# CS/INFO 431
# Web Information Systems
# Spring 2007

Carl Lagoze
Theresa Velden

# What is a *Web Information System*?

- The WWW provides an *interoperability layer* for a general-purpose networked information environment.
- A *web information system* exploits this architecture and extends it for targeted applications
  - Education
  - Scholarship
  - Cultural heritage
  - Commerce
  - …

# Lot's of people calls these *Digital Libraries*

- Indeed some of the things we'll be talking about are digital libraries
- Used by funding agencies (e.g., NSF) to promote research in this area
- Lots of "digital library" conferences
- Problems
  - Library is somewhat of a reserved word
    - Professional ownership
  - Library has good and bad connotations
  - Not all information environments should be tagged as libraries

# Beyond Access and Search

- The web provides a basic organize (link) and access (HTTP) architecture
- Google, etc. provides a great general-purpose search service over that architecture
- But there's got to be more!!
- The web is really just one giant database on which to develop and *organize*:
  - information
  - knowledge
  - wisdom

# Layering over the data

- Mashing up
- Aggregating
- Organizing and relating
- Integrating the data with computational services
- Integrating data with social behavior and networks
  - Capture reuse, and benefit from the "wisdom of crowds"
  - Personalize information at various levels of community

# And let’s not forget that libraries do more…

- Preservation
- Privacy
- Integrity and authenticity
- Selection

# Library Tradition

- Functions
  - Selection
  - Collection
  - Organization
  - Reference
  - Preservation
- Characteristics
  - Standardized
  - Professionalized
  - Service-oriented
  - In it for the long-haul
  - Conservative


# Web “Tradition”

- Decentralized/Anarchic/Illegal
- Agreements are technical (at best)
- Roles are undefined and fluid
- You don’t have to be an expert (or “no one knows you are a dog”)
- Immediate
- Ephemeral

