

SIFTER: A Content-based Information Filtering System



**Indiana University Purdue University Indianapolis
Indiana University Bloomington**

**Mathew J. Palakal
Rajeev R. Raje
Snehasis Mukhopadhyay
Javed Mostafa**

<http://sifter.indiana.edu>

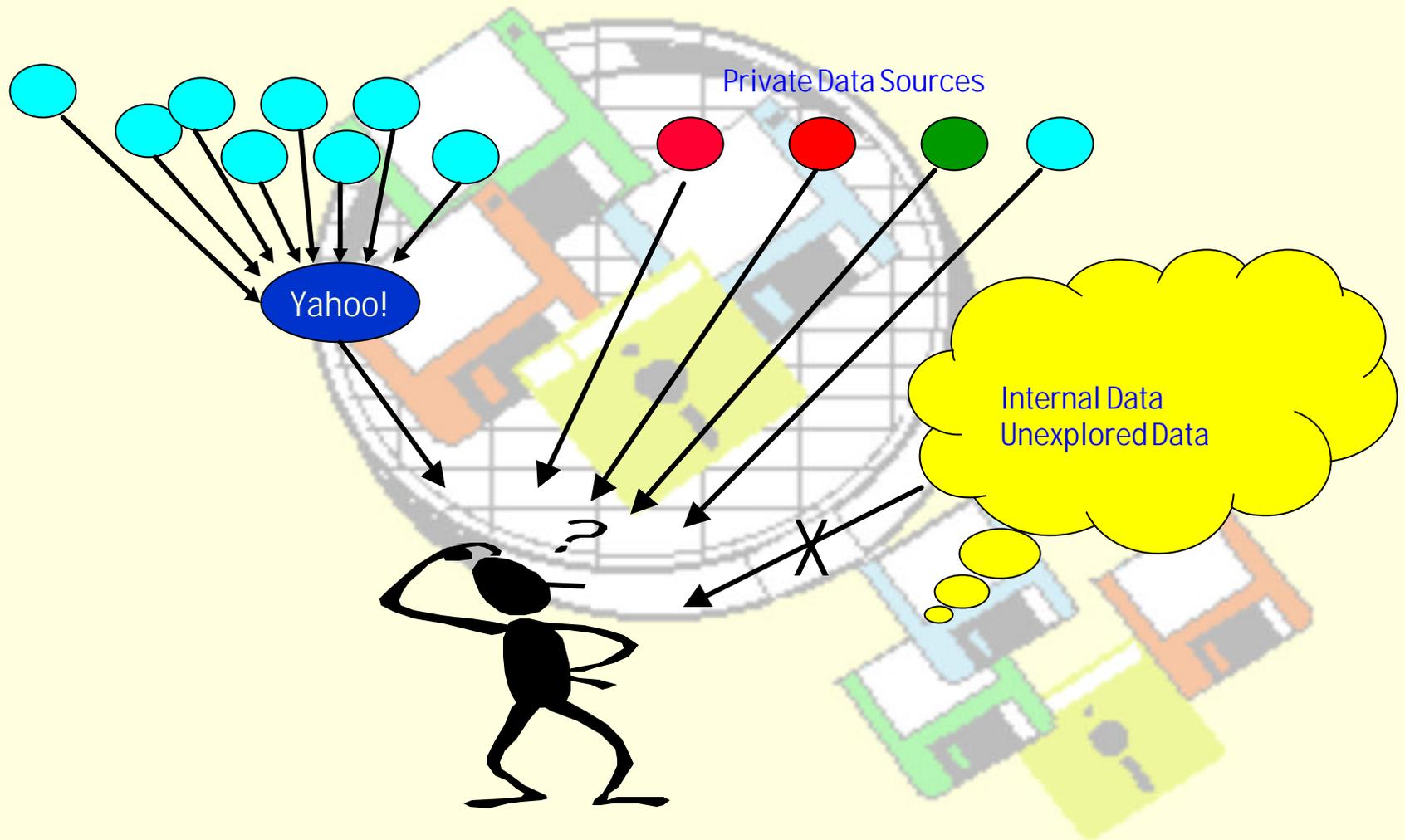
<http://sifter.cs.iupui.edu>



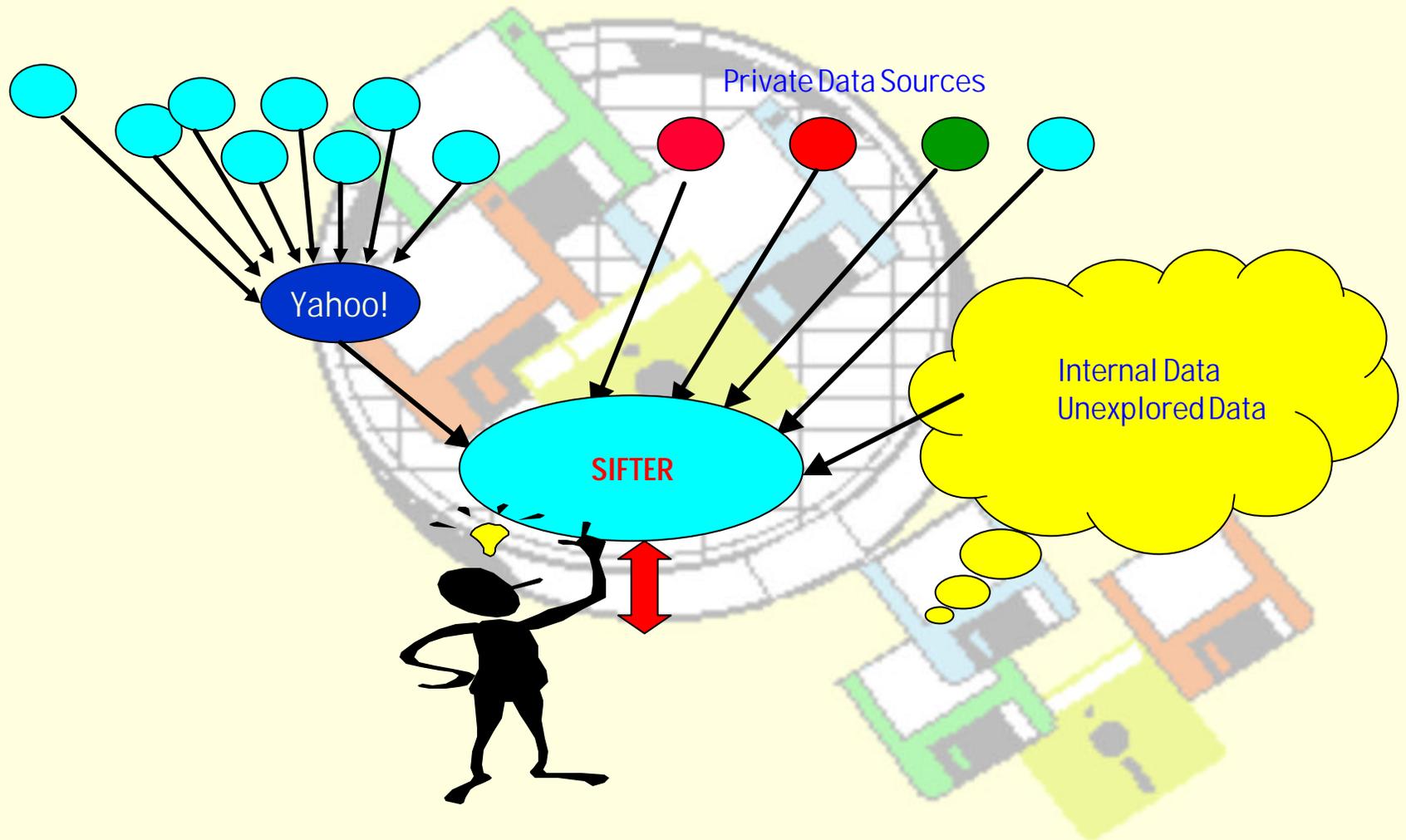
SIFTER: Motivation

- ***Information Overload* -- Reality in Today's World**
 - **A Need for locating highly 'Relevant Information'**
 - **A Necessity for an existence of a single tool for accessing multiple data sources and formats**
 - **Continuous Updating of relevant information**
 - **Privacy**
 - **Collaborative Work environments**
- 

