AI Methods in Data Warehousing

A System Architectural View

Walter Kriha

Business Driver: Customer Relationship Management (CRM)

- learn more about your Customer
- Provide personalized offerings (cheaper, targeted)
- Make better use of in-house information (e.g. financial research)
- Somehow use all the data collected

The web is accelerating the problems (terabytes of clickstream data) and provides new solutions: Web-mining, the Web-House)

CRM: Simulate Advisor Functions

Client oriented:

- Know interests and hobbies
- Know personal situati
- Know situation in life
- Know plans and hopes

Bank oriented:

- Know where to find information and what applications to use
- Know how to translate, summarize and prepare for customer
- Know who to ask if in trouble

Plus: new ideas from automatic knowledge discovery etc. that even a real advisor can't do!

Overview

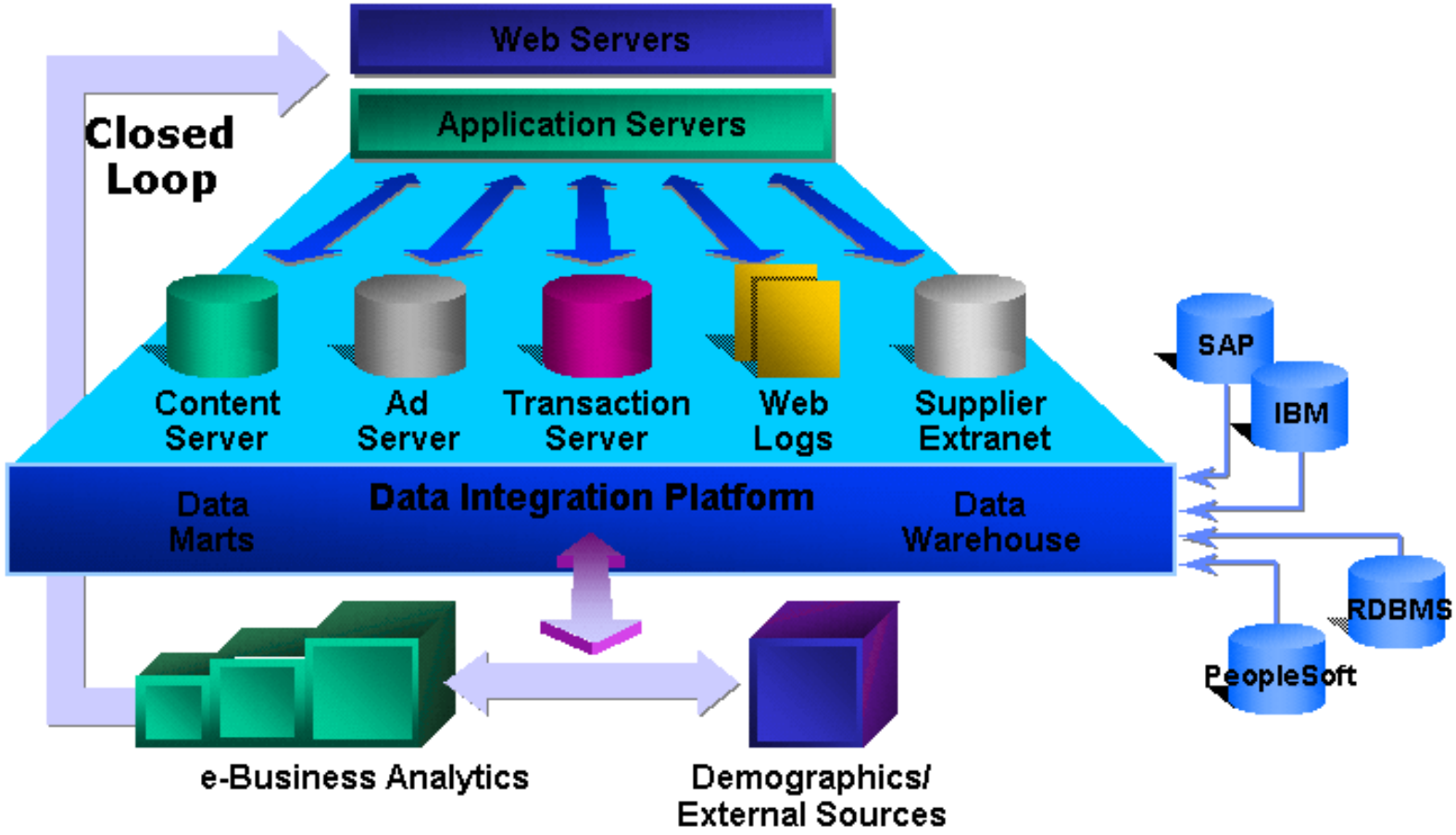
- Requirements coming from a dynamic, personalized Portal Page
- Data Collection and DW Import
- AI Methods used to solve requirements
- How to flow the results back into the portal

