greatlearning

Machine Learning

Text Mining

greatlearning

- Nature of Corporate Data
  - 20-75-5 Rule
  - Structure of Data
- Text as Data
  - Loose Structure
  - Poor Spellings
  - Non traditional g
  - Multi-lingual

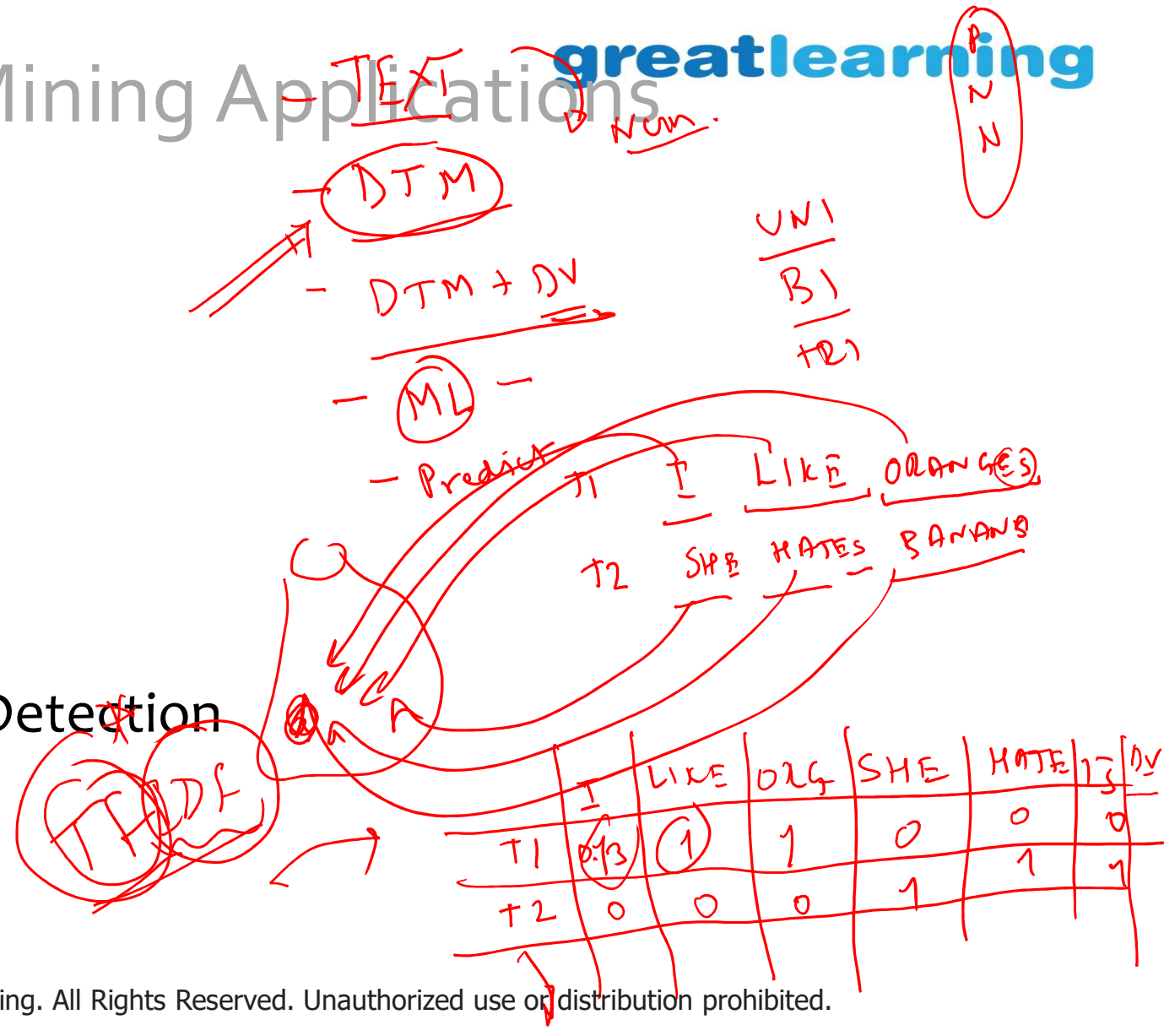
Why is it Hard?

greatlearning

- Ambiguity:
  - "I put my bag in the car. It is large and blue"
  - "It" = bag? "It" = car?
- Context:
  - Homonyms, metaphors
  - Sarcasm

Text Mining Applications

- Web and Social Media
  - Discovery of:
    - Sentiments
    - Opinions
    - Emotions
    - Topics
- Fraud and Irregularity Detection
- Spam identification
- Business Intelligence
- Content Enrichment



Process of TM

greatlearning

- Generic Process
  - Problem or Challenges
  - Collecting data
  - Human input- Average
  - Text mining packages
  - Creating Corpus
  - Processing Text- upper, lower, stop words, stem
  - Document Term Matrix
- Using Methods of Analytics