The Current Model



The New Model

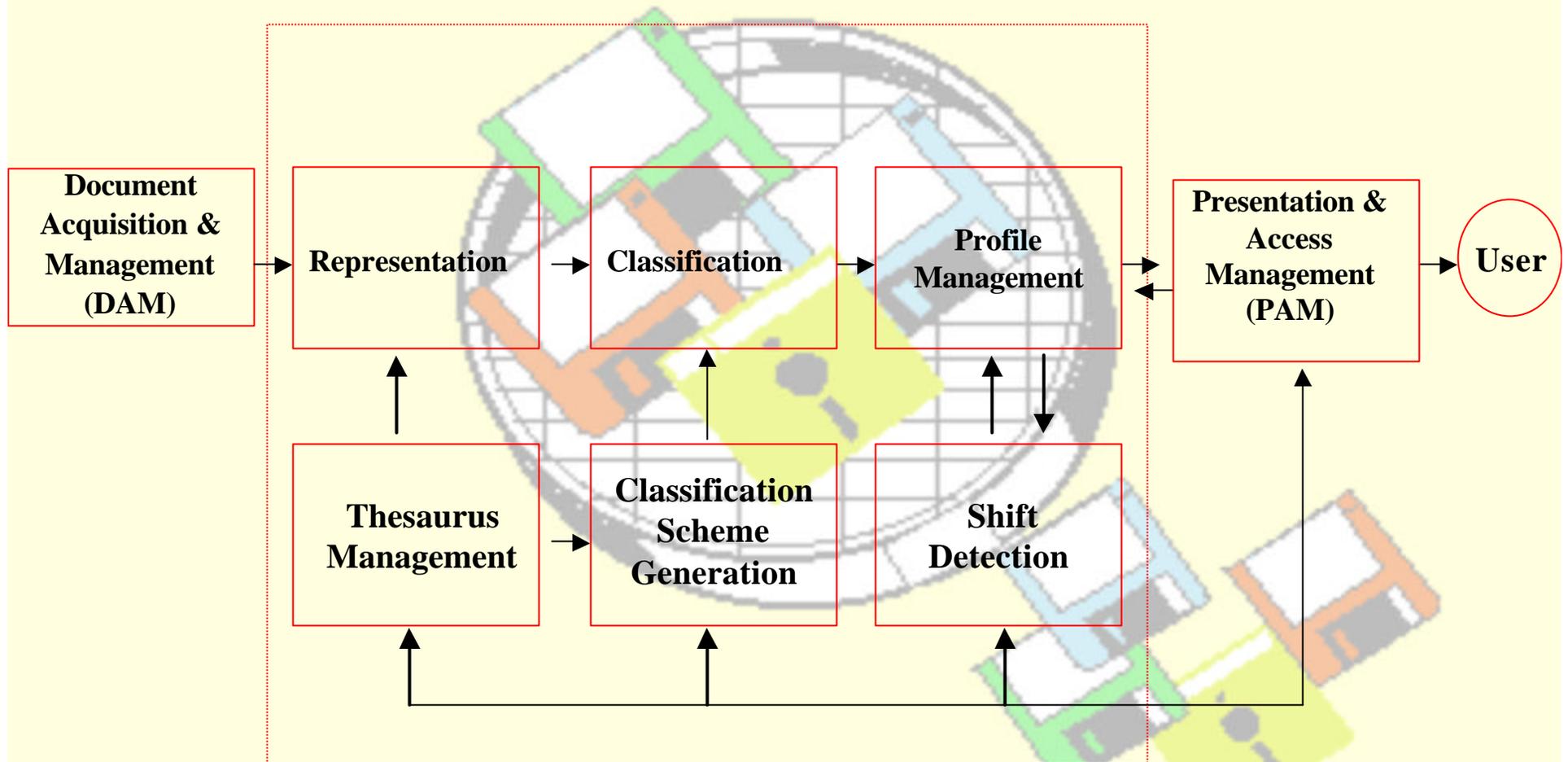


Usages of SIFTER

- A **Personalized** Content Manager
- A **Productivity Enhancement** Tool
- A **Nth Degree Information Personalization** Tool
- A **Collaborative** Research Tool
- A **Private Information** Tool

