Data Mining Meets E-Business: Opportunities and Challenges

Umeshwar Dayal
(with colleagues from the Data Mining Solutions and
E-Business Process Management research groups)
Hewlett-Packard Labs.
Palo Alto, CA
dayal@hpl.hp.com

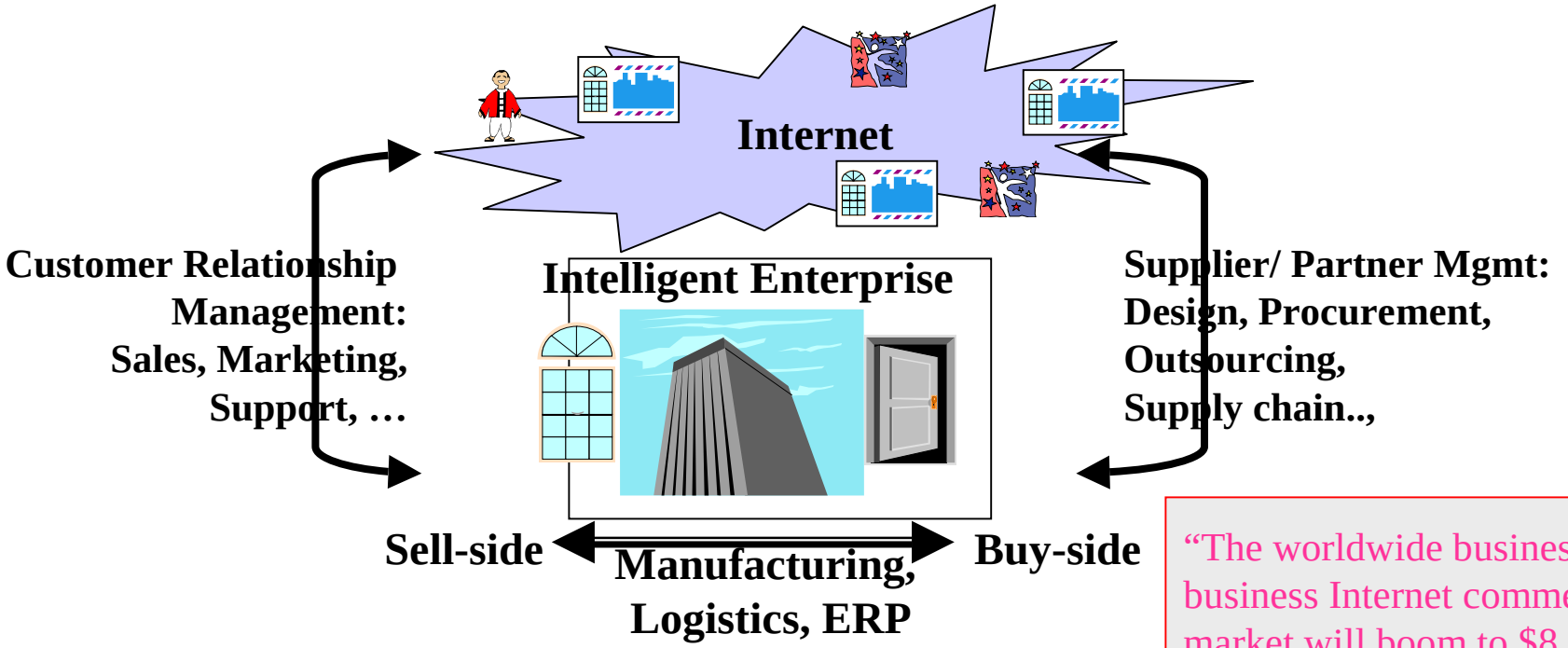
e

Outline

- Context: The E-Business Landscape and Data Mining Opportunities
- Four Cases
  - Customer Relationship Management
  - Catalog Creation and Service Discovery
    - Text Categorization
    - Information Extraction from Semi-structured Text
  - Business Process Intelligence
- Conclusions

e

The E-Business Landscape



An Intelligent Enterprise in the E-Services Marketplace must achieve Automation, Integration, and Optimization across all customer relationship, supply chain, and internal business processes by:
gathering, managing, and analyzing large amounts of data on its customers, products, services, operations, suppliers, and partners, and all the transactions in between.

"The worldwide business-to-business Internet commerce market will boom to $8.5 trillion in 2005 despite economic slowdowns. B-to-B Internet commerce sales totaled more than $433B in 2000, up 189% from 1999, and are expected to more than double to $919B this year." [Gartner Report].

e

Data Mining Landscape

- Commercial activity: Has shifted from horizontal software and toolkits to vertical applications, system integration, and services.
- Many data mining opportunities exist for the intelligent enterprise in the e-business marketplace
  - Intelligent customer relationship management: segmentation, personalization, marketing, support
  - Supply chain management: procurement, dynamic discovery & bundling of services, pricing
  - End-to-end optimization of business processes: customer demand through ERP & manufacturing to procurement
- Research: Must shift from obsession with algorithms to developing solutions enriched by data mining (“invisible, embedded data mining”, “closing the loop”).

What Industry Analysts Are Saying

- **Top CIO Priorities 1999** (*Gartner Group*)

Business	Technical
Improve Customer Service Capabilities	Build Intranet & Extranet Capabilities
Develop New Distribution Channels	Exploit Data Warehousing & Data Mining
Improve Targeted Marketing Abilities	Implement E-Commerce
Enable Knowledge Transfer	Build IT Infrastructure
Streamline Internal Business Processes	Improve network and system security

- **Market demand is very large**
  - E-Intelligence spending in 2003 estimated to be $31B (*IDC*)
  - It is the next wave in IT spending…will eventually reach or exceed the ERP market (*Merrill Lynch*)
  - CRM analytic application market forecast to grow at 54.1% per year through 2003 (*IDC*)
  - By 2002, the number of data mining projects will grow more than 300% to improve customer relationships and help enterprises listen to their customers (*Gartner Group, 1999*)
    - Interactive personalization
    - Text mining
    - Resource optimization
  - By 2003, at least 90% of all consumer-intensive industries with e-point-of-service/sales will utilize data mining models to predict customer preferences (*Gartner Group, 1999*).

e

Challenges

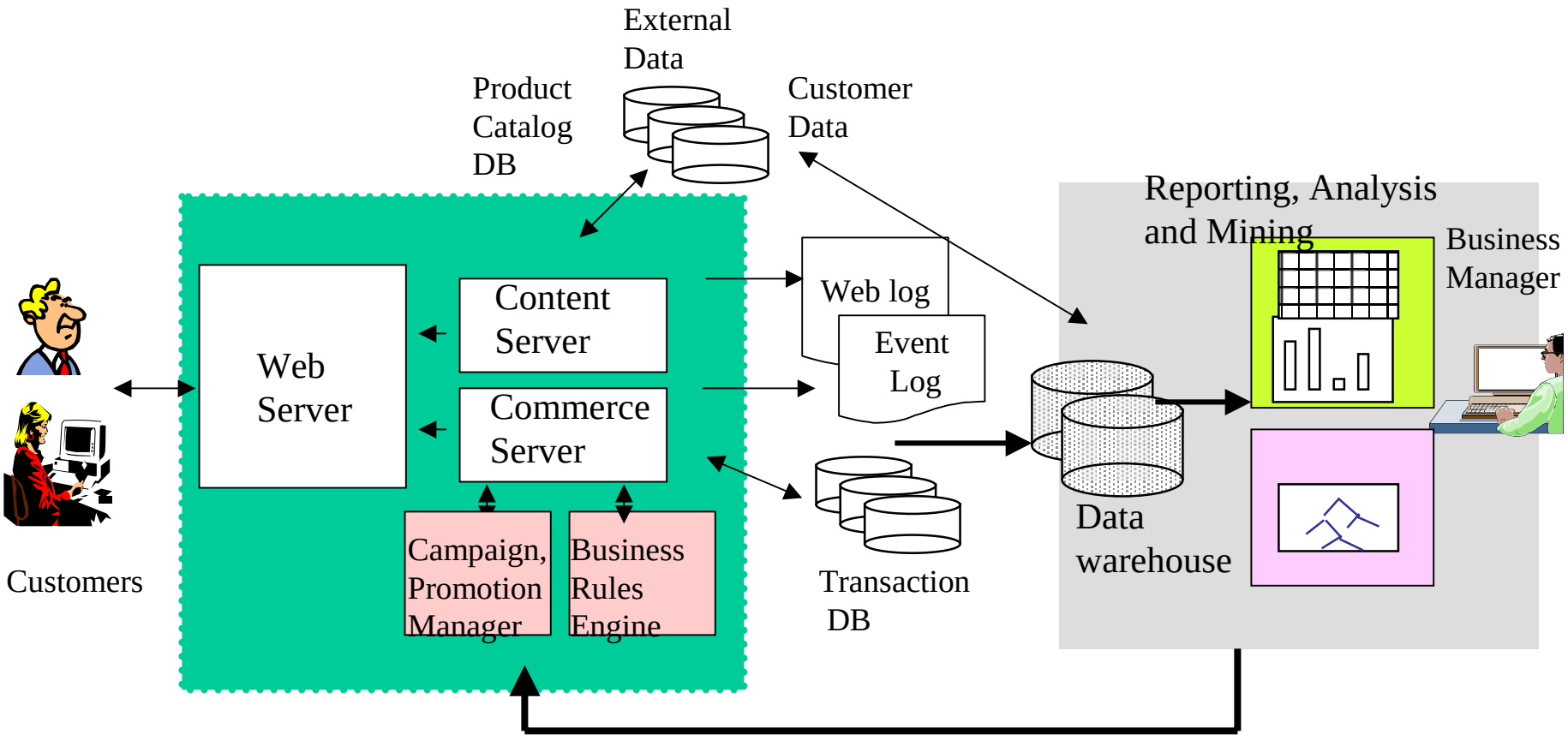
- Scalability: Very high data volumes and data flow rates
  - Large retail site: 35000 products, 4.2 billion transactions, tens to hundreds of TBs per year
  - Have to consider scalability of the whole architecture
- Complex, structured, semi-structured, and unstructured data
- Data extraction, cleaning, and consolidation from many sources
  - Integrate data warehousing, on-line analytical processing (OLAP), and data mining.
- Interactive, on-line mining
  - Incorporate real-time data streams, "live" updates, user interactions
  - Incremental analysis
  - Interactive visualization
- Integrate into complete solutions
  - Use results of analysis and mining for decision making, e.g., marketing campaigns, adapting business processes, supply chain optimization

e

Outline

- Context: The Intelligent Enterprise, E-Business, and Data Mining Opportunities
- Four Cases
  - Customer Relationship Management
  - Catalog Creation and Service Discovery
    - Text Categorization
    - Information Extraction from Semi-structured Text
  - Business Process Intelligence
- Conclusions

