

Outline

- Introduction to indexing
 - Document preparation
 - The vector space model
 - Indexing
 - Evaluation

Document Preparation

- Standardize
 - file formats
 - labels/names/categories (metadata)
 - determine useful content
- Also called normalization
- Real world documents have unexpectedly varied contents...
 - e.g., <html>, winzip.exe, madonna.mp3

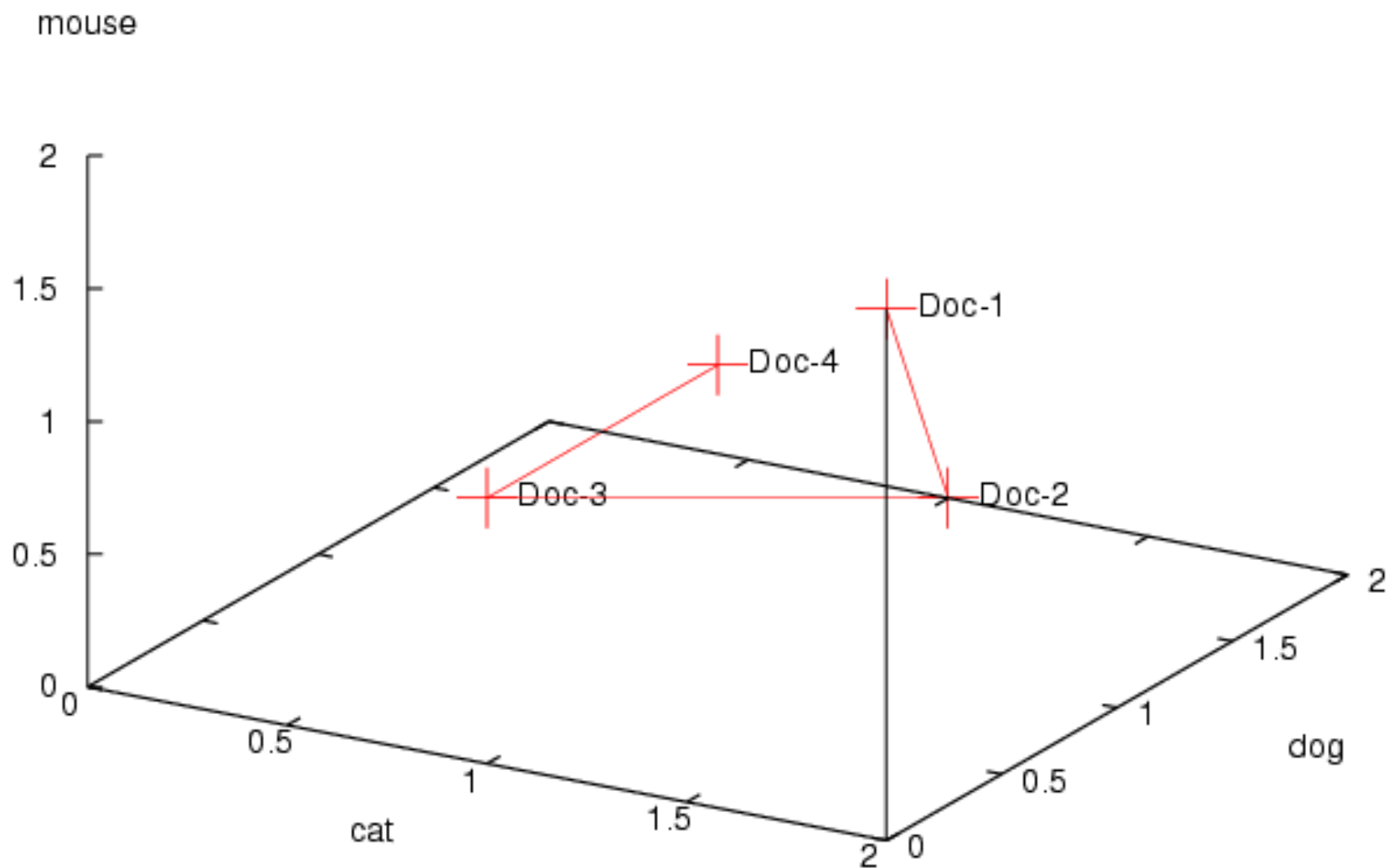
Vector Space Modeling

- Terms and documents represented by vectors of numbers
- Component numbers (i.e., each dimension) can represent presence or quantity of terms, concepts, or keywords
- Relevance is measured by similarity (distance) between vectors
- Set of document vectors form a matrix
 - e.g., a Term-Document matrix

