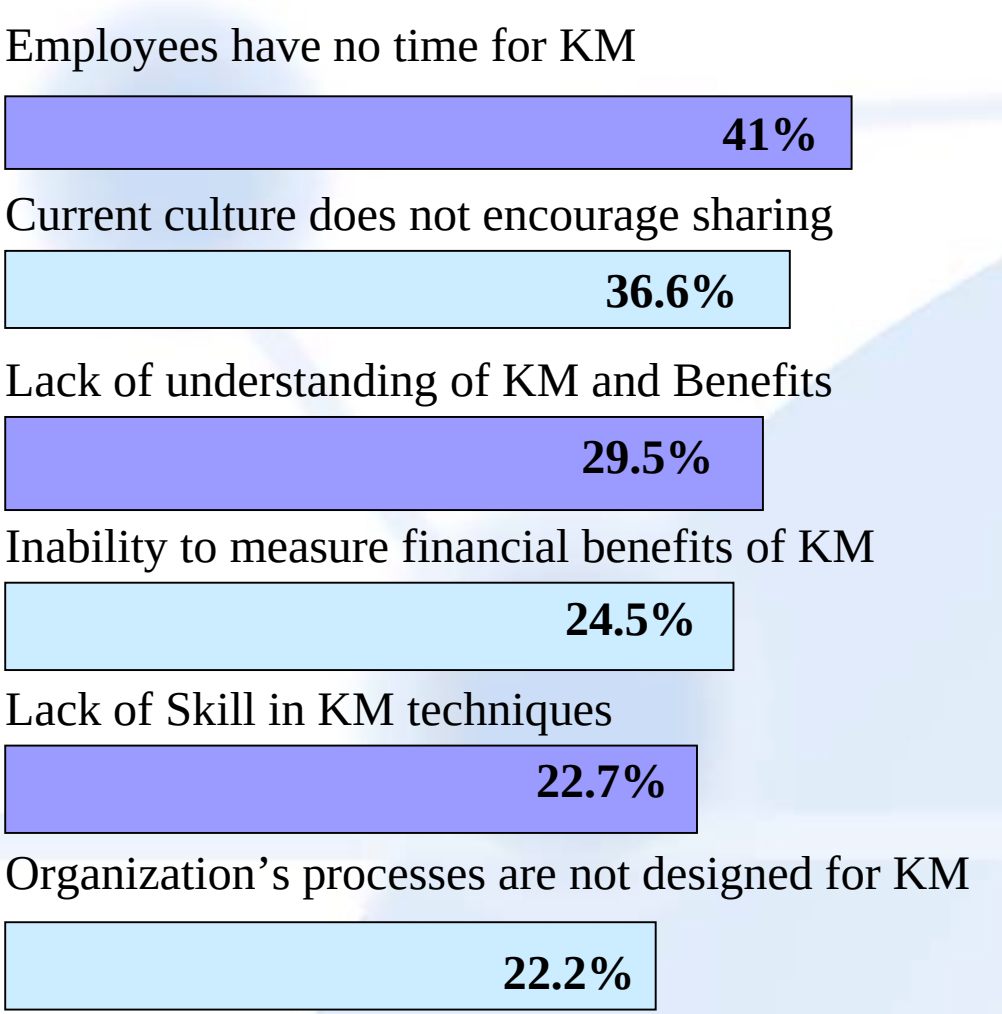
Knowledge Management and IDC May 2001

Planned Length Of Project



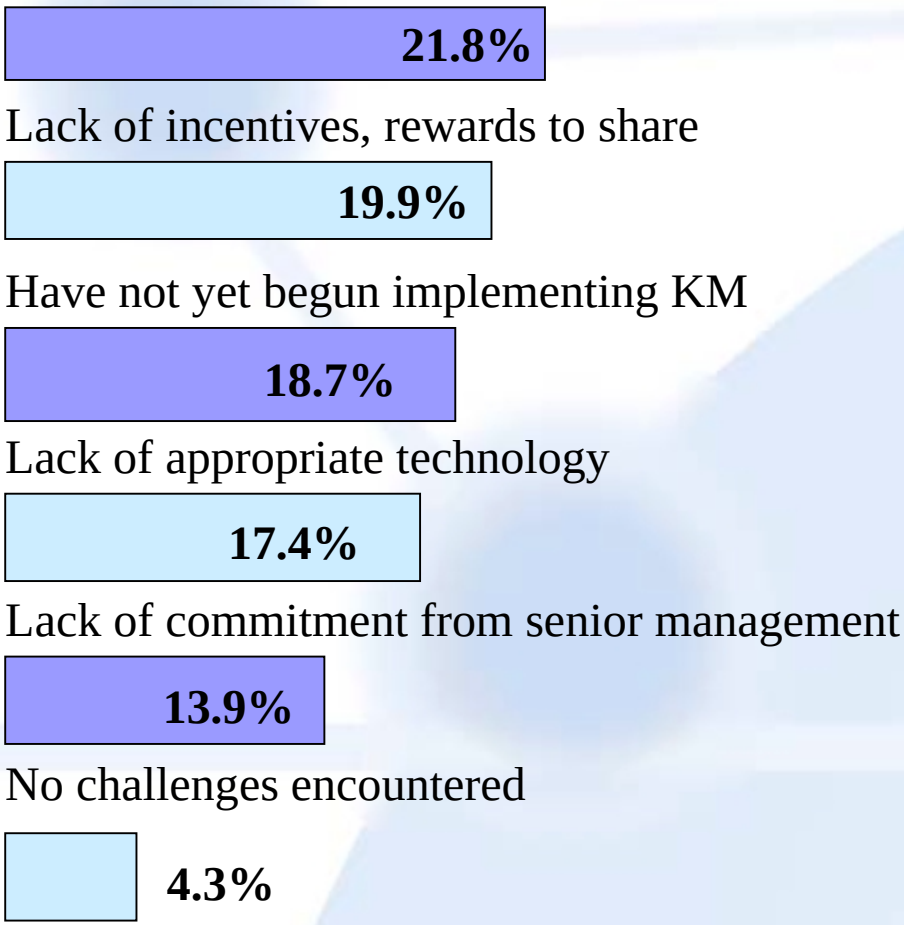
Knowledge Management and IDC May 2001

Implementation Challenges



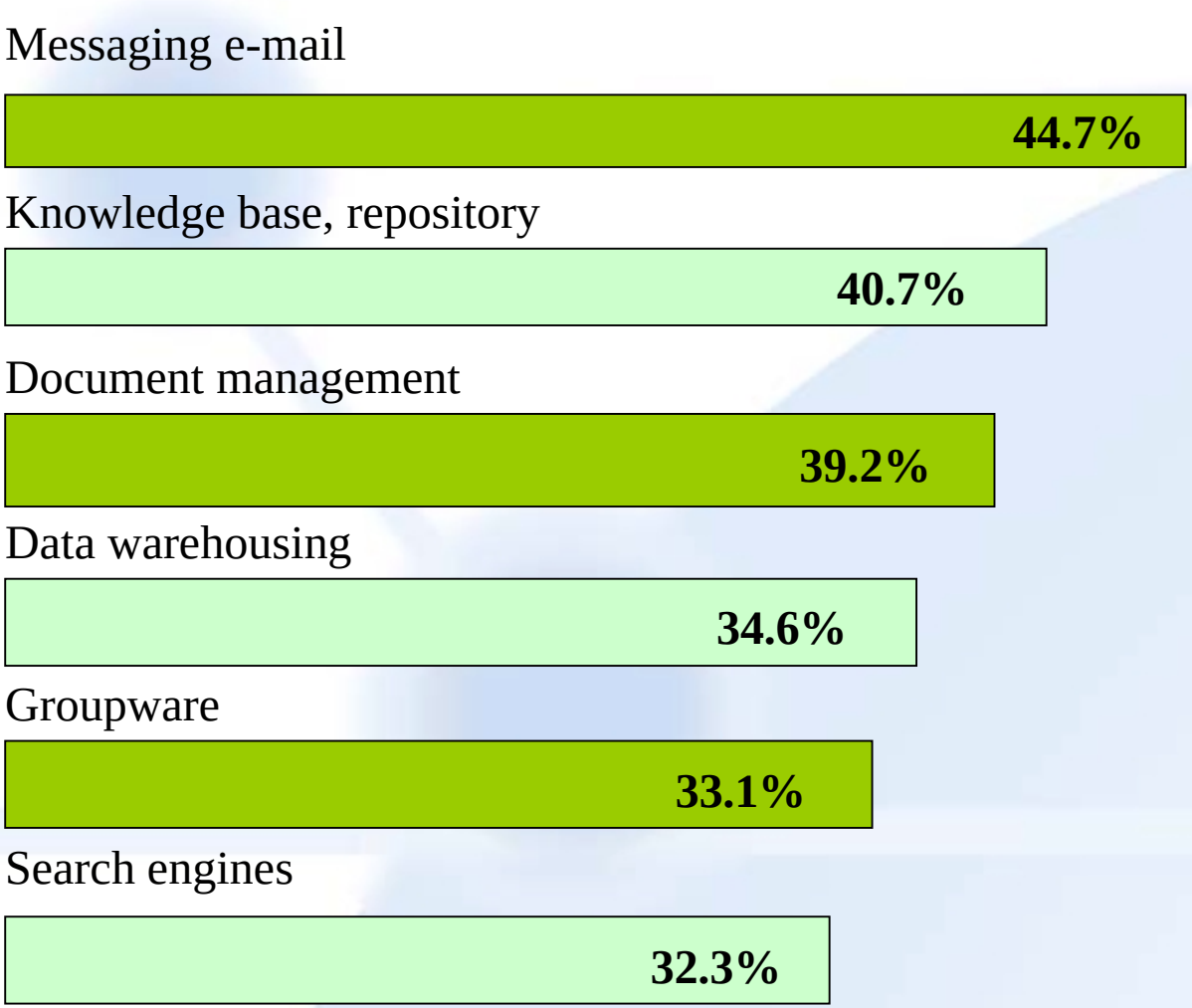
Continue

Implementation Challenges



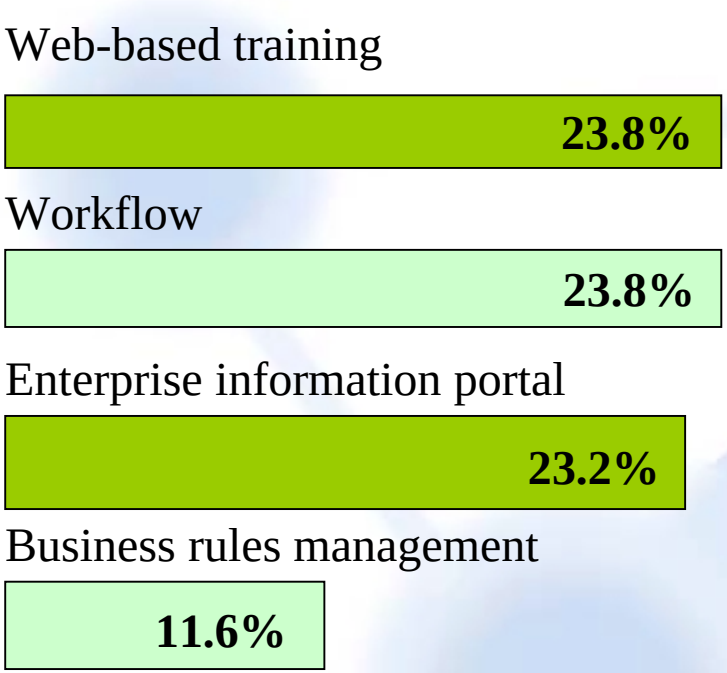