A Portal: A self-adapting System

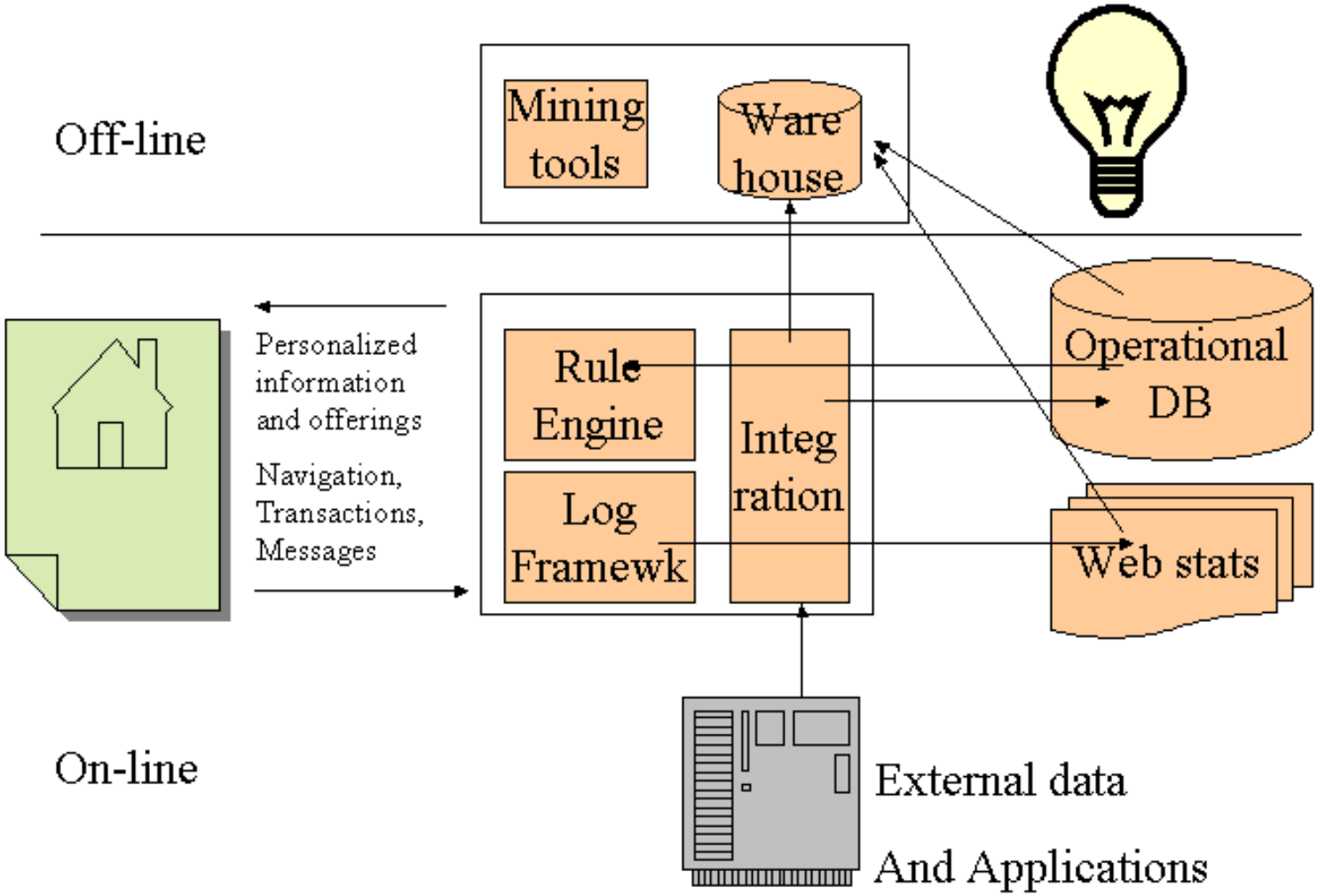
- Collect information for and about customers
- Learn from it
- Adapt to the individual customer by using the "lessons learned"

The problem: a portal does not have the time to learn. This needs to happen off-line in a warehouse!

DW Integration: Sources



DW Integration: Structure



What information do we have?

- The pages the customer selected (order, topics etc.)
- Customer interests from homepage self-configuration
- Customer transactions
- Customer messages (forum, advisor)
- Internal financial information

The data collection and import process needs to preserve the links between different information channels (e.g. order of customer activity)

Common: customi

Interest in our services (homepage config)

transactions

Velcom

ld like t

to our

E-Banking: balance

Interest in shares etc.

me that fits nicely

rent investment strategy.

Portfolio: Siemens, Swisskom, Esso,

Message activity

Messages: 3 new
From foo: hi Mrs. Rich

mon: Banner

News: IBM invests in company Y

Quotes: UBS 500, ARBA 200

Special interest (filters selected)

forum activity

Links: myweather.c
UBS glossary etc.

asian

Charts: Sony

Forum: art banking, 12

What do we want to know?

