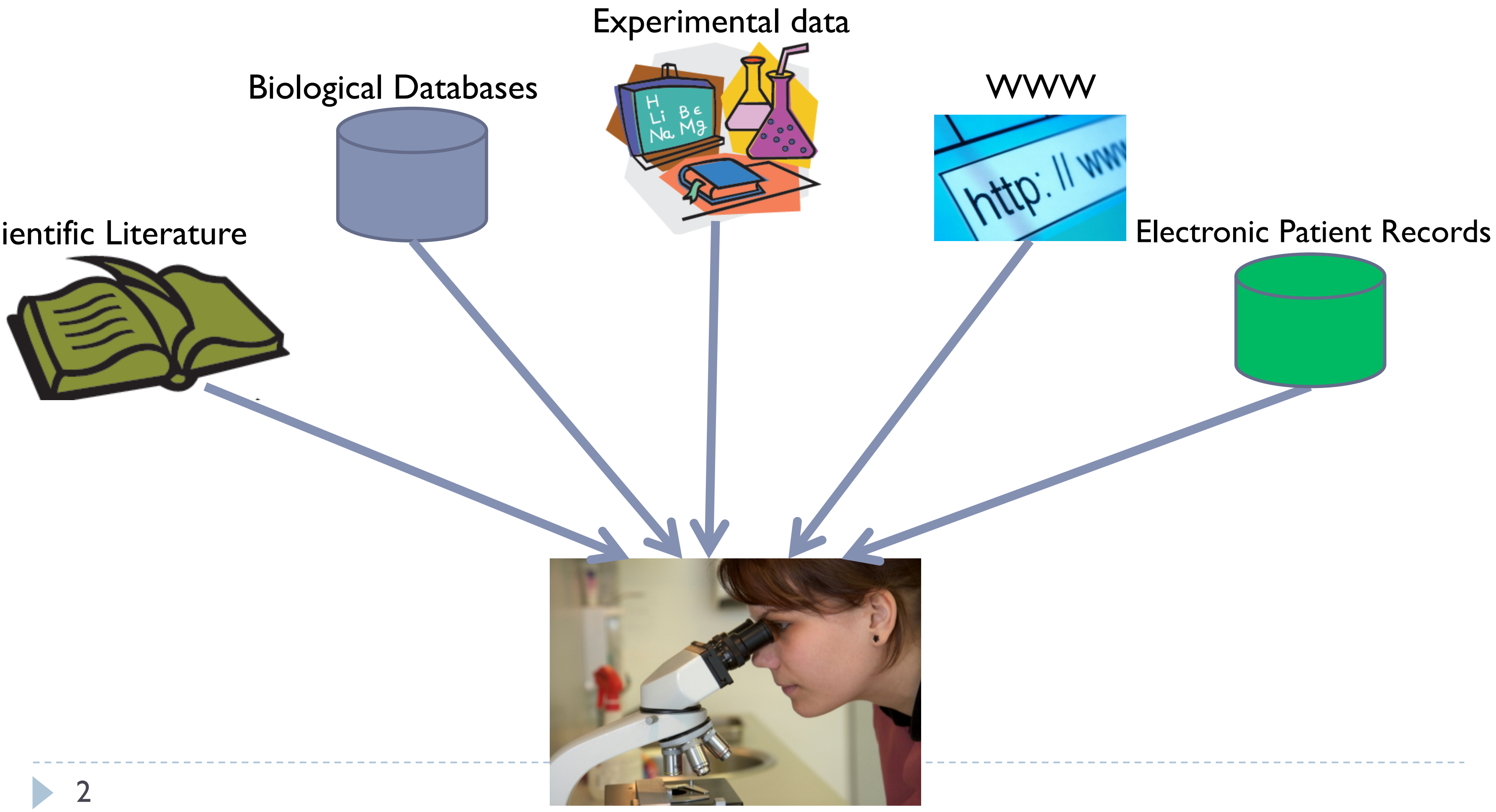
MEBI 591C/598 – Data and Text Mining in Biomedical Informatics

