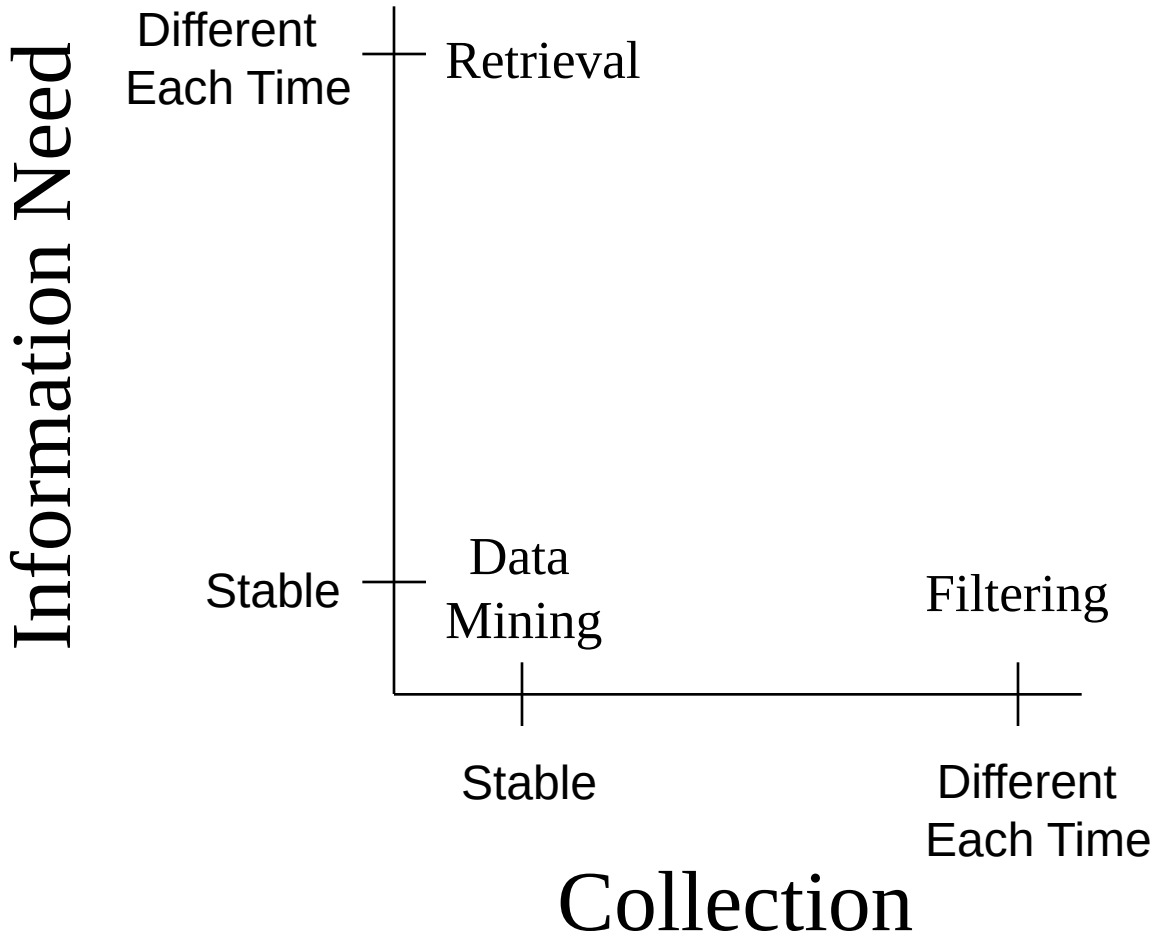
Information Filtering

LBSC 796/INFM 718R

Douglas W. Oard

Session 10, April 13, 2011

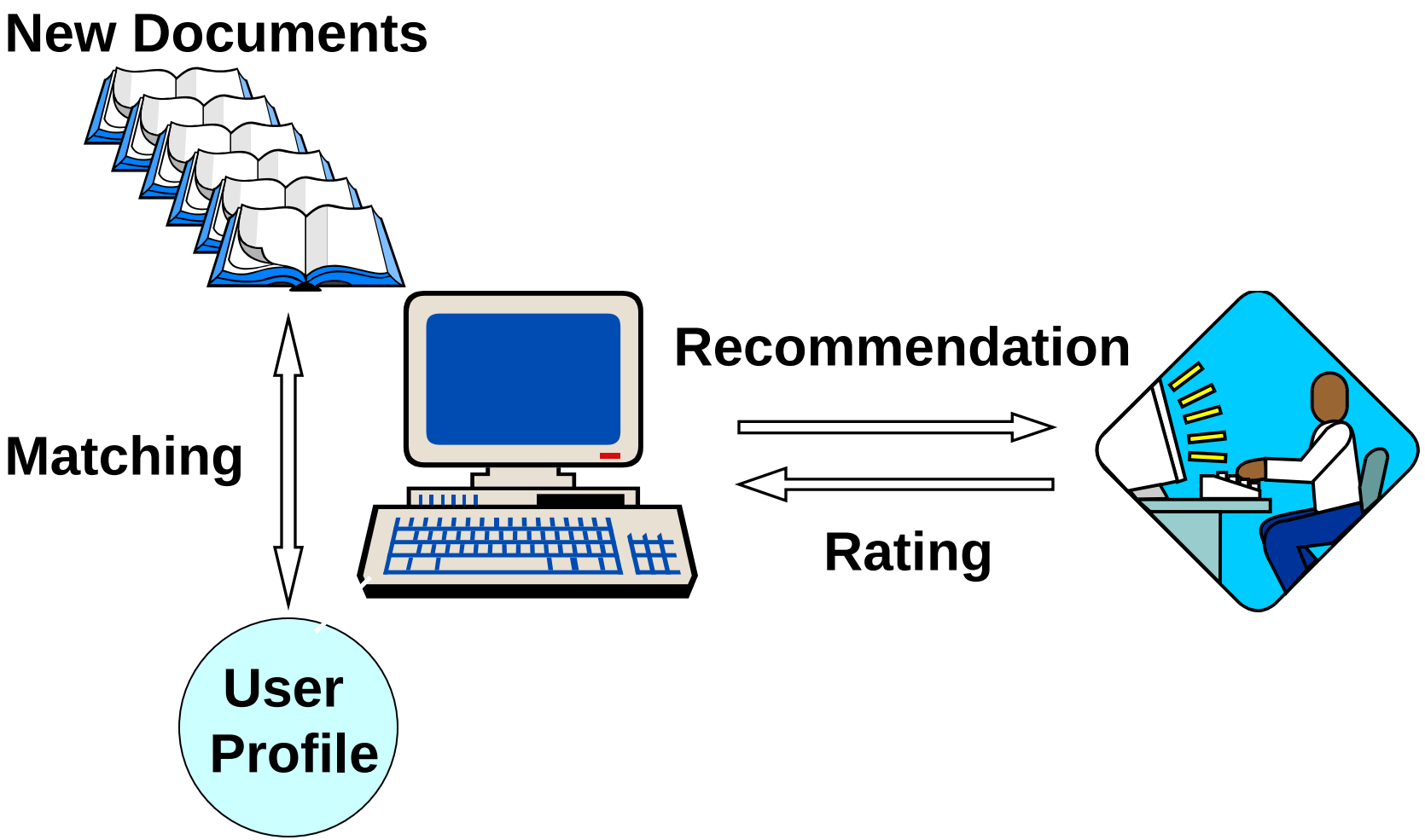
Information Access Problems



Information Filtering

- An abstract problem in which:
  - The information need is stable
    - Characterized by a “profile”
  - A stream of documents is arriving
    - Each must either be presented to the user or not
- Introduced by Luhn in 1958
  - As “Selective Dissemination of Information”
- Named “Filtering” by Denning in 1975

Information Filtering



Standing Queries

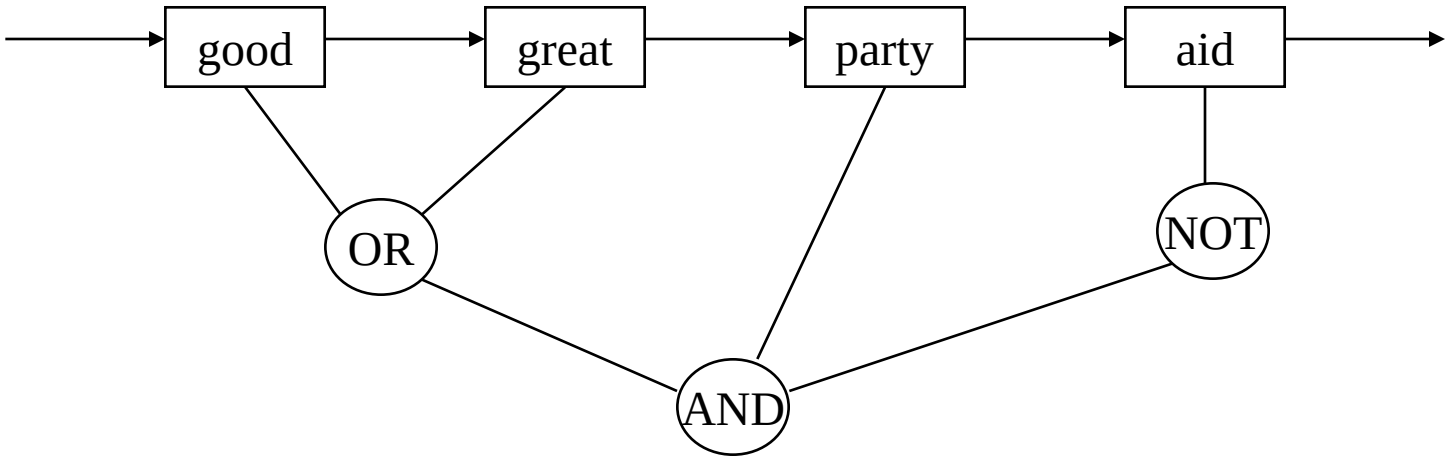
- Use any information retrieval system
  - Boolean, vector space, probabilistic, …
- Have the user specify a “standing query”
  - This will be the profile
- Limit the standing query by date
  - Each use, show what arrived since the last use

What's Wrong With That?

