



Data-Mining, MetaData, and Digital Libraries

Elizabeth D. Liddy

**Director, Center for Natural Language Processing
Professor, School of Information Studies
Syracuse University**

May 17, 2001



Data-Mining

- **The computational process of extracting useful information from massive amounts of digital data**



Data-Mining

- **The computational process of extracting useful information from massive amounts of digital data**
 - **mapping low-level data into a richer, more abstract representation**



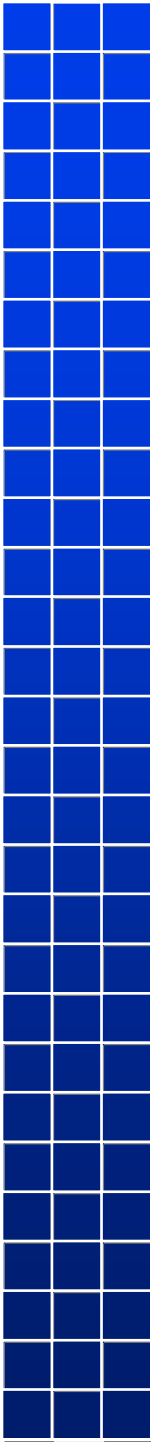
Data-Mining

- **The computational process of extracting useful information from massive amounts of digital data**
 - **mapping low-level data into a richer, more abstract representation**
 - **detecting meaningful patterns implicitly present in the data**



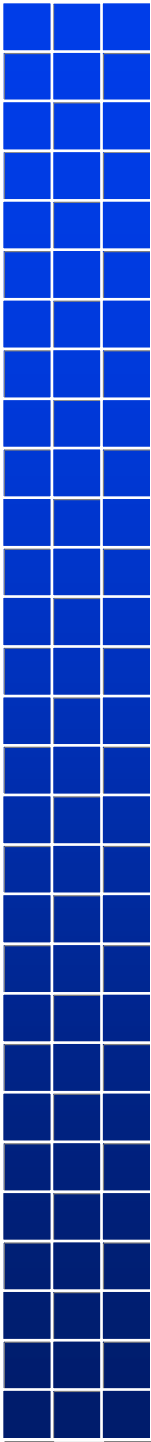
Data-Mining

- **The computational process of extracting useful information from massive amounts of digital data**
 - **mapping low-level data into a richer, more abstract representation**
 - **detecting meaningful patterns implicitly present in the data**
 - **typically conducted on structured databases**



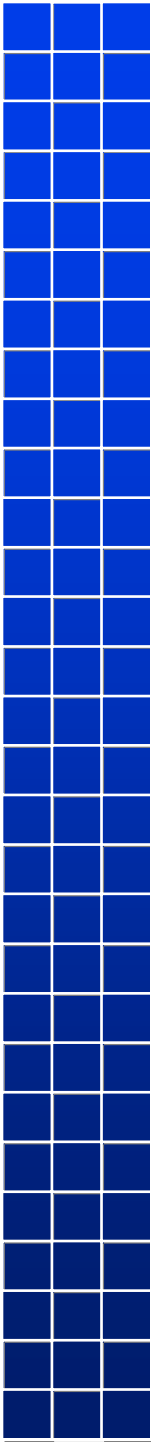
Data-Mining → Text Mining

- **no need to limit mining to information available in structured databases**



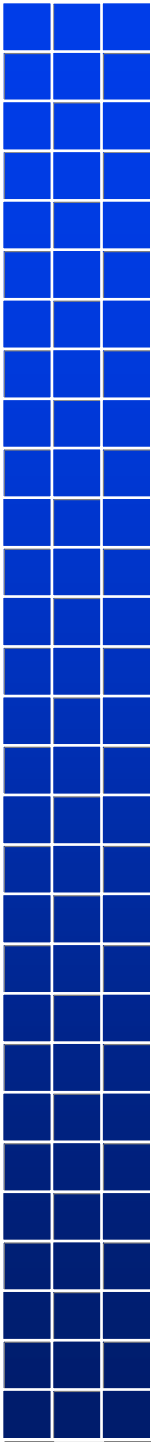
Data-Mining → Text Mining

- **no need to limit mining to information available in structured databases**
- **knowledge bases need not be manually constructed**



Data-Mining → Text Mining

- **no need to limit mining to information available in structured databases**
- **knowledge bases need not be manually constructed**
- **much of the knowledge of interest resides in unstructured, naturally occurring texts**