Meliha Yetisgen-Yildiz

Information Overload Problem



Experimental data

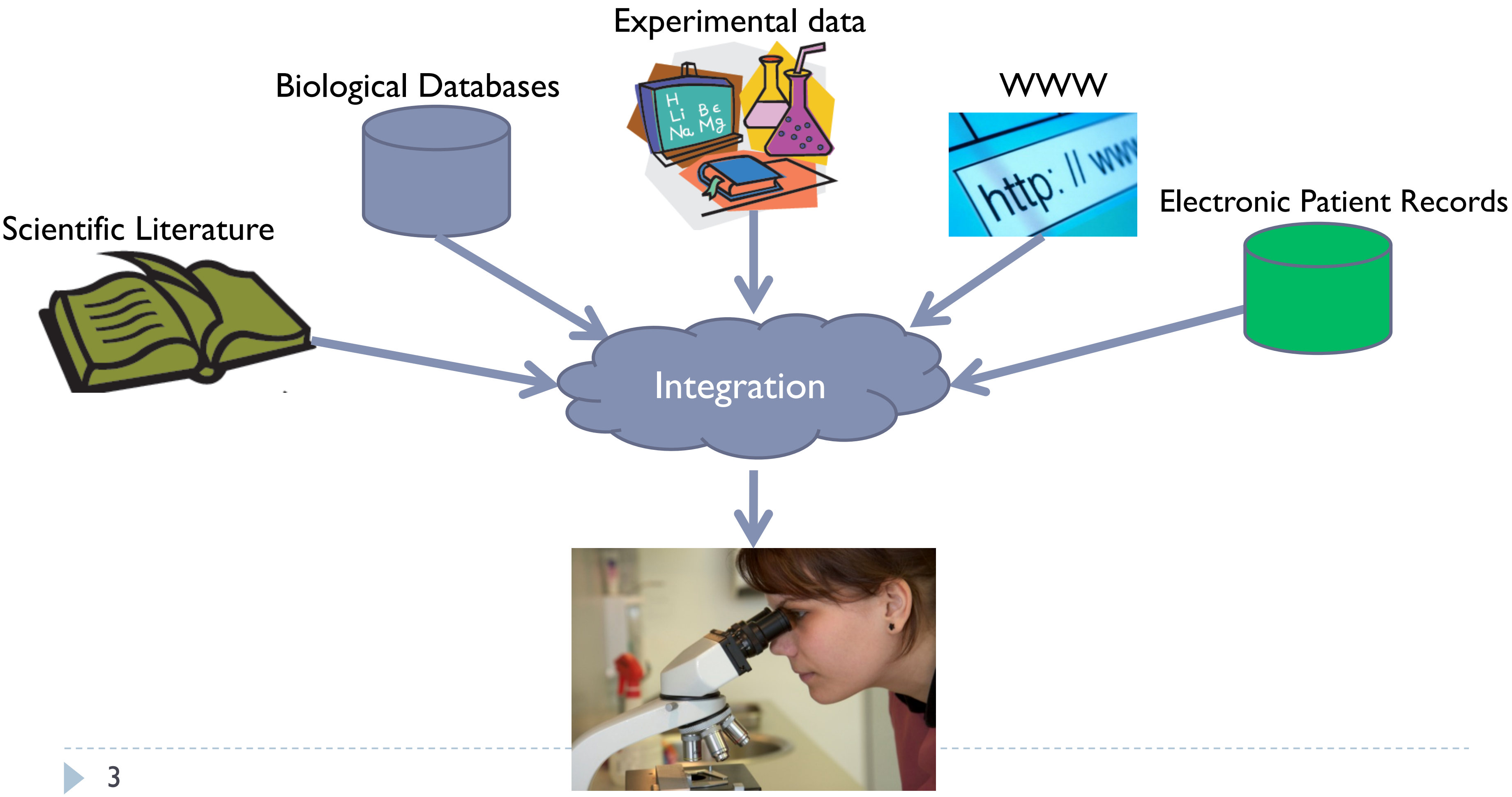
Biological Databases

WWW

Scientific Literature

Electronic Patient Records

Integration



Experimental data

Biological Databases

WWW

Electronic Patient Records

Scientific Literature

Integration

Integration

- Requires translation of information available in text resources to computable forms
  - Bridge the gap between basic biomedical research and clinical research
  - Translate both types of research into practice
- Core Technologies
  - Data/Text Mining
  - Natural language processing

Definition - Data Mining

- Development of methods and techniques for making sense of data – Pattern discovery and extraction in structured data.
  - mapping low-level data (which are typically too voluminous to understand and digest easily) into other forms that might be
    - more compact (i.e., a short report)
    - more abstract (i.e., a descriptive approximation or model of the process that generated the data)
    - more useful (i.e., a predictive model for estimating the value of future cases)