- Unnecessary indexing overhead
  - Indexing only speeds up retrospective searches
- Every profile is treated separately
  - The same work might be done repeatedly
- Forming effective queries by hand is hard
  - The computer might be able to help
- It is OK for text, but what about audio, video, …
  - Are words the only possible basis for filtering?

Stream Search: “Fast Data Finder”

- Boolean filtering using custom hardware
  - Up to 10,000 documents per second (in 1996!)
- Words pass through a pipeline architecture
  - Each element looks for one word



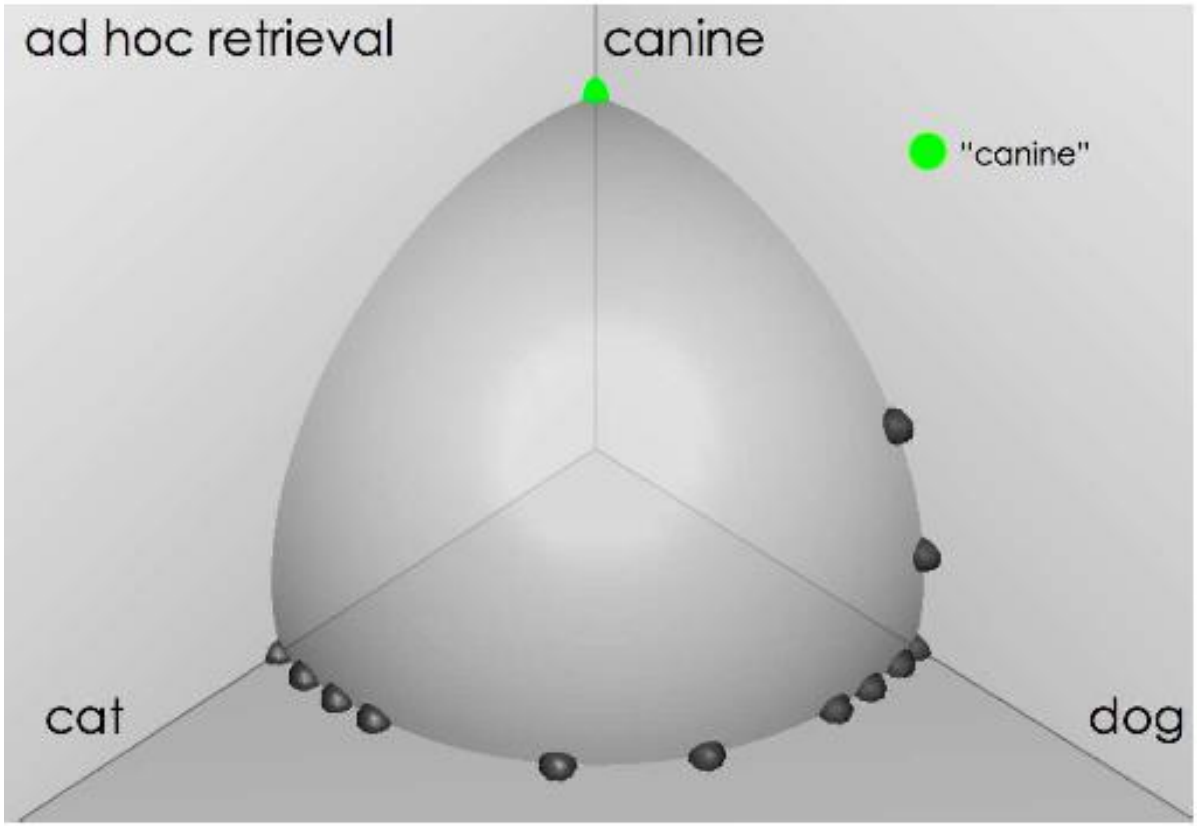
Profile Indexing (SIFT)

- Build an inverted file of profiles
  - Postings are profiles that contain each term
- RAM can hold 5 million profiles/GB
  - And several machines could run in parallel
- Both Boolean and vector space matching
  - User-selected threshold for each ranked profile
    - Hand-tuned on a web page using today's news

Profile Indexing Limitations

- Privacy
  - Central profile registry, associated with known users
- Usability
  - Manual profile creation is time consuming
    - May not be kept up to date
  - Threshold values vary by topic and lack “meaning”

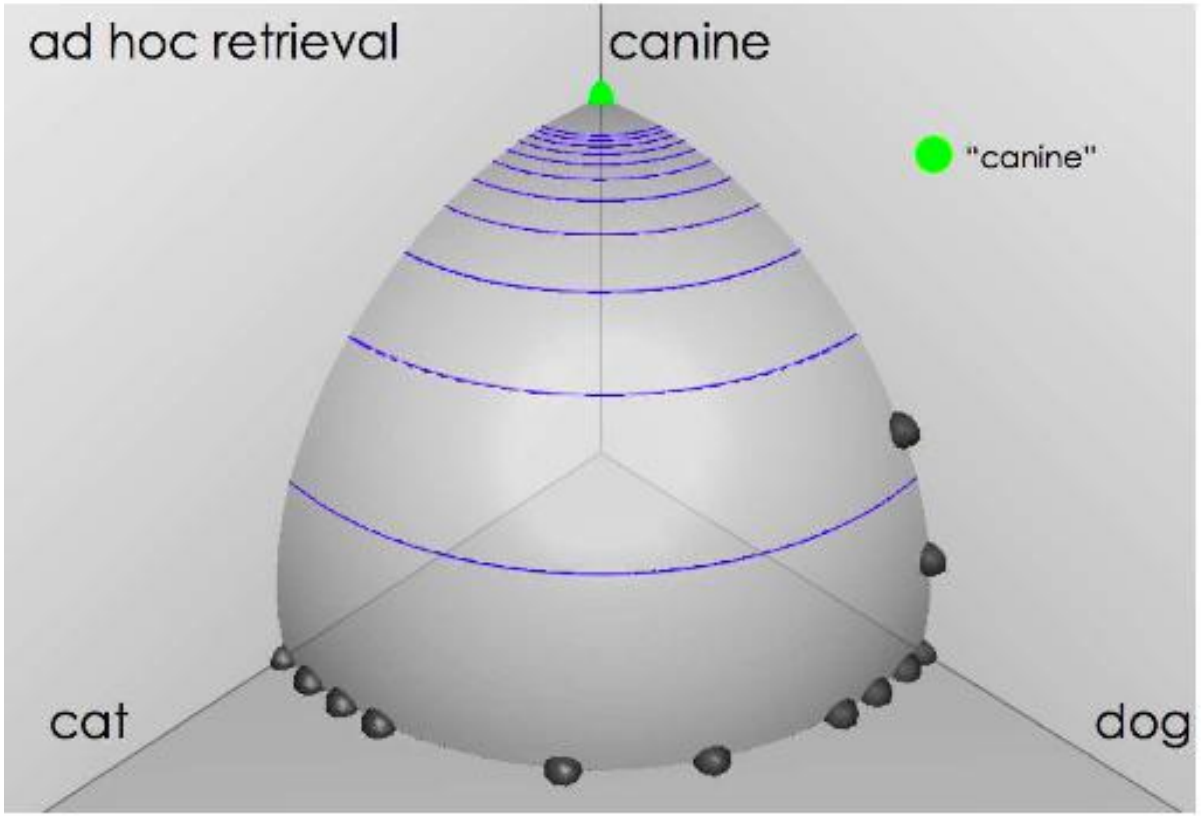
Vector space example: query “canine” (1)



Source:

Fernando Díaz

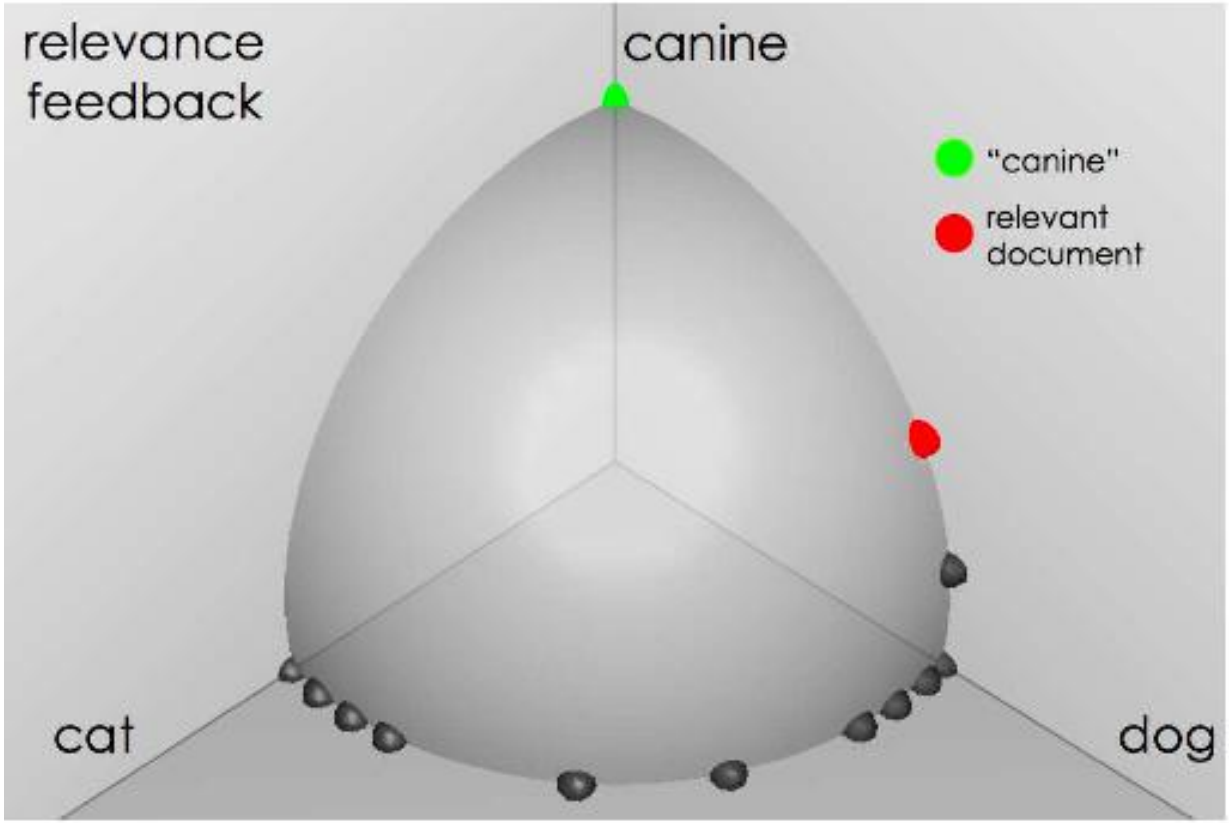
Similarity of docs to query “canine”



Source:

Fernando Díaz

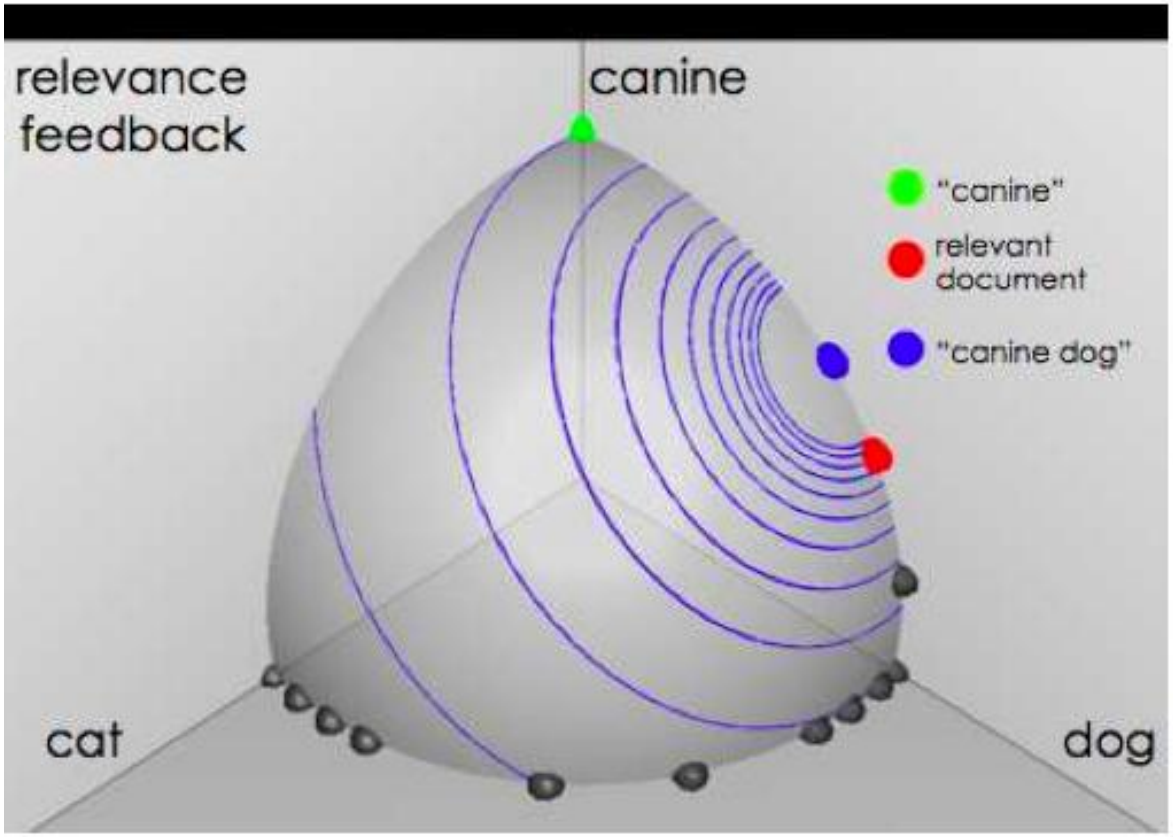
User feedback: Select relevant documents



Source:

Fernando Díaz

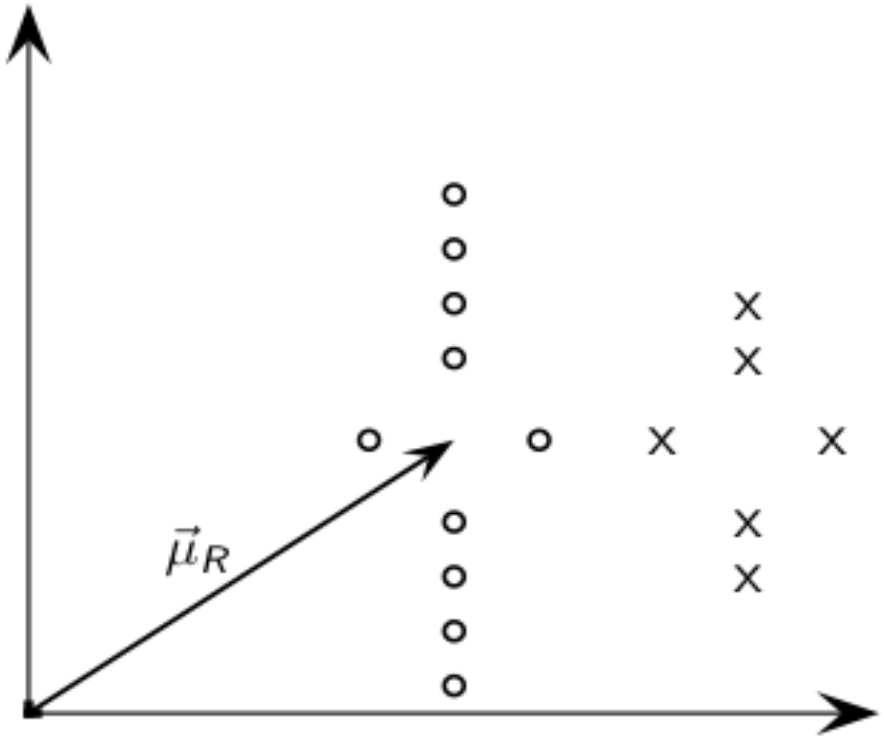
Results after relevance feedback



Source:

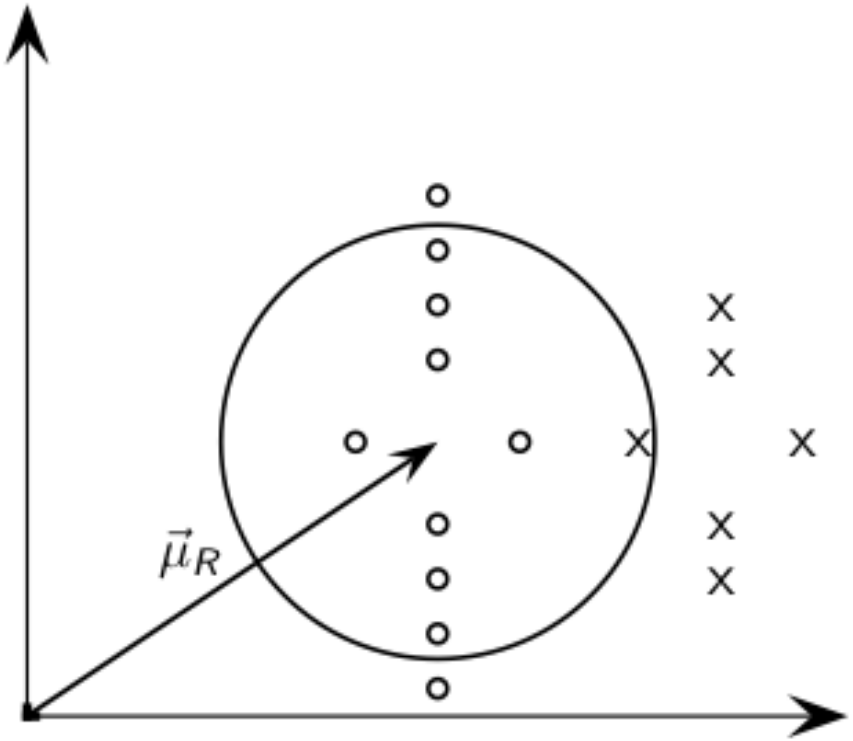
Fernando Díaz

Rocchio' illustrated



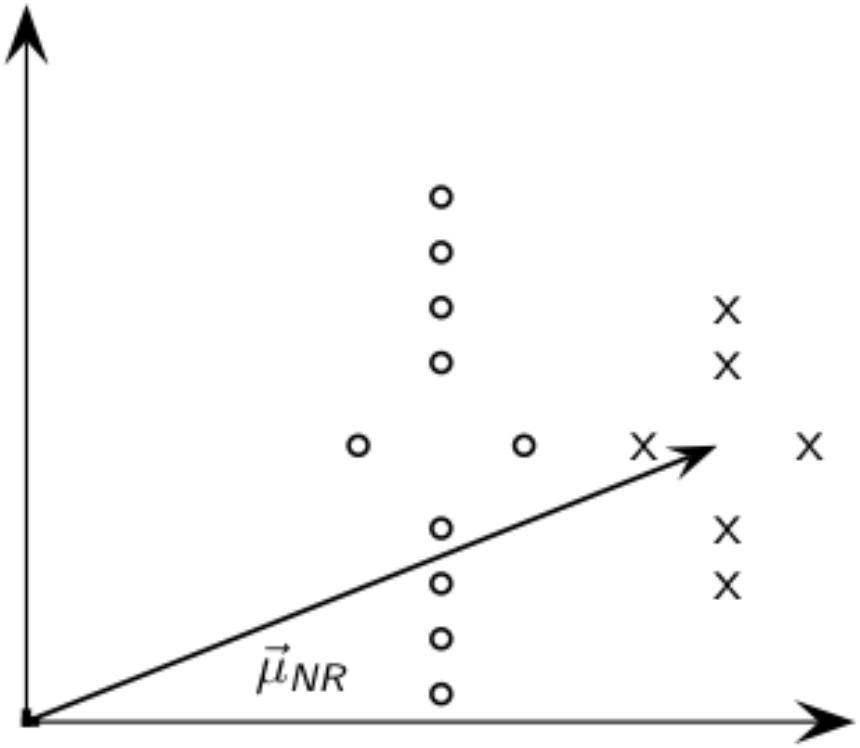
$\vec{\mu}_R$: centroid of relevant documents

Rocchio' illustrated



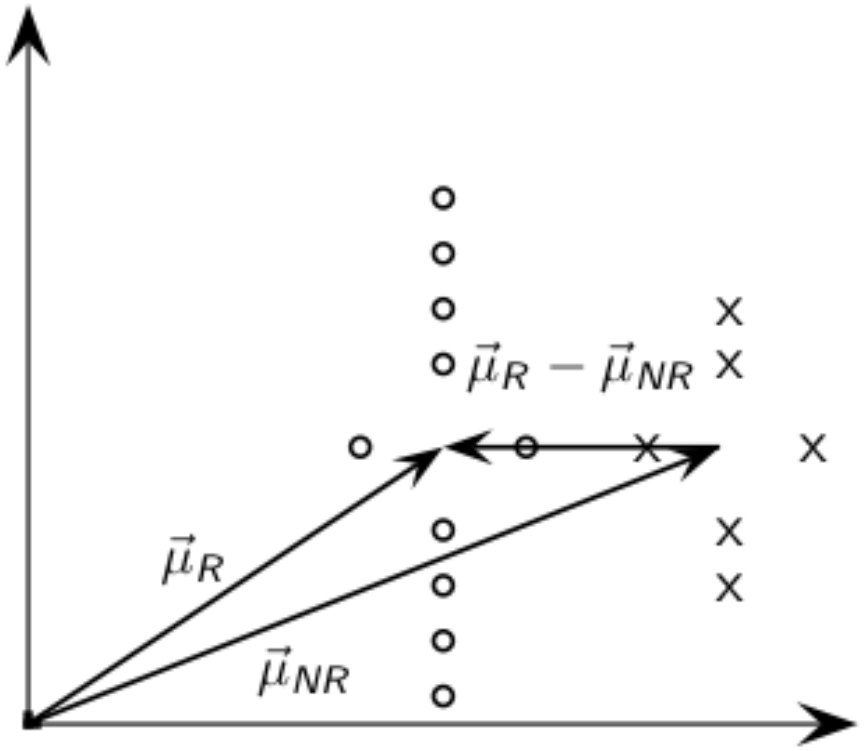
$\vec{\mu}_R$ does not separate relevant / nonrelevant.

Rocchio' illustrated



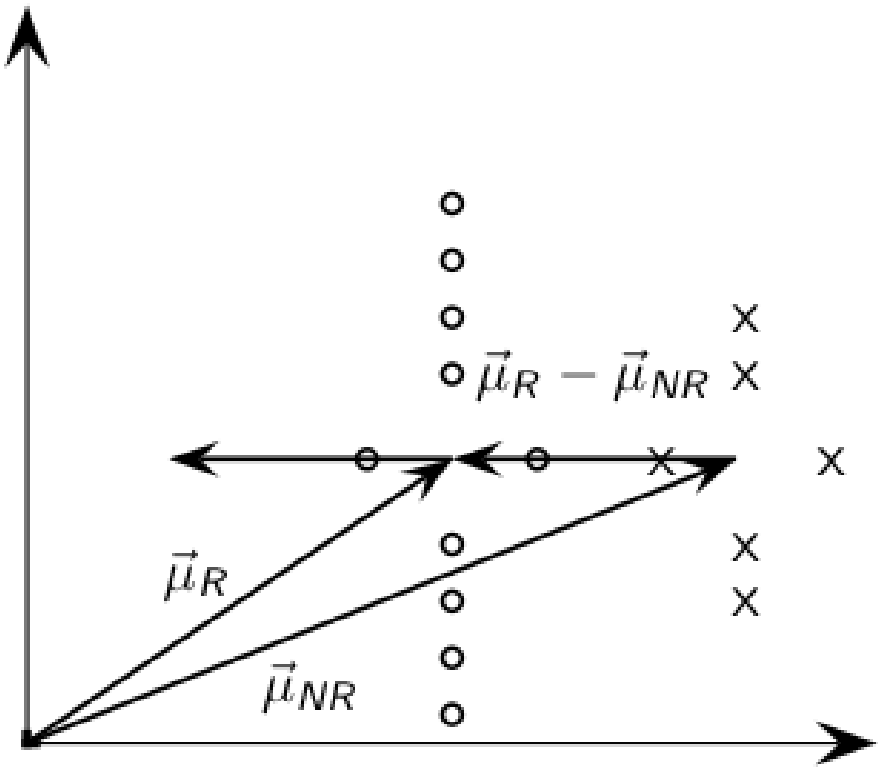
$\vec{\mu}_{NR}$: centroid of nonrelevant documents.

Rocchio' illustrated



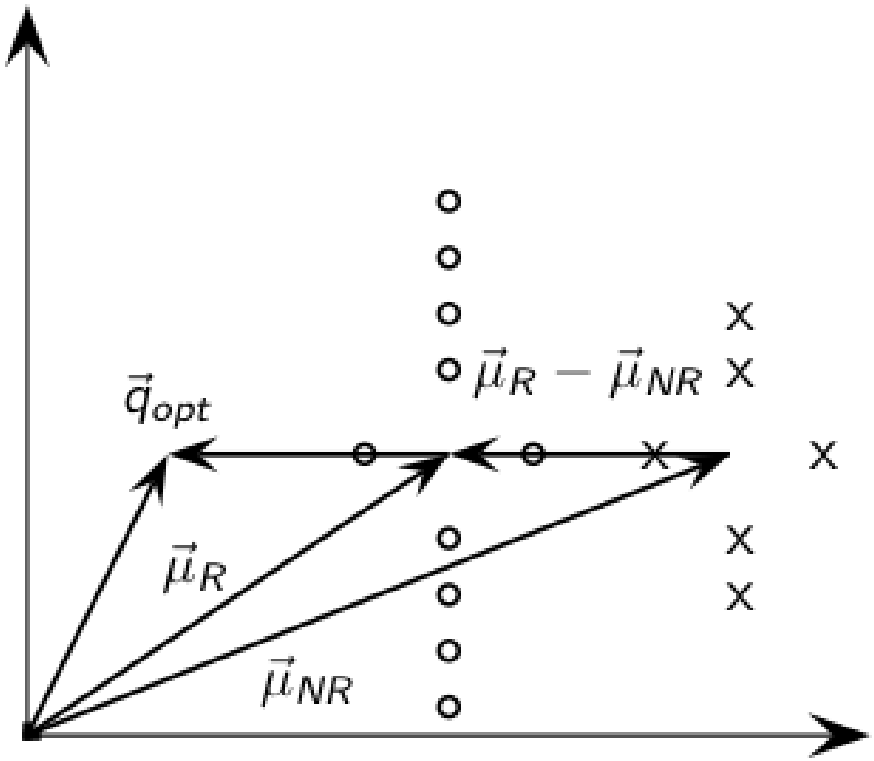
$\vec{\mu}_R$ $\vec{\mu}_{NR}$: difference vector