Knowledge Management and IDC May 2001

Types of Software Purchased



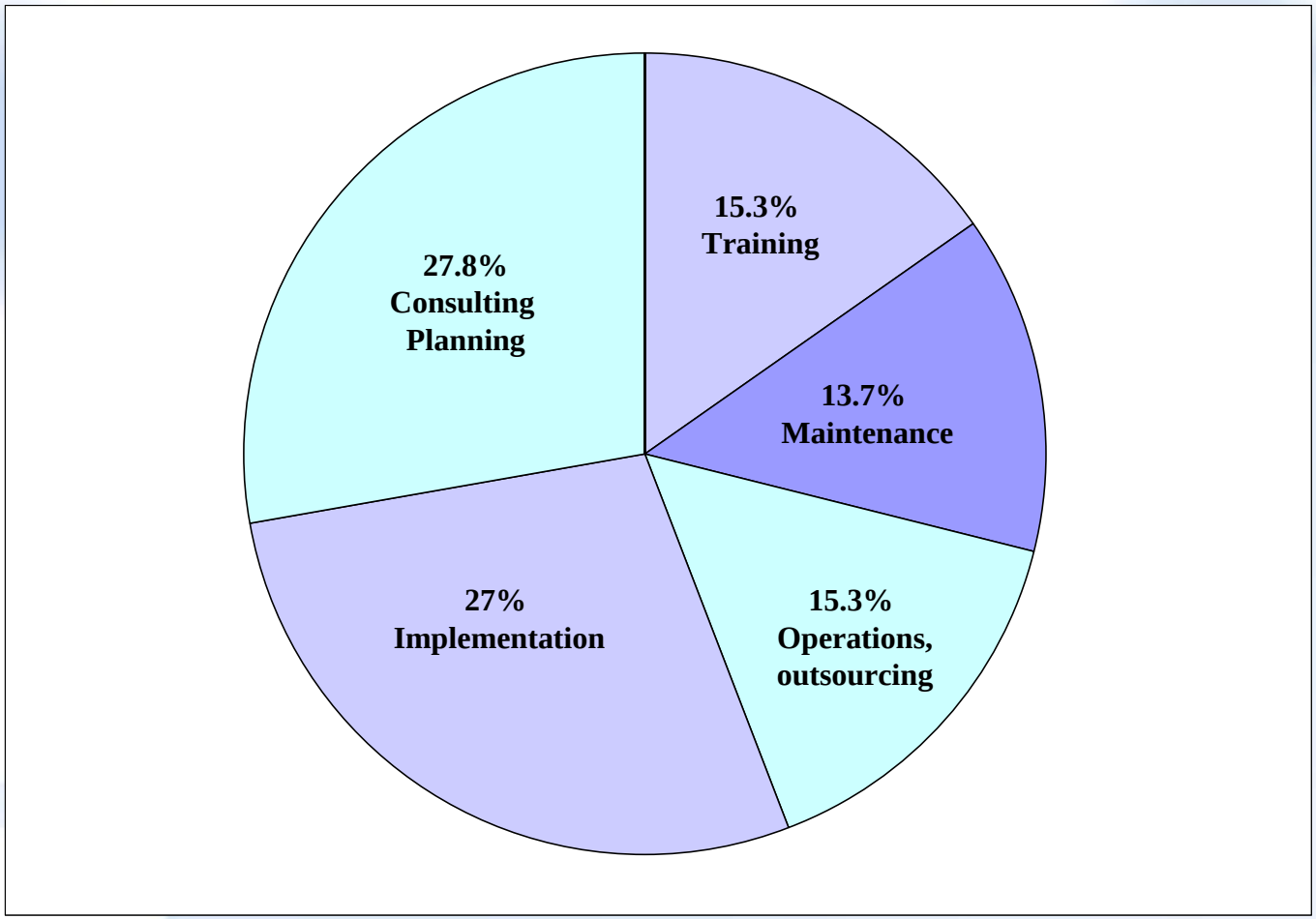
Continue

Types of Software Purchased



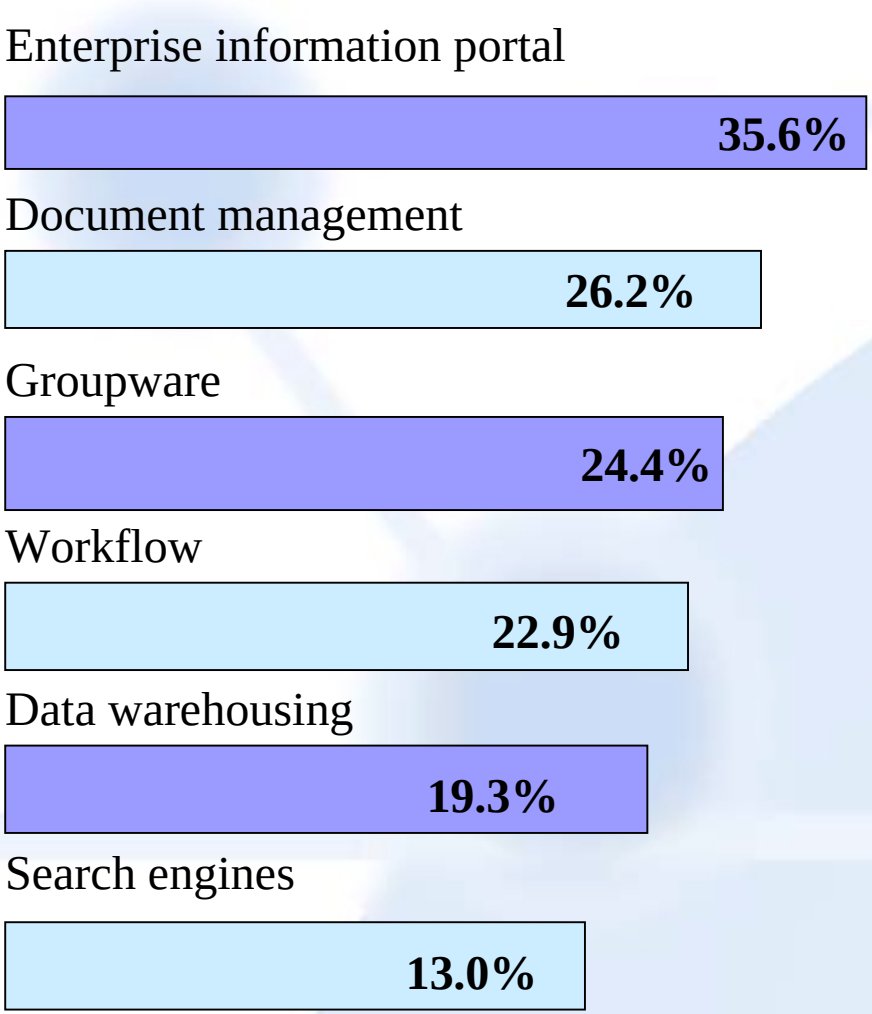
Knowledge Management and IDC May 2001

Spending On IT Services For KM



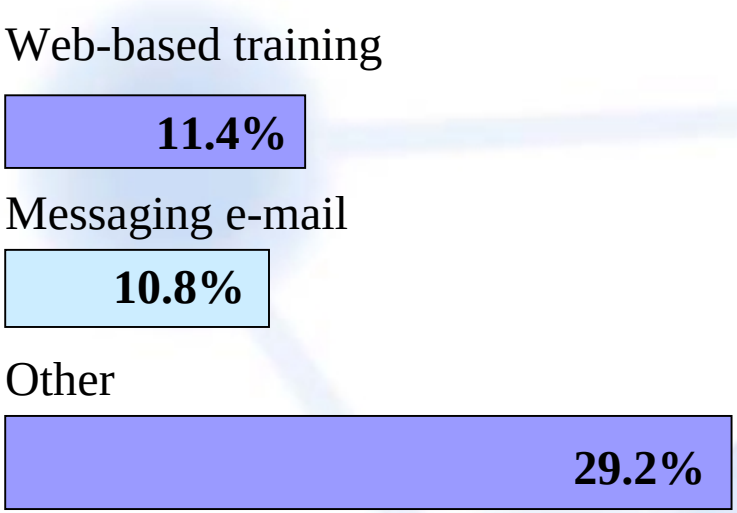
Knowledge Management and IDC May 2001