Data-Mining → Text Mining

- **no need to limit mining to information available in structured databases**
- **knowledge bases need not be manually constructed**
- **much of the knowledge of interest resides in unstructured, naturally occurring texts**
- **NLP can provide necessary technology for Text Mining**



Text Mining Defined

- **Process of analyzing any naturally occurring text**
- **For the purpose of discovering, and then ‘extracting’ or ‘tagging’ vital elements of information in the text**
- **Either, for:**
 - storage in a database for exploration using standard data-mining algorithms
 - OR --
 - automatic metatagging of documents



MetaData



MetaData

- **Originally based on MARC Format**
 - **Record-oriented language with rigorous formats tailored for bibliographic entries**
 - **Does not support many of the features needed for full-text documents**



MetaData

- **Originally based on MARC Format**
 - Record-oriented language with rigorous formats tailored for bibliographic entries
 - Does not support many of the features needed for full-text documents
- **Moved to SGML, a more flexible standard**
 - A syntax and a philosophy
 - Information about an object “is to be tagged meaningfully” and tags are to be consistent



‘Dublin +’ Metadata Schema

Dublin Core Metadata Elements

- **Contributor**
- **Coverage**
- **Creator**
- **Date**
- **Description**
- **Format**
- **Identifier**
- **Language**
- **Publisher**
- **Relation**
- **Rights**
- **Source**
- **Subject**
- **Title**
- **Type**



'Dublin +' Metadata Schema

Dublin Core Metadata Elements

- Contributor
- Coverage
- Creator
- Date
- Description
- Format
- Identifier
- Language
- Publisher
- Relation
- Rights
- Source
- Subject
- Title
- Type

GEM Metadata Elements

- Audience
- Cataloging
- Duration
- Essential Resources
- Grade Level
- Pedagogy
- Quality
- Standards



Educational Object

Stream Channel Erosion Activity

Student/Teacher Background Information

Rivers and streams form the channels in which they flow. A river channel is formed by the quantity of water and debris that is carried by the water in it. The water carves and maintains the conduit containing it. Thus, the channel is self-adjusting. If the volume of water, or amount of debris is changed, the channel adjusts to the new set of conditions.

...

Student Objectives

The student will discuss stream sedimentation that occurred in the Grand Canyon as a result of the controlled release from Glen Canyon Dam.

...



NLP Processing Example

Input:

The student will discuss stream sedimentation that occurred in the Grand Canyon as a result of the controlled release from Glen Canyon Dam.



NLP Processing Example

Input:

The student will discuss stream sedimentation that occurred in the Grand Canyon as a result of the controlled release from Glen Canyon Dam.

Morphological Analysis:

The student will discuss stream sedimentation that occurred in the Grand Canyon as a result of the controlled release from Glen Canyon Dam.



NLP Processing Example

Input:

The student will discuss stream sedimentation that occurred in the Grand Canyon as a result of the controlled release from Glen Canyon Dam.

Morphological Analysis:

The student will discuss stream sedimentation that occurred in the Grand Canyon as a result of the controlled release from Glen Canyon Dam.

Lexical Analysis (part-of-speech Tagging):

The|DT student|NN will|MD discuss|VB stream|NN sedimentation|NN that|WDT occurred|VBD in|IN the|DT Grand|NP Canyon|NP as|IN a|DT result|NN of|IN the|DT controlled|JJ release|NN from|IN Glen|NP Canyon|NP Dam|NP .|.



NLP Processing Example (cont'd)

Syntactic Analysis (phrase identification):

The|DT student|NN will|MD discuss|VB <CN> stream_sedimentation
</CN> that|WDT occurred|VBD in|IN the|DT <PN> Grand_Canyon
</PN> as|IN a|DT result|NN of|IN the|DT <CN> controlled_release
</CN> from|IN <PN> Glen_Canyon_Dam <PN> .|.

NLP Processing Example (cont'd)

Syntactic Analysis (phrase identification):

The|DT student|NN will|MD discuss|VB <CN> stream_sedimentation </CN> that|WDT occurred|VBD in|IN the|DT <PN> Grand_Canyon </PN> as|IN a|DT result|NN of|IN the|DT <CN> controlled_release </CN> from|IN <PN> Glen_Canyon_Dam <PN> .|.