- Does a customer know how to work the system (site usability)?
- Does a customer voice dissatisfaction with company (customer retention)
- If new financial information enters the system – which customers might be interested in it (content extraction, customer notification)?

Which AI techniques might answer those questions?

What do we want to provide?

- A personalized homepage that adapts itself to the customers interests (from self-customization to automatic integration)
- An early warning system for disgruntled customers or customers that have difficulties working the site
- An ontology for financial information
- An integrated view of the company and its services and information ("electronic advisor")

See: "Finance with a personal touch", Communications of the ACM Aug.2000/Vol.43 No.8

Dynamic, personalized and **INTEGRATED** homepage

Common: customize,

Personal "touch"

Welcome M
We would like to y ou to our
New Instrument X hat fits nicely
To your current investment strategy.

Portfo
add X?

Messages: 3 new
From advisor: about **X** inv.

Common: Banner about **X**

Quotes: UBS 500,
X 100

News: IBM invests in
X now listed on NA

Research: **X** future p
asian equity update

Connect communities and site content

Links: **X** homepage
myweather.com,.

Forum: **X** is discussed here

Charts: **X**

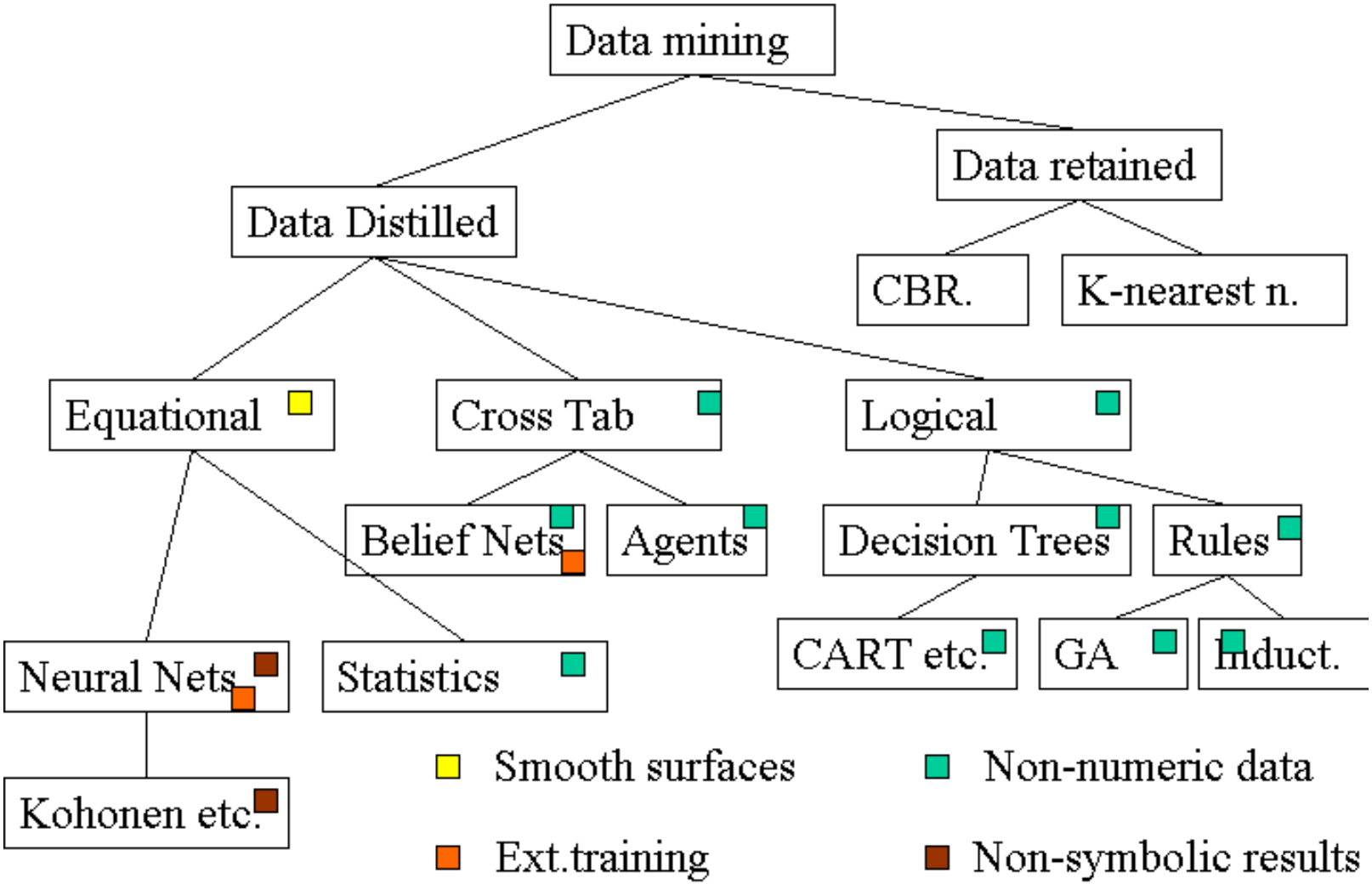
Data Mining

- The automatic extraction of hidden predictive information from large databases
- An AI-technique: automated knowledge discovery, prediction and forensic analysis through machine learning