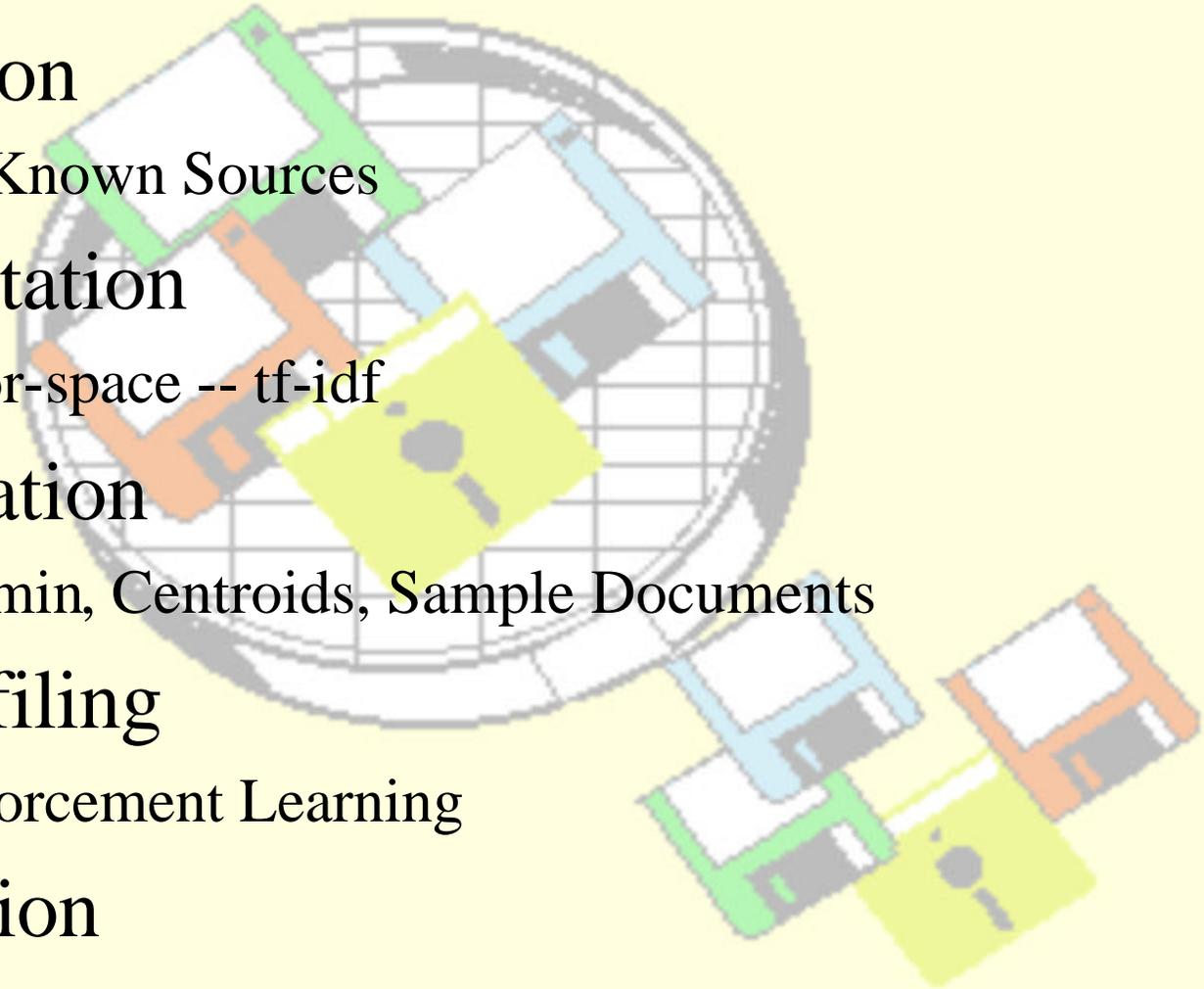


Single Agent Filter (SIFTER)



SIFTER

- Acquisition
 - File, Known Sources
- Representation
 - Vector-space -- tf-idf
- Classification
 - Maximin, Centroids, Sample Documents
- User Profiling
 - Reinforcement Learning
- Presentation
 - GUI



Document Representation and Vector Space Model

- Identify the concepts that describe the content of the given document
- Convert a document to a numeric or symbolic form
- Documents are vectors of weighted terms, defined in a *thesaurus* -- *How to generate?*
- Weights -- *tf* (term frequency) and *idf* (inverse document frequency) -- Simple and effective

Classification

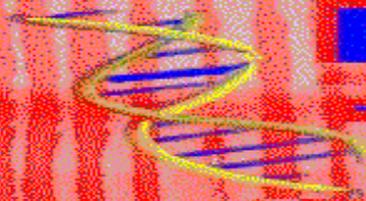
- Maximin-Distance: unsupervised clustering algorithm based on the document set
- Distance Metric: Cosine similarity measure (Salton)
- A point is chosen that has the largest distance from the centroids and is added as a new centroid if this distance is larger than a threshold

User Profiling

- Learn user interest levels for given categories
- Relies on relevance feedback from user
- Uses a simple reinforcement learning algorithms (known as Pursuit Learning)
 - maintains an action probability vector and a estimated relevance probabilities vector
 - both these vectors are updated continuously

SIFTER BioSifter