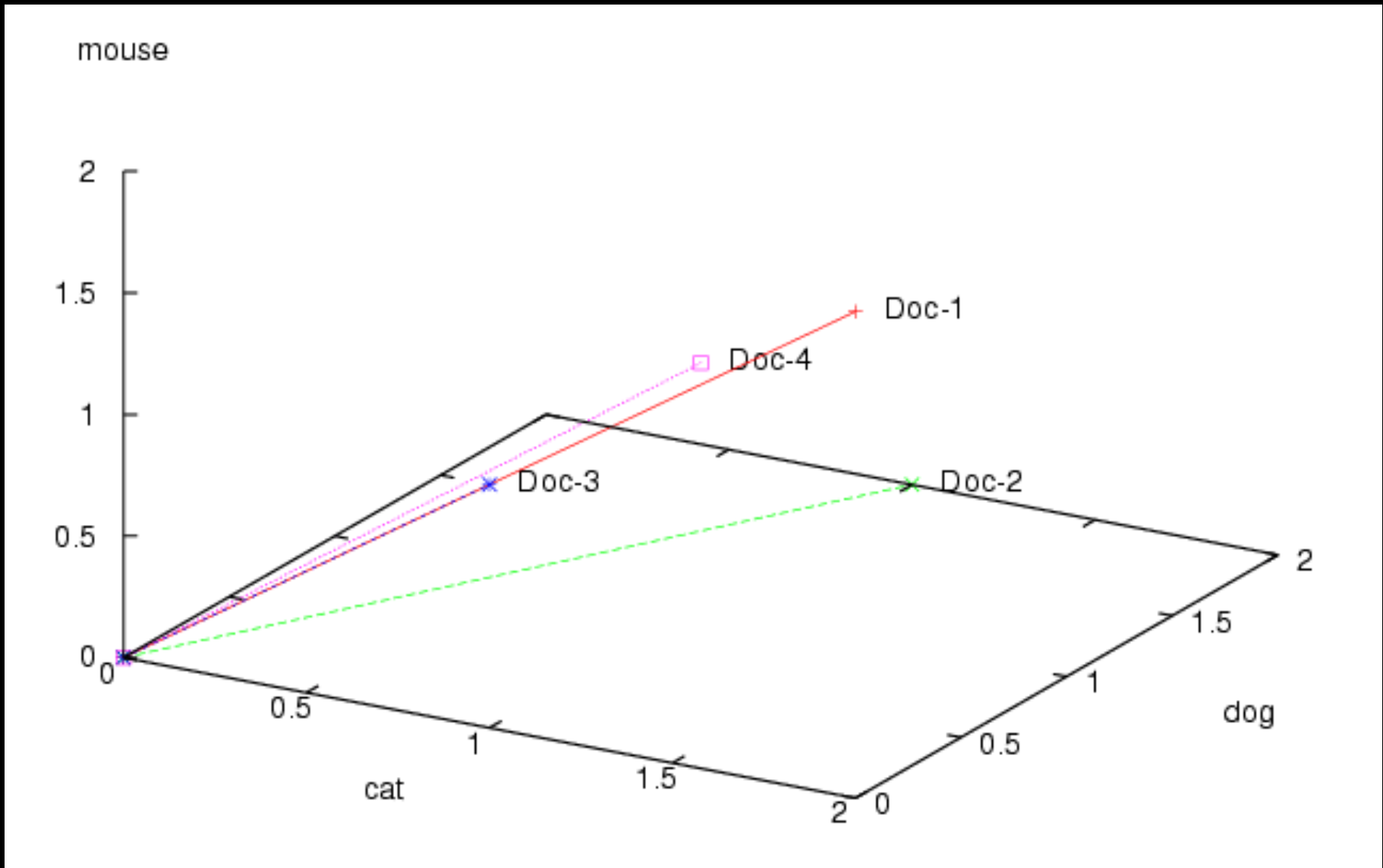
Matrix Example

Term	Doc 1	Doc 2	Doc 3	Doc 4
cat	2	1	1	1
dog	0	2	0	1
mouse	2	0	1	1

Vector Space Plot



Vector Space Plot



Vector Space Issues

- Many distance (similarity) measures possible.
- Various term weightings and normalizations.
- Term-by-document matrices are large.
- Dimension reduction techniques helpful.
- Matrices sparse with many zero elements (99%).
- Queries are also documents in this space.
- Calculating distances to every vector expensive.