Web Mining

- Adds text-mining, ontologies and things like xml to the above

Data Mining Methods



Data Preparation

- Catch complete session data for a specific user
- Store meta-information from content with behavioral data
- Create different data structures for different analytics (e.g. Polygenesis)

Use a special log framework! Make sure there are meta-data for the content available (e.g. dynamically generated page content)

Data Analysis

Content Mining (e.g Segmentation of Topics)

- Cluster Analysis
- Classification

Usage Mining (e.g. Segmentation of Customers)

- Pattern detection
- Association rules

Problem: How to express similarity and distance

- Linguistic analysis, statistics (k-nearest-neighbours)
- Machine learning (Neuronal nets, decision trees)

Problem: How to create a user profile e.g from navigation data

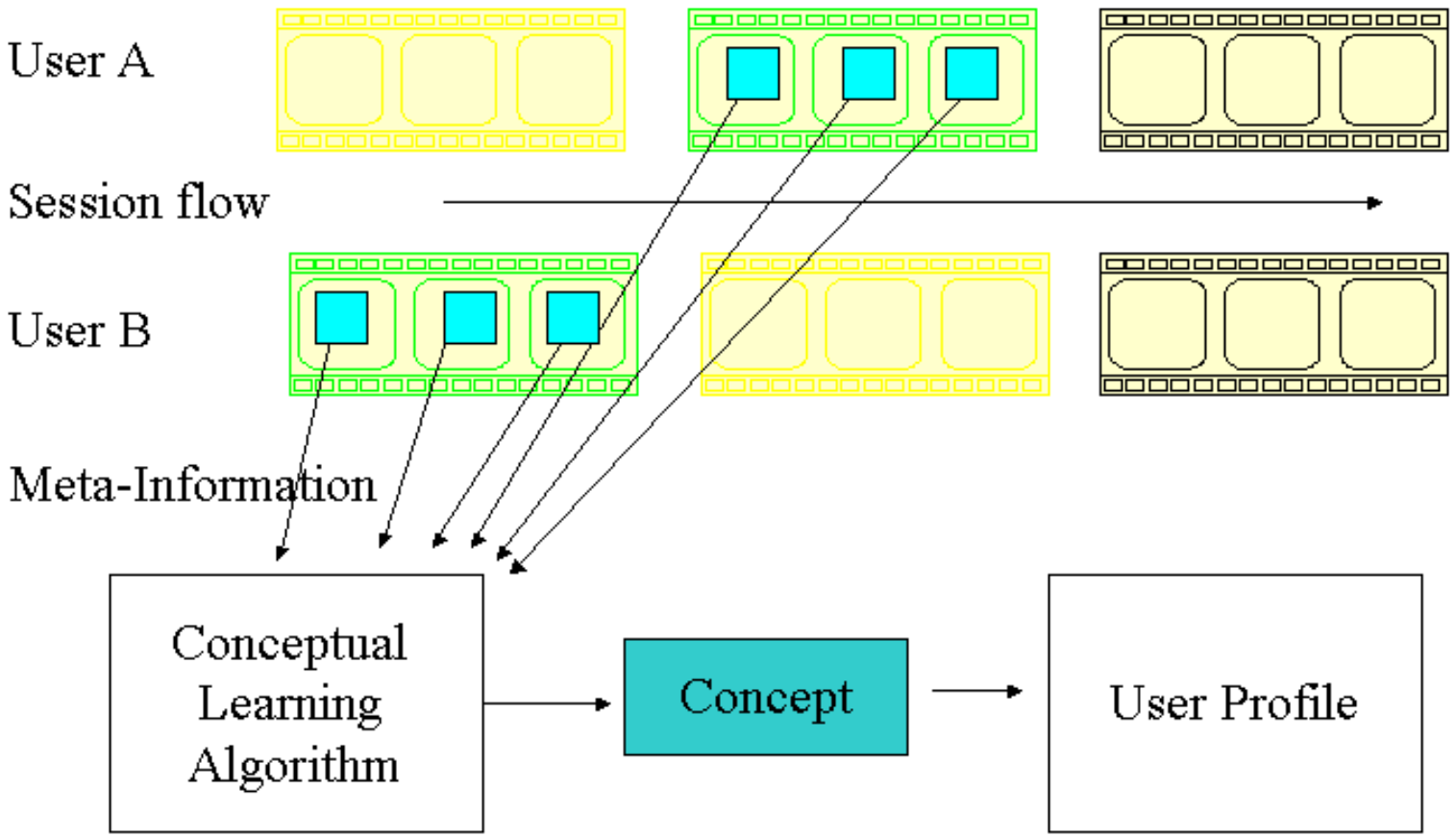
collaborative filtering: derive content similarities from behavioral similarities

