Document Classification

Miroslav Halas
(mhalas@smu.edu)
CSE 8331 Data Mining
Southern Methodist University

Paper vs Digital

- Shift towards electronic documents, such as invoices, bank statements...
- Advantages: storage, delivery, reliability, cost...
- Disadvantages: copying, authenticity, experience (=trust)...
- Facilitate the shift
  - Improve handling of increased amount of paper
  - Process it electronically
  - Computers needs to understand the documents

Definitions

- Paper document – convey information as iconic symbols (no photographs)[34]
- Electronic document = document – collections of images (pages)
- Form – Manhattan type layout[18], horizontal and vertical lines, preprinted and user filled data (fields)
- Focus – fill-in forms, business and scientific documents –> benchmark - University of Washington image database I – III [9, 14, 17, 19, 20, 21...]