Semantic Analysis Phase One (proper name interpretation)

The|DT student|NN will|MD discuss|VB <CN> stream_sedimentation </CN> that|WDT occurred|VBD in|IN the|DT <PN cat=geography/location> Grand_Canyon </PN> as|IN a|DT result|NN of|IN the|DT <CN> controlled_release </CN> from|IN <PN cat=buildings&structures> Glen_Canyon_Dam </PN> .|.



Extraction for ML-based Classification

Types of Features:

- Non-linguistic
 - Length of a document
- Linguistic
 - Root forms of a words
 - Part-of-speech tags:
 - Noun, Verb, Proper Noun, and Numeric Concept phrases
 - Proper Name & Numeric Concept categories
 - Semantic Relations
 - Discourse Level Characteristics (e.g.genre features)
 - Concepts (sense disambiguated words/phrases)

ML-based Text Classification:

- Regression model
- Nearest neighbor classifier
- Bayesian probabilistic classifier
- Decision trees



Automatic Metadata Generation

Automatic Metadata Generation

Title (SIE):	Grand Canyon: Flood! - Stream Channel Erosion Activity
Grade Levels (SIE):	6, 7, 8
GEM Subjects (TC):	Science--Geology Mathematics--Geometry Mathematics--Measurement Science--Process Skills Science--Instructional Issues
Keywords (TIE):	
Proper Names:	Colorado River (river), Grand Canyon (geography/location), Glen Canyon Dam (buildings& structures)
Subject Keywords:	channels, conduit, controlled release, dam, reservoir, rivers, sediment, streams, volume of flow
Material Keywords:	cookie sheet, roasting pan, cup, sand, clayboard, water, paper towel, pencil, paper
Procedure Keywords:	poke a hole, divide, take, hold, pour, make drawing, identify areas, diagram, compare

Automatic MetaData Generation (cont'd)

Pedagogy (TC)	Collaborative learning Hands on learning
Tool For (SIE/TIE):	Teachers
Resource Type (TC):	Lesson Plan
Format (SIE):	text/HTML
Placed Online (SIE):	1998-09-02
Name (SIE):	PBS Online
Role (SIE):	onlineProvider
Homepage (SIE):	http://www.pbs.org

Metadata Generation Methods:

- Structured Information Extraction (SIE)**
- Textual Information Extraction (TIE)**
- Text Categorization (TC)**