Software Budget Allotments



Continue

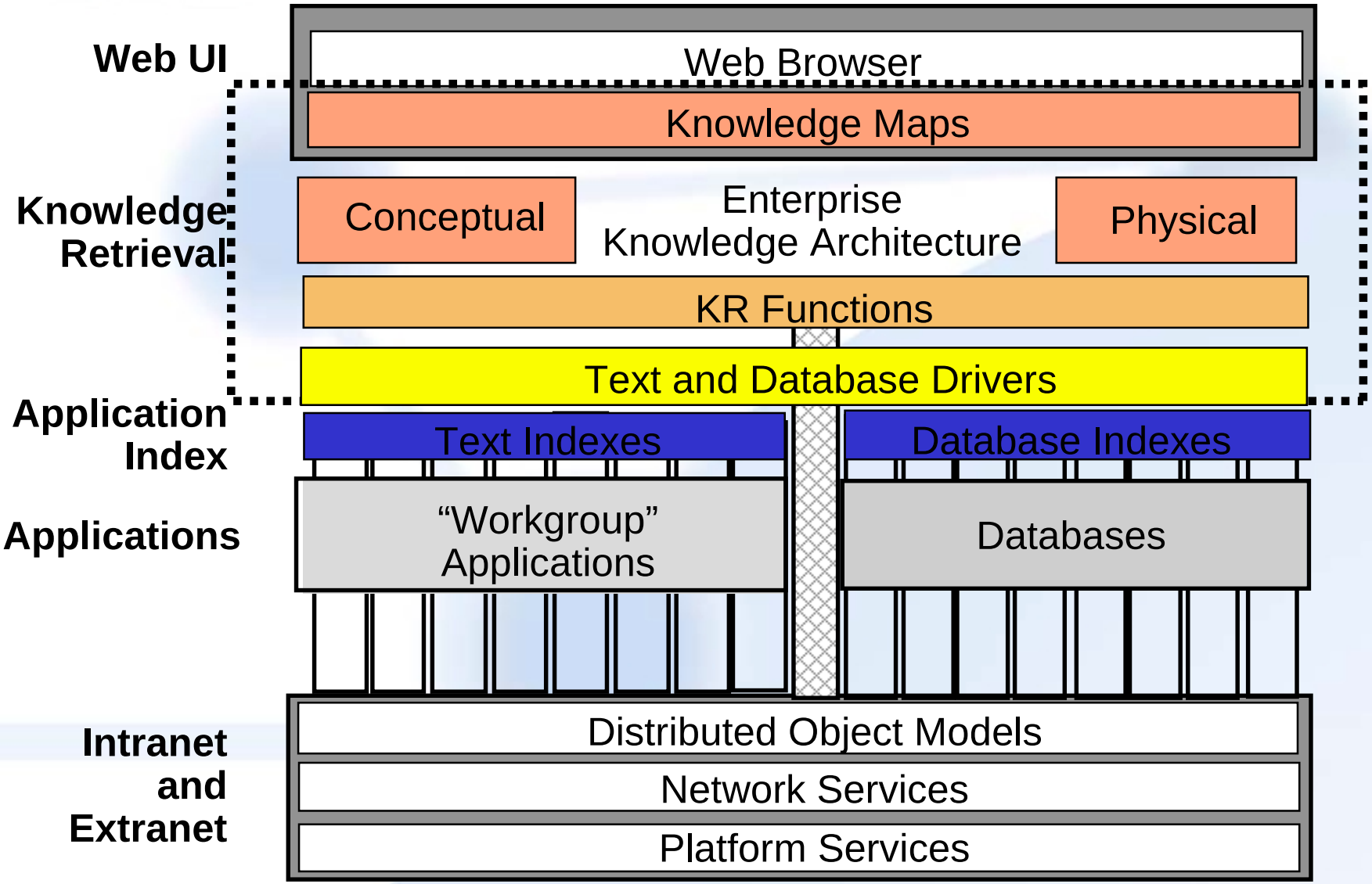
Software Budget Allotments



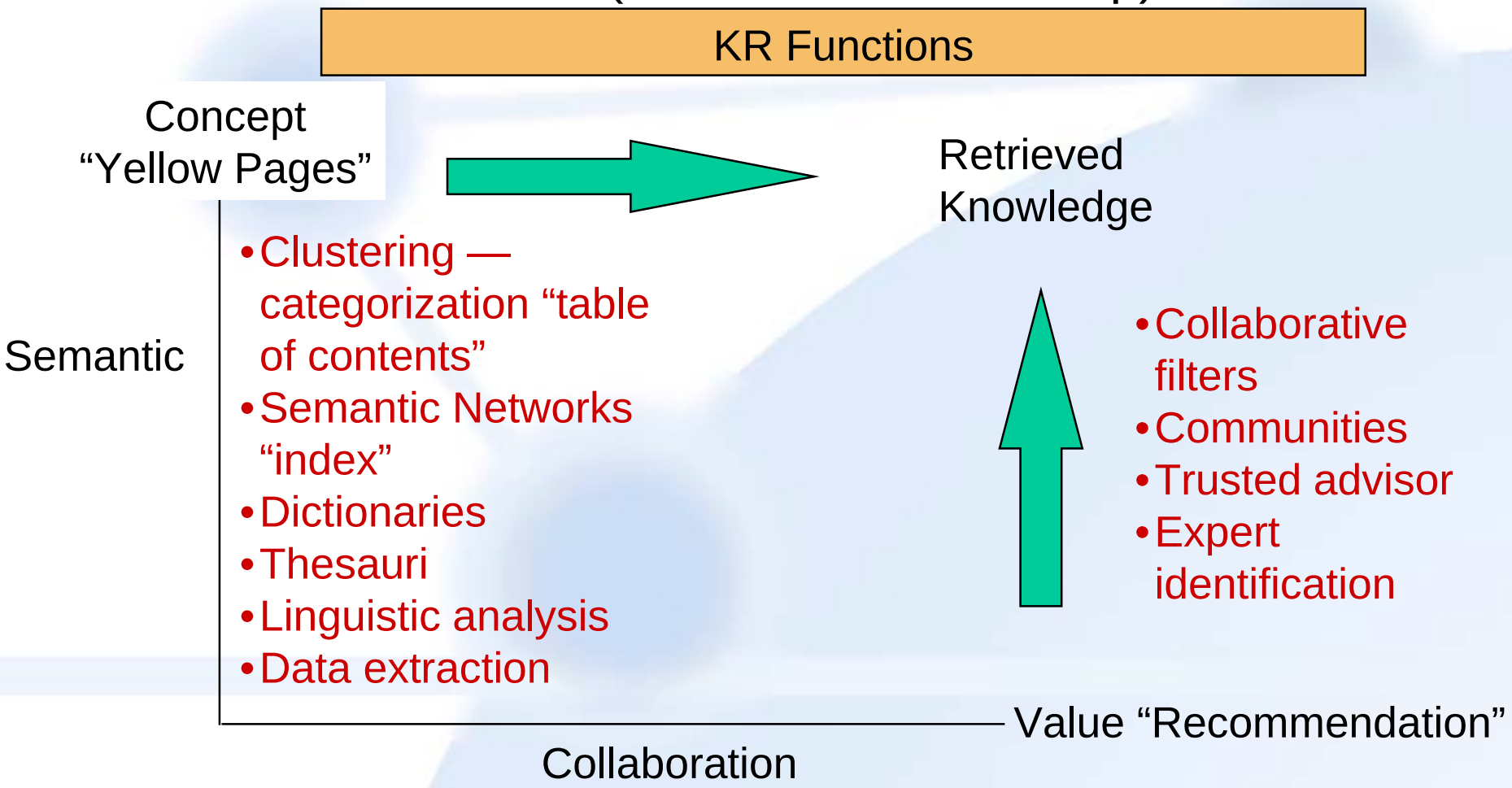
Knowledge Management Systems (KMS)

- Characteristics of KMS
- The Industry and the Market
- Major Vendors and Systems

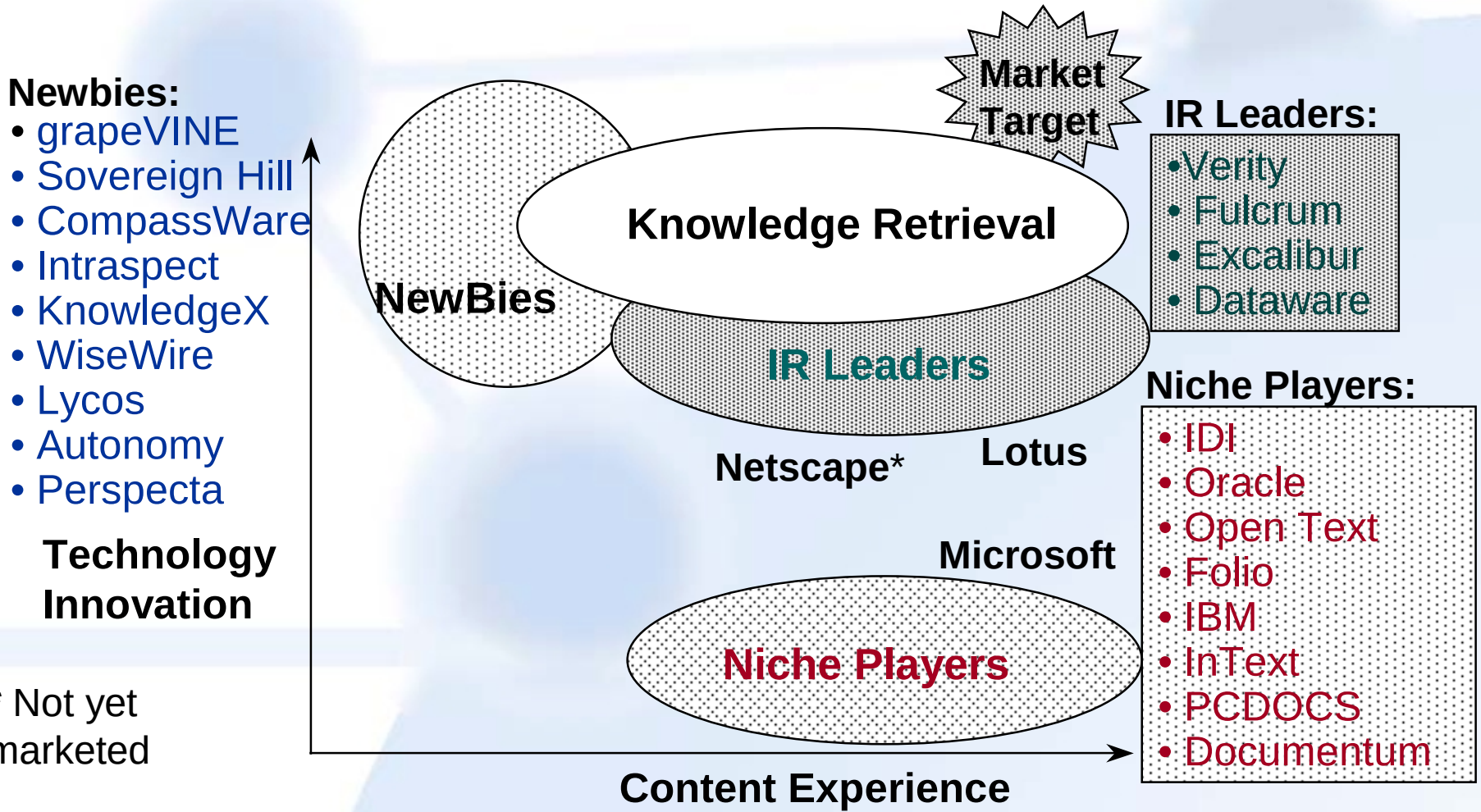
KM Architecture (Source: GartnerGroup)