Case 1: Intelligent Customer Relationship Management



Product/page recommendations
Target marketing, promotions

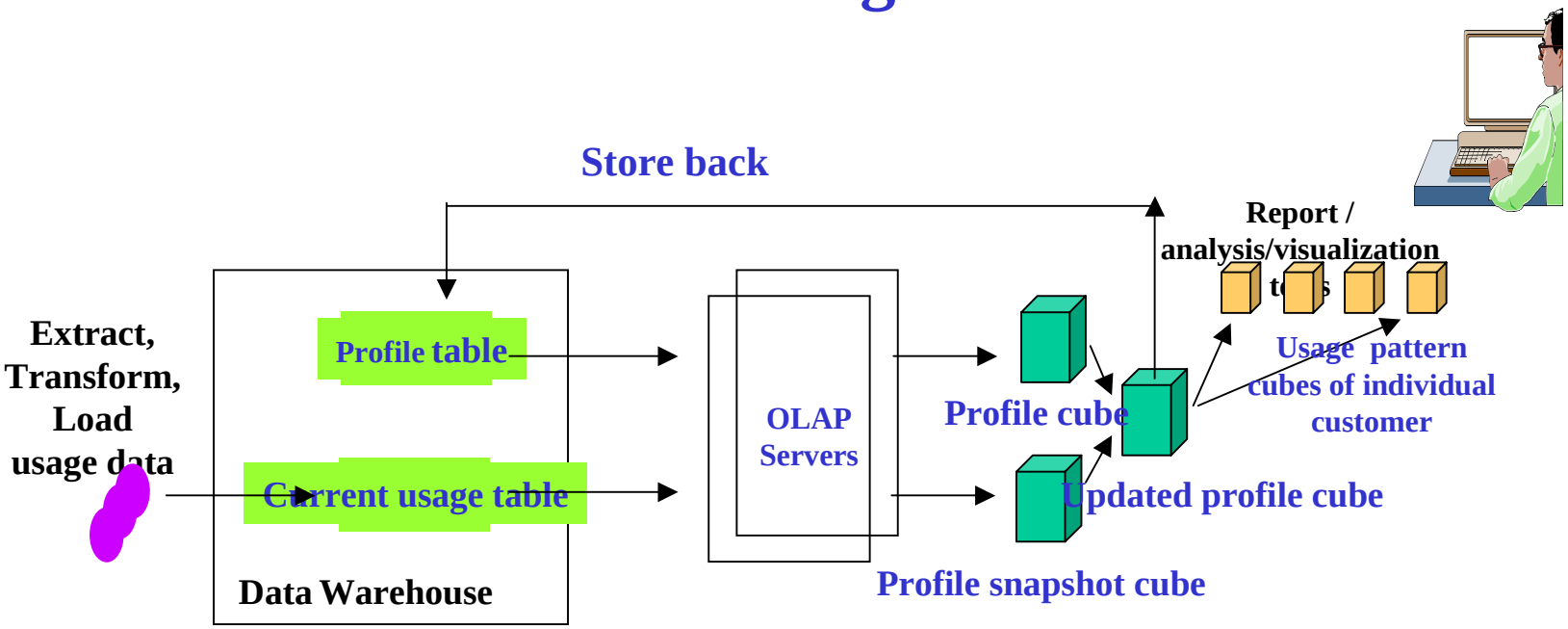
Customer profiling
Customer/market segmentation
Product affinity analysis

e

Data Mining for Intelligent CRM

- Data Sources:
  - web logs: page accessed, IP address, time, referring site, bytes, …
  - event logs: ads seen, products seen, products added to shopping cart, products bought, abandoned shopping carts, …
  - transaction database: customer id, products ordered, time, quantity, price, …
  - query logs: search terms used, documents returned, …
- Types of analysis
  - Multidimensional analysis (profiling)
  - Association rules (product affinities)
  - Clustering, classification (segmentation)
  - Similarity (collaborative filtering)

OLAP-Based Profiling Architecture



- Typically, OLAP (On-Line Analytical Processing) is used as a front-end tool for analysis.
- OLAP servers provide memory mgmt, efficient computation over data cubes.
- Traditionally, intended for relatively static operation: periodic batch refresh of the warehouse, re-compute data cubes, re-evaluate queries and reports.
- We use OLAP servers as data summarization engines in a computational pipeline.

Q. Chen, M. Hsu, U. Dayal "OLAP-Based Scalable Profiling of Customer Behaviour", First Intl. Conf. On Data Warehousing and Knowledge Discovery (DAWAK) 1999.

e

OLAP: Operations on Data Cubes

- Represent data by multidimensional cubes: (hierarchical) dimensions and measures

- Dice, slice: Select a sub cube , e.g., sales where city = LA & month = Jan98

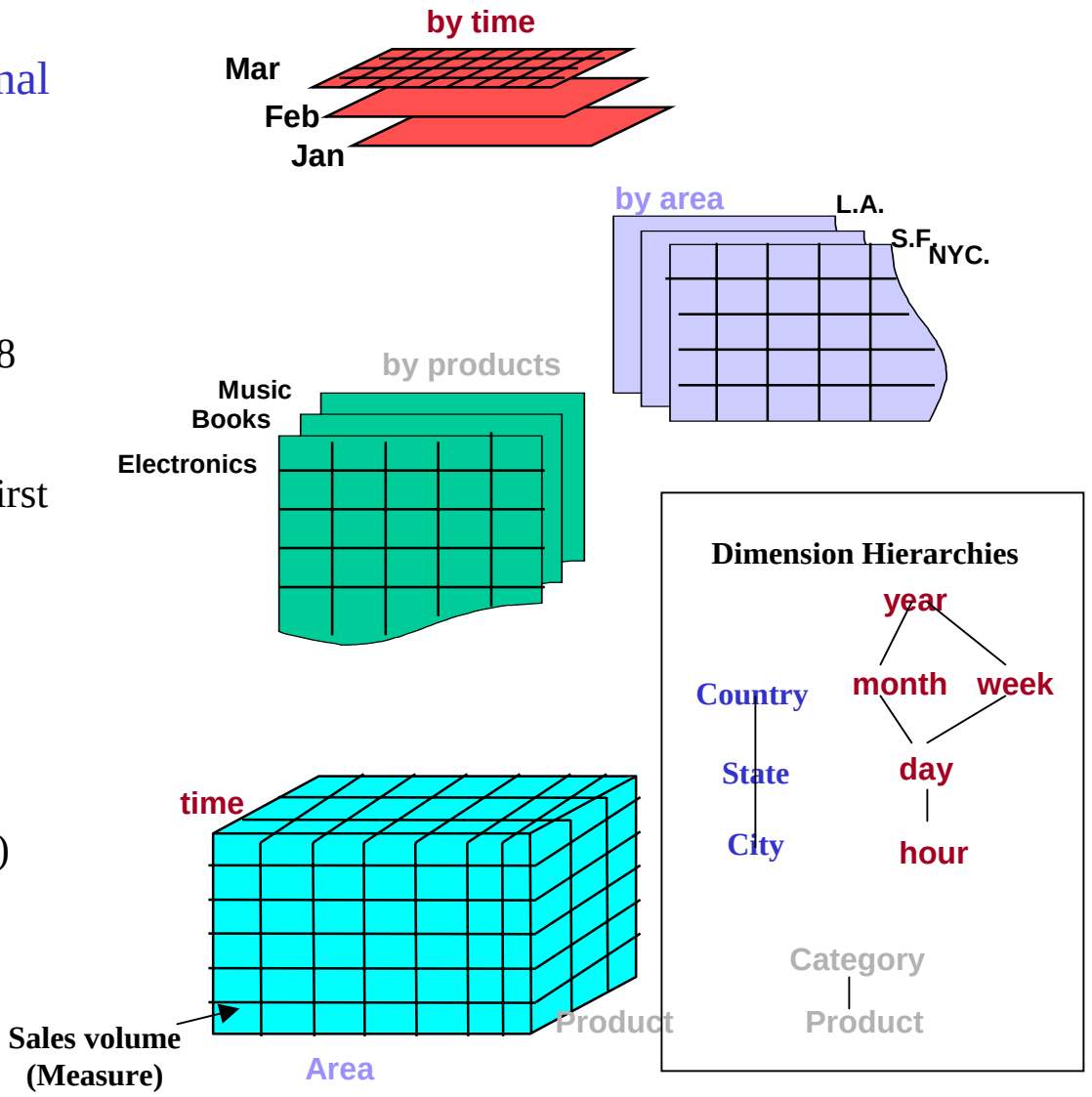
- Roll-up (summarize), drill-down (detail): e.g., Total sales of books for first quarter ‘98 in CA

- Ad-hoc queries

- Flexible report types

- Powerful derivations: Get derived measures, e.g., profit = (sales - expense) across all dimensions

- Ranking: e.g. top 10% of cities by average quarterly sales of books



e

OLAP-Based Mining

- Enables powerful analysis and multi-level summarization of e-commerce data.
- Scalable to large data volumes and data flow rates.
- Supports continuous, incremental analysis:
  - Use OLAP server as a compute engine: create only those cubes that are needed (can think of cubes as materialized views over data in the warehouse); use only those dimensions that are needed for particular analyses; use binning to reduce the cardinality of the dimensions.
  - Store back results persistently in the data warehouse (RDB) to overcome data size limitations.
- OLAP scripts as high-level language for multi-dimensional, multi-level data mining.
- Model customer profiles, patterns, similarity measures, association rules as cubes
  - compute efficiently using cube operations in the OLAP server
  - evolve incrementally in real-time as new data flows in
  - multi-dimensional, multi-level analysis over cubes provides enhanced expressive power (e.g., richer association rules) by integrating OLAP style drill down, rollup operations with data mining tasks.