- Aimed at Customizing and Adapting SIFTER to Biological Domain
 - **Successfully Customized**
 - **PubMed as the Document Source**
 - **Documents and Thesaurus for Type II Diabetes**
 - **Stand-alone Version in Java and HTML**
 - **Tested and Deployed at Eli Lilly & Co.**



BioSifter

The Future in Information
Filtering Technology

SIFTER



Documents

View

Delete

50% Diffuse-type giant cell tumor: clinicopatholo
 50% Histologic Evidence of Foreign Body Granulat
 50% Expression of Myostatin Gene in Regeneratin
 50% Early mechanisms of renal injury in hyperchc
 50% Efficient vaccination by intradermal or intrar
 50% The function of alpha-crystallin in vision.
 50% Bundle Formation of Smooth Muscle Desmin
 50% Alport syndrome and diffuse leiomyomatosis
 50% Calpain-I induced alterations in the cytoskel
 N Human Uterine Myocytes Retain Their Energy C
 N The formation of Uranus and Neptune in the Ju
 N ICP measurement accuracy; the effect of tempe
 N [Results of stomach resection by Billroth II with

- Very Interested Interested
 Neutral Little Interest
 Not Interested

Domains

Biology

Thesauruses

Diabetes
GeneList

Add Details Delete

Subscriptions

Add New Delete

BioSifter

The Future in Information Filtering Technology

SIFTER



See www.sifter.salk.edu for more information

Word to be added:

Home

Add

Delete

Re-Train

Thesaurus

Classes and Relevance

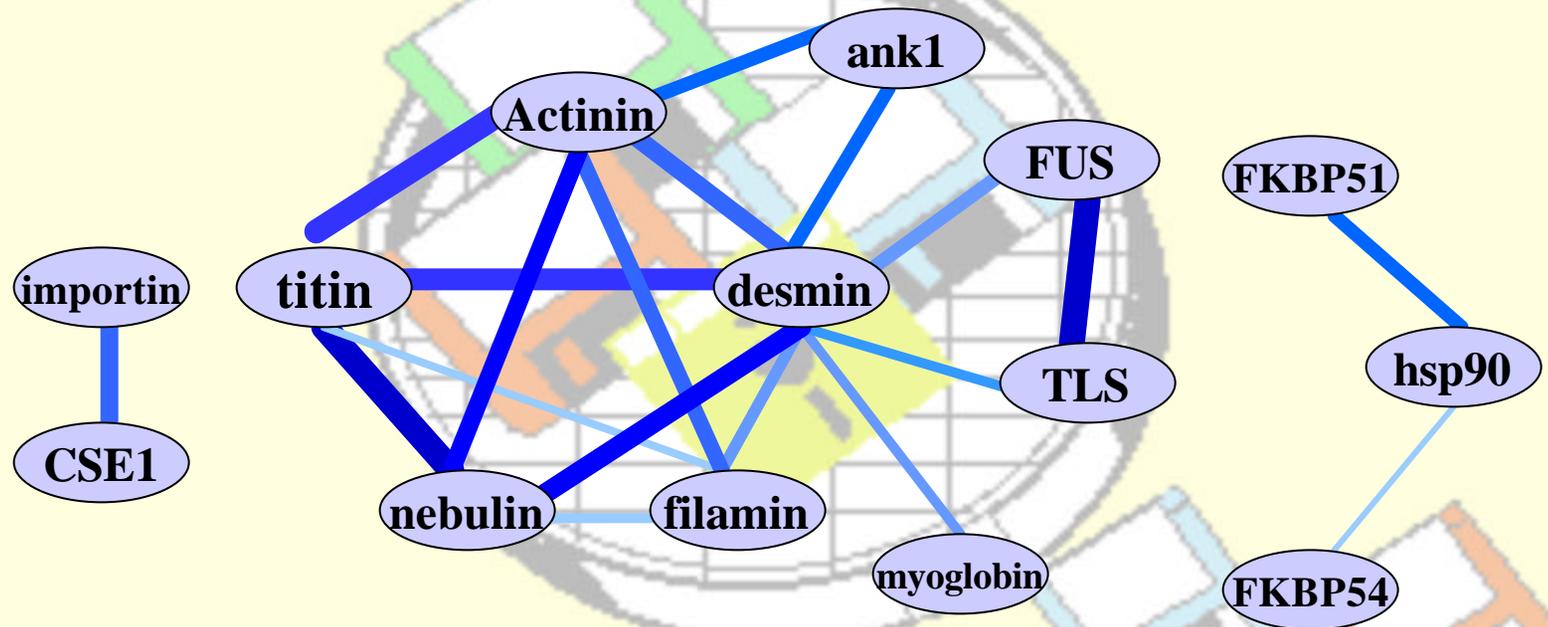
PHARMACOGENOMIC	CHROMOSOME, LOCUS, GENOME, VARIATIONS
SNP, SINGLE NUCLEOTIDE	HYPERINSULINEMIA, HYPERINSULINEMIC, HYPERTENSION
AFFECTED SIB PAIR	HYPOGLYCEMIA, INSULIN SECRETION, BLOOD GLUCOSE
AFFECTED MEMBER	PANCREATIC ISLET, INSULIN-DEPENDENT
ASSOCIATION ANALYSIS	INSULIN RECEPTOR, IFRS, INSULIN SENSITIVITY, NON-INSULIN-DEPENDENT, INSULIN-SECRETION
GENOME SCAN, GENOMICS	AUTOIMMUNE, ANTIBODY
CHROMOSOME	ODDS RATIO, T-CELL
INHERITANCE, HERITABILITY	ALLELIC, POLYMORPHISM
LOCUS, LOCI	BETA CELL, HYPERGLYCEMIA
ALLELIC, ALLELE, ALLELES	INSULIN-RESISTANT, GLUCOSE INTOLERANCE
POSITIONAL CLONING	IDDM, PROINSULIN
GENOME	INSULIN SENSITIVITY, INSULIN THERAPY
VARIATIONS, VARIATIONS	GENOTYPE, BLOOD GLUCOSE
#MUTATION	TWIN STUDY, INSULIN SECRETION
FAMILIAL	GLUCOKINASE, BLOOD GLUCOSE
CHROMOSOME SCAN	INSULIN-RESISTANT, INSULIN SENSITIVITY, HETEROZYGOUS, HOMOZYGOUS
CROSS-SECTIONAL STUDY	GLUCAGON, INSULIN-RESISTANT
LOD SCORE	FAMILIAL, HYPERTENSION
GENETIC MODEL	NON-INSULIN-DEPENDENT, BLOOD GLUCOSE
POLYMORPHISM, POLYMORPHISMS	INSULIN RECEPTOR, GLUCOSE TRANSPORTER
ODDS RATIO	CROSS-SECTIONAL STUDY, BLOOD GLUCOSE