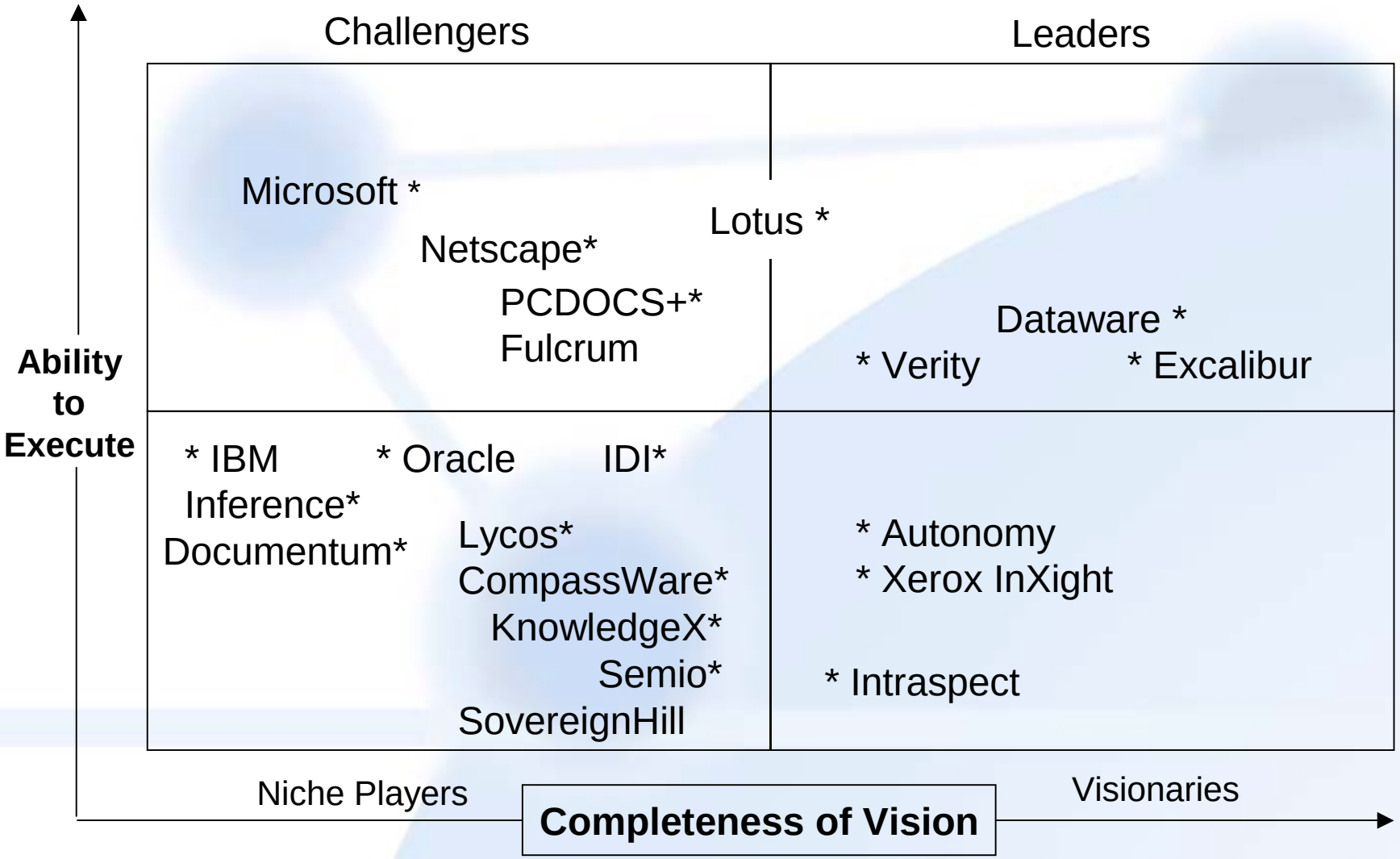
Knowledge Retrieval Level (Source: GartnerGroup)



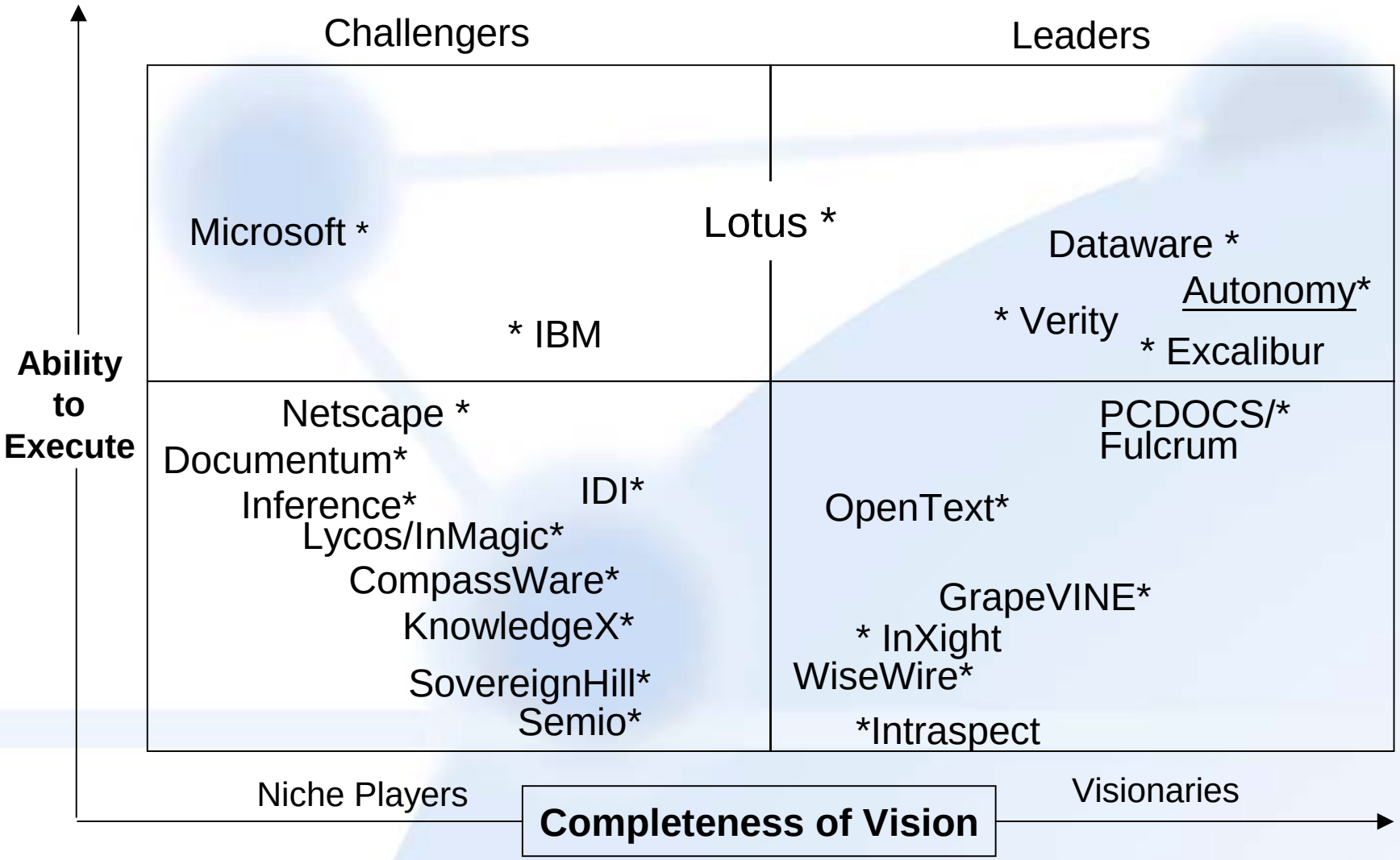
Knowledge Retrieval Vendor Direction (Source: GartnerGroup)



Knowledge Retrieval Vendors



KM Software Vendors

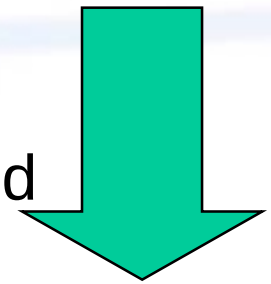


From Federal Research to Commercial Start-ups

• U. Mass:	Sovereign Hill
• MIT Media Lab:	Perspecta
• Xerox PARC:	InXight
• Batelle:	ThemeMedia
• U. Waterloo:	OpenText
• Cambridge U.	Autonomy
• U. Arizona:	Knowledge Computing Corporation (KCC)

