The Magic of Models
A Case of Vector Space Model in Web Search

Venkatesh Vinayakarao
venkateshv@cmi.ac.in
http://vvtesh.co.in

Research Science Initiative, Chennai (RSIC)
Chennai Mathematical Institute

Simplicity boils down to two steps. Identify the essential. Eliminate the rest. **–Leo Babauta.**

How to tame complexity?

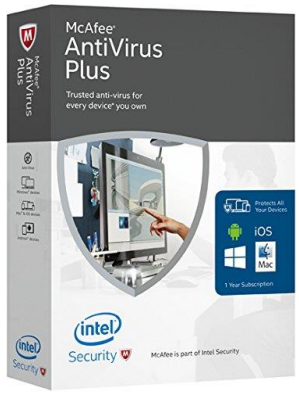
Can we learn from the world around us?

How to get All India Rank 1 in JEE?

*Am not AIR 1 in JEE. Just for fun.

How did we humans build this?

We could build these too!



How to tame complexity?

Can we learn from the world around us?

Taming Complexity

Contents

Taming Complexity

Key Principles

1. Hierarchy
2. Abstraction
3. Keeping Related Things Together
4. …

Models

- A model is a representation of an *idea*, an *object*, a *process* or even a *system*
- Used as tools to understand (define, quantify, visualize, …) the real world.

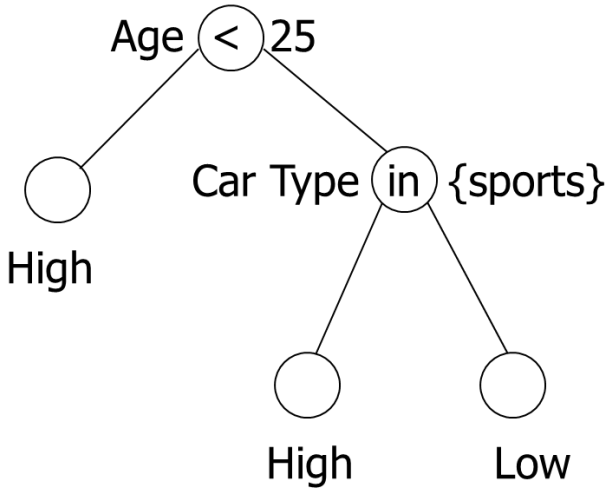
Decision Tree Model

Data Set

Age	Car Type	Risk
23	Family	High
17	Sports	High
43	Sports	High
68	Family	Low
32	Truck	Low
20	Family	High