Indexing

- Indexing is the process of assigning keywords to all documents of a corpus.
- It is a relation mapping each doc to set of keywords that it is about:

$$\text{Index} : \text{doc}_i \xrightarrow{\text{about}} \{\text{kw}_j\}$$

- The inverse mapping captures, for each keyword, the document it describes:

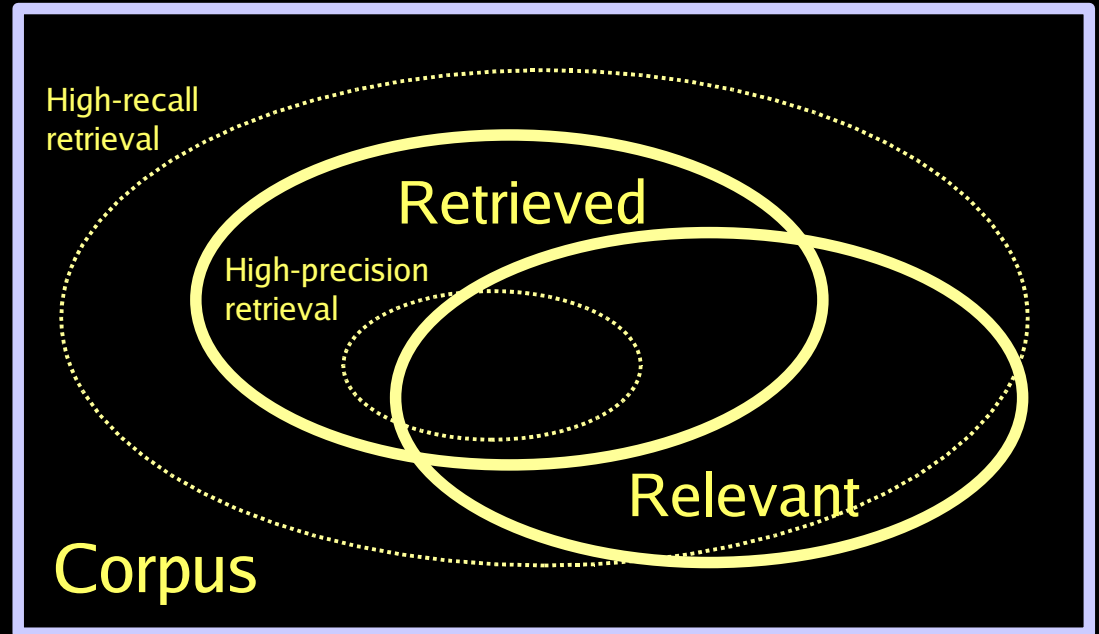
$$\text{Index}^{-1} : \{\text{kw}_j\} \xrightarrow{\text{describes}} \text{doc}_i$$

What to Index?

- Metatext
 - Categories
 - Keywords
 - Titles, authors, publication data
- All text?
 - Phrases?
 - Punctuation?
 - Hyphenation?
 - Word roots (stems)?

Evaluating Success

- Was the search engine helpful?
- How many relevant documents did the search engine find?



- What fraction of all relevant documents did it find?
 - $\text{Recall} = |\text{Retrieved AND Relevant}| / |\text{Relevant}|$
- What fraction of retrieved documents were relevant?
 - $\text{Precision} = |\text{Retrieved AND Relevant}| / |\text{Retrieved}|$

Homework

- Visit searchenginewatch.com/webmasters/features.html
 - learn which engines use
 - stop lists, metatags, alt text, etc.
 - and lots more
- Keep reading texts (see syllabus)