Definition - Text Mining

- Development of methods and techniques for making sense of data, but data is embedded in FREE-FROM TEXT!
- Challenges:
  - Natural language permits an enormous amount of expressiveness, variety, ambiguity, and vagueness
    - Easy for humans
      - □ Common sense knowledge
      - □ Reasoning capacity
    - Difficult for computers
      - □ No common sense knowledge
      - □ No reasoning capacity

Definition - Natural Language Processing

- Ultimate Goal: To build computer systems that perform as well at using natural language as humans do
- Immediate Goal: To build computer systems that can process text and speech more intelligently
- Areas:
  - Linguistics
  - Rule/grammar based approaches
  - Machine learning and statistics

In this seminar series:

- Lectures/presentations to very briefly introduce
  - Text mining/NLP sub problems:
    - Part-of-speech tagging
    - Parsing
    - Word-sense disambiguation
  - Machine learning techniques for text/data mining
  - Other data resources
    - Medical Knowledge-bases: i.e. UMLS
    - Corpora and datasets
  - Open source libraries
    - i.e. weka, minorthird, …

Lojistics

- Two class codes:
  - MEBI 591C – 1 credit
  - MEBI 598 – 3 credits
- Webpage: http://faculty.washington.edu/melihay/MEBI591C.htm
  - Slides + suggested reading list for the week
  - Related References
- Email List: mebi591c_sp10@u.washington.edu
- Time: Wednesdays, 3:30-4:20 p.m.
- Location: Health Sciences, Room E-212
- Office Hour: TBD
  - Monday-Tuesday-Thursday 10:00-12:00
- Instructor: Meliha Yetisgen-Yildiz
  - Email: melihay@u.washington.edu

Requirements

- Presentation (Required for 598 & 591)
  - 50 minutes presentation+discussion+question answering
  - Content
    - Research/Project Idea
      - Motivation + Problem + Potential Solution
    - Survey or literature review
      - A general area
        - Text mining: named entity recognition - gene name identification
        - Data Mining: classification, clustering
      - Available resources for a given area
        - Open source libraries
        - Data resources
    - Paper
      - Conference or journal article
  - Preparation:
    - Email the plan + reading list at least 3 days prior to class

Requirements

- System Design – i2b2 Challenge (Required for 598 – Optional for 591)

- The fourth i2b2 challenge is a three tiered challenge that studies:
  1. extraction of medical problems, tests, and treatments
  2. classification of assertions made on medical problems
  3. relations of medical problems, tests, and treatments

2010 - I2b2 Challenge

- Important Dates:
  - March 5th – Registration opens
  - April 15th – Commitment to Participate in Challenge & Training Data Release
    - E-mail me if you are interested in participating in this challenge!
  - July 15th – Test Data Release
  - September 1st – Short papers due
  - October 1st – Invitations to present at the Workshop
  - November, 2010 - Workshop

Benefits

- Hands-on programming experience with clinical text
- 1 workshop paper + 1 JAMIA paper (if invited)
  - Check class website for the links to JAMIA papers of previous challenges

System Design

- We will discuss in detail:
  - Problem
  - Corpus
  - Systems submitted to previous i2b2 challenges\
    - 2009 – Obesity Challenge
    - 2008 – Smoking Challenge
- Development Environment
  - OS:
    - Linux
  - Server:
    - patas at LING
  - Programming Language:
    - Java
  - Style:
    - Java Code Conventions: http://java.sun.com/docs/codeconv/
    - JavaDoc: http://java.sun.com/j2se/javadoc/
  - Editors:
    - IntelliJ: http://www.jetbrains.com/idea/
    - Netbeans: http://netbeans.org/

Email Me by Monday (April 5th)

- Deadline to fax signed data use agreement: April 15th

Tentative Schedule

- Week #1 - 03/31: Introduction and planning – melihay
- Week #2 - 04/07: Text Mining/NLP Sub-problems – melihay
- Week #3 - 04/14: Machine learning in Data/TextMining/NLP – melihay
- Week #4 - 04/21: presentation - TBD
- Week #5 - 04/28: presentation - TBD
- Week #6 - 05/05: presentation - TBD
- Week #7 - 05/12: presentation – TBD
- Week #8 – 05/19: presentation - TBD
- Week #9 - 05/26: i2b2 - Solution Proposals
  - extraction of medical problems, tests, and treatments
  - classification of assertions made on medical problems
- Week#10 - 06/02: i2b2 - Solution Proposals
  - relations of medical problems, tests, and treatments

Questions