

DATE: 7.3.2005

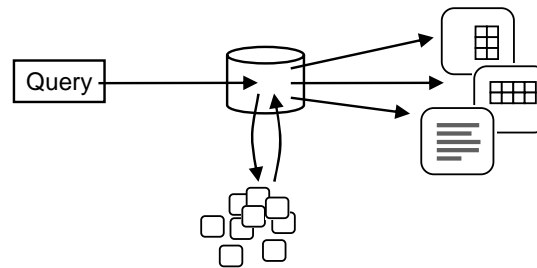
Short Introduction to Web Information Extraction

Whereas IR retrieves documents out of a collection of documents, IE extracts relevant information from such documents

Information Retrieval vs. Information Extraction

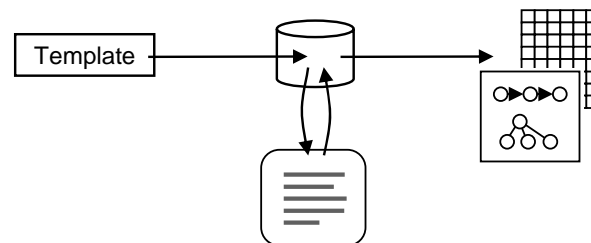
DATE: 7.3.2005

IR



- Retrieve relevant documents from collections of documents
- Probability Theory
- Statistics

IE

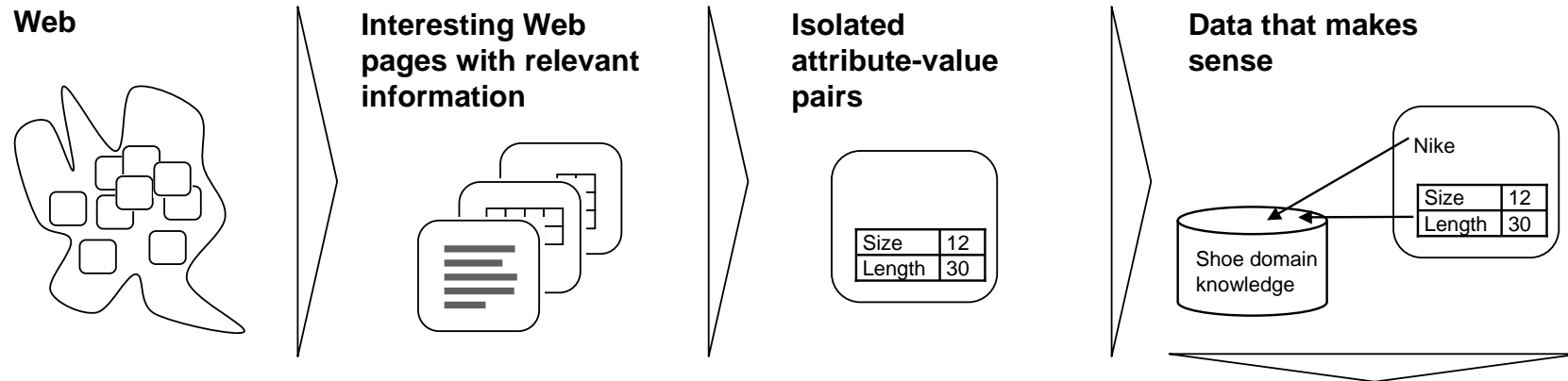
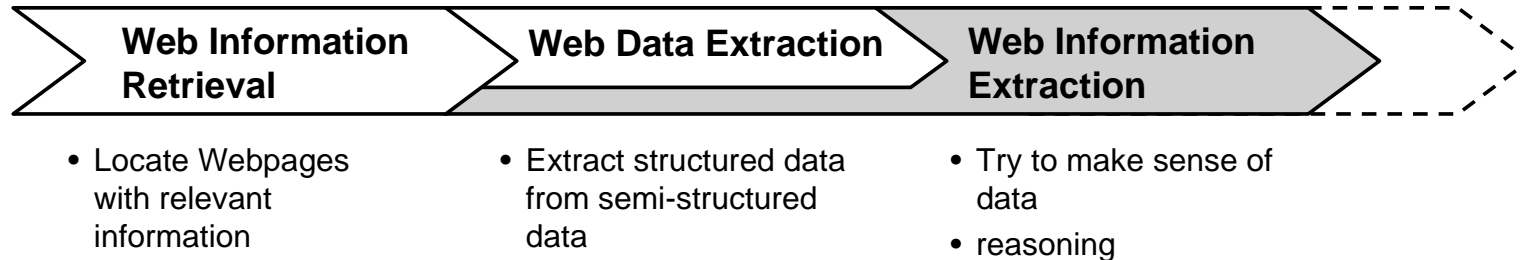


- Extract relevant information from documents
- Rule-based systems in computational linguistics
- NLP

