

Introduction to **Information Retrieval**

Structured vs. Unstructured Data

IR vs. databases:

Structured vs unstructured data

- Structured data tends to refer to information in “tables”

Employee	Manager	Salary
Smith	Jones	50000
Chang	Smith	60000
Ivy	Smith	50000

Typically allows numerical range and exact match (for text) queries, e.g.,

Salary < 60000 AND Manager = Smith.

Unstructured data

- Typically refers to free text
- Allows
 - Keyword queries including operators
 - More sophisticated “concept” queries e.g.,
 - find all web pages dealing with *drug abuse*
- Classic model for searching text documents

Semi-structured data

- In fact almost no data is “unstructured”
- E.g., this slide has distinctly identified zones such as the *Title* and *Bullets*
 - ... to say nothing of linguistic structure
- Facilitates “semi-structured” search such as
 - *Title* contains data AND *Bullets* contain search
- Or even
 - *Title* is about Object Oriented Programming AND *Author* something like stro*rup
 - where * is the wild-card operator

Introduction to **Information Retrieval**

Structured vs. Unstructured Data