Example: Find Session Topics automatically

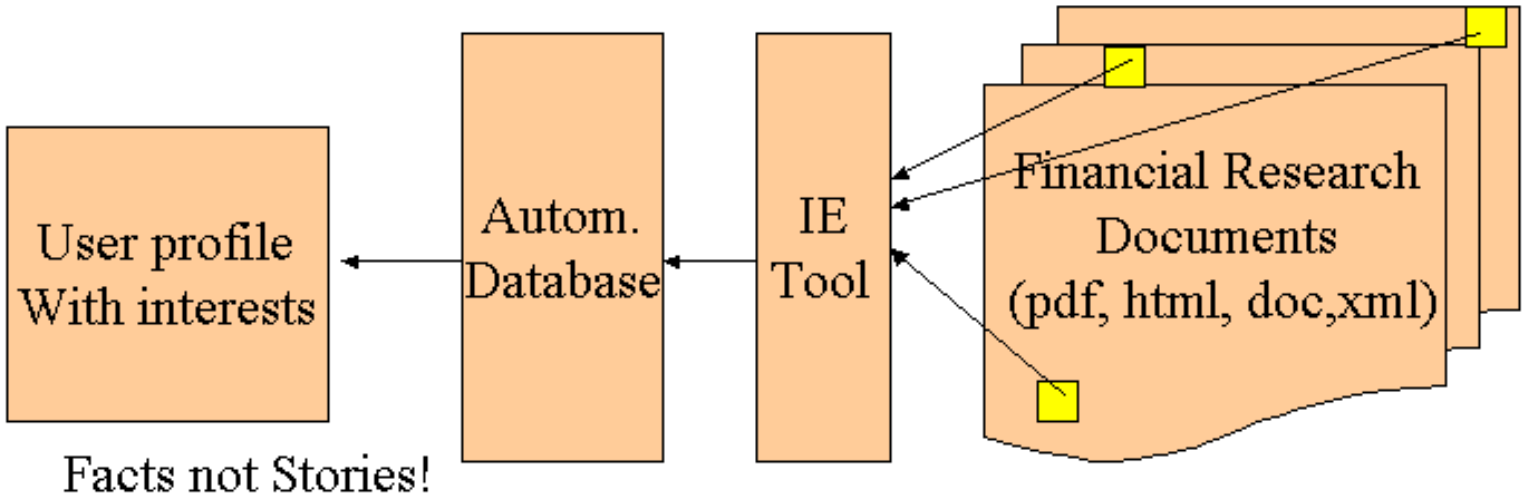
(Combined content and behavioral analysis)

- Use statistical cluster mining to extract page-views that co-occur during sessions (visit coherence assumption)
- Use a concept learning algorithm that matches the clusters (of page-views) with the meta-information of the pages to extract common attributes
- Those common attributes form a “concept”

Learning Concepts



The Text-Warehouse: Information Extraction



Facts not Stories!

Serving personalized information requires fine-grained extraction of interesting facts from text bodies in various formats

Methods for Information Extraction

Natural Language Processing

- Analyze Syntax to derive Semantics
- Context changes break algorithm

Wrapper Induction

- Use contextual features to infer semantics (e.g. html tags)
- Very brittle in case of source changes

Both methods use extraction patterns that were acquired through machine learning based on training documents.

More textual methods

- Thematic Index: Generate the reference taxonomy from training documents (linguistic and statistic analysis)
- Clustering: group similar documents with respect to a feature vector and similarity measure (SOM and other clustering technologies)

Automatic Text Classification

Case: Building a directory for an enterprise portal

Rule based: Experts formulate rules and vertical vocabularies (Verity, Intelligent Classifier)

Example-Based: A machine learning approach based on training documents and iterative improvement (e.g Autonomy, using Bayesian Networks)

Fully automated text classification is not feasible today. Cyborg classification needed. More tagged data needed.