Rocchio' illustrated



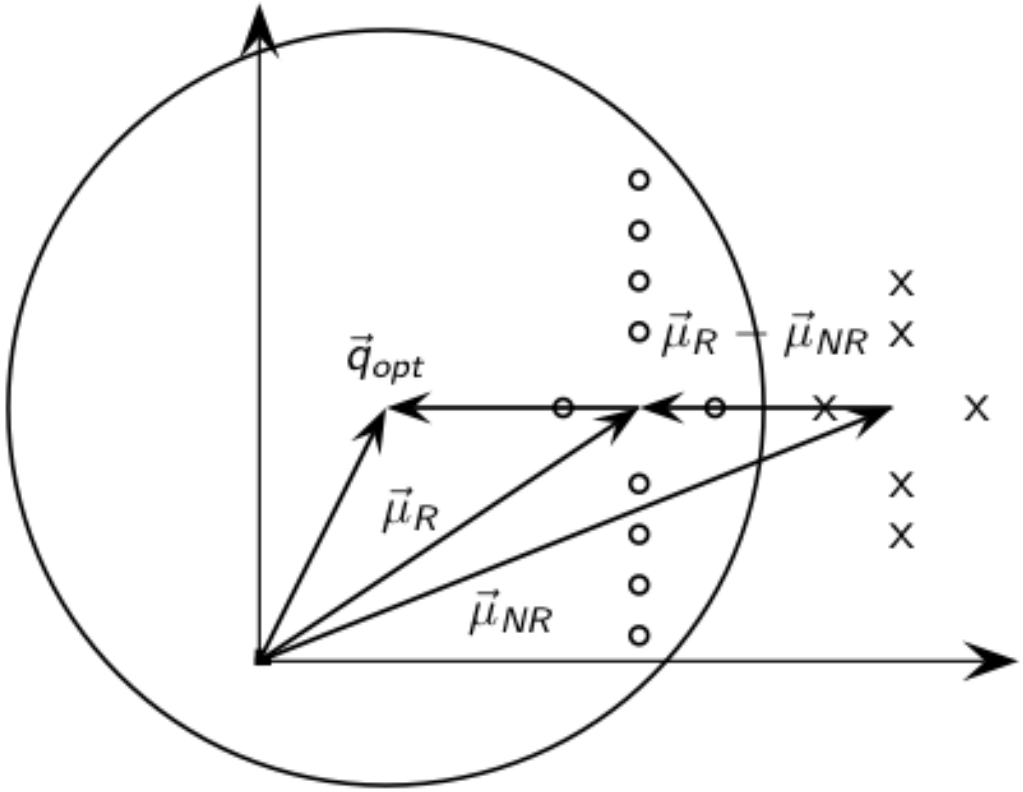
Add difference vector to $\vec{\mu}_R$ …

Rocchio' illustrated



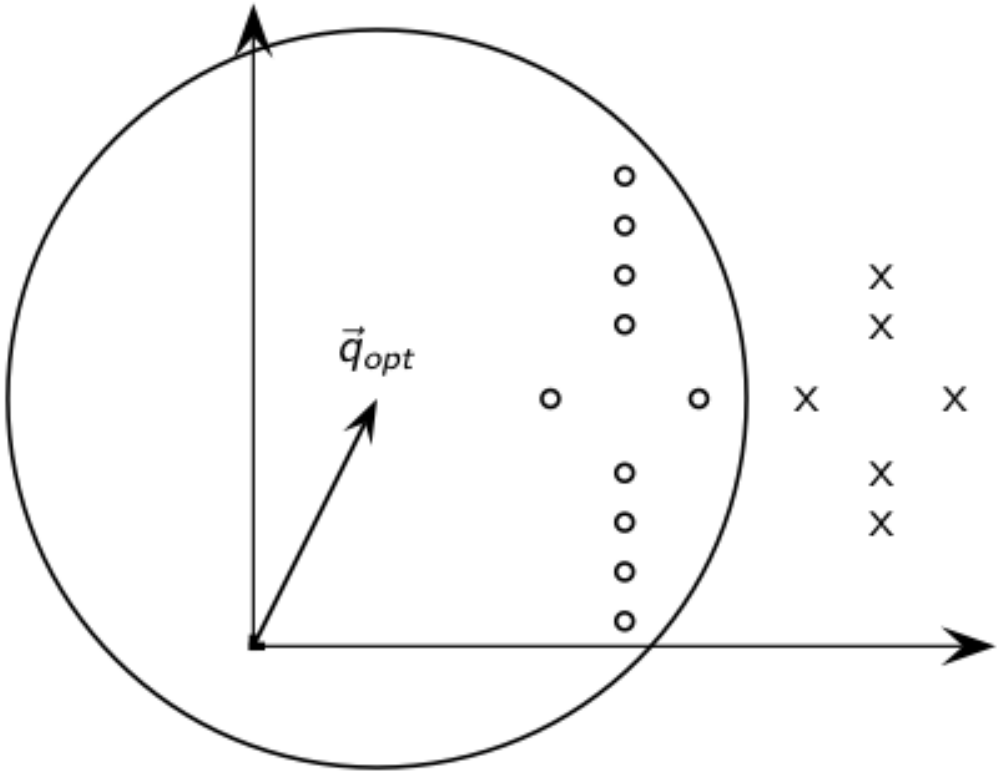
… to ge $\vec{q}_{opt}$

Rocchio' illustrated



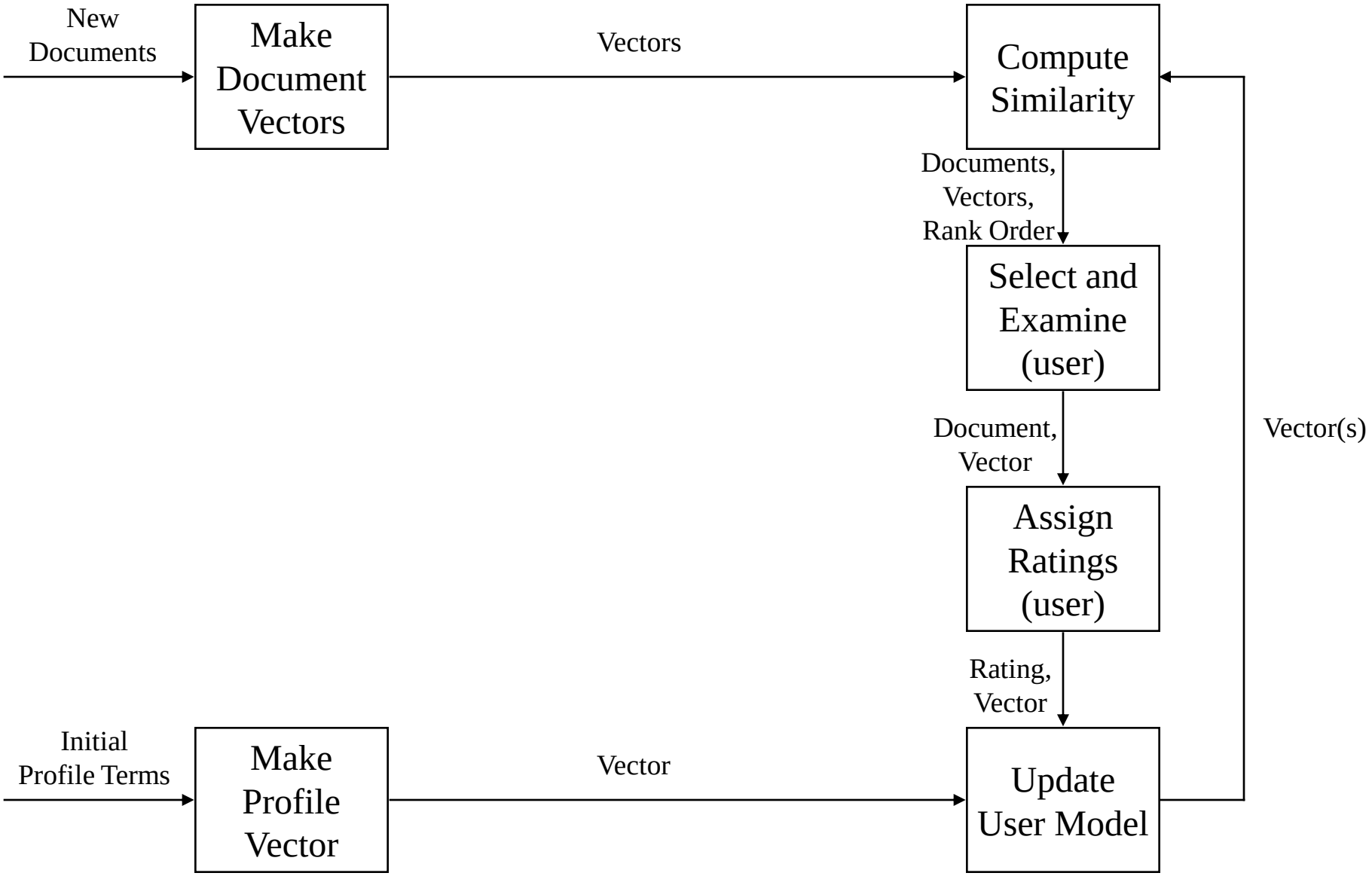
$\vec{q}_{opt}$ separates relevant / nonrelevant perfectly.

