

Outline

- Supervised learning
 - Introduction

Why supervised learning?

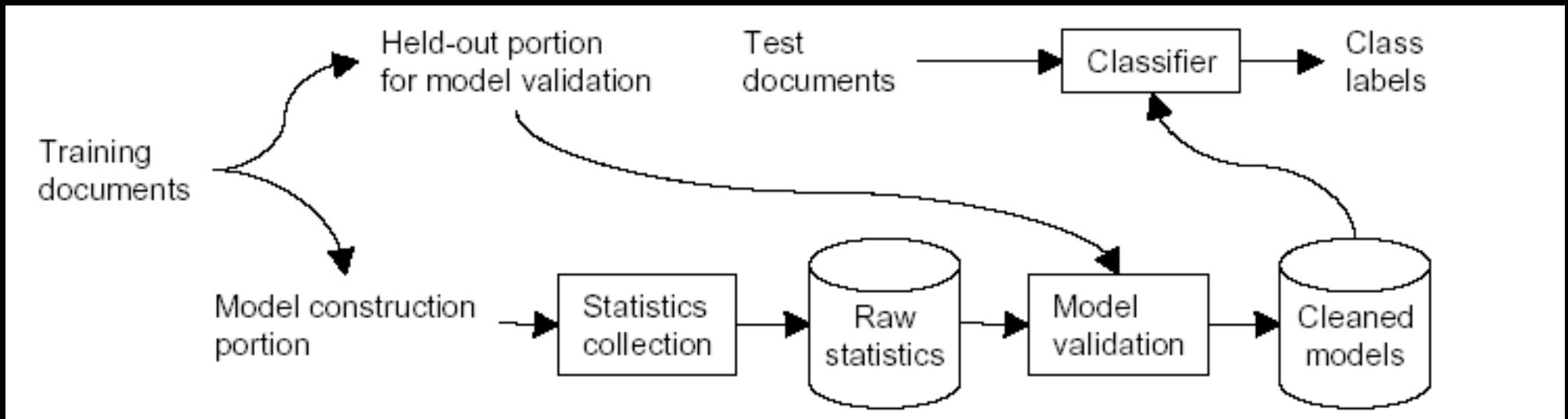
- Many text classification problems
 - E-mail filtering
 - Bookmark organization
 - News story tracking
 - Author identification
 - e.g., Federalist Papers
 - Assign topic labels to documents
 - Hyperlink classification
- In general – to learn to organize knowledge

Organizing knowledge

- Systematic knowledge structures
- Ontologies
 - Dewey decimal system
 - Library of Congress catalog
 - AMS Mathematics Subject Classification
 - US Patent subject classification
- Web catalogs
 - Yahoo & dmoz ODP
- Problem: Manual maintenance

Supervised learning

- Learning to assign objects to classes given examples
- Learner (classifier)



A typical supervised text learning scenario.

Differences with text

- Machine learning classification techniques used for structured data
 - Well-studied
 - Typically limited set of features
- Text
 - Lots of features and lots of noise
 - No fixed number of columns
 - No categorical attribute values
 - Data scarcity
 - Larger number of class labels