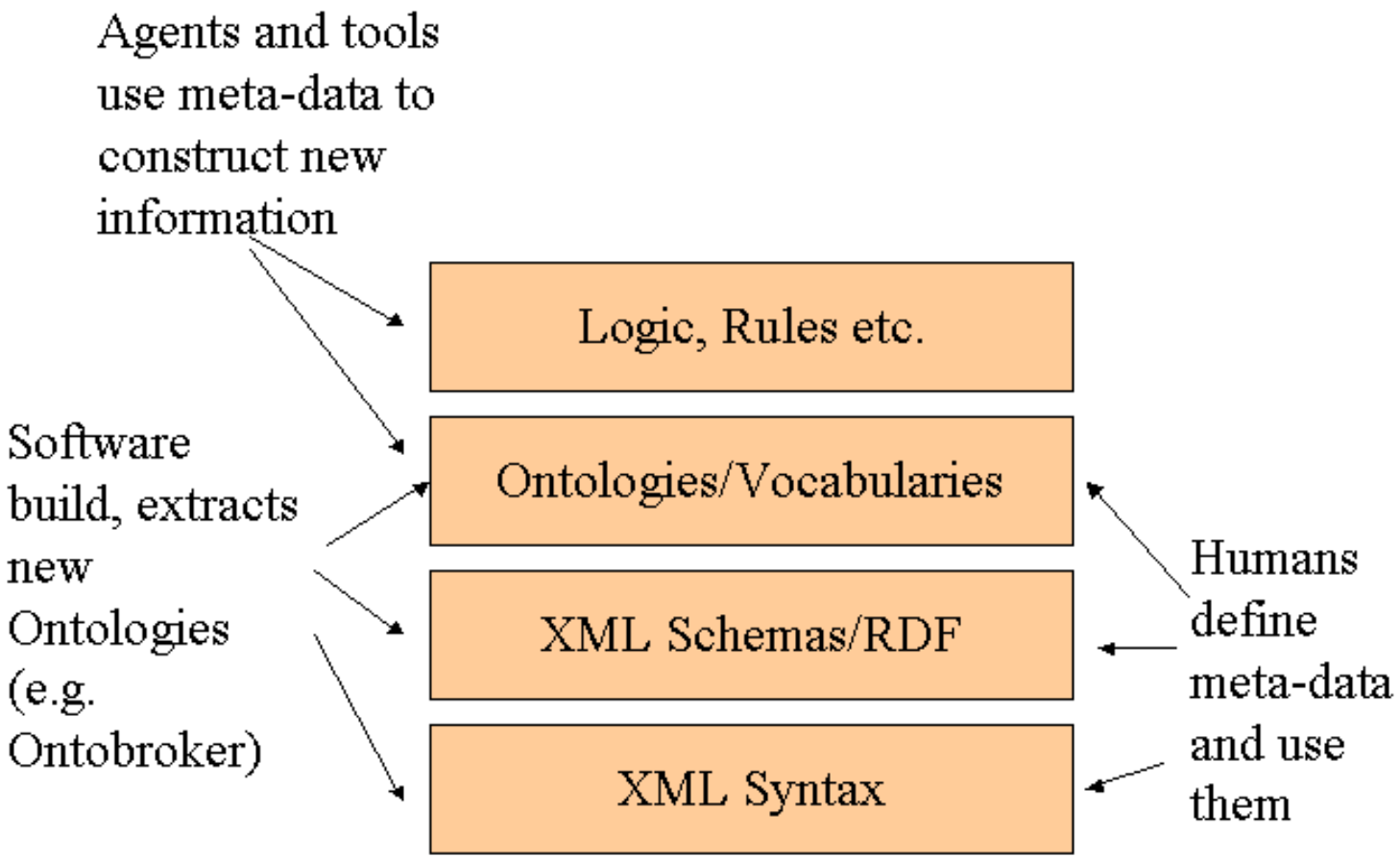
The Meta-data/Ontology Problem

"The key limiting factor at present is the difficulty of building and maintaining ontologies for web use"

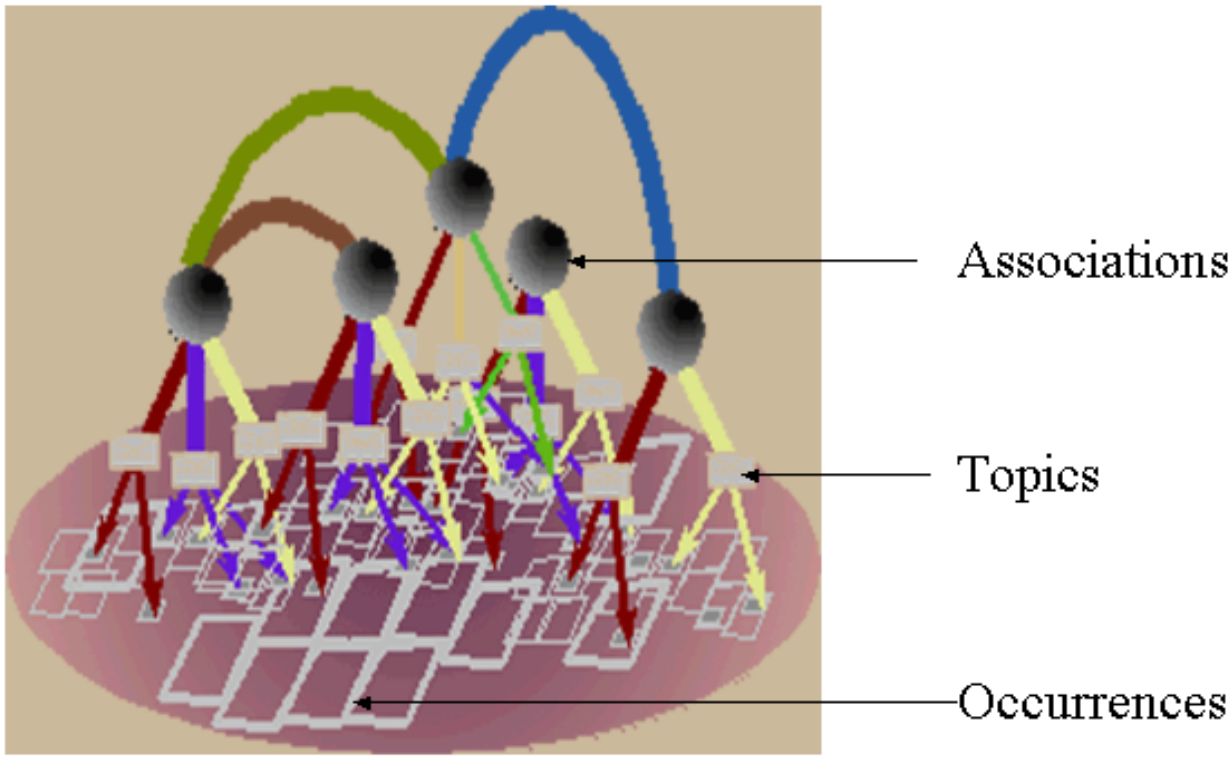
J.Hendler, Is there an Intelligent Agent in your future?