How BioSifter help Pharmaceutical Researchers?

- Reducing the *Information Overhead*
- Rapidly *Adapting to User Interests and New Sources*
- Detecting *New Information Sources*
- Discovering *Novel Correlations*
- Identifying *Internal/External Collaborators -- Acquiring/Selling In/Out-house Knowledge*
- Creating a *Dynamic Web of Intelligent Filters*

Knowledge Discovery

Gene-Pair Relationship:



Data based on 5000 PubMed documents. Thesaurus consists of 67 Gene Terms. The thickness & color of lines indicate relative strengths of associations.

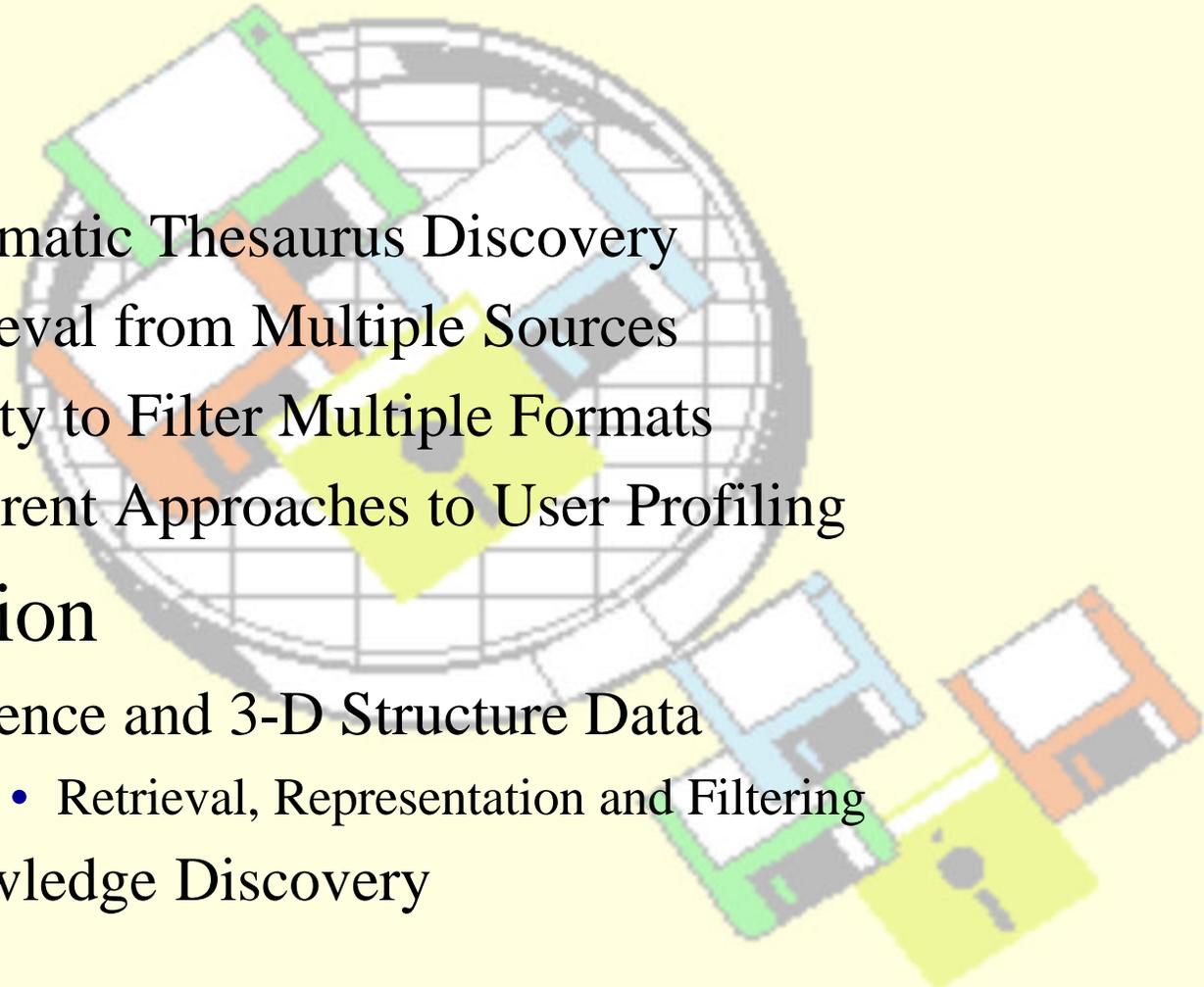
Future Plans

- System

- Automatic Thesaurus Discovery
- Retrieval from Multiple Sources
- Ability to Filter Multiple Formats
- Different Approaches to User Profiling

- Application

- Sequence and 3-D Structure Data
 - Retrieval, Representation and Filtering
- Knowledge Discovery



D-SIFTER and SIFTER II

- **D-SIFTER**

- Distributed Filtering System
- Homogeneous
- Classification/Profiling
- Collaboration Models

- **SIFTER II**

- Uniform Structure of an Agent
- Multiple and Heterogeneous Agents
- Collaboration Models



	Sifter	Bull's Eve	Mv Yahoo	Mv Lvcos	Northern Lights	Purple Yogi
Search Method	Any – Profile-based, User Specified, Discovery	Searching 700 Sources	Internal Sources – large number	Internal Sources – large number	Select a Source out of a list and/or all internal sources	Fixed Number of Popular sites