Rocchio' illustrated

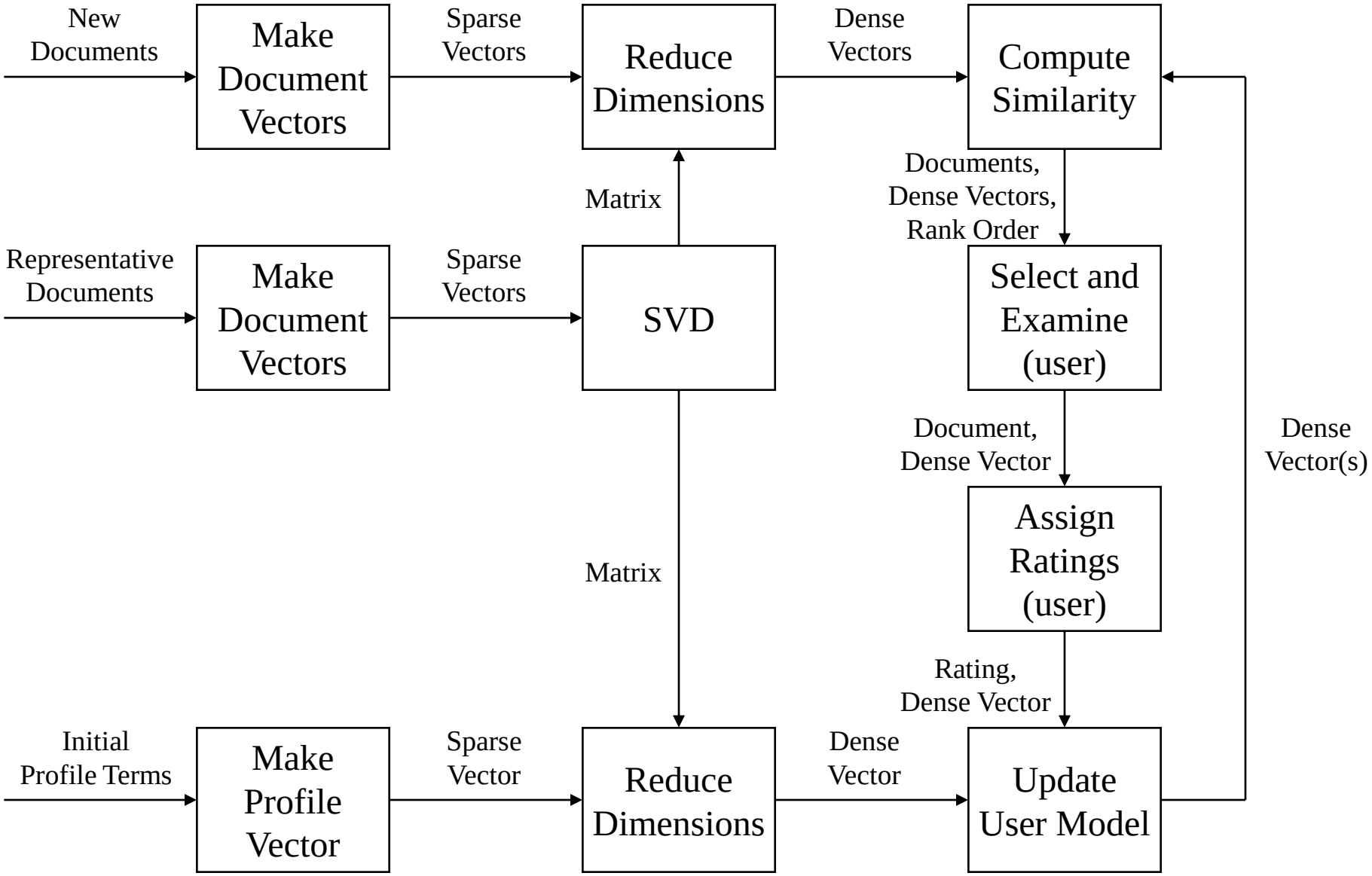


$\vec{q}_{opt}$ separates relevant / nonrelevant perfectly.

Adaptive Content-Based Filtering



Latent Semantic Indexing



Content-Based Filtering Challenges

- IDF estimation
  - Unseen profile terms would have infinite IDF!
  - Incremental updates, side collection
- Interaction design
  - Score threshold, batch updates
- Evaluation
  - Residual measures

Machine Learning for User Modeling

- All learning systems share two problems
  - They need some basis for making predictions
    - This is called an “inductive bias”
  - They must balance adaptation with generalization

Machine Learning Techniques

- Hill climbing (Rocchio)
- Instance-based learning (kNN)
- Rule induction
- Statistical classification
- Regression
- Neural networks
- Genetic algorithms

Rule Induction

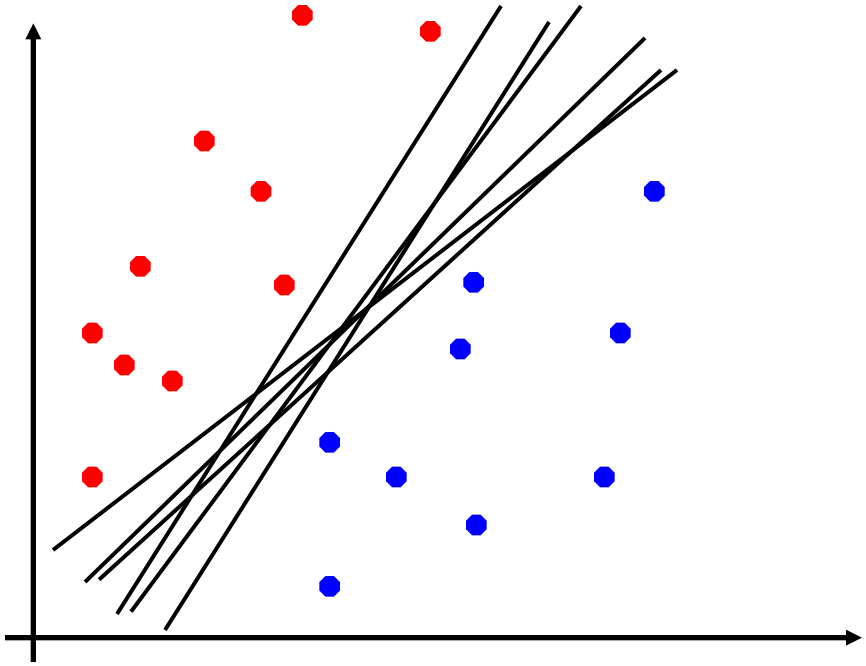
- Automatically derived Boolean profiles
  - (Hopefully) effective and easily explained
- Specificity from the "perfect query"
  - AND terms in a document, OR the documents
- Generality from a bias favoring short profiles
  - e.g., penalize rules with more Boolean operators
  - Balanced by rewards for precision, recall, …

Statistical Classification

- Represent documents as vectors
  - Usual approach based on TF, IDF, Length
- Build a statistical models of rel and non-rel
  - e.g., (mixture of) Gaussian distributions
- Find a surface separating the distributions
  - e.g., a hyperplane
- Rank documents by distance from that surface

Linear Separators

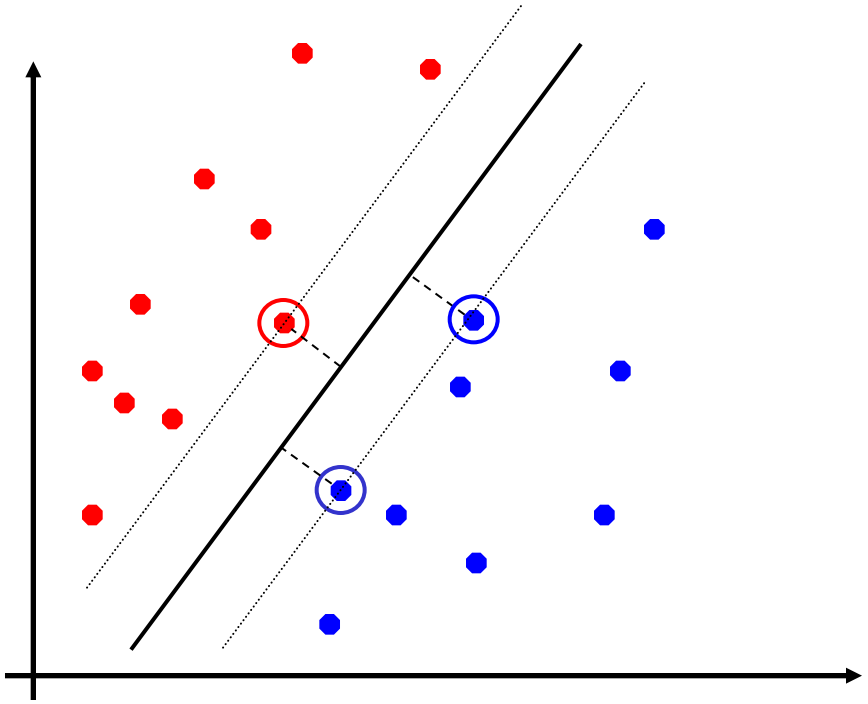
- Which of the linear separators is optimal?