This is also true for all kinds of information integration e.g. financial research

The Solution: Semantic Web?

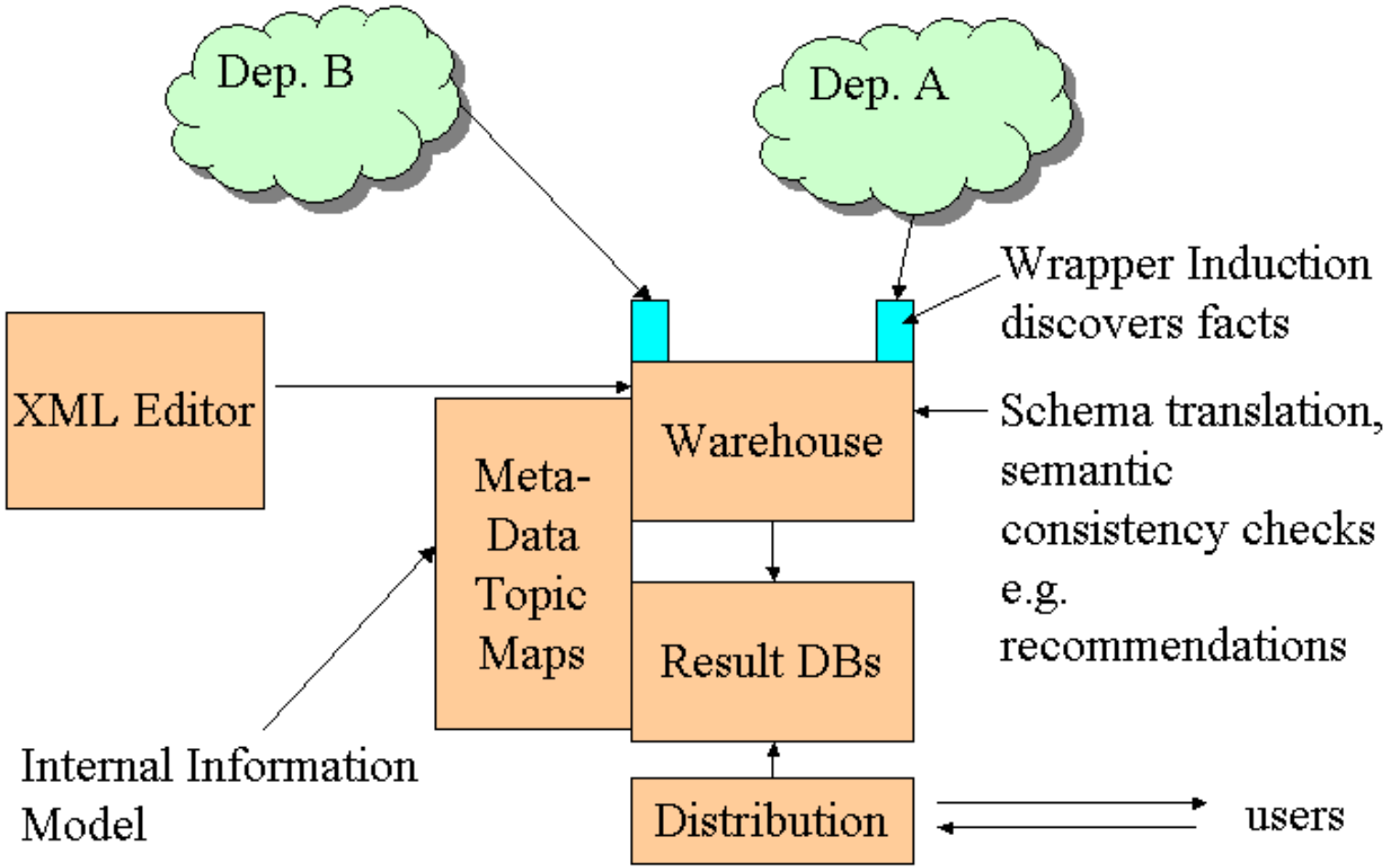


AI on Topic Maps?