In the context of the Web, IR and IE can be seen as consecutive processes

Web Information Extraction – a personal view

DATE: 7.3.2005



- Web page talks about shoes
- Size 12 & Length 30 -> Web page uses rather rare USA Children Size
- ? American Webstore

Defintion by Meadow*:

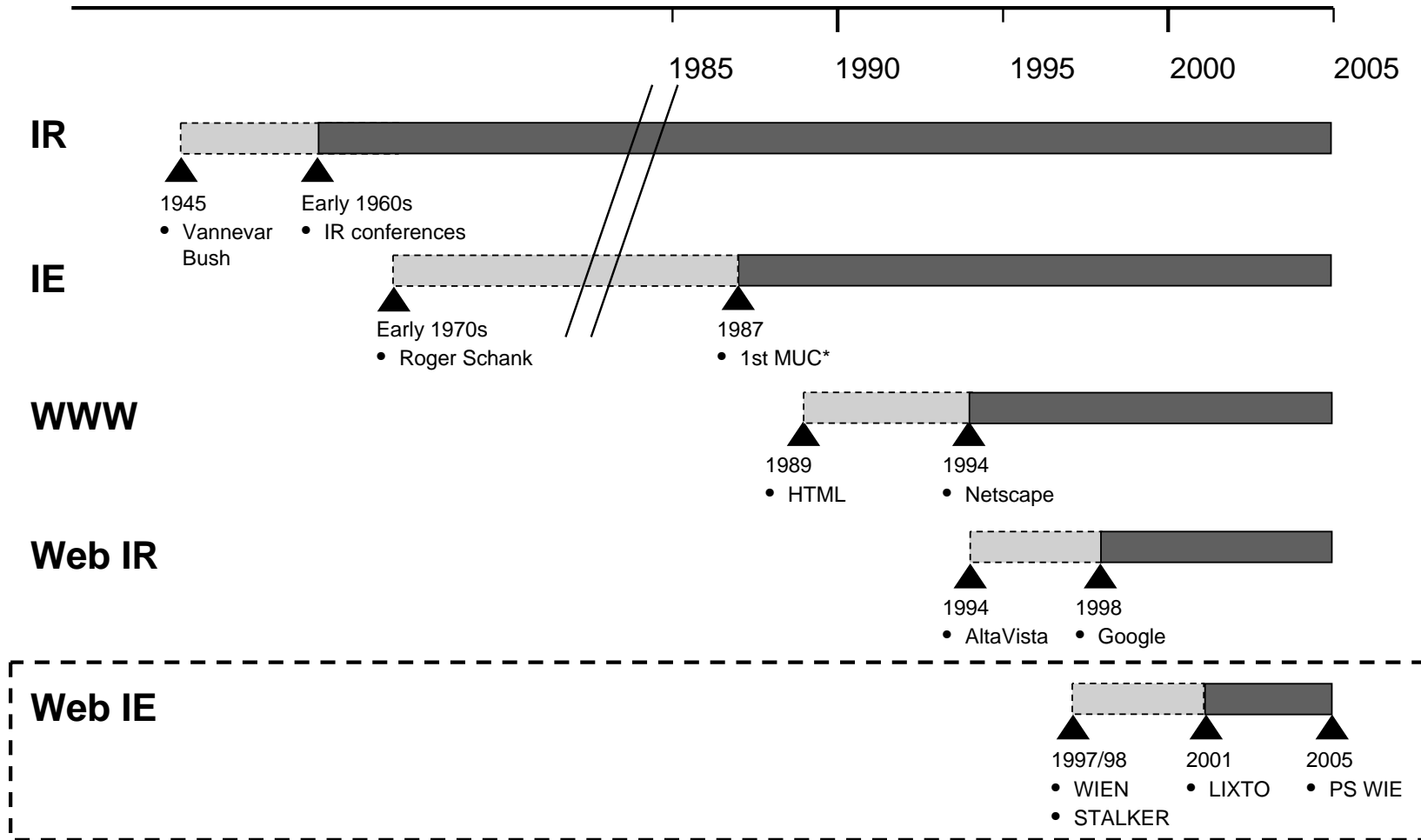
- Data: isolated attribute-value pairs
- Information: data in a conceptual framework

* C.T. Meadow, Text Information Retrieval Systems, Academic Press, 1992
Source: Wolfgang Gatterbauer, www.dbai.tuwien.ac.at/education/wie

This Proseminar is going to focus on Web Information Extraction (WIE), a rather young research area

Web Information Extraction – Timeline

DATE: 7.3.2005



* MUC (Message Understanding Conferences): “Analyzing free text, identifying events of a specified type, and filling a data base template with information about each such events.”

Source: Wolfgang Gatterbauer, www.dbai.tuwien.ac.at/education/wie