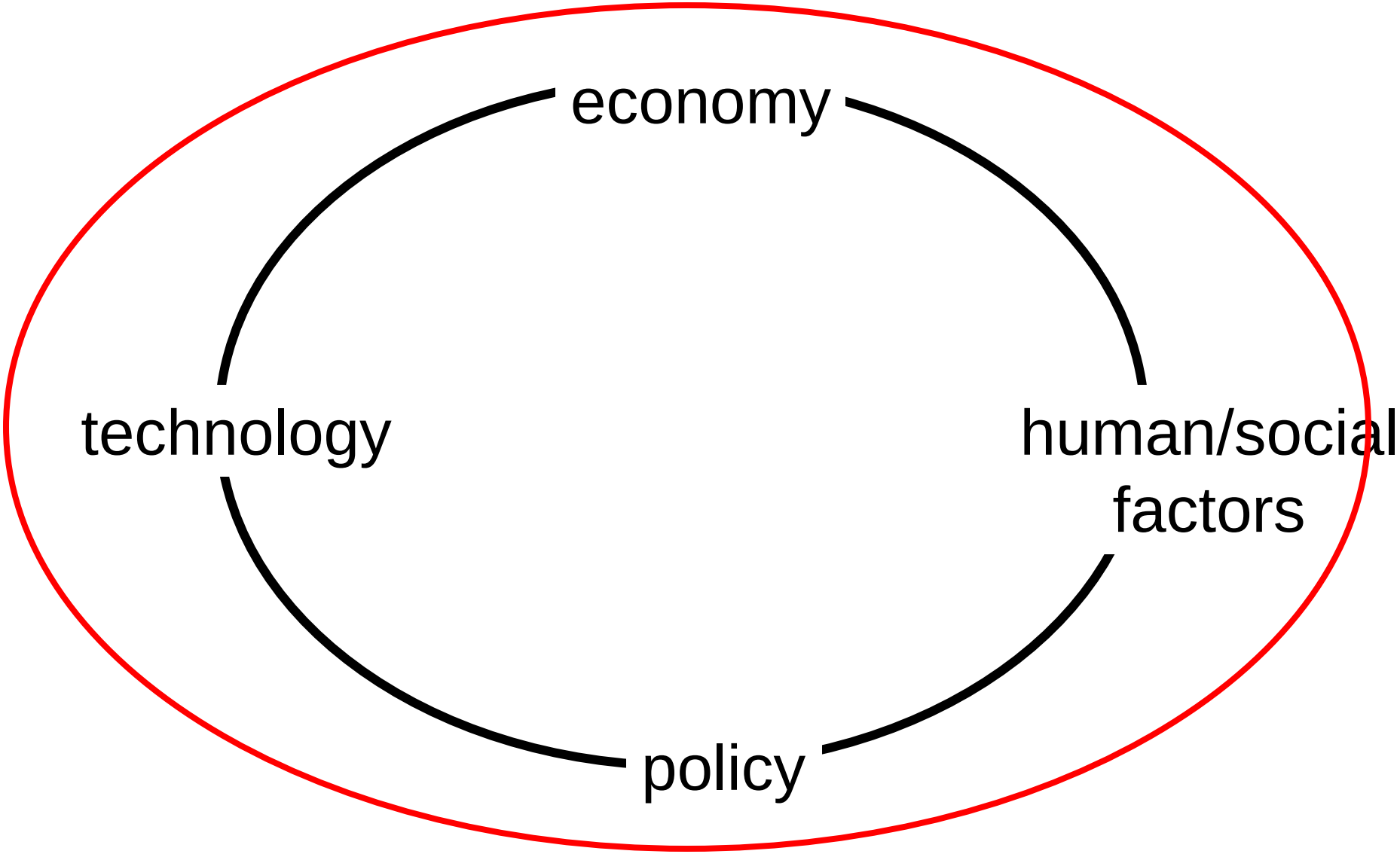
# Finding the Appropriate Blend


- **There are many points on this spectrum**
  - Evolutionary perspective: preserve traditional information institutions such as libraries but adapt them to digital context
  - Revolutionary perspective: ubiquitous computing and networks render many traditional practices irrelevant

# Building systems to add value

- At their core libraries add value to content (organize, select, preserve)
- The Web and Internet is the largest collection of data known to humans
  - Traditional "library-like" content
  - Informal content
  - Artifacts of social interaction
  - Integration of content (data) and services (computation)
- How can we build relationships and integrate this content to add value to it
  - create that data->information->knowledge->wisdom continuum
  - information network overlay

