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**

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

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

Source: Wolfgang Gatterbauer, www.dbai.tuwien.ac.at/education/wie
Web Information Extraction – a personal view

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

Web Information Retrieval
- Locate Webpages with relevant information

Web Data Extraction
- Extract structured data from semi-structured data

Web Information Extraction
- Try to make sense of data
- reasoning

--

Definition by Meadow*:
- Data: isolated attribute-value pairs
- Information: data in a conceptual framework

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

* 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