Cube-based Associations

- Association rules are represented as cubes
  - can be generated by cube operations
  - can be maintained as cube cells
  - Scalable to large data sets
- Allows definition of new kinds of multilevel, multidimensional association rules with enhanced expressive power
  - scoped association rules based on different elements

    cross-sale rule based on transactions (traditional shopping basket analysis)

    $x \in Transactions: contain\_product(x, A) \Rightarrow contain\_product(x, B)$

    cross-sale rule based on customers (regardless of whether purchased in the same transaction)

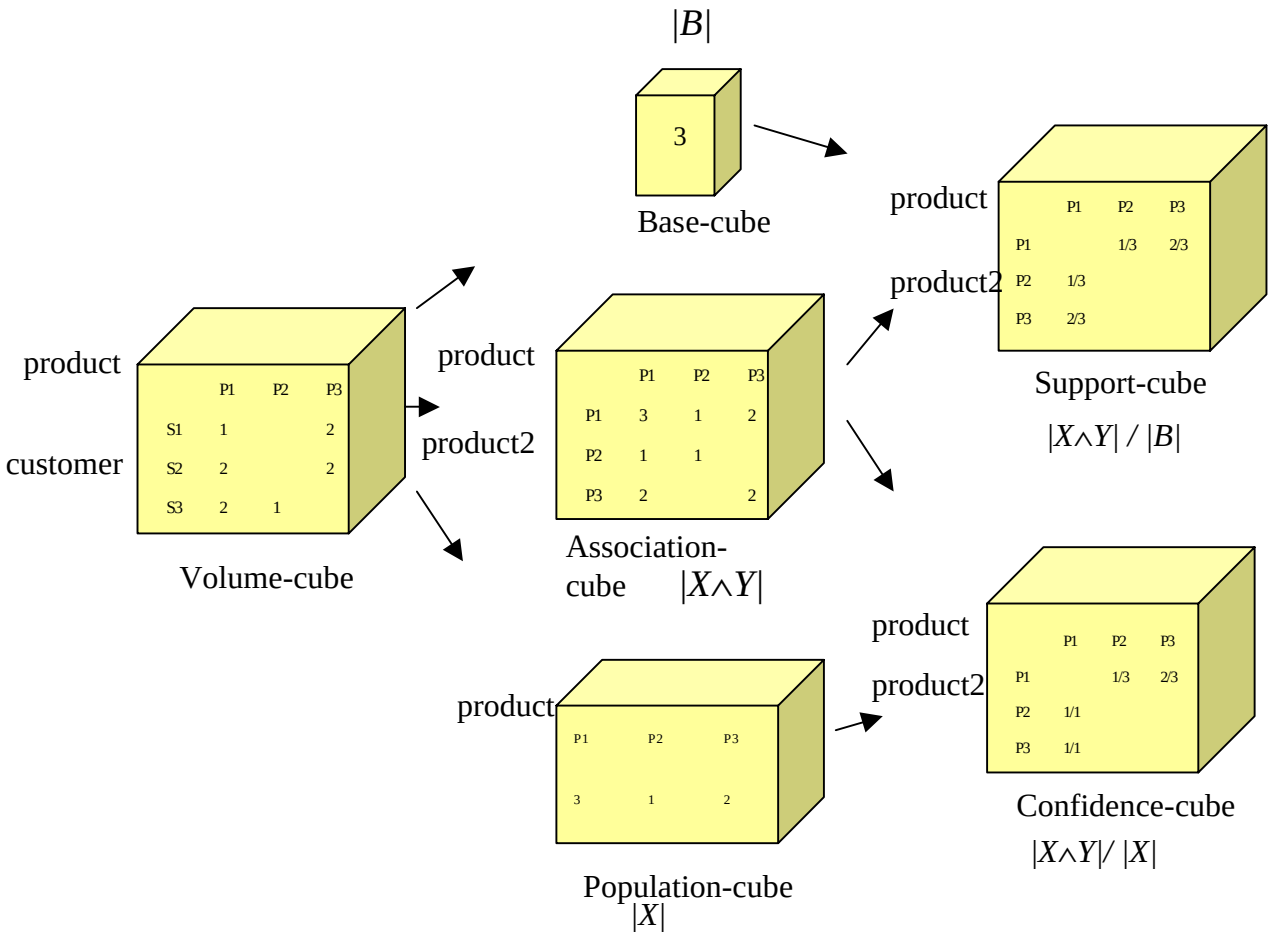
    $x \in Customers: buy\_product(x, A) \Rightarrow buy\_product(x, B)$
  - **multidimensional rule**

    $[x \in Customers: buy\_product(x, 'A') \Rightarrow buy\_product(x, 'B')] \mid customer\_group = 'engineer', area = 'Los Angeles', time = 'Jan98'$
  - **high-level rule**

    $[x \in Customers: buy\_product(x, 'A') \Rightarrow buy\_product(x, 'B')] \mid customer\_group = 'engineer', area = 'California', time = 'Year98'$

e

Cube-Based Association Rule Mining

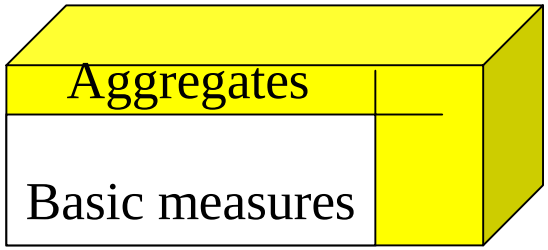


e

OLAP-based Profiling

- Scalability challenges
- Huge data volumes and data flow rates: a busy e-commerce site can generate hundreds of millions of events per day.
  - Solution: Scale using parallel loading and analysis
- Fine-grained analysis (e.g., individual customer profiling) requires very large, very sparse cubes
  - Example: a newspaper web site had 48,128 customers * 10,432 referring sites * 18,085 pages * 24 hours per day => ~200 trillion cells!
  - Compressed for storage, but cube rollup operation very slow (~10,000 hours!)
  - Solution: careful design + optimizations yielded 3-4 orders of magnitude improvement.

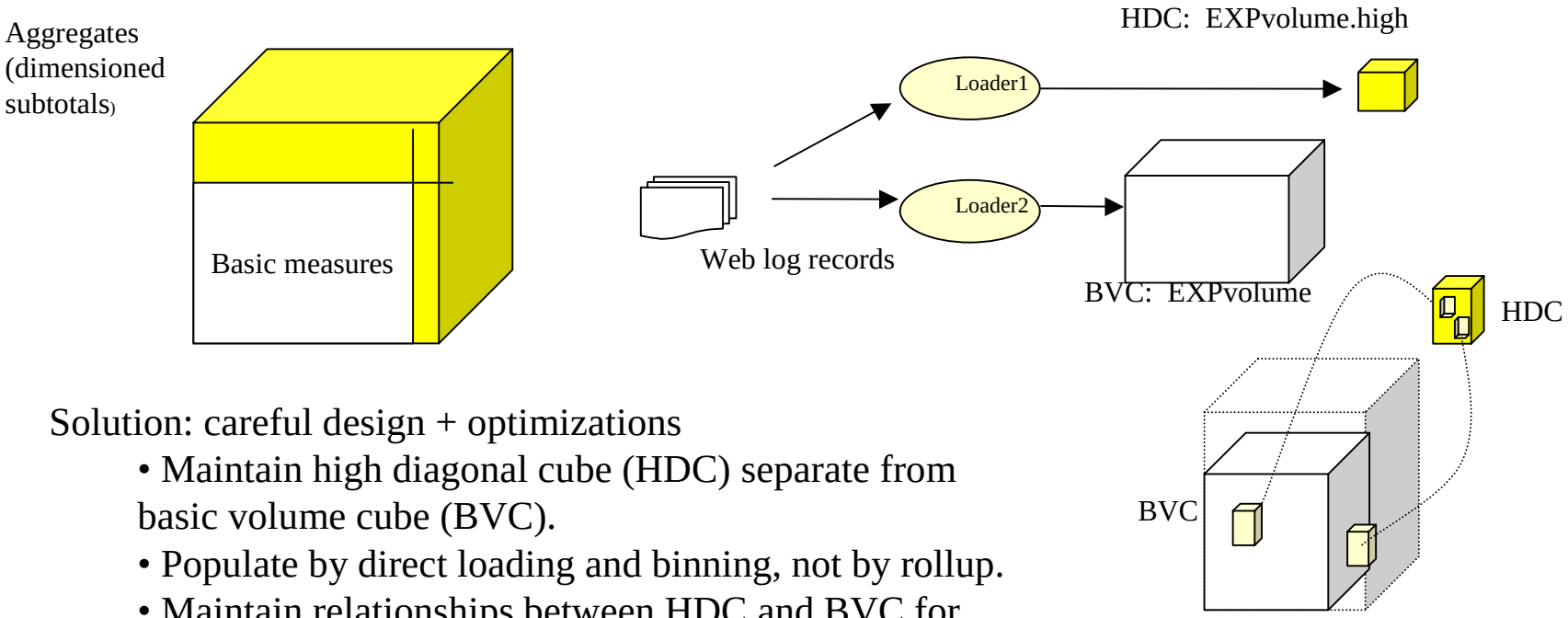
Scalability of Cube Rollup



- Dimension hierarchies
  - ip : 63.211.140.164 →origin : CA
  - uri: exp.com/TODAY/topstory.html →subject: exp.com/TODAY/
- Typical cube rollup operation (embedded total)
  - When original cube has multiple large-sized dimensions, a large number of additional cells are needed to hold the embedded-total.
  - In the above example, these sub-totals occupy approximately 50 trillion cells in the rolled up cube, out of a total of 267 trillion cells.
  - While the OLAP engine compresses sparse cubes for efficient storage, the cells containing nulls must be checked in some way during the rollup operation.
- Rolling up such a cube as a whole is impractical.

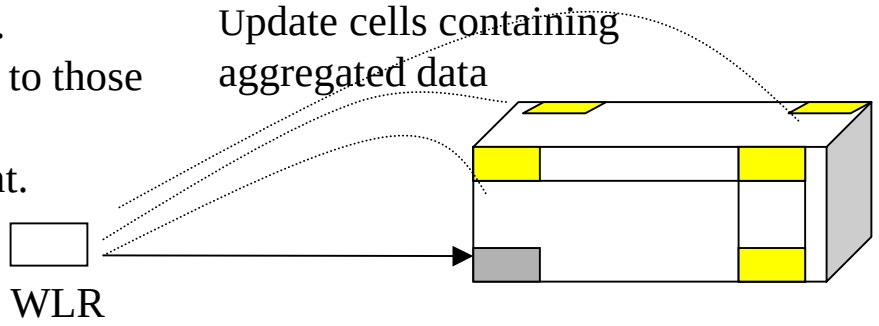
e

Scaling: Huge, Sparse Cubes



Solution: careful design + optimizations

- Maintain high diagonal cube (HDC) separate from basic volume cube (BVC).
- Populate by direct loading and binning, not by rollup.
- Maintain relationships between HDC and BVC for drilldown.
- .Compute intermediate aggregates on demand.
- High-profile cubes: limit dimension elements to those corresponding to cells with large counts.
- Yielded 3-4 orders of magnitude improvement.