# Many facets of the problem/solution



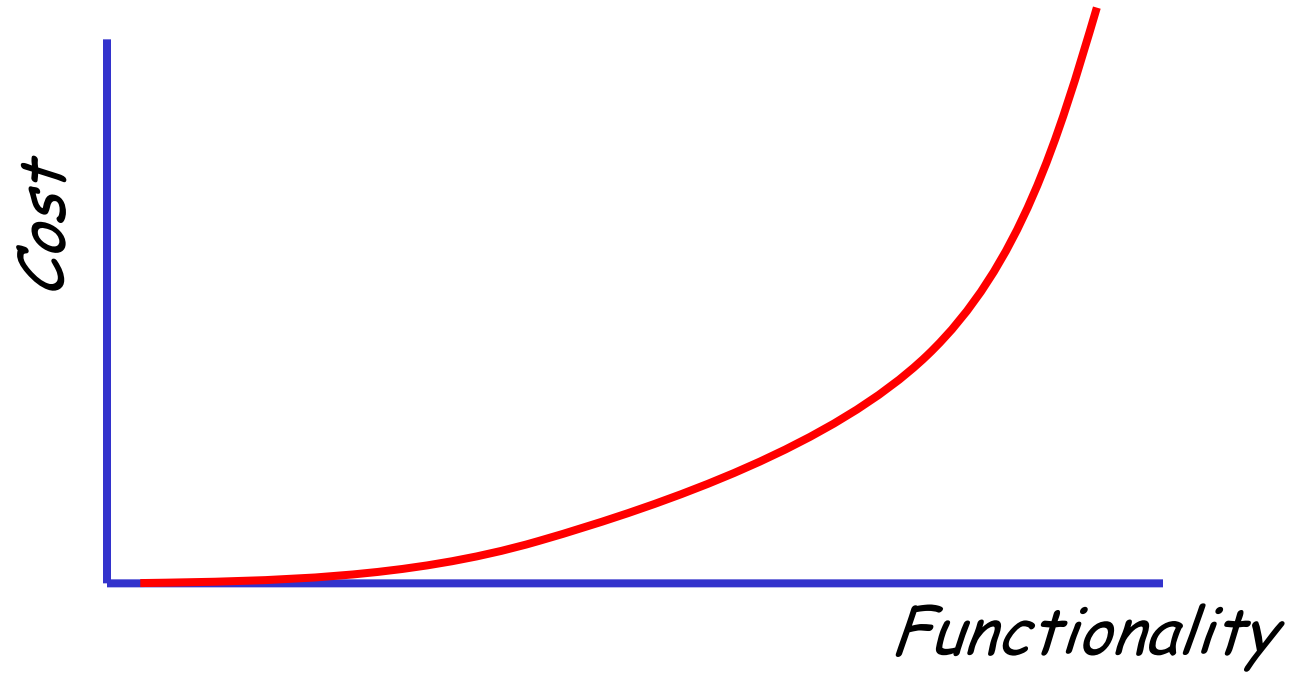
# What we'll talk about in this course

- Web Architecture
  - URIs
  - Web graph structure (Resources, Representations, Links)
  - Web protocols (HTTP)
- Information Units
  - Identity
  - Types
- Relating/Organizing Units
  - Cataloging
  - Metadata
  - Information Models
- Semi-structured data
  - XML
  - Schema
  - XSLT
- Semantic Web
  - RDF
  - Ontologies
- Service-oriented architectures
  - Integrating content and computation
  - Workflows
- Information Preservation
  - Traditions
  - New Models
- Intellectual Property
  - Copyright
  - Rights Management
  - Privacy
- Scholarly Publishing

# Interoperability

- Allowing distributed heterogeneous systems to work together
- Exists on many dimensions
- The web is an interoperability layer

# Interoperability Cost/Benefit Curve



# Course Web Resources

http://www.cs.cornell.edu/Courses/cs431/2007sp

# Code of Academic Integrity

http://cuinfo.cornell.edu/Academic/AIC.html

# Some Pet Peeves

# Lagoze's general course philosophy

- A course is a collaborative experience
- Instructor provides the structure and foundation for learning
- Student engages, contributes, challenges
- We learn from each other

And now for some history…

# Library of Alexandria

- Established by Ptolemy II in 290 BC
- 532K papyrus rolls
- Acquisition by copying mandate
- Destroyed in 490 AD, lots of theories and stories:
  - Burning alive of Hypatia, the last keeper of the library
- http://www.bibalex.org/English/index.aspx

# Melvil Dewey

- "Father of modern librarianship"
- Frustrated by dedicated shelving method
- Invented method of classifying into 10 categories
  - Note cultural artifacts
- 21st edition of Dewey Classification system now published
- Started ALA

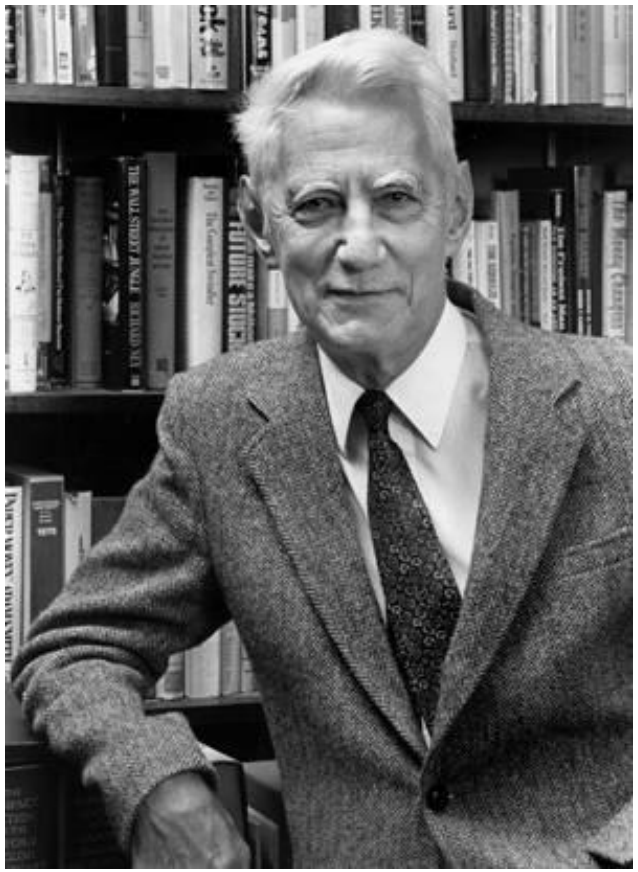
# S. R. Ranganathan

- Colon Classification System
- 42 main classes
- Subject classification by appending facets within class: who, what, when, where