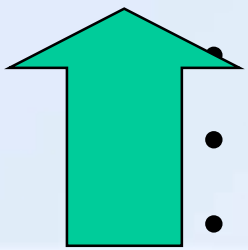
Two Approaches to Codifying Knowledge

Top-Down Approach



- Structured
- Manual
- Human-driven

Bottom-Up Approach

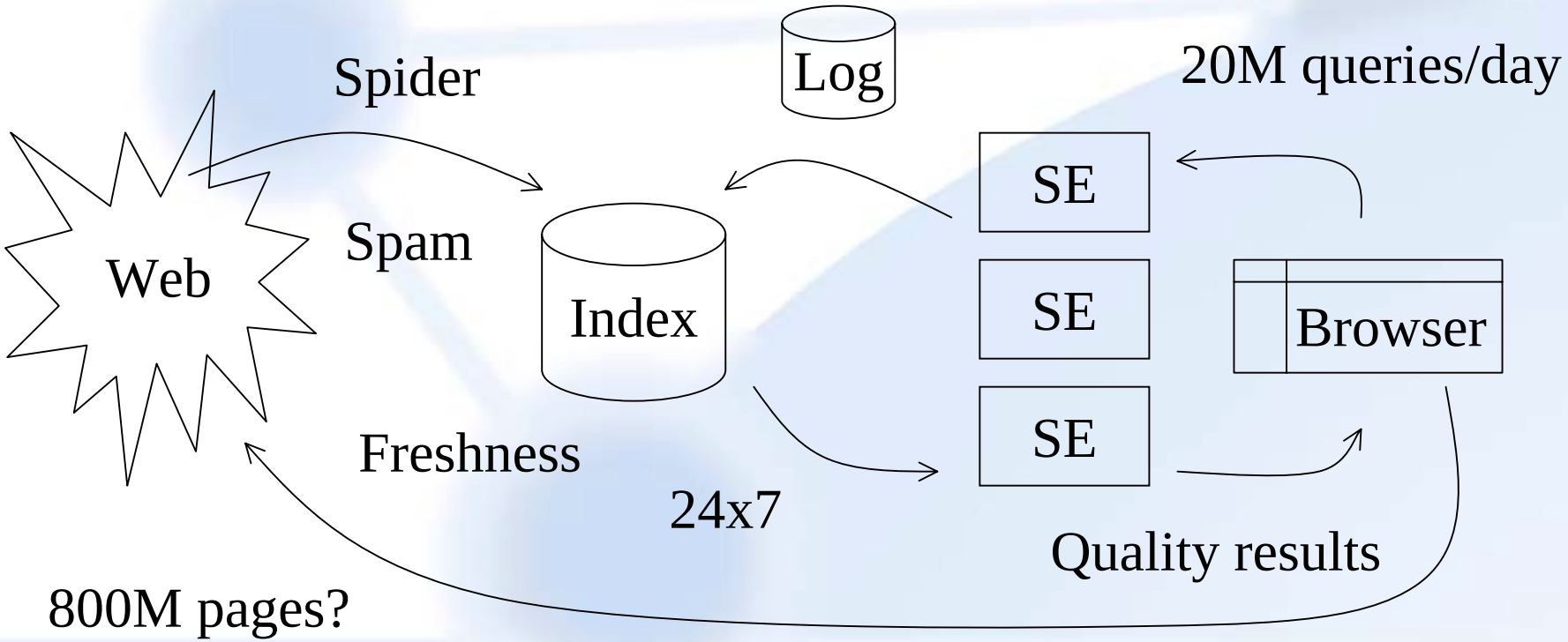


- Unstructured
- System-aided
- Data/Info-driven

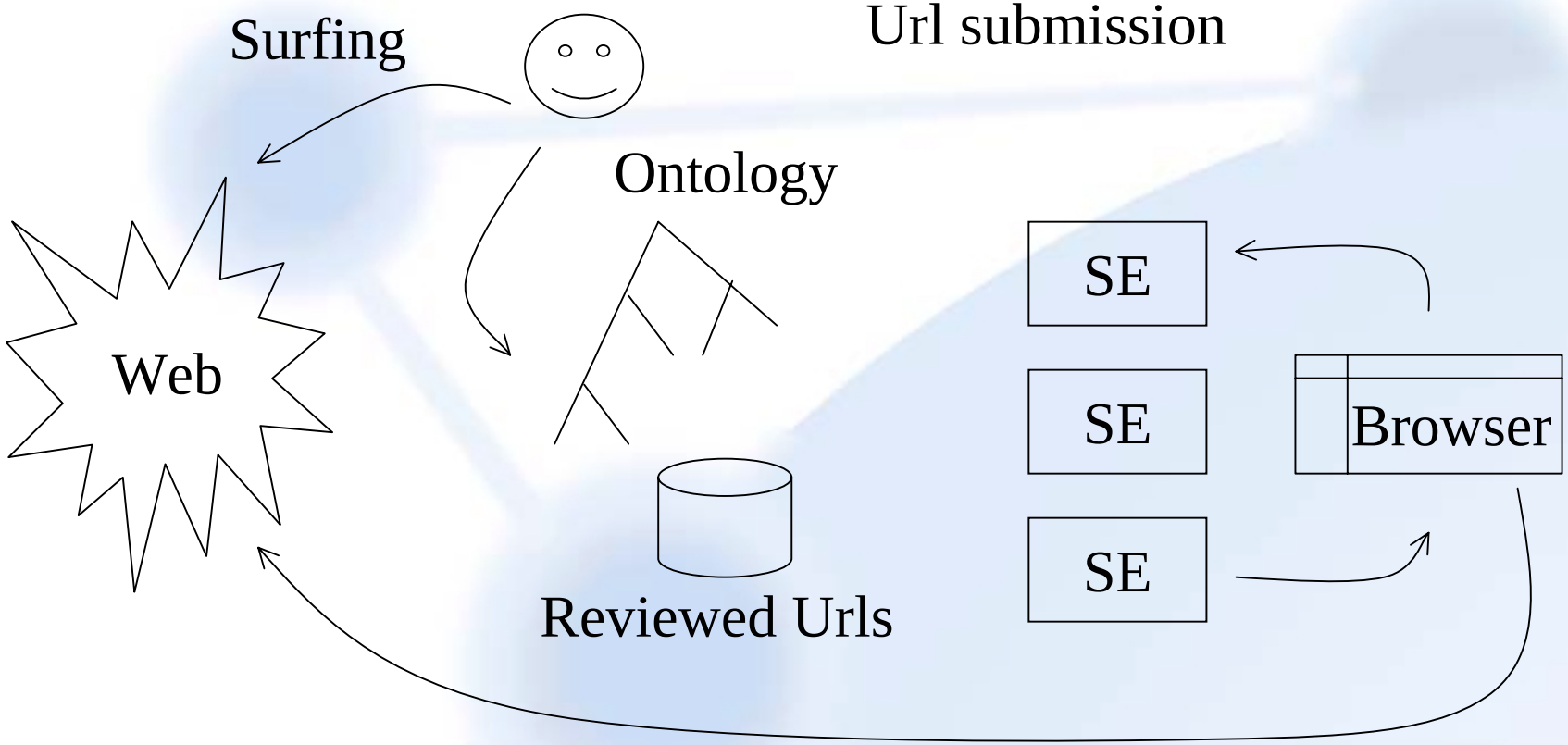
Knowledge Management Related Field: Search Engine

(Source: Jan Peterson and William Chang, Excite)

Basic Architectures: Search



Basic Architectures: Directory



Spidering

- Web HTML data
  - Hyperlinked
  - Directed, disconnected graph
  - Dynamic and static data
  - Estimated 800M indexible pages
- Freshness
  - How often are pages revisited?

Indexing

- Size
  - from 50 to 150M urls
  - 50 to 100% indexing overhead
    - 200 to 400GB indices
- Representation
  - Fields, meta-tags and content
  - NLP: stemming?

Search

- Augmented Vector-space
  - Ranked results with Boolean filtering
- Quality-based re-ranking
  - Based on hyperlink data
  - or user behavior
- Spam
  - Manipulation of content to improve placement

Queries

- Short expressions of information need
  - 2.3 words on average
  - Relevance overload is a key issue
    - Users typically only view top results
- Search is a high volume business
  - Yahoo! 50M queries/day
  - Excite 30M queries/day
  - Infoseek 15M queries/day

Alta Vista: within site search, machine translation

Directory

- Manual categorization and rating
  - Labor intensive
    - 20 to 50 editors
  - High quality, but low coverage
    - 200-500K urls
- Browsable ontology
- Open Directory is a distributed solution

Yahoo: manual ontology (200 ontologists)

Business Model

- Advertising
  - Highly targeted, based on query
  - Keyword selling; Between $3 to $25 CPM
- Cost per query is critical
  - Between $.5 and $1.0 per thousand
- Distribution
  - Many portals outsource search

Web Resources

- Search Engine Watch
  - www.searchenginewatch.com
- "Analysis of a Very Large Alta Vista Query Log"; Silverstein et al.
  - www.research.digital.com/SRC
- "The Anatomy of a Large-Scale Hypertextual Web Search Engine"; Brin and Page
  - google.stanford.edu/long321.htm
- WWW conferences: www10.org

Special Collections

- Newswire
- Newsgroups
  - Specialized services (Deja)
- Information extraction
  - Shopping catalog
  - Events; recipes, etc.

The Hidden Web

- Non-indexible content
  - Behind passwords, firewalls
  - Dynamic content
  - Often searchable through local interface
- Network of distributed search resources
  - How to access?
  - Ask Jeeves!

Spam

- Manipulation of content to affect ranking
  - Bogus meta tags
  - Hidden text
  - Jump pages tuned for each search engine
- Add Url is a spammer's tool
  - 99% of submissions are spam
- It's an arms race

The Role of NLP

- Many Search Engines do not stem
  - Precision bias suggests conservative term treatment
- What about non-English documents
  - N-grams are popular for Chinese
  - Language ID anyone?

Link Analysis

- Authors vote via links
  - Pages with higher inlink are higher quality
- Not all links are equal
  - Links from higher quality sites are better
  - Links in context are better
- Resistant to Spam
  - Only cross-site links considered

Page Rank (Page'98)

- Limiting distribution of a random walk
  - Jump to a random page with Prob. ε
  - Follow a link with Prob. 1- ε
- Probability of landing at a page D:
  - $\varepsilon/T + \Sigma P(C)/L(C)$
  - Sum over pages leading to D
  - L(C) = number of links on page D

HITS (Kleinberg'98)

- Hubs: pages that point to many good pages
- Authorities: pages pointed to by many good pages
- Operates over a vincity graph
  - pages relevant to a query
- Refined by the IBM Clever group
  - further contextualization

Evaluation

- No industry standard benchmark
  - Evaluations are qualitative
  - Excessive claims abound
  - Press is not be discerning
- Shifting target
  - Indices change daily
  - Cross engine comparison elusive

Who asks What?

- Query logs revisited
- Query-based indexing – why index things people don't ask for?
- If they ask for A, give them B
- From atomic concepts to query extensions
- Structure of questions and answers
  - Shyam Kapur's chunks

Futures

- Vertical markets – healthcare, real estate, jobs and resumes, etc.
- Localized search
- Search as embedded app
- Shopping 'bots
- Open Problems
- Has the bubble burst?

Acquisition of Communities

- Email, killer app of the internet
  - Mailing lists
- Usenet Newsgroups
- Bulletin boards
- Chat rooms
- Instant messaging
  - buddy lists, ICQ (I Seek You)

The New Networks

- A consumer revolution
  - The community makes the brand
  - Winning brands empower consumers, embrace the internet's viral efficiency
- Media is at the core of brand marketing
- From portals to networks
  - navigation, advertising, commerce

The New Networks

- Ingredients:
  - Search engine audience
  - Ad agency
  - Old media
  - Verticals
  - Bank
  - Venture capital
  - Access, technology, and services providers

Keiretsu

- SoftBank
  - YAHOO!, Ziff-Davis, NASDAQ?
- Kleiner Perkins
  - AOL, Concentric, Sun, Netscape, Intuit, Excite
- Microsoft
  - MSN, MSNBC, NBC, CNET, Snap, Xoom, GE
- AT&T
  - TCI, AtHome, Excite

From SE to ePortal

- Spidering: Intranet and Internet crawling
- Integration: legacy systems and databases
- Content: aggregation and conversion
- Process: Collaboration, chat, workflow management, calendaring, and such
- Analysis: data and text mining, agent/alert

Knowledge Management Related Field: Data Mining

(Source: Michael Welge
Automated Learning Group, NCSA)

Why Data Mining? -- Potential Applications

- Database analysis, decision support, and automation
  - Market and Sales Analysis
  - Fraud Detection
  - Manufacturing Process Analysis
  - Risk Analysis and Management
  - Experimental Results Analysis
  - Scientific Data Analysis
  - Text Document Analysis