Q. Chen, U. Dayal, M. Hsu, "An OLAP-Based Scalable Web Analysis Engine", Proc. 2nd Intl. Conf. on Data Warehousing and Knowledge Discovery (DAWAK) 2000.

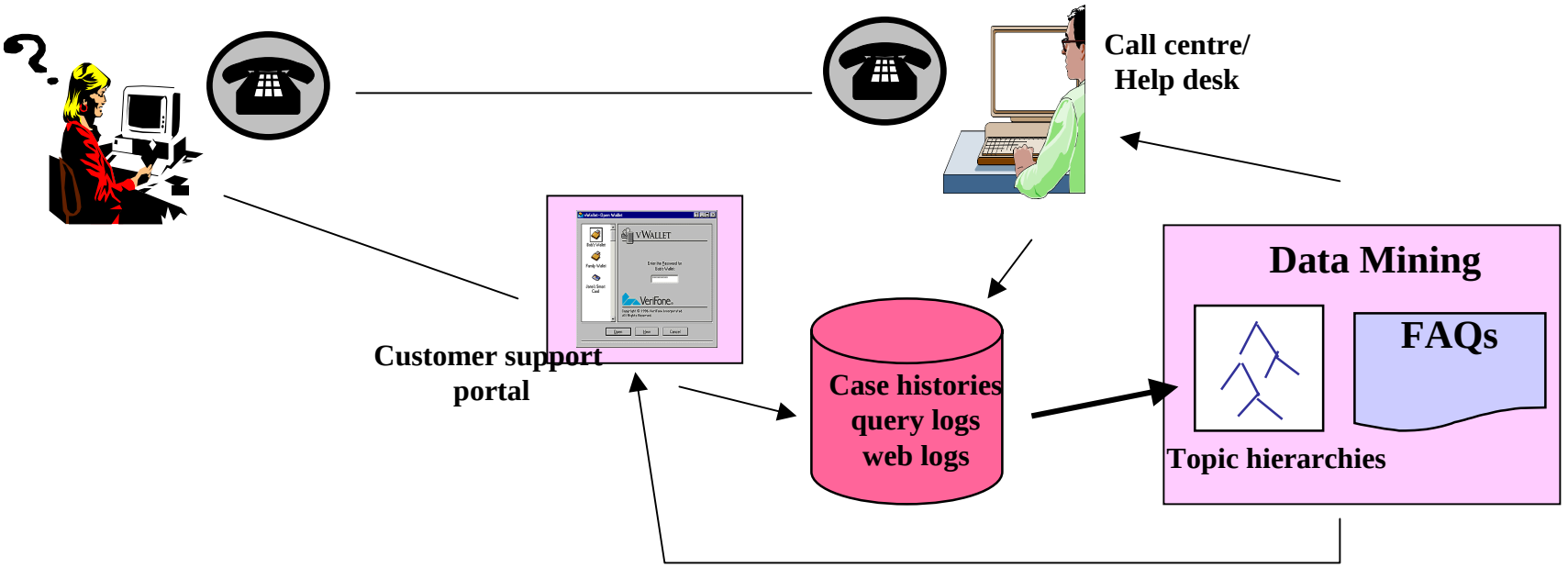
e

Outline

- Context: The Intelligent Enterprise, E-Business, and Data Mining Opportunities
- Four Cases
  - Customer Relationship Management
  - Catalog Creation and Service Discovery
    - Text Categorization
    - Information Extraction from Semi-structured Text
  - Business Process Intelligence
- Conclusions

e

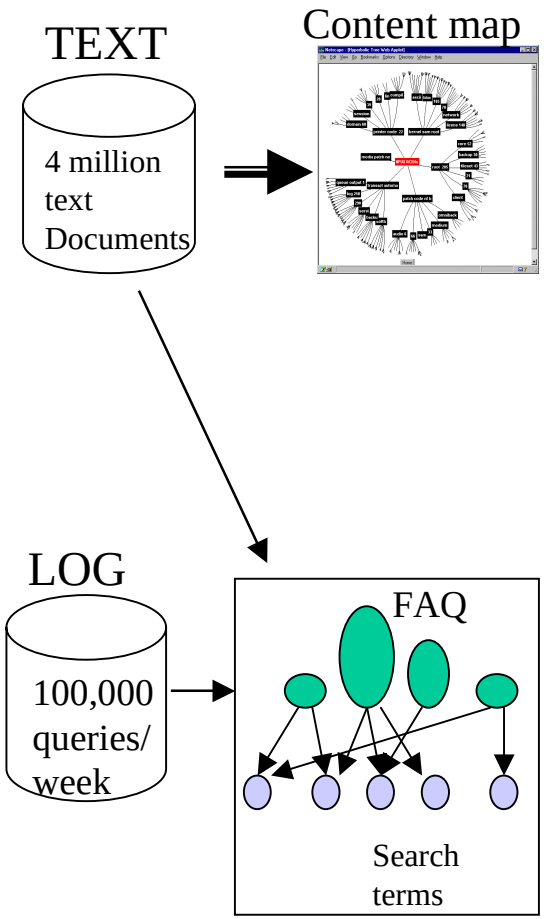
Case 2: Text Categorization

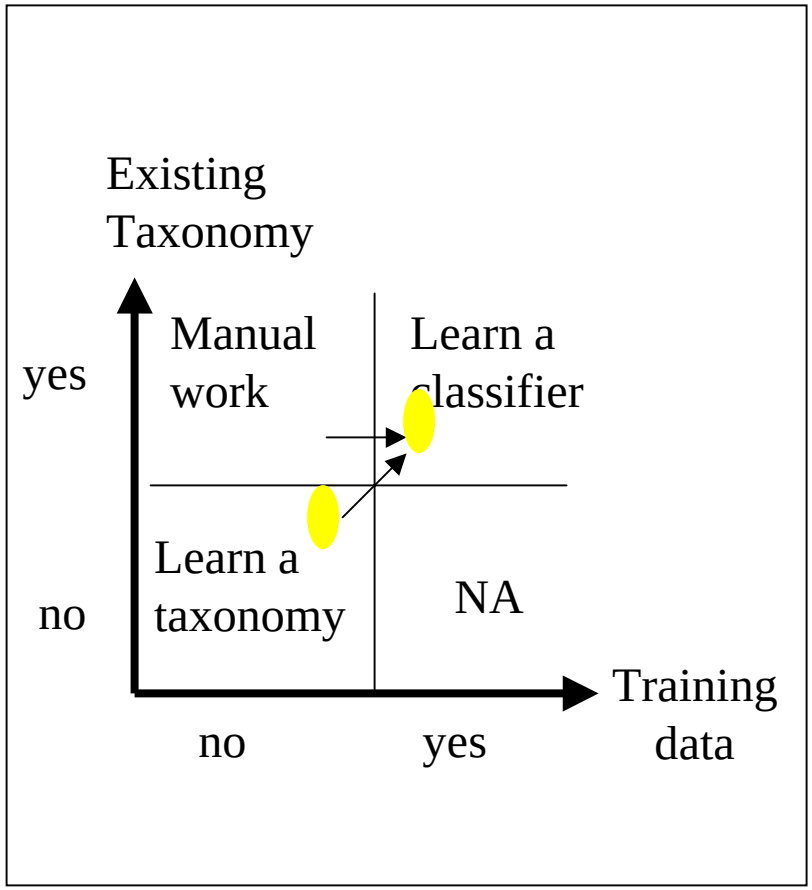


- Mine content and usage data
  - Automatically build topic hierarchy and categorize documents to assist in search.
  - Extract problems/ FAQs, and recommend relevant documents.

e

Text Categorization Framework

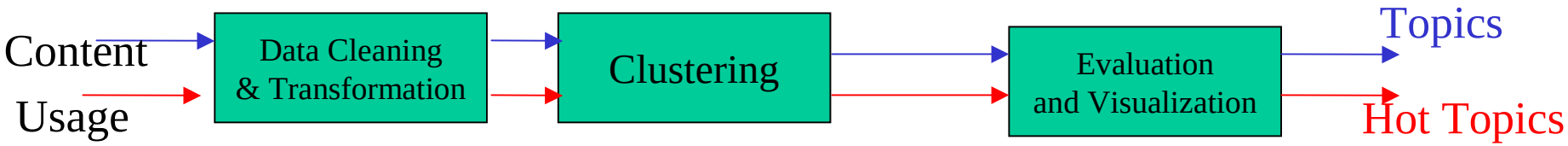




e

Topic Hierarchy Creation & Text Categorization

- Mine content and usage (query logs) data
  - Automatically build topic hierarchy and categorize documents to assist in search.
  - Extract problems/ FAQs and relevant solution documents, and place them on topic hierarchy.



- Extract key words and phrases
- Transform documents and query log records into vectors
- Cluster hierarchically
- Label each cluster with significant words, phrases
- Visualize as hyperbolic tree for navigation/browsing

e

Challenges in Text Categorization

- Problem: Docs are noisy, conversational, not well structured, replete with typos, abbreviations, jargon, unconventional text (e.g., code fragments, tables)
- Difficult issues:
  - Normalization and cleaning
  - Sentence boundary detection & extraction of most significant sections of the document
  - Feature selection
  - Scalable, incremental, robust clustering algorithms
  - Clustering techniques were effective in producing leaf nodes of the taxonomy
  - Hierarchical clustering to produce higher nodes of the taxonomy proved very difficult
  - Labeling the nodes of the taxonomy (with terms that are semantically meaningful to humans) proved very difficult
- Data mining as an aid to human experts, e.g., suggestions for expanding or modifying a taxonomy, generating “hot topics” for placement in a taxonomy, generating cross-index terms.

e

Toolkit for Normalization and Summarization

	Anomaly	Effect	Functionality Required	Tools**
Stage 1 (General Cleaning)	Typos Misspellings Abbreviations	False word occurrences	Unify representation of words	- Thesaurus Assistant -Normalizer
Stage 2 (Task-specific Cleaning)	Code Dumps Cryptic tables	Complicate sentence identification possibly w/o adding value	Removal* of code, dumps and tables	- Code Remover - Table Remover
Stage 3 (Extrac-tion)	---	---	Obtain summary	- Sentence Identifier - Sentence Scorer

