ECE8813
Statistical Language Processing

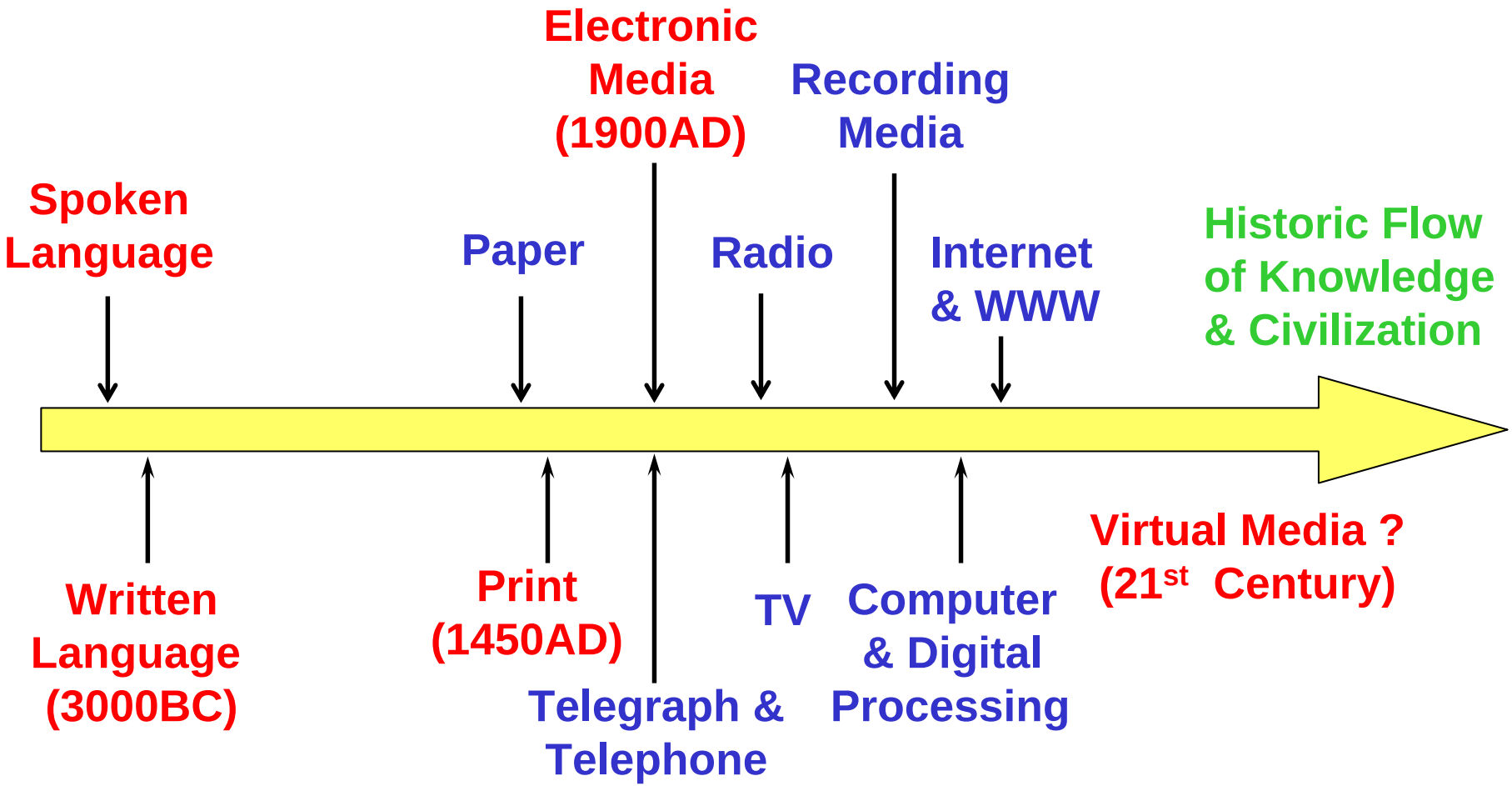
Lecture 1: Introduction

Chin-Hui Lee
School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, GA 30332, USA
chl@ece.gatech.edu