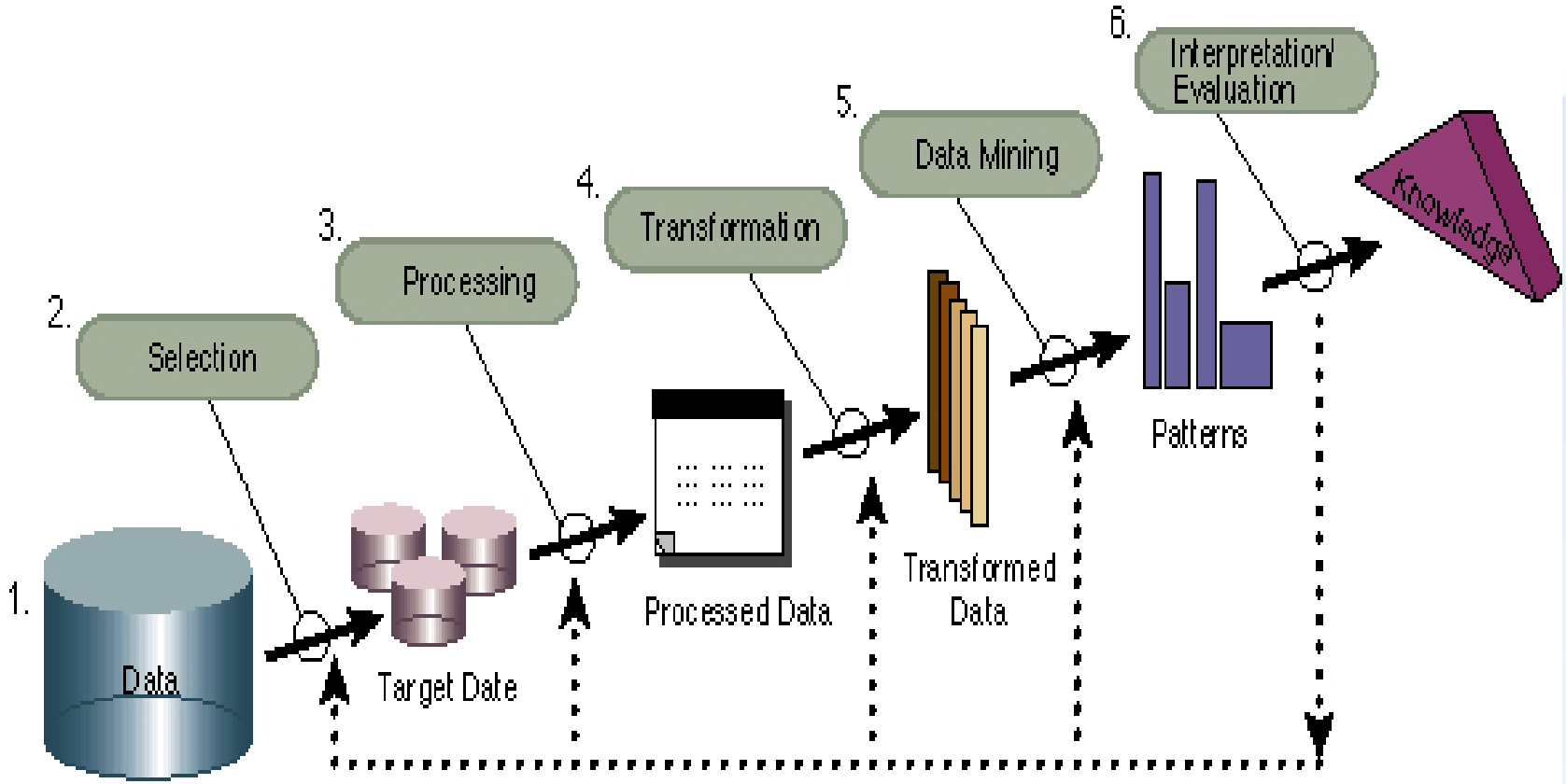
Data Mining: Confluence of Multiple Disciplines

- Database Systems, Data Warehouses, and OLAP
- Machine Learning
- Statistics
- Mathematical Programming
- Visualization
- High Performance Computing

Data Mining: On What Kind of Data?

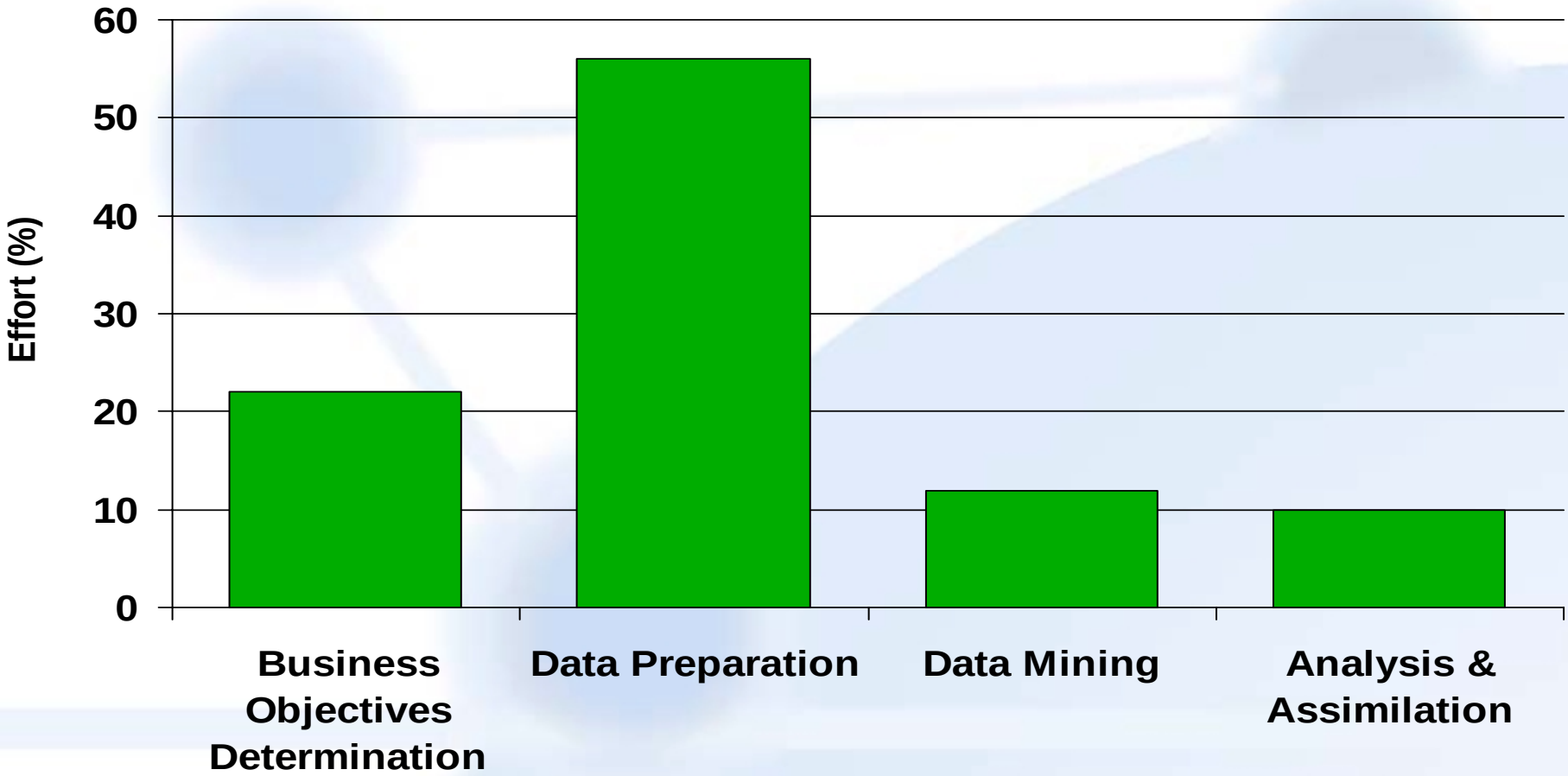
- Relational Databases
- Data Warehouses
- Transactional Databases
- Advanced Database Systems
  - Object-Relational
  - Spatial
  - Temporal
  - Text
  - Heterogeneous, Legacy, and Distributed
  - WWW (web mining)

Data Mining: A KDD Process



An Overview of the Steps That Compose the KDD Process

Required Effort for Each KDD Step



Data Mining Models and Methods

Predictive Modeling

- Classification
- Value prediction

Database Segmentation

- Demographic clustering
- Neural clustering

Link Analysis

- Associations discovery
- Sequential pattern discovery
- Similar time sequence discovery

Deviation Detection

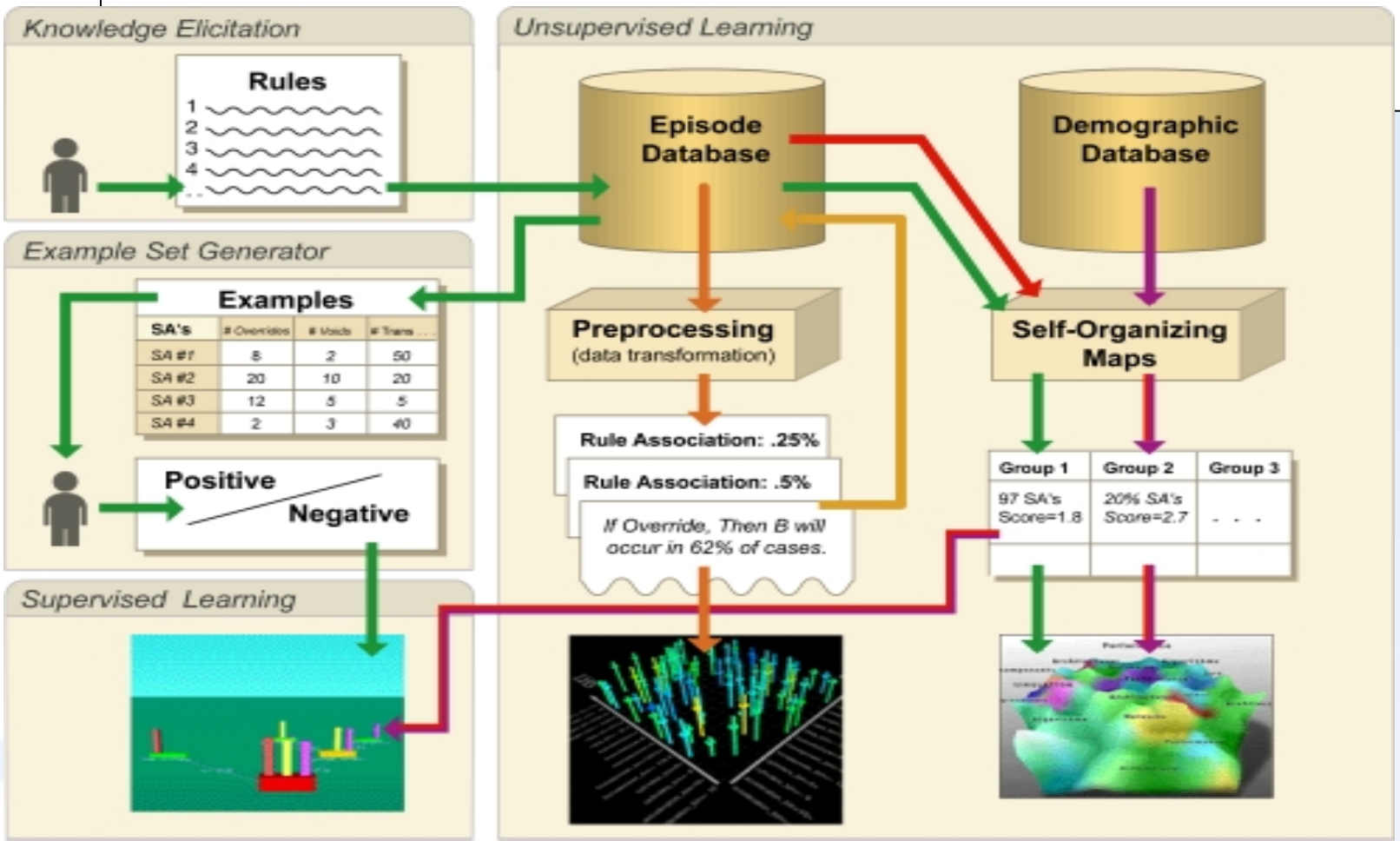
- Visualization
- Statistics

Fraud and Inappropriate Practice Prevention

Business Objective:

The focus of this project was on the recent and steady 12% annual rise in *overrides*. The overall business objective of the project was to find a way to ensure that the *overrides* were appropriate without negatively affecting service provided by the SAs.