M. Castellanos, J. Stinger: "A Practical Approach to Extracting Relevant Sentences in the Presence of Dirty Text", SIAM Data Mining Workshop on Text Mining, April 2001.

e

Thesaurus Generation for Feature Engineering

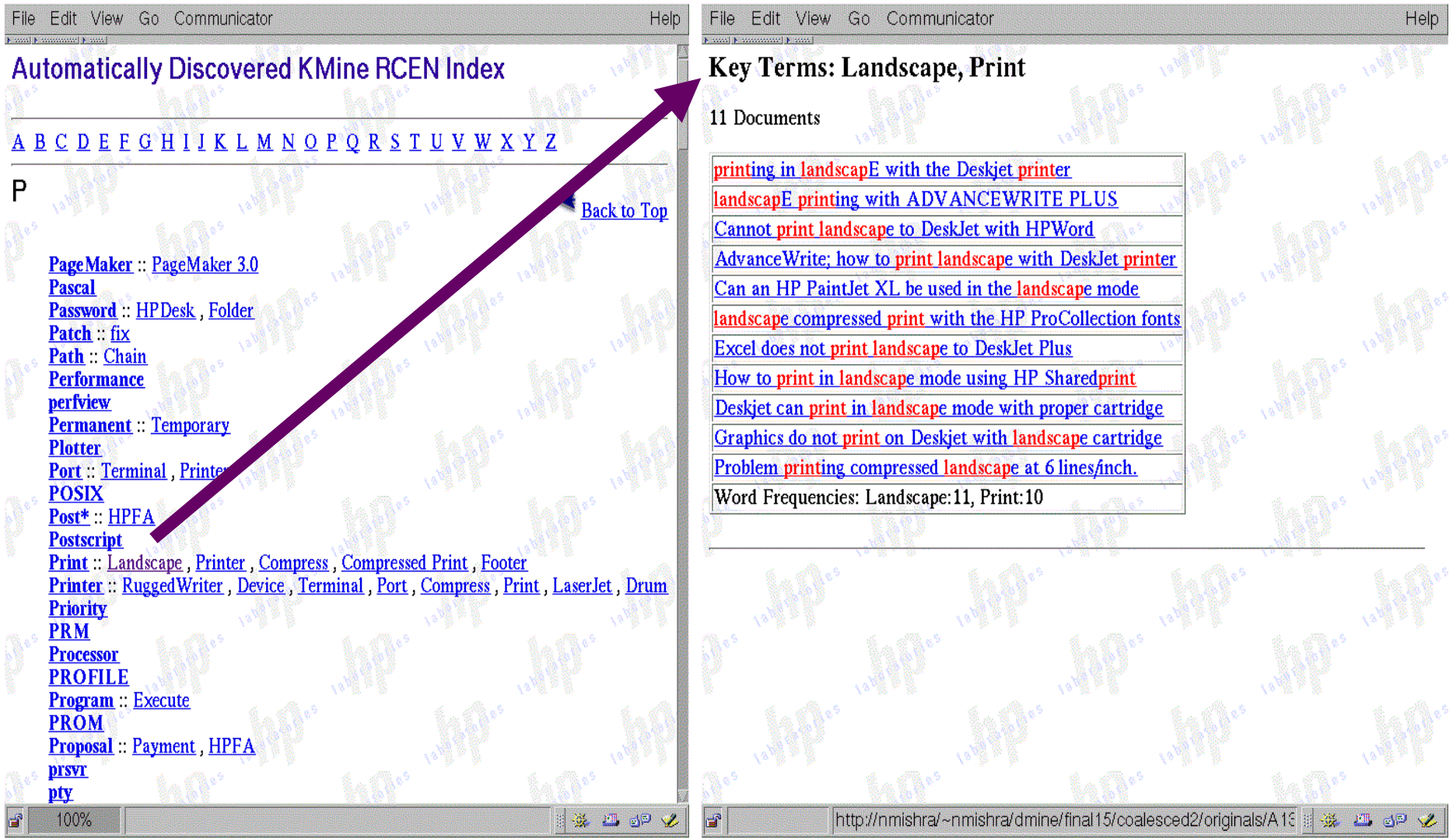
- In many text mining techniques, the basic ingredient is the frequency of occurrence of words
- Typos, misspellings, abbreviations mislead the results
  - different orthographic representations for same “word” will be taken as different words
- unless… we add a “clean-up” preprocessing step to the text mining task: *normalization*

omniback
omni back

desc omniback
11.0 omniback
omniback
omni back
omniback 3.0
10.20 omniback
omniback 3.00
omniback 3.1
omniback 3.10
omniback ii
omnibackii
omniback2
omniback gui
omniback db
omniback emer
omnibook
omniback 2.55

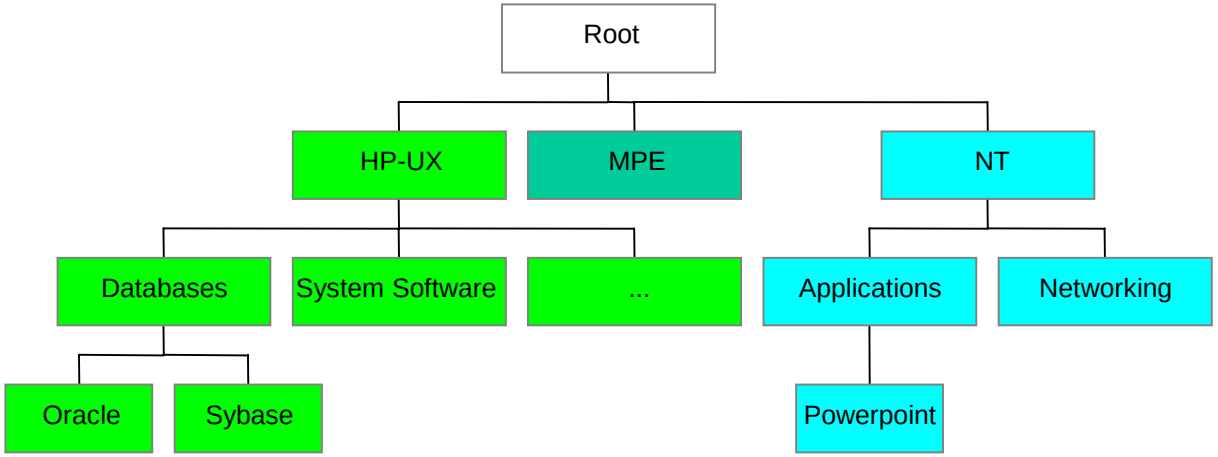
e

Automatically Indexing Document Collections



e

Hierarchical Classification



- Goal:
  - Given a clean document, find the best class for it in the topic hierarchy
  - If you misclassify a document, at least have it be somewhere reasonable
  - Some human verification / correction / training is available
    - Ideally, automate this (4,000,000 documents)
- Challenges:
  - How wrong is wrong? Evaluating coherence of the hierarchy
  - Unbalanced datasets
  - Taking advantage of the hierarchy
  - Can we avoid enormous training sets (co-training)
  - Evolution of the hierarchy

e

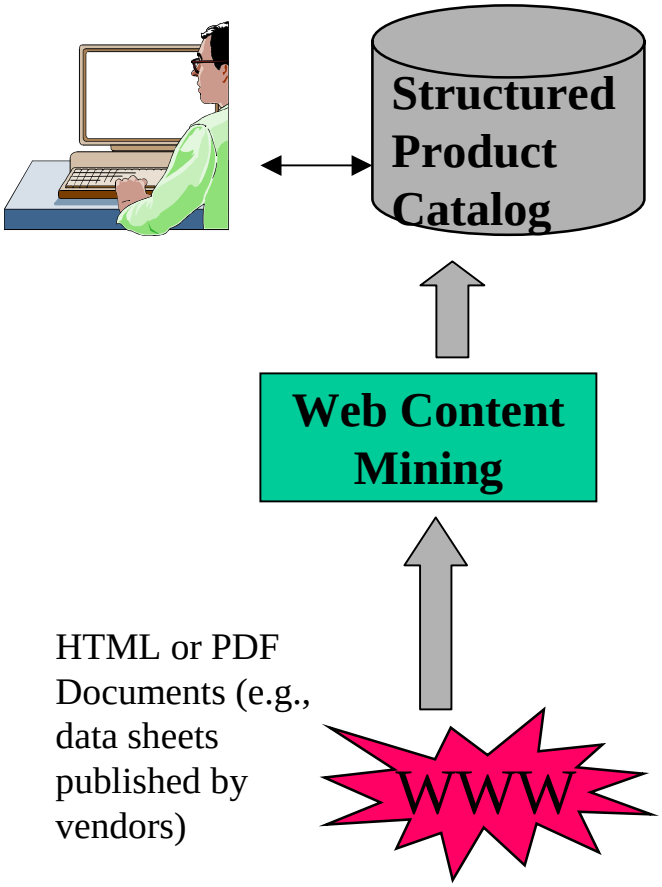
Outline

- Context: The Intelligent Enterprise, E-Business, and Data Mining Opportunities
- Four Cases
  - Customer Relationship Management
  - Catalog Creation and Service Discovery
    - Text Categorization
    - Information Extraction from Semi-structured Text
  - Business Process Intelligence
- Conclusions

e

Case 3: Information Extraction for Catalog Creation, Service Discovery

Parametric Search, Supply Chain applications, Service Discovery



" Find processor with low power consumption @ 3.3V & operating at clock speed > 50 MHz & leadtime < 6 weeks with cost < $35@qty=10000 "

- Web navigation
- Document structure recognition (e.g. table recognition in pdf)
- Attribute extraction and tagging
- XML formulation

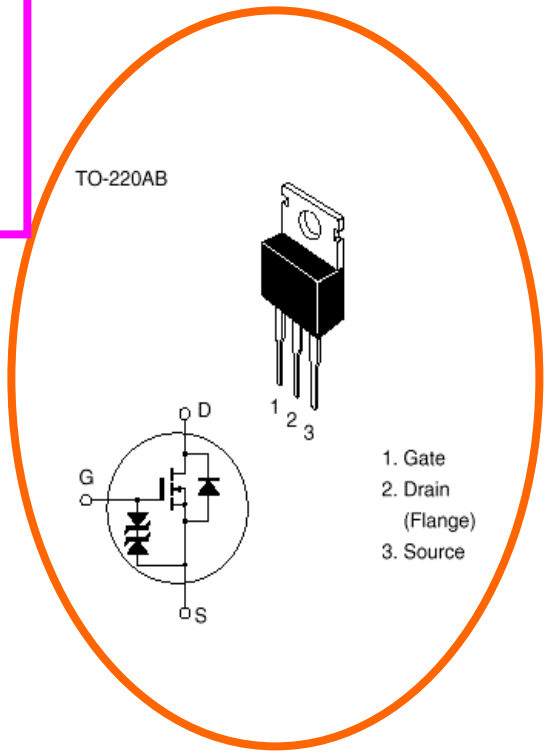
e

Problem: Attribute Values May Be Found in Free Text, Lists, Tables, Diagrams

The HB56RW832DZJ is a 8M X 32 dynamic RAM Small Outline Dual In–line Memory Module (S.O.DIMM), mounted 16 pieces of 16–Mbit DRAM (HM51W17400) sealed in TCP package. An outline of the HB56RW832DZJ is 72–pin Zig Zag Dual tabs socket type compact and thin package. Therefore, the HB56RW832DZJ makes high density mounting possible without surface mount technology. The HB56RW832DZJ provides common data inputs and outputs. Decoupling capacitors are mounted on the module board.