Digital Library

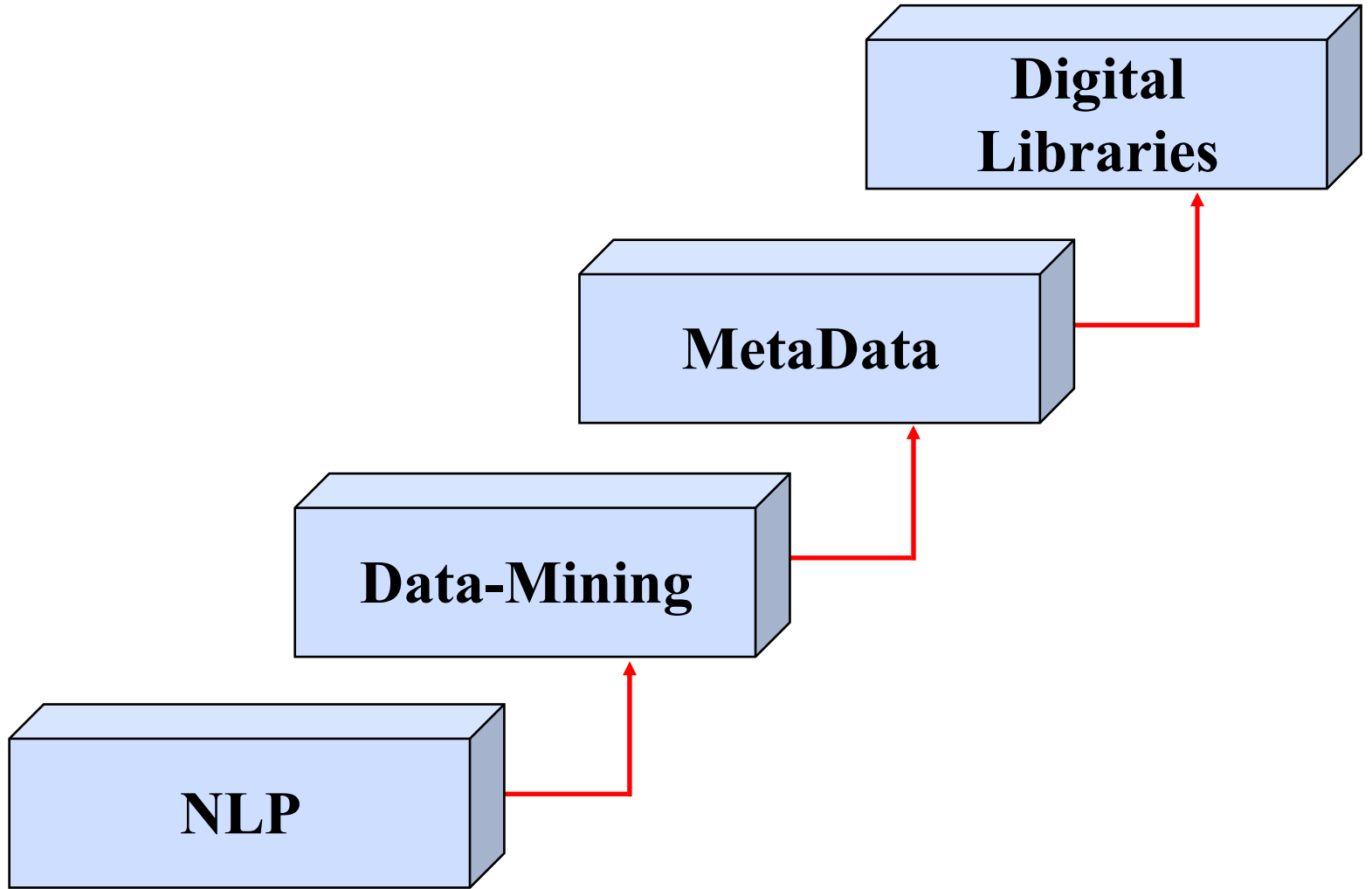
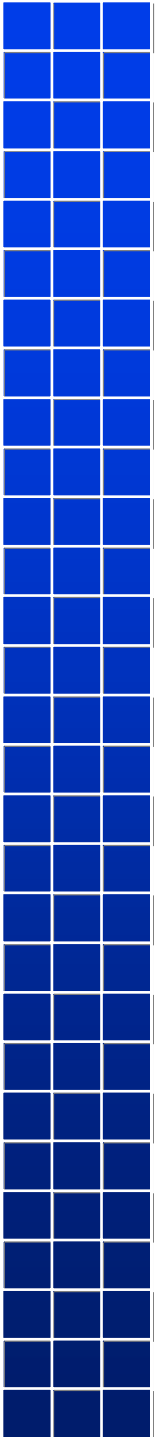


Digital Library

A Digital Library is:

- a collection of digital objects (**Repository**),
- with descriptions of these objects (**Metadata**),
- which provides search, browsing, and retrieval (**Services**),
- to a distributed set of users (**Community**).

Based on Modern Information Retrieval





Research Issues – Era 1

- **What is the appropriate set of metatags to be assigned to objects to make them optimally accessible for a range of users?**



Research Issues – Era 1

- **What is the appropriate set of metatags to be assigned to objects to make them optimally accessible for a range of users?
How can we determine what this set is?**



Research Issues – Era 1

- **What is the appropriate set of metatags to be assigned to objects to make them optimally accessible for a range of users? How can we determine what this set is?**
- **Since Controlled Vocabularies were shown in the past to be less useful than free-text indexing, what retrieval performance can Digital Libraries expect from them?**



Research Issues – Era 1

- **What is the appropriate set of metatags to be assigned to objects to make them optimally accessible for a range of users? How can we determine what this set is?**
- **Since Controlled Vocabularies were shown in the past to be less useful than free-text indexing, what retrieval performance can Digital Libraries expect from them?**
- **Can NLP be relied on to automatically produce adequately useful meta-data tags for all Digital Library services?**



Research Issues – Era 2



Research Issues – Era 2

- **How can we extend Data-Mining applications to understand both:**
 - **the content of Digital Libraries**
 - **the uses of Digital Libraries**



Research Issues – Era 2

- **How can we extend Data-Mining applications to understand both:**
 - the content of Digital Libraries
 - the uses of Digital Libraries
- **What MetaData would need to be generated?**
 - **Can we generate it automatically?**