Question: **What is the risk (high or low) if age is below 25?**

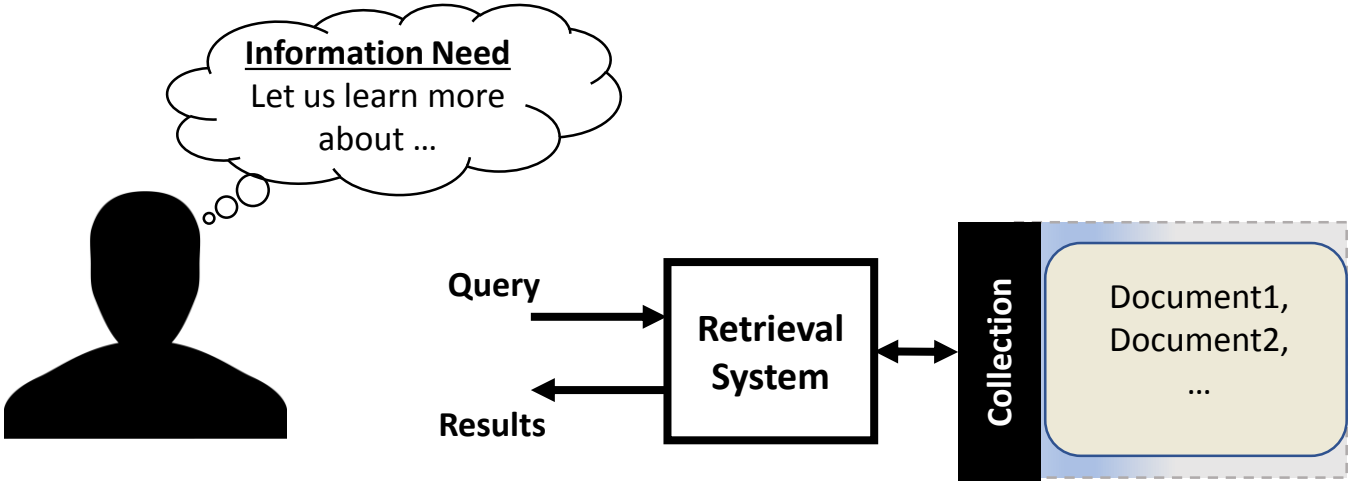
Decision Tree



Magic of Models

A **Bag of Words** Model for Search Engines

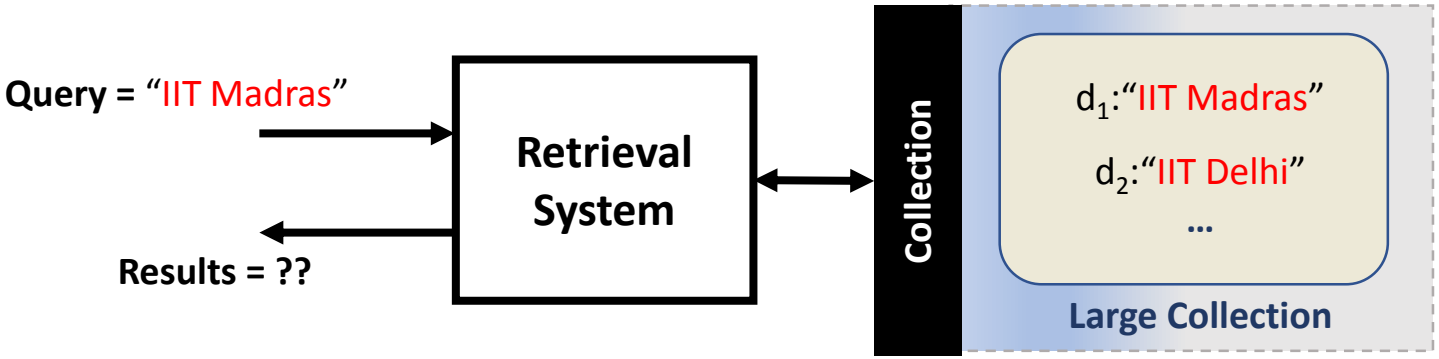
Search Engines



Simple Retrieval Problem

- A **collection** with 5 **documents** having the following contents
  - d1: IIT Madras
  - d2: IIT Delhi
  - d3: IIT Kanpur
  - d4: IIT Goa
  - d5: IIT Bombay
- **Query** is
  - IIT Madras
- Which **document** will you match and why?

The Problem: How to Build a Retrieval System?



One (bad) Approach

- First match the **term** IIT.
  - Filter out documents that contain this term.
- Next match the **term** Madras.
  - Filter out documents that contain this term.

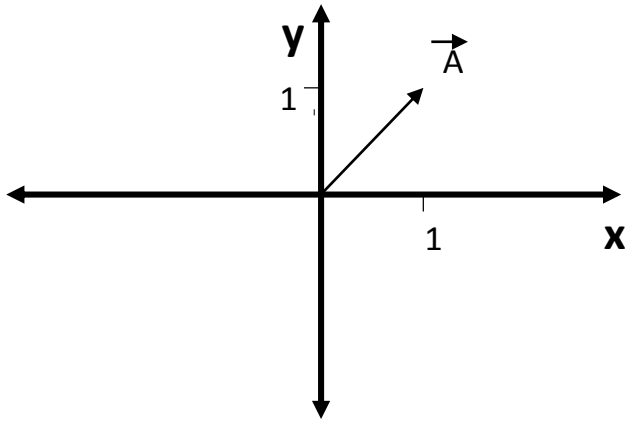
Multiple iterations!
Quiz: Can we do better?

A Better Approach

Revisiting Vectors

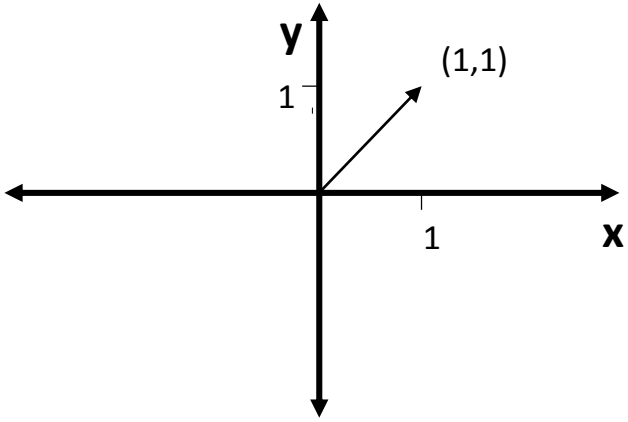
Vectors