- Visible light output : lp = 633nm typ. (equal to He–Ne laser)
- Optical power output : 5mW CW
- Low operating voltage : 2.7V Max.
- Single longitudinal mode.
- Built–in photodiode for monitoring laser output.

Items	Symbols	Min	Typ	Max	Units	Test Conditions
Optical output power	P_O	5	––	––	mW	Kink free
Threshold current	I_{th}	20	45	70	mA	––
Operating current	I_{op}	––	55	85	mA	P_O = 5 m
Operating voltage	V_{OP}	––	––	2.7	V	P_O = 5 m
Lasing wavelength	I_p	625	633	640	nm	P_O = 5 mW

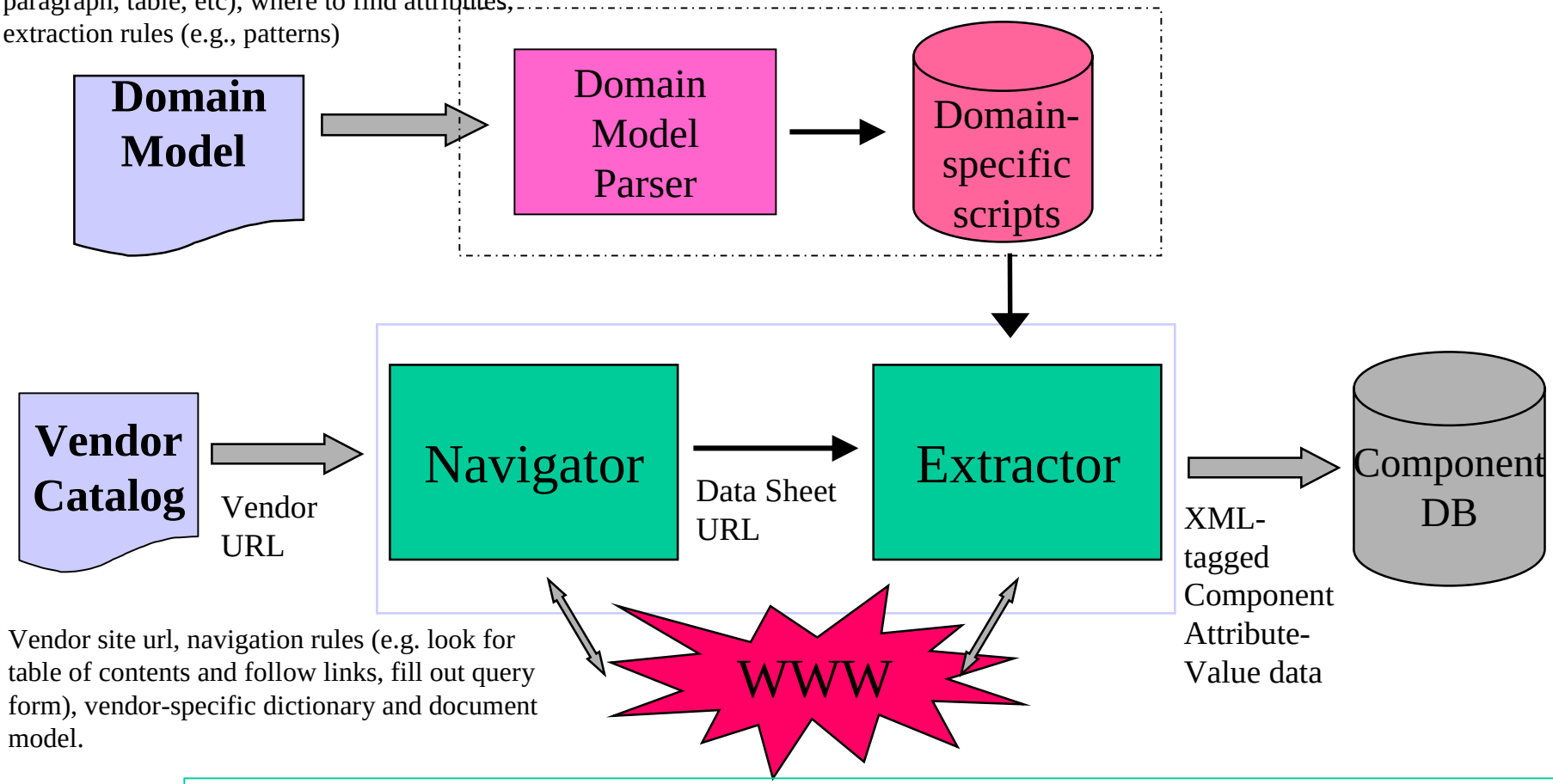


e

Solution: Model-Driven Content Mining Agents

Product concept model: Product family hierarchy, applicable attributes, thesaurus (e.g., synonyms, units, conversions)
Document model: Document structure (section, paragraph, table, etc), where to find attributes, extraction rules (e.g., patterns)

Alternative approach: wrapping web sites.
Does not work well for very heterogeneous web sites; more sensitive to restructuring of the pages; does not work with PDF content.



Vendor site url, navigation rules (e.g. look for table of contents and follow links, fill out query form), vendor-specific dictionary and document model.

M. Castellanos, J. Stinger, M. Lemon, M.Hsu, U. Dayal, P.Siegel "Component Advisor: a tool for automatically extracting electronic component data from Web datasheets." WWW7 Workshop on Reuse of Web-based Information, April 1998.

e

Extraction from Data Sheets -- Problems

- First identify hidden structures (tables, lists, paragraphs) in the data. For HTML tagged documents, this is easier than for PDF documents. But 95% of the data sheets are in PDF.
- Existing PDF to HTML/XML conversion tools have font and formatting problems, and do not handle tables.
- Content mining agent combines several heuristics
  - Font analysis: exploit cues inherent in font usage to detect potential section headings, row and column labels in tables, etc.
  - Image analysis: histograms of pixel density
  - Geometric analysis: spacing between words on a line, lining up of words in columns, etc.

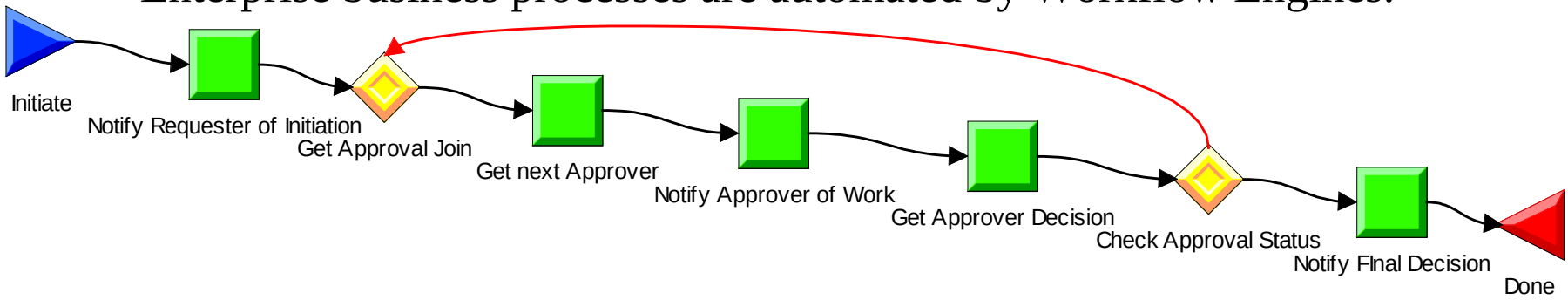
Outline

- Context: The Intelligent Enterprise, E-Business, and Data Mining Opportunities
- Four Cases
  - Customer Relationship Management
  - Catalog Creation and Service Discovery
    - Text Categorization
    - Information Extraction from Semi-structured Text
  - Business Process Intelligence
- Conclusions

e

Case 4: Business Process Intelligence

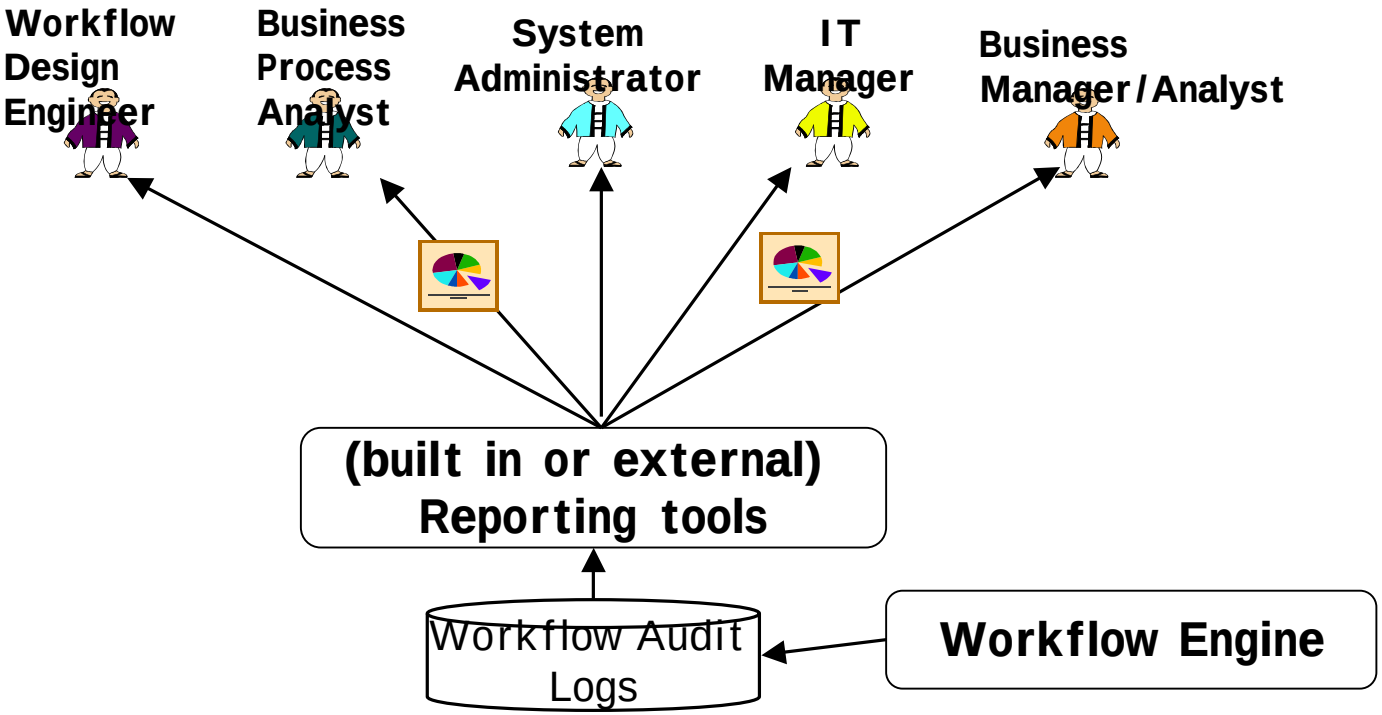
- Goal: improving the quality of enterprise business processes & services
  - *Internal* quality, as perceived by the service provider (e.g. reduced operating costs)
  - *External* quality, as perceived by the user (e.g., better service)
- Enterprise business processes are automated by Workflow Engines.