Course Information

- **Subject:** Statistical Language Processing
- **Prerequisite:** ECE3075, ECE4270
- **Background Expected**
  - Basic Mathematics and Physics
  - Digital Signal Processing
  - Basic Discrete Math, Probability Theory and Linear Algebra
- **Tools Expected:**
  - MATLAB and other Programming Tools
  - Language-specific tools will be discussed in Class
- **Teaching Philosophy**
  - Textbooks and reading assignments: your main source of learning
  - Class Lectures: exploring beyond the textbooks
  - Homework: hand-on and get-your-hands-dirty exercises
  - Class Project: a good way to go deeper into a particular topic
- **Website:** http://users.ece.gatech.edu/~chl/ECE8813.sp09

Evolution of Language and Media



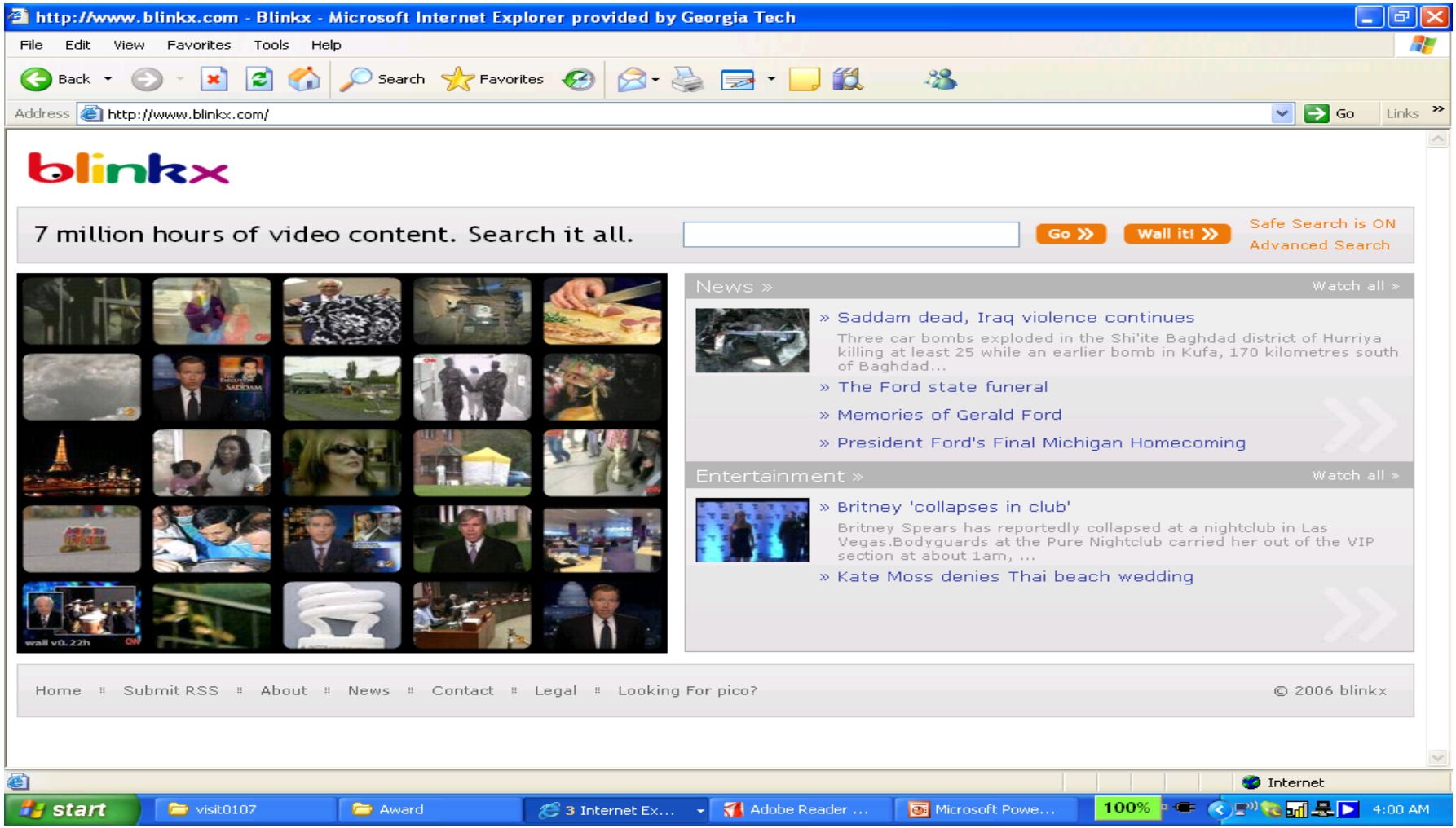
Information in Language

- **Human Language Communication**
  - Taxonomy, speech and language
  - Information: topics, intention, summarization, writer, etc.
- **Linguistic Information**
  - Words, sentences, paragraphs, documents, etc.
  - Language-specific characteristics
- **Language information in media**
  - Speaker profile: gender, age, accent, size, education…
  - Other factors: speaking rate, emotion state, etc.

Computational linguistics to extract language information !!

First Google, Now Google News

Blink-X: Video Search with ASR Output



Information Explosion on the Web (Part of "The World is Flat" by Friedman)

- **Over two billion of English web documents in 2005**
  - Still growing exponentially in most languages (>1B for Chinese)
  - From text only to multimedia and multilingual documents
- **Global data storage explosion (HD+network storage)**
  - In 2002, global information increased by 5 billion GBs, about 800MB per person, enough to fill up 500,000 Library of Congress
- **Memory explosion on PC and portable devices**
  - From 512 KB for PC (Gates) to portable Library of Congress
- **Cell phone sales close to 500M units in 2003**
  - Mobile information appliances are commodities and fashions
  - **A million iPhone sold in 74 days (2 years for iPod)**
- **From Command/Control to Connect/Collaborate**
  - **World leaders & executives access Google, Blackberry, iPhone**

Study Topics and Applications

- Introduction
- Mathematical foundations
- Linguistics essentials
- Corpus-based linguistic analysis
- Word collocation
- Statistical *n*-gram, Markov models
- Part-of-speech tagging
- Word sense disambiguation
- Probabilistic parsing
- Information retrieval
- Text categorization and topic identification
- Parallel alignment and statistical machine translation
- Other Related Topics

Grading Policy

- Homework: 25%
- Examinations: 40% (to be discussed)
  - Quiz#1 (10%), Quiz#2 (10%), Final (20%)
- Class Project: 35%
  - On language processing related topics
  - A list of possible projects will be posted and discussed
  - MATLAB and C/C++ tools are essential
  - Project term papers are due before the Finals Week

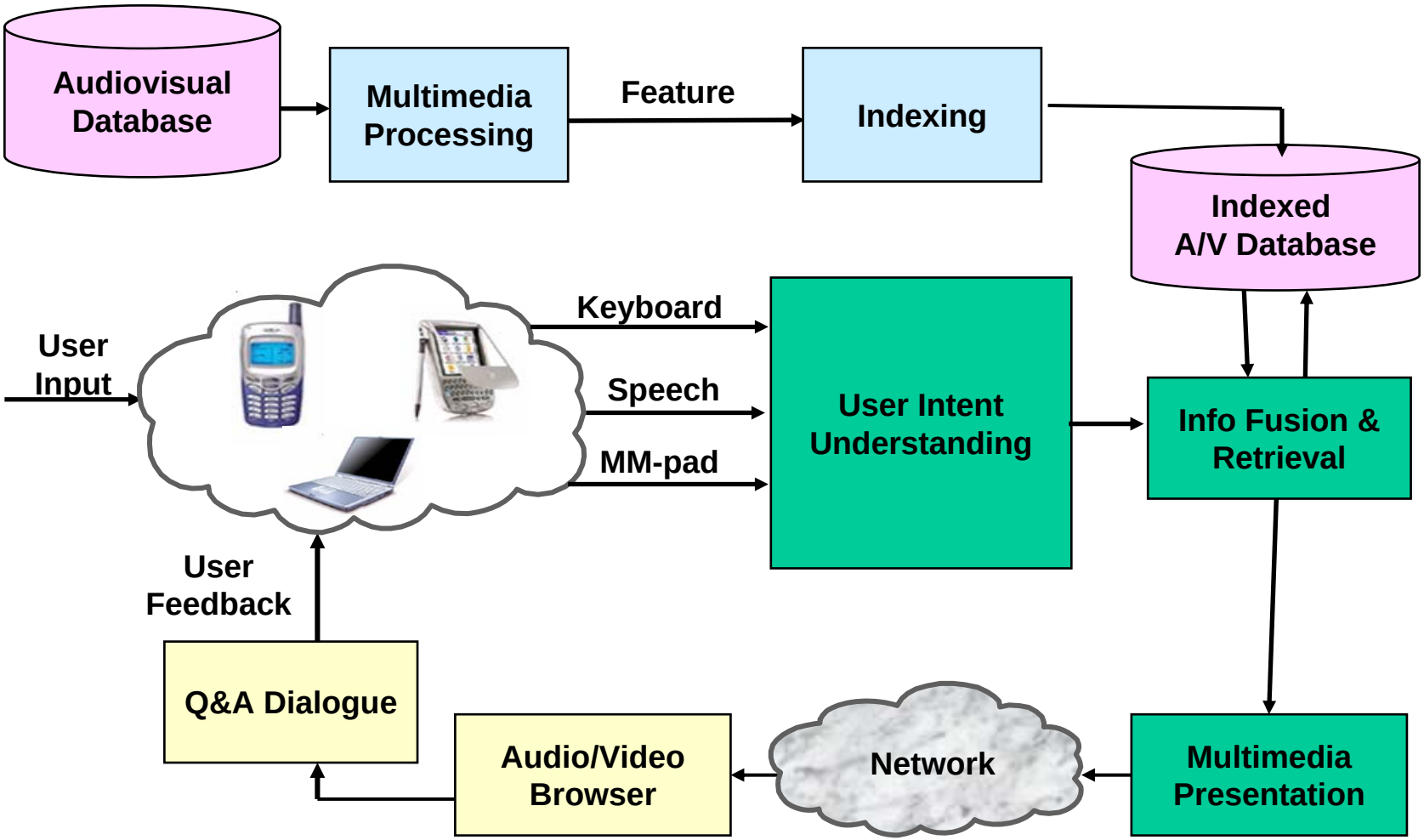
Support Information

- Lecture: TR 12:05-13:25
- Office Hours: TR 10:30 – 11:30 or appointment
- Textbook:
  - Manning and Schutze: *Foundations of Statistical Natural Language Processing*, MIT Press, 2001
- Reading:
  - C. Cherry, *On Human Communications*, MIT Press, 1968
  - D. G. Stork (ed.), *HAL's Legacy,* MIT Press, 1997
- Contact Information
  - Centergy One Rm 5180, 404-894-7468, chl@ece.gatech.edu

Web Information Access & Presentation

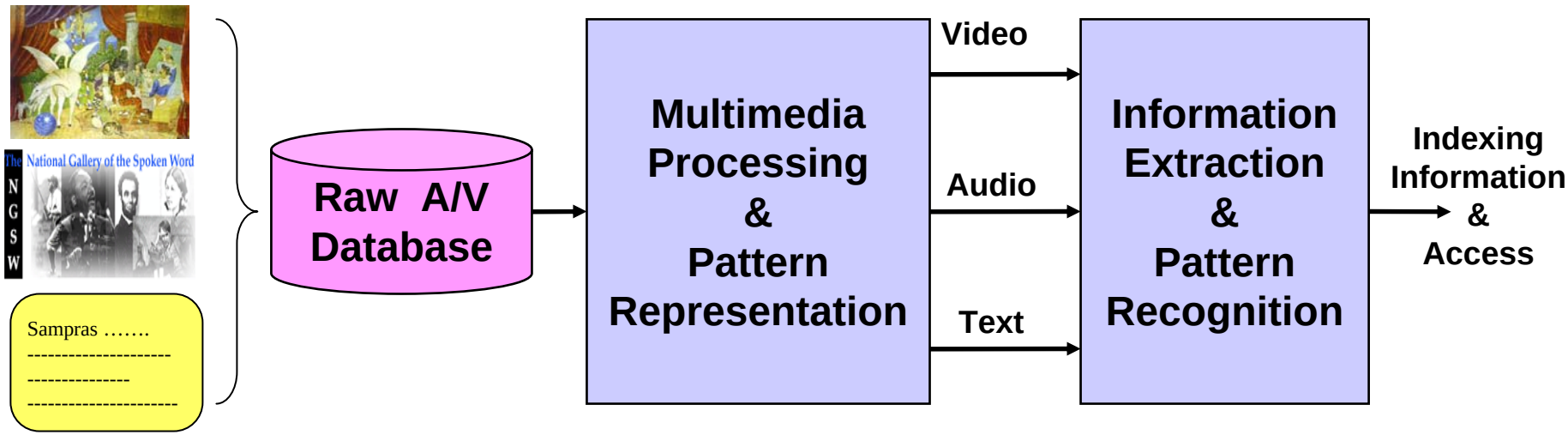
- Web data mining
- Web content extraction
- Topic detection and automatic summarization
- Information rendering and presentation
- Q&A construction for natural interface

12. A Grand View: Multimodal Access of Multilingual Multimedia Information



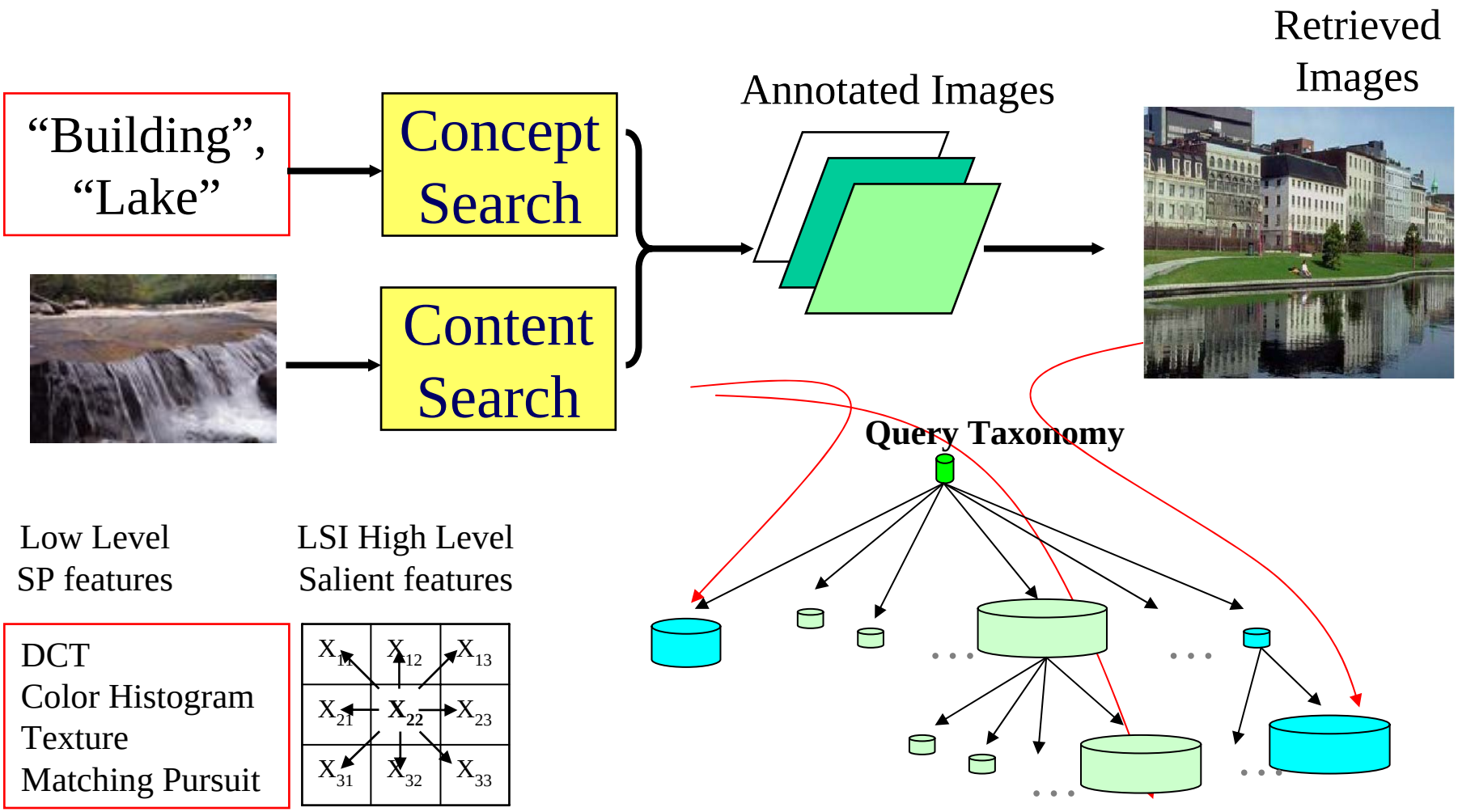
Information Extraction from Video

- Video: speech, audio, image, text, and others

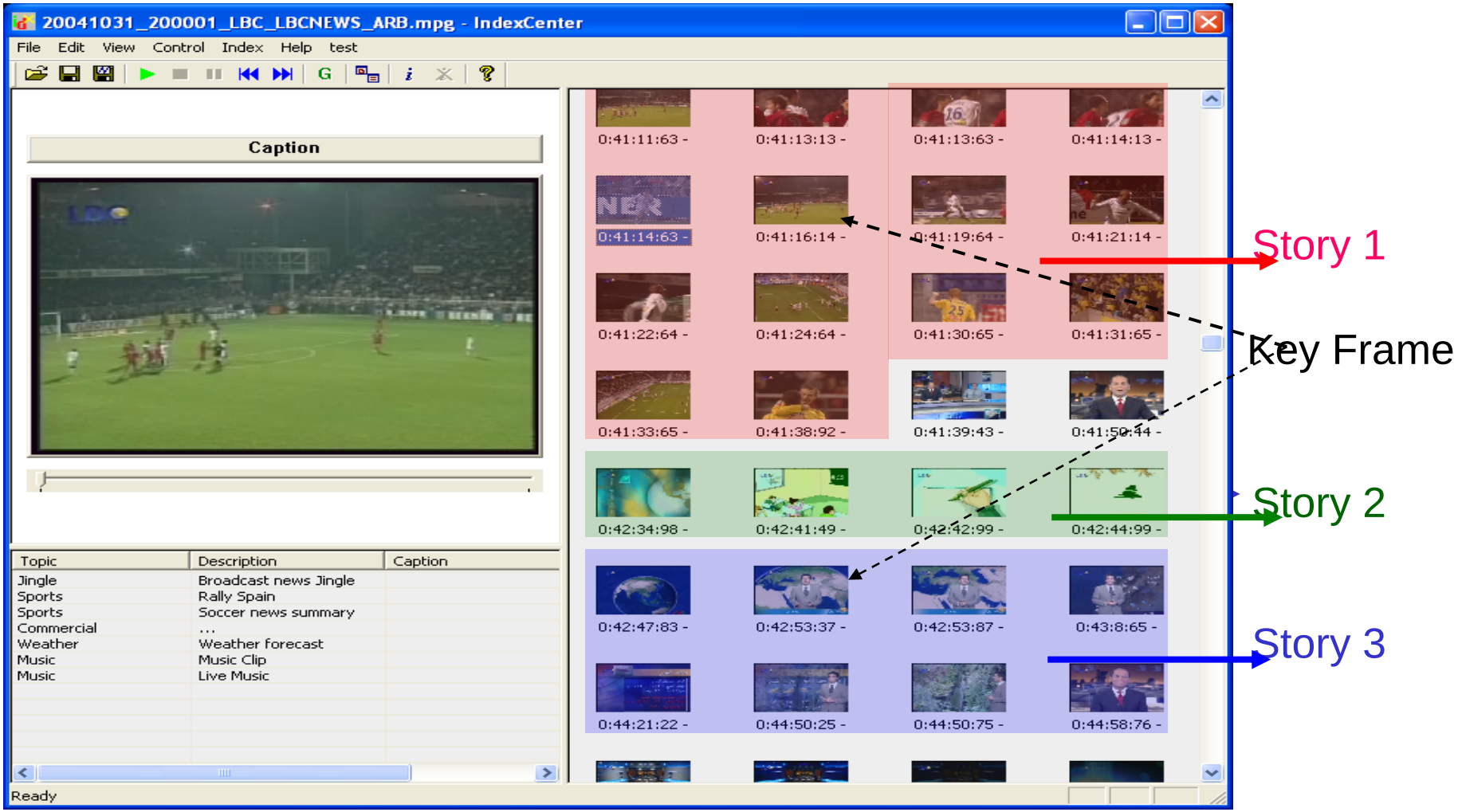


Info extraction through media pattern recognition: text information may come from close caption, video text, speech recognition or image annotation

Concept vs. Content Based Image Search



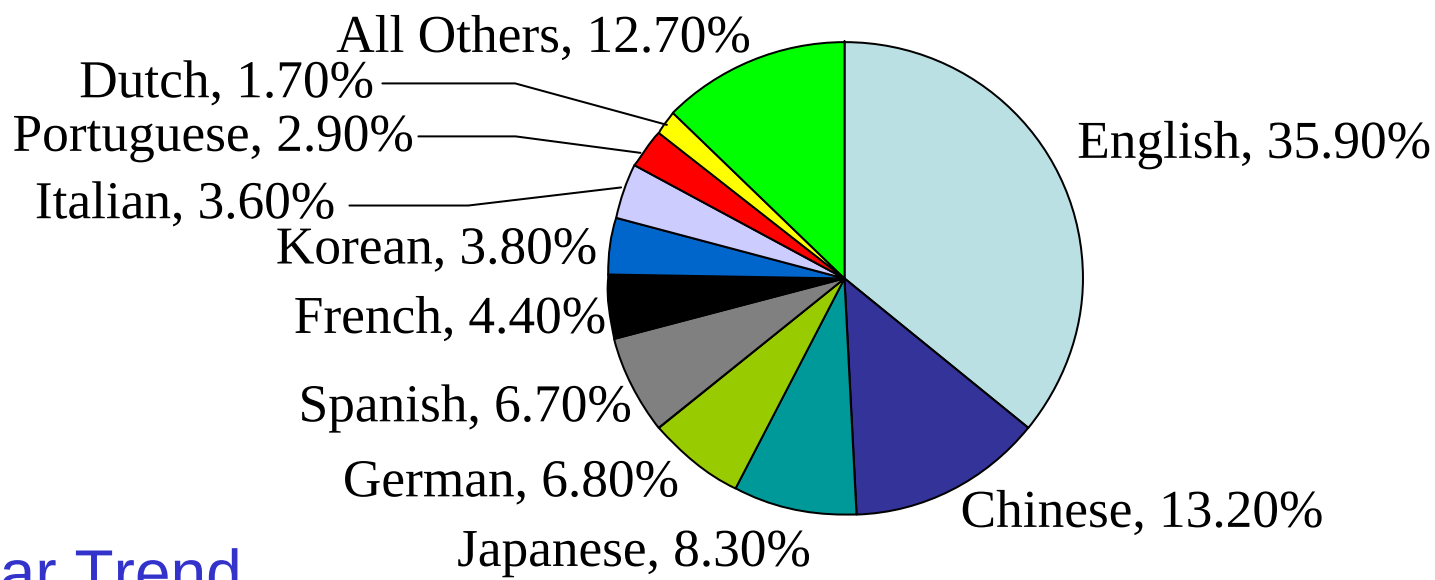
Video Story Segmentation



Video Clip Browsing over IP on 3G

Multilingual Web (Content is King)

- Internet Users by Language (end of 2004, 800 million)



- Clear Trend
  - Non-English users continuously increasing
  - Japanese and Chinese are currently (and may continuously be) the two largest non-English groups (provided by L.-s. Lee, NY+TU)

Example: Google (Translate This!)

- “Que ce soit à l’Halloween ou dans ta vie de tous les jours, ils y a des règles à suivre lorsque tu dois traverser la rue. »

➔

- “That it is in Halloween or in your life of the every day, they has rules there to follow when you must cross the street.”

Google’s statistical machine translation system outperformed all other state-of-the-art competitors using billions of words as training data running on thousands of machines constantly updating *n*-grams LM

Projects for My Statistical NLP Course

1. *N*-gram for ranking word order from a bag of words
2. Document and key term clustering
3. Text categorization and topic identification
4. IR: document indexing and Retrieval (Google Text?)
5. Part-of-speech tagging
6. Domain-specific message understanding
7. Language identification of encrypted documents
8. Automatic image annotation
9. Spoken language identification
10. Video story segmentation
11. Transliteration and other multilingual applications
12. Music genre classification
13. Any others?

Summary

- Today's Class
  - Information about ECE8813, Spring 2009
  - Web: http://www.ece.gatech.edu/~chl/ECE8813.sp09
  - Class web page and data will be ready soon
- Next Class
  - Mathematical Foundations on Jan. 8
- Reading Assignments
  - Manning and Schutze, Chapters 1 & 2