User Feedback	Explicit	No	No	No	No	Based on capturing user behavior
Personalization	Profile-based, Thesaurus Adaptation, Exact Profile	No	Login Information (Add Categories)	Login Information (Drag/Drop Categories)	No	Based on User's browsing habits, Approximate Profile
Privacy	Profile at user's site – Password Protected	No	On Server	On Server	No	Profile at user's site – Password Protected, Site History not Maintained
Learning	Multi-level	No	No	No	No	Yes
Multi-domain	Yes	Yes	Yes	Yes	Yes	Yes
Web Accessibility	Yes	Yes	Yes	Yes	Yes	Yes
Portability	Yes	Windows	Any Browser	Any Browser	Any Browser	Any Browser
Collaboration	Yes	No	No	No	No	No
Agent-based	Yes	No	No	No	No	Yogi-based
Keyword or Concept-based	Concept (tf/idf)	Keyword	Keyword	Keyword	Keyword	?
Size	~200KB	~300KB	N/A	N/A	N/A	?
Additional Features	GUI, Plots for Learning and Clustering, Handle Abrupt Interest Changes	GUI, Analyze Scores (how?)	GUI, Multiple Contents, Tools	GUI, Multiple Contents, Tools	GUI, Ranking (how?) and Custom Folders	Cannot download the client for testing
Target Audience	Individuals, Portals, Organizations, Groups	Individuals	Individuals	Individuals	Individuals	Individuals



Thank You

{mpalakal, rraje, smukhopa}@cs.iupui.edu

jm@indiana.edu

sifter@cs.iupui.edu