- These engines monitor many aspects of process execution and service delivery
  - Who does what, when, how long do they take
- Record data in audit logs that can be used to analyze, understand, and optimize processes.

e

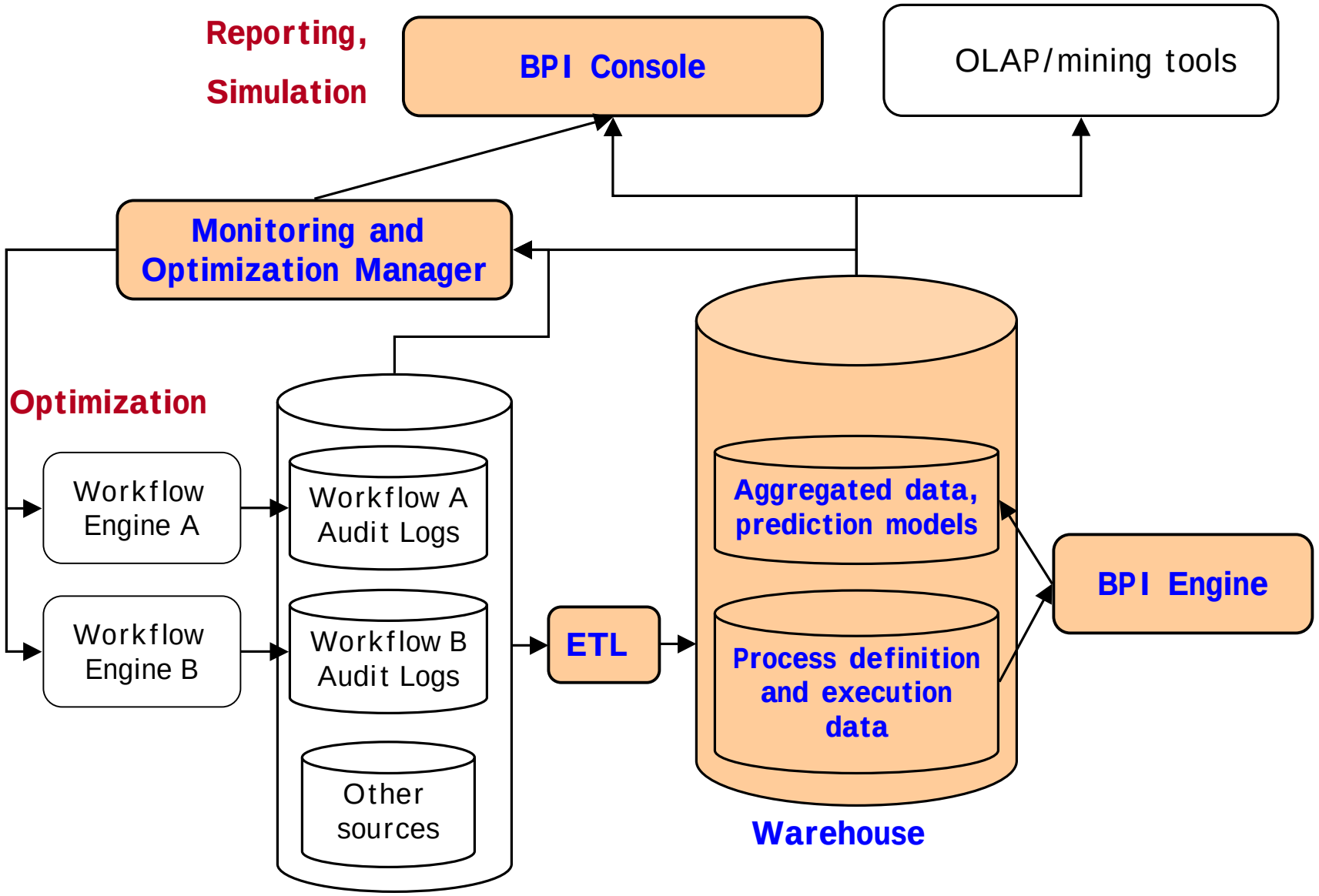
Current Situation: Reporting Tools



- Writing the “right” queries is very difficult and time-consuming
  - What is the performance and outcome of activities executed on Fridays?
  - Which resources perform best for a given activity?
  - How does the relative performance of a resource change as a function of time?
- Dirty data, missing values, special codes
- Query performance is poor: complex queries involving joins and aggregation
- Little support for integrating other data sources or multidimensional analysis
- No support for understanding the causes of problems, predicting problems, or optimizing processes.

e

Business Process Intelligence



e

Example Application: Exception Analysis, Prediction, and Prevention

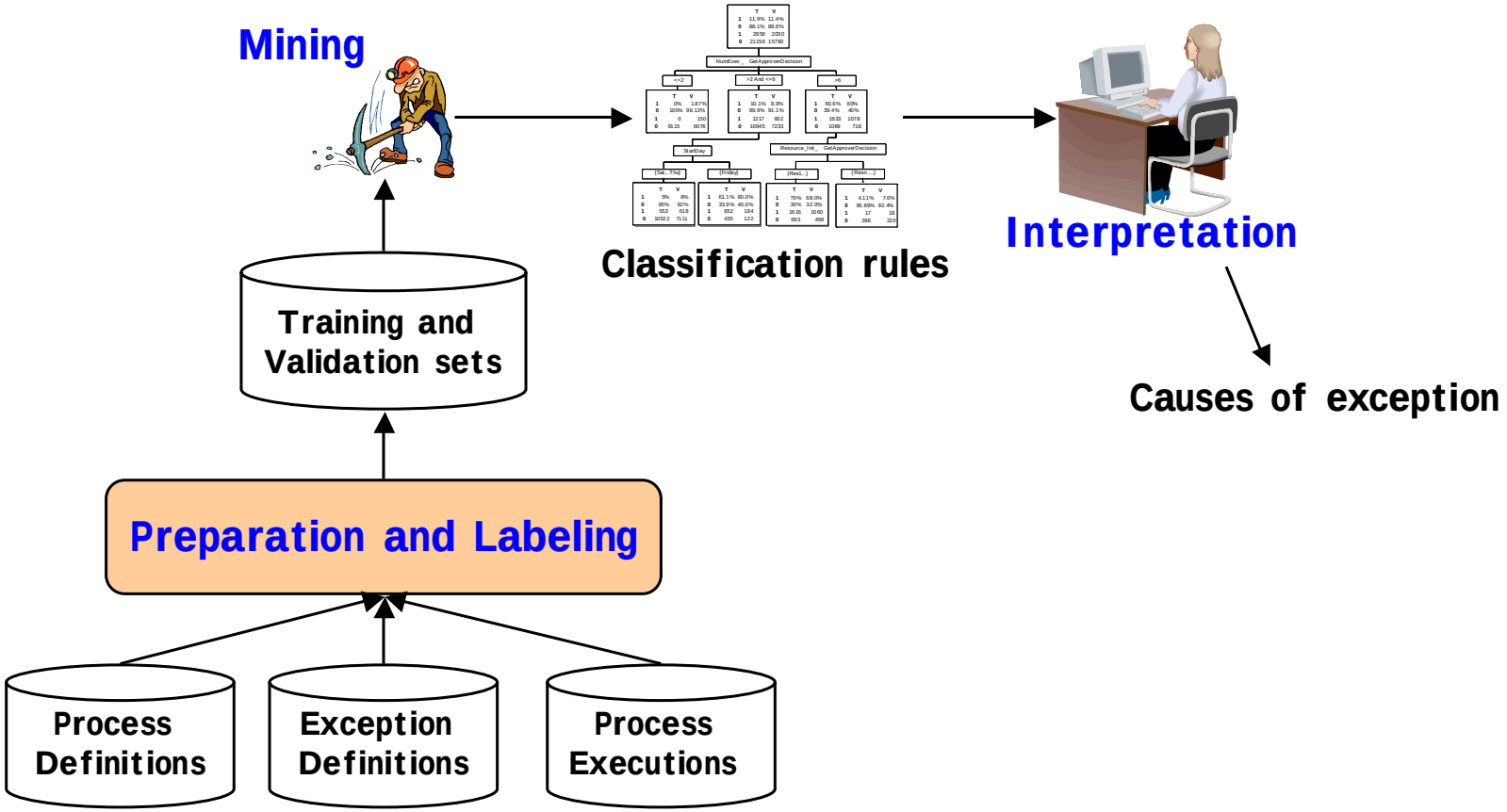
- Service providers need to deliver services (execute processes) with high and predictable quality.
- A key issue is reducing the occurrence of exceptions.
  - Exception: a deviation from the optimal (or acceptable) execution. It is a high-level, user-defined, subjective concept.
- To help reduce the occurrence of exceptions, support:
  - Exception Analysis: identify the causes of exceptional behaviors.
  - Exception Prediction: predict the occurrence of exceptions as early as possible during process execution.
  - Exception Prevention: take actions to avoid (when possible and convenient) the occurrence of the exceptional situation.