Fraud and Inappropriate Practice Prevention



Link Analysis (Rule Association)

- Given a database, find all associations of the form:

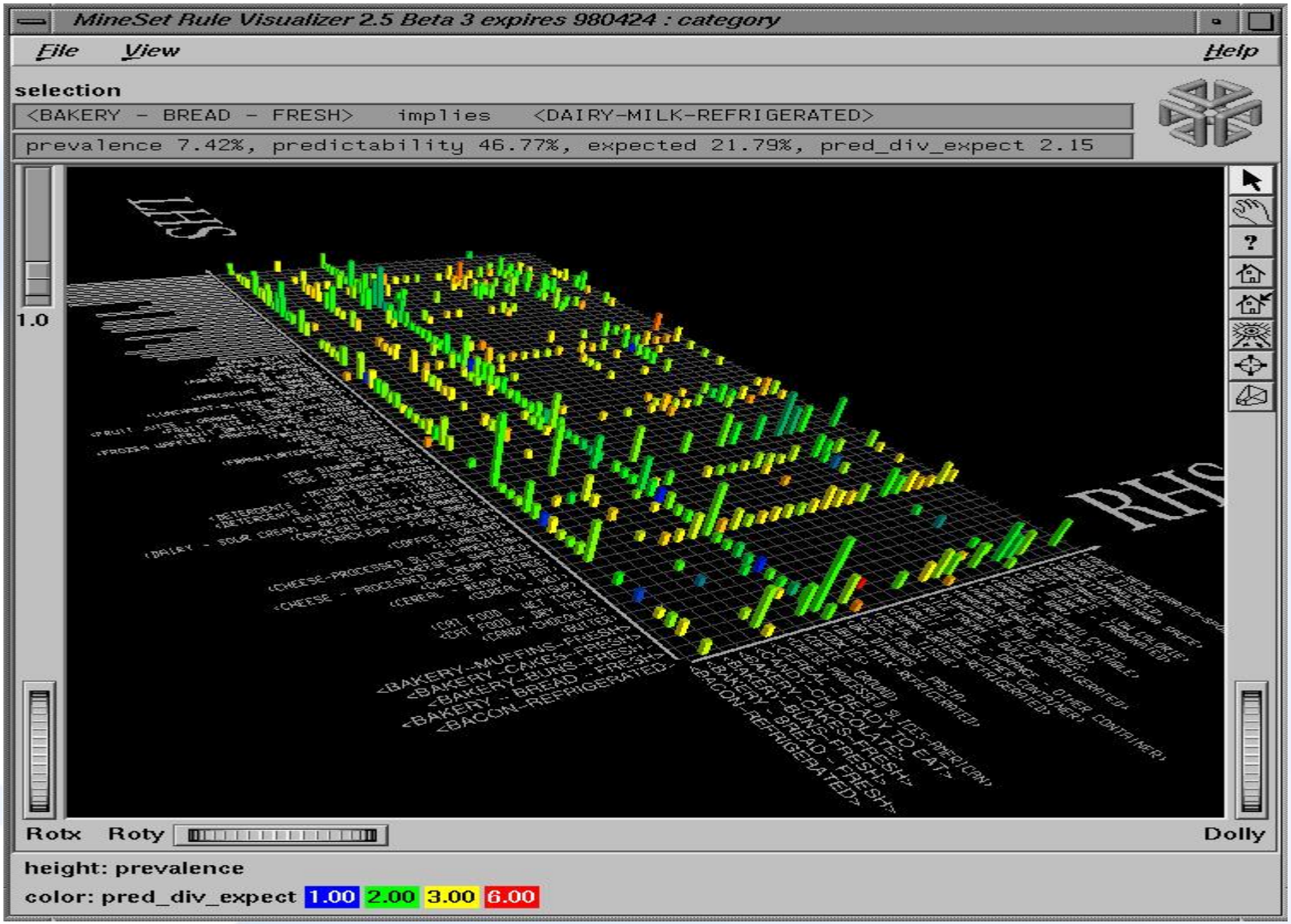
  IF < LHS > THEN <RHS >

  Prevalence = frequency of the LHS and RHS occurring together

  Predictability = fraction of the RHS out of all items with the LHS

  e.g., Beer and diaper

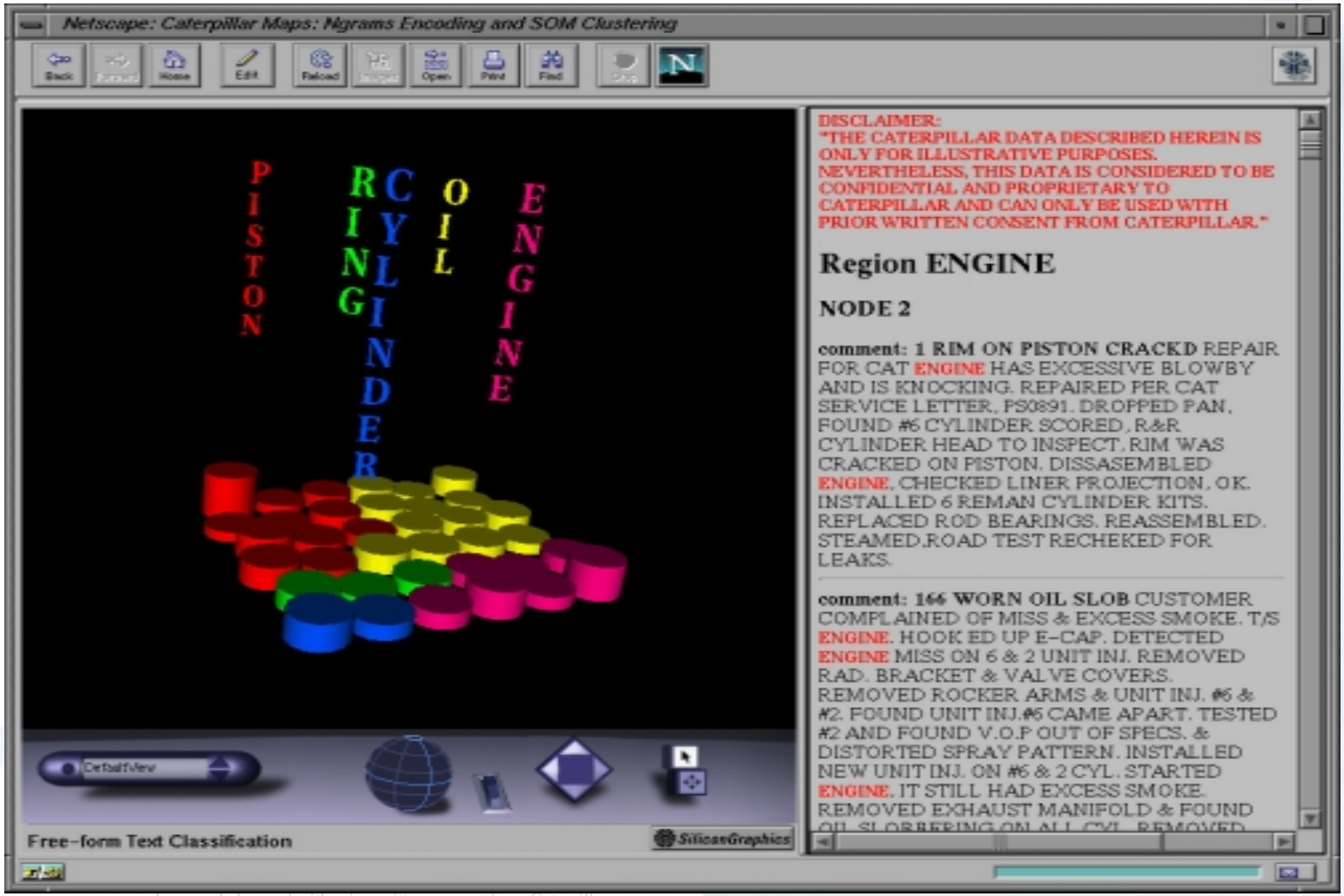
Rule Association - Basket Analysis



Data/Information Visualization

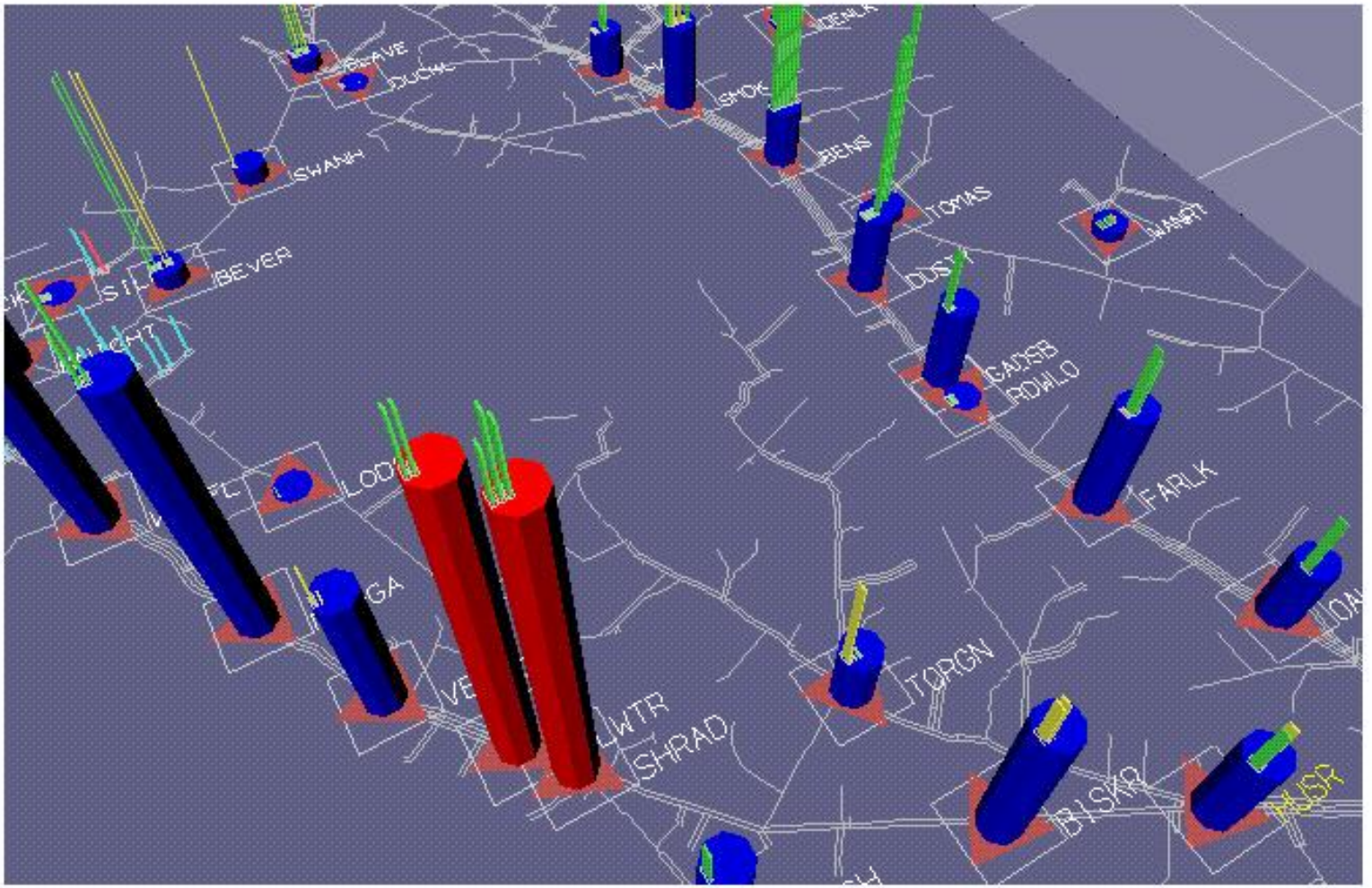
- Gain insight into the contents and complexity of the database being analyzed
- Vast amounts of under utilized data
- Time-critical decisions hampered
- Key information difficult to find
- Results presentation
- Reduced perceptual, interpretative, cognitive burden

Text Mining Visualization

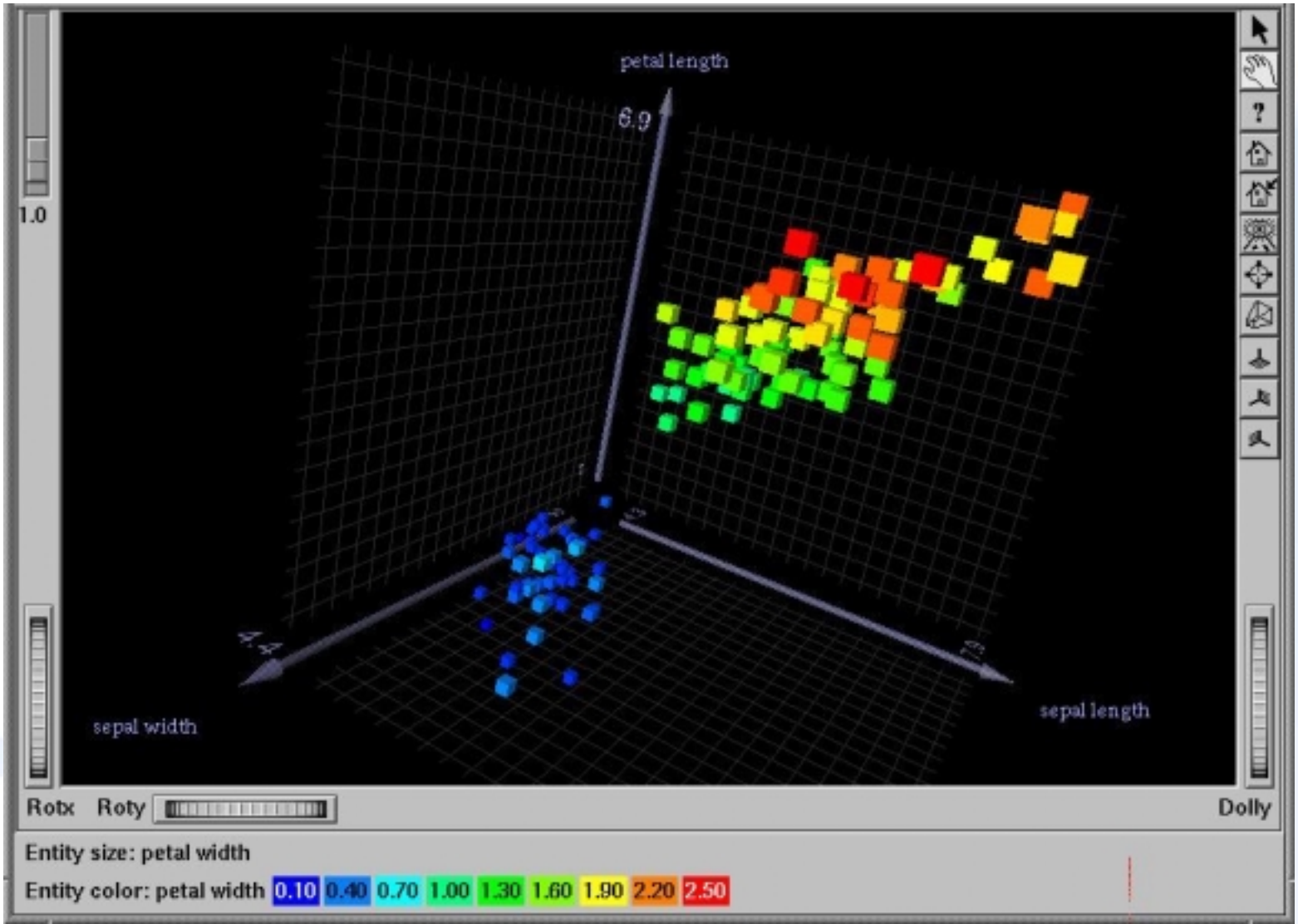


Industrial Process Control

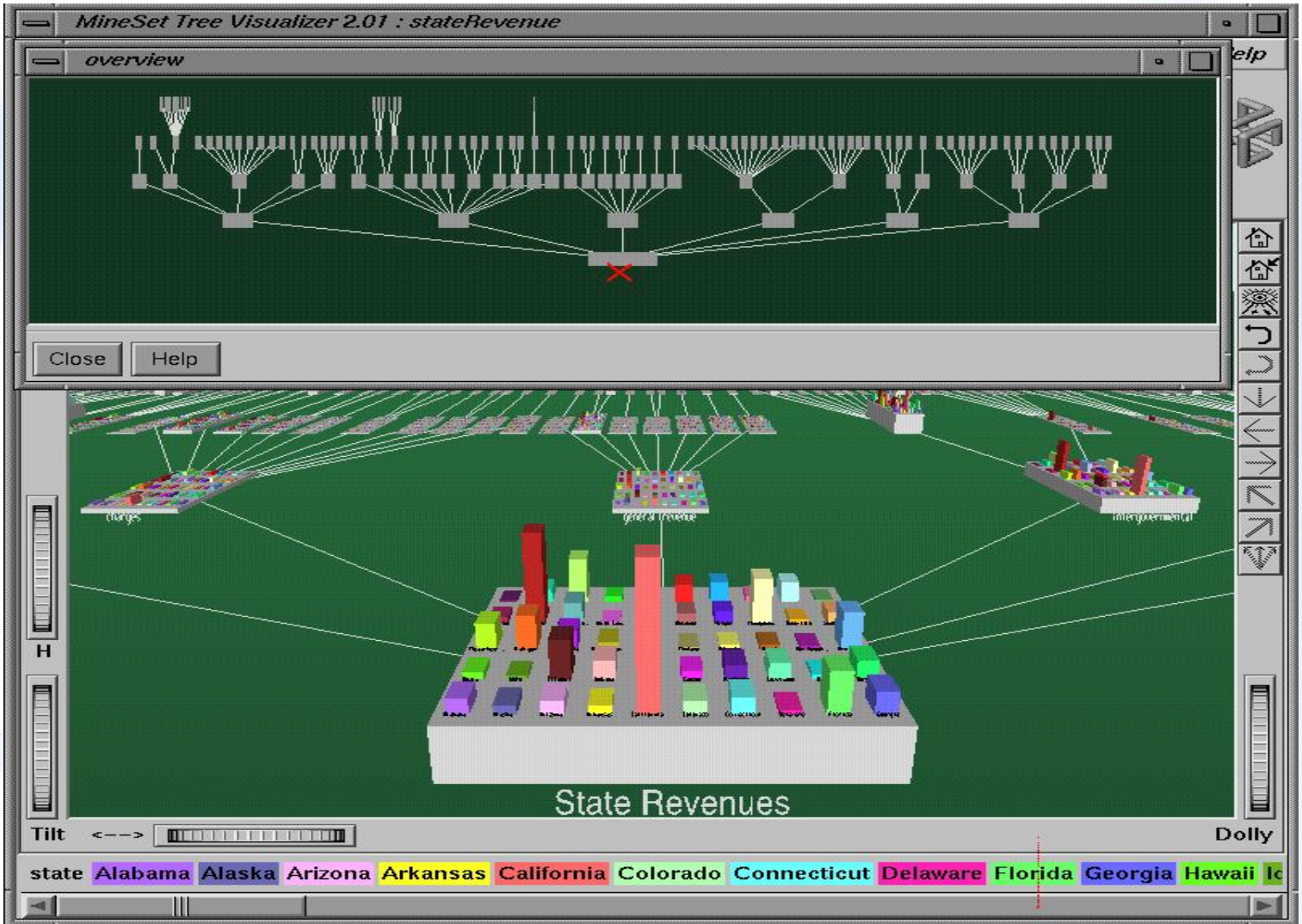
Pipeline Monitor



Scatter Visualizer



Decision Tree Visualizer



Requirements For Successful Data Mining

- There is a sponsor for the application.
- The business case for the application is clearly understood and measurable, and the objectives are likely to be achievable given the resources being applied.
- The application has a high likelihood of having a significant impact on the business.
- Business domain knowledge is available.
- Good quality, relevant data in sufficient quantities is available.

Requirements For Successful Data Mining

- The right people – business domain, data management, and data mining experts. People who have “been there and done that”

For a first time project the following criteria could be added:

- The scope of the application is limited. Try to show results within 3-6 months.
- The data source should be limited to those that are well known, relatively clean and freely accessible.

From Data Mining to Text Mining

- Techniques: linguistics analysis, clustering, unsupervised learning, case-based reasoning
- Ontologies: XML/RDF, content management
- P1000: A picture is worth 1000 words
- Formats/types: email, reports, web pages, etc.
- Integration: KMS and IT infrastructure
- Cultural: rewards and unintended consequences