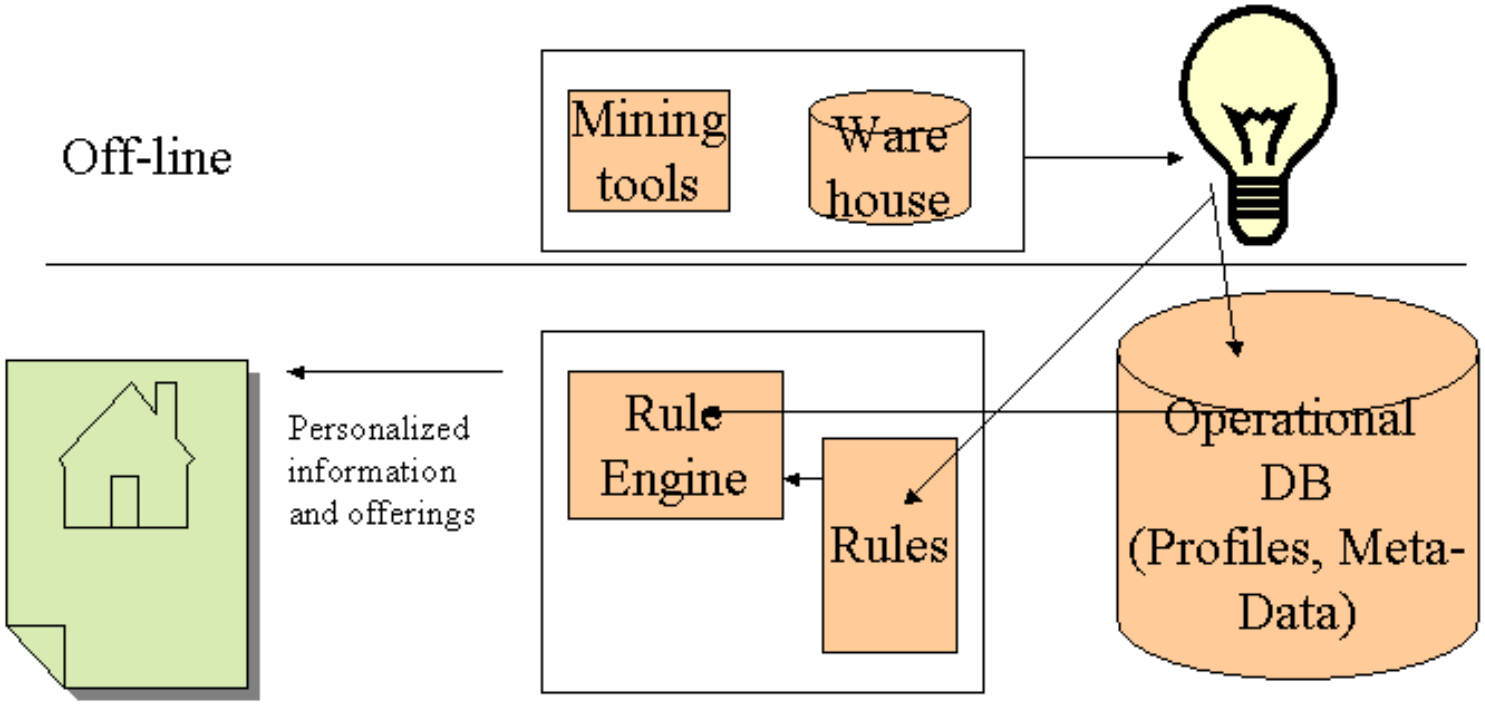


See: James D.Mason, Ferrets and Topic Maps, Knowledge Engineering for an Analytical Engine

Financial Research Integration



Deployment


Off-line
Mining tools
Ware house
Personalized information and offerings
Rule Engine
Rules
Operational DB (Profiles, Meta-Data)
On-line

The Main Problems for the "Web-house"

- Portal architecture must be designed to collect the proper information and to **use** the results from the web-house easily
- Portal content is at the same time customer offer as well as customer measuring tool
- Few people understand both the portal system aspect and the warehouse analytical aspect.

Resources

- Katherine C.Adams, Extracting Knowledge (www.intelligentkm.com/feature/010507/feat.shmtl)
- Dan Sullyvan, Beyond The Numbers (www.intelligententerprise.com/000410/feat2.shtml)
- Communications of the ACM, August 2000/Vol.43 Nr. 8
- Information Discovery, A Characterization of Data Mining Technologies and Process (www.datamining.com/dm-tech.htm)
- Dan R.Greening, Data Mining on the Web (www.webtechniques.com/archives/2000/01/greening.html)

Data Mining Tools (examples)

- IBM Intelligent Miner
- SPSS, Clementine
- SAS
- Netica (Belief Nets)