D. Grigori, F. Casati, U. Dayal, M-C. Shan: "Improving Business Process Quality through Exception Understanding, Prediction, and Analysis." Proc. Intl. Conf. on Very Large Data Bases, Sept. 2001.

e

Approach to Exception Analysis

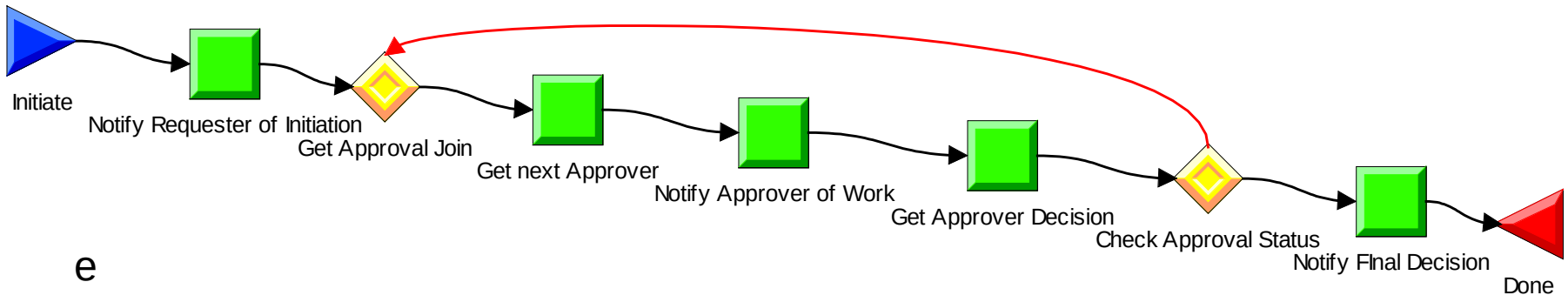
- Mine process definition and execution data
  - We treat exception analysis as a classification problem



e

Experimental Results: Analysis

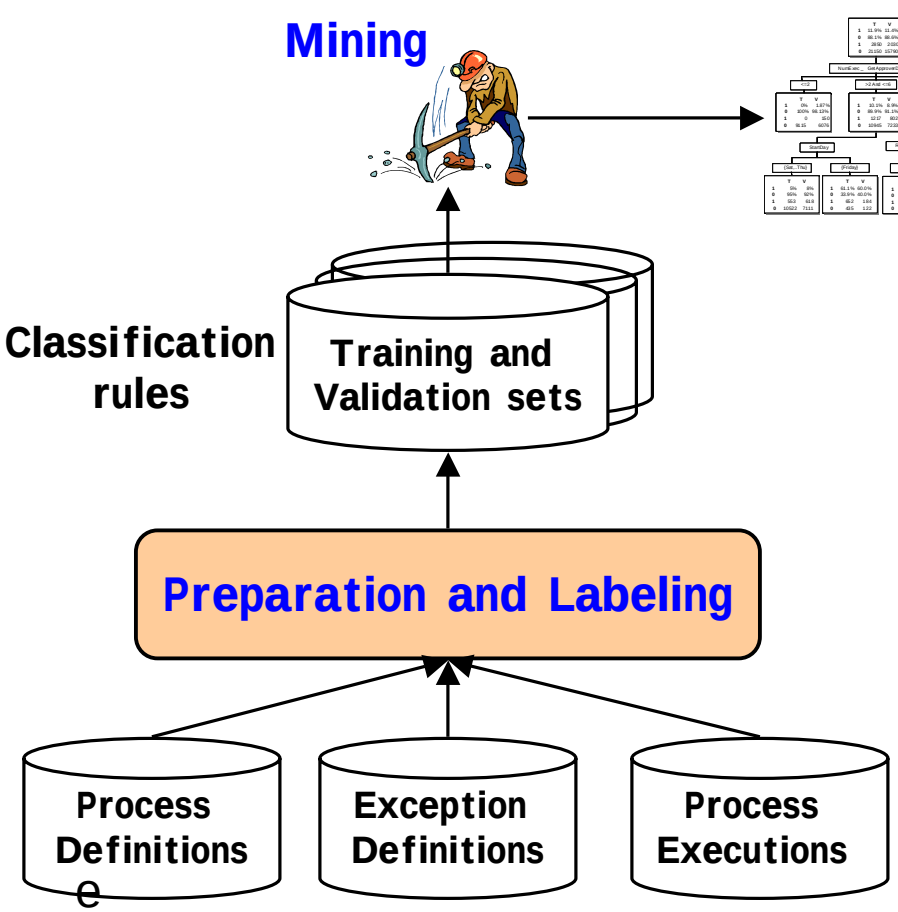
- We applied the techniques to Administrative processes to analyze *process duration* exceptions
  - Process considered “long” when over 20 days
  - On average, 15% of instances were exceptional
- Analysis:
  - When a certain node were executed by resources in group A, 70% of the instances was exceptional.
  - When the node was executed by resources in group B, 5% of the instances were exceptional



e

Exception Prediction

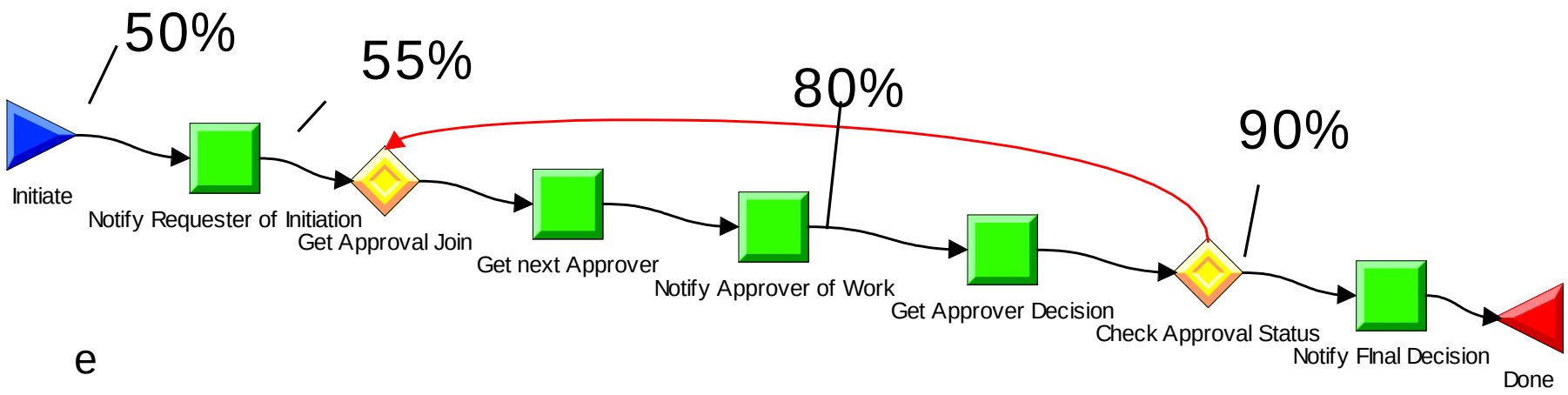
- Goal: predict occurrence of exception as early as possible
  - Prediction accuracy increases as process execution progresses



Several Training/Validation sets were prepared (one for each execution stage). Each set only includes process execution attributes defined at that stage. A predictive model was generated for each stage.

Experimental Results: Prediction

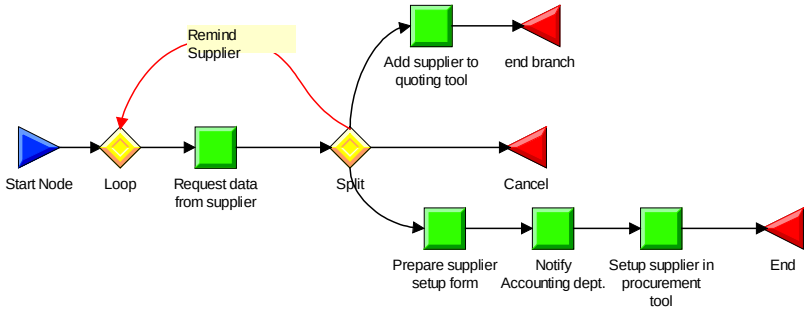
- Good predictions at the very start of the process
  - A process input variable determines the number of loops, and therefore was correlated to the process duration
    - For some other combination of input data, as high as 50% exception probability
- After the execution of a “critical” node, prediction accuracy increased substantially.
- A lot more work needs to be done to prevent exceptions.

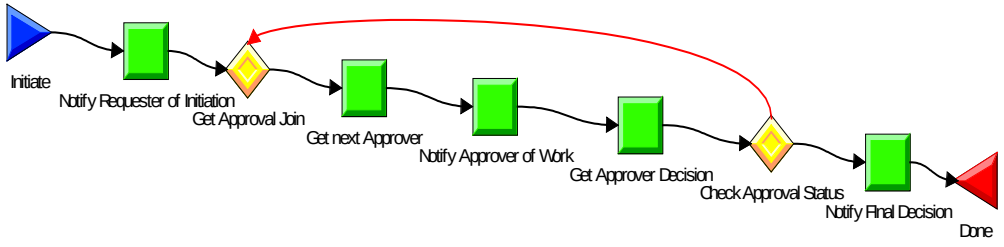


e

Process Improvement

- Designing processes is challenging
  - Difficult to know the process (even for the people involved in it)
  - Difficult for the modeler to ask the right questions, get the right answers
- Business Process Intelligence supports process (re)design, by emphasizing problems and inefficiencies





e

Outline

- Context: The Intelligent Enterprise, E-Business, and Data Mining Opportunities
- Four Cases
  - Customer Relationship Management
  - Catalog Creation and Service Discovery
    - Text Categorization
    - Information Extraction from Semi-structured Text
  - Business Process Intelligence
- Conclusions

Conclusions

- Commercial Landscape: Shift from horizontal software, toolkits to vertical applications, system integration, and services.
- Research: Must shift from obsession with algorithms to developing solutions enabled by data mining ("invisible, embedded data mining").
- Many applications of usage mining and content mining, and combinations of these, for e-business.
- Use many different techniques drawn from different disciplines:
  - For usage mining: OLAP, clustering, association rules, classification, …
  - for content mining: clustering, classification, information retrieval, linguistic analysis, …
- Have to address end-to-end scalability of the whole solution architecture.
- Data preparation and cleaning are still an art.
- Important to close the loop: use the results of mining for decision making and optimization of business processes.

e