Original from Ray Mooney

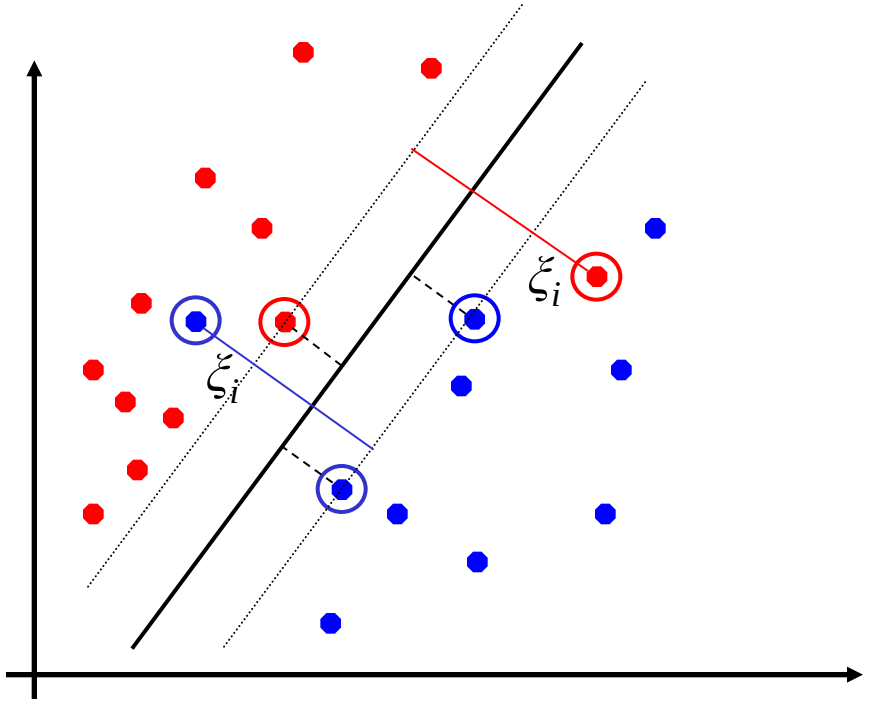
Maximum Margin Classification

- Implies that only support vectors matter; other training examples are ignorable.