# Vannevar Bush

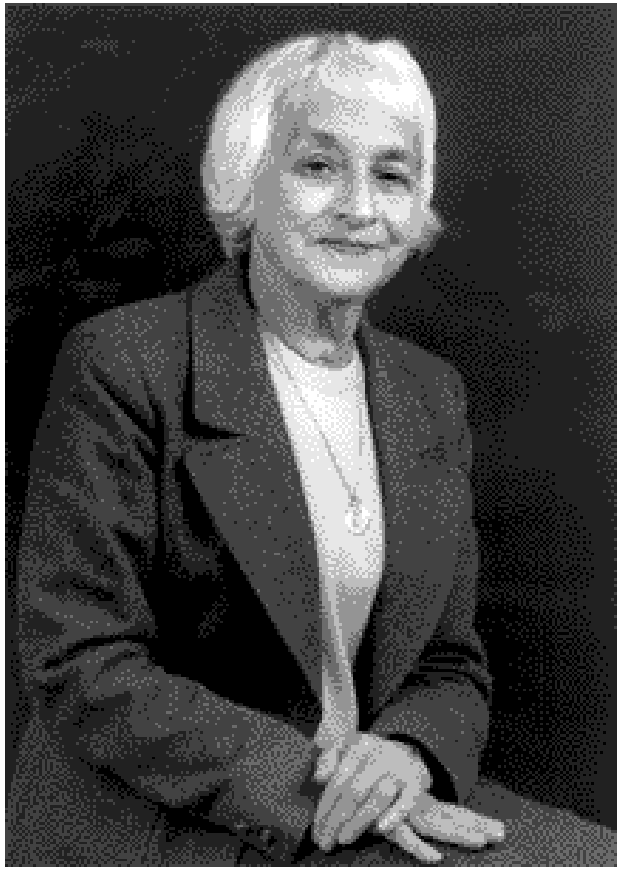
- "As We May Think" Atlantic Monthly 1945
- Pivotal landmark in hypertext research
- "This is the essential feature of the *memex*. The process of tying two items together is the important thing"

# Claude Shannon



- "Father of Information Theory"
- Seminal "The Mathematical Theory of Communication"
- Data vs. Information

# Henriette Avram



- "Mother of MARC", "Melvil Dewey of the 20th Century"
- Developed MAchine Readable Cataloging (MARC)
- Allows standardization and sharing of bibliographic records

# J.C.R. Licklider

- “Man-Computer Symbiosis”
- Developed the idea of the “universal network” and interactive computing
- Developed and led ARPANET funding initiative

# Inventors of Internet

- Cerf, Kahn, Metcalfe, etc.
- Packet rather than circuit switching
- Layered protocols (TCP/IP, telnet, ftp…)

# Ted Nelson

- Inventor of the notion of “non-sequential writing” and term “hyptertext” and “hypermedia” circa 1960
- Founder of Project Xanadu

# Gerard Salton

- Preeminent figure in modern information retrieval
- SMART information retrieval system: basis of many well-known IR concepts
- Among founders of Cornell CS department

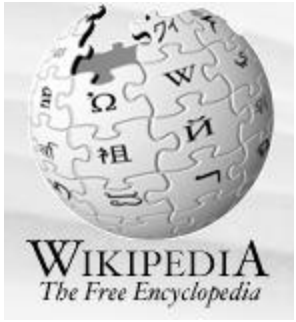
# Tim Berners-Lee

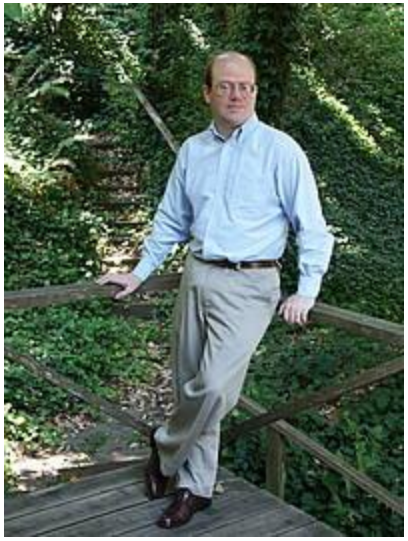

- Inventor of the World Wide Web – CERN 1989
- First client and server 1990
- Director of World Wide Web Consortium and faculty at MIT

# Sergey Brin and Larry Page

- Two Stanford students who failed to get their Ph.Ds.

# “Jimbo” Wales and Larry Sanger



- Made the wisdom of crowds an amazing reality.

# CS 431 Student

# Who am I?

- Member of Information Science Program
- Research areas: interoperability architecture, metadata, document architecture, Scholarly Publishing
- Publications, Personal, etc.
  - http://www.cs.cornell.edu/lagoze/