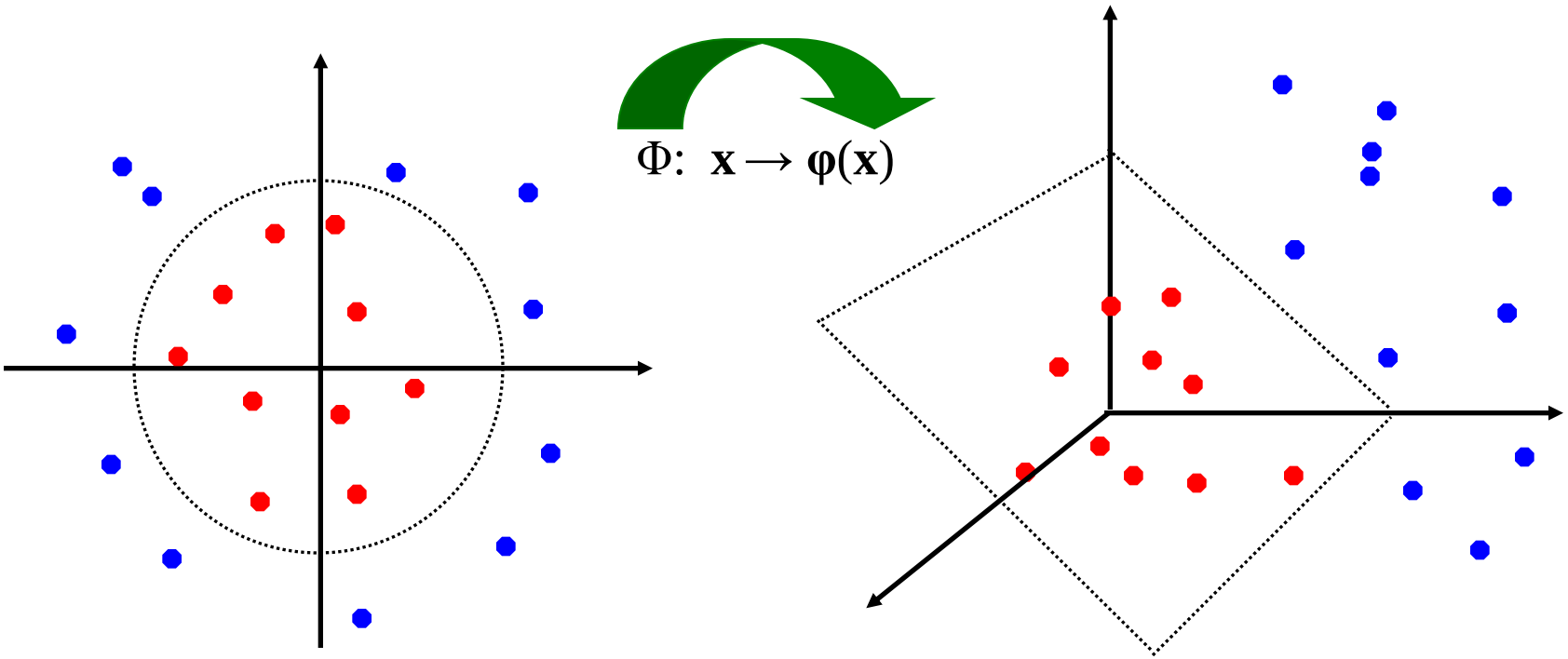
Original from Ray Mooney

Soft Margin SVM



Original from Ray Mooney

Non-linear SVMs

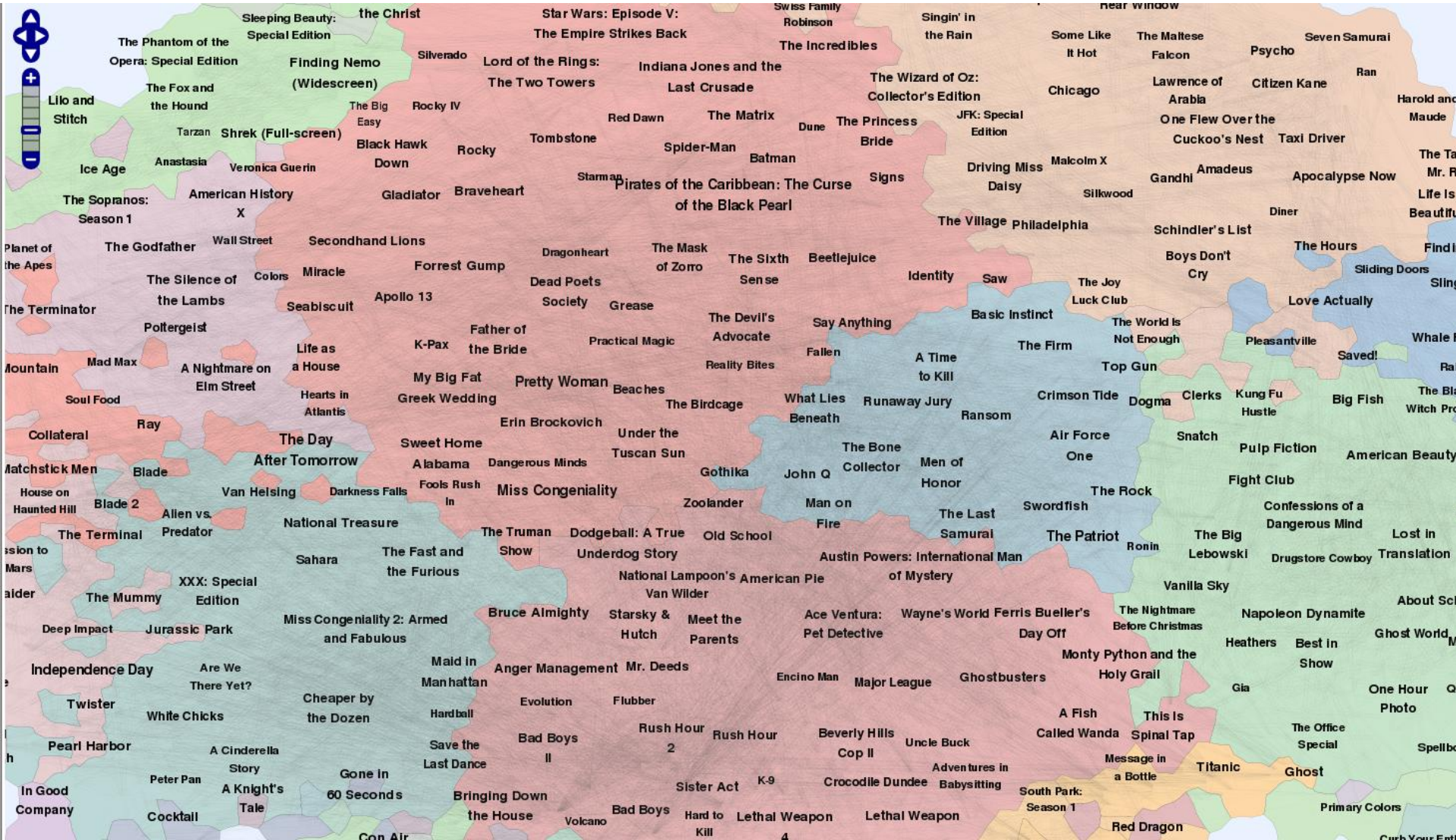


Original from Ray Mooney

Training Strategies

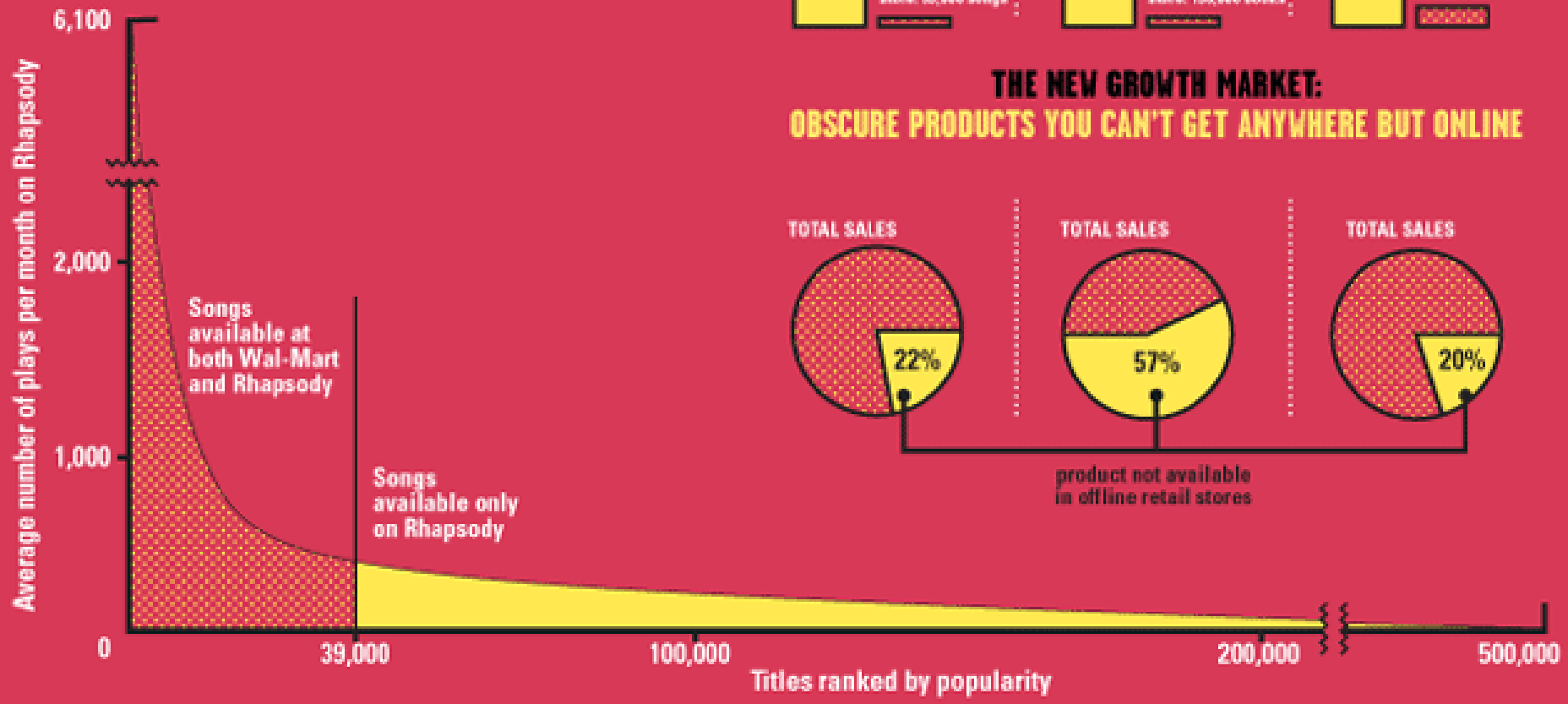
- Overtraining can hurt performance
  - Performance on training data rises and plateaus
  - Performance on new data rises, then falls
- One strategy is to learn less each time
  - But it is hard to guess the right learning rate
- Usual approach: Split the training set
  - Training, DevTest for finding “new data” peak

NetFlix Challenge



ANATOMY OF THE LONG TAIL

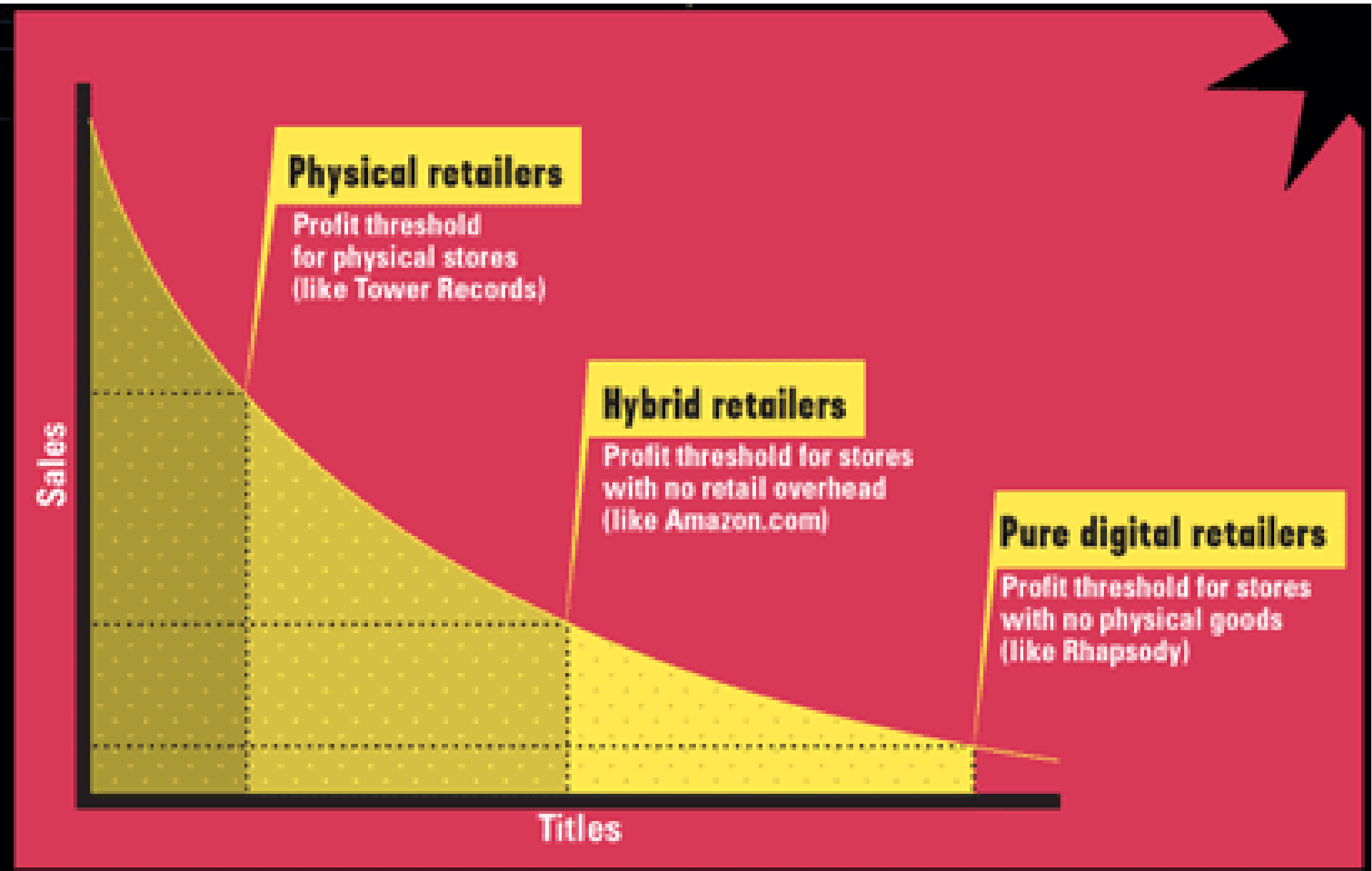
Online services carry far more inventory than traditional retailers. Rhapsody, for example, offers 19 times as many songs as Wal-Mart's stock of 39,000 tunes. The appetite for Rhapsody's more obscure tunes (charted below in yellow) makes up the so-called Long Tail. Meanwhile, even as consumers flock to mainstream books, music, and films (right), there is real demand for niche fare found only online.



THE NEW GROWTH MARKET:
OBSCURE PRODUCTS YOU CAN'T GET ANYWHERE BUT ONLINE

Sources: Erik Brynjolfsson and Jeffrey Hu, MIT, and Michael Smith, Carnegie Mellon; Barnes & Noble; Netflix; RealNetworks

Effect of Inventory Costs



Spam Filtering

- Adversarial IR
  - Targeting, probing, spam traps, adaptation cycle
- Compression-based techniques
- Blacklists and whitelists
  - Members-only mailing lists, zombies
- Identity authentication
  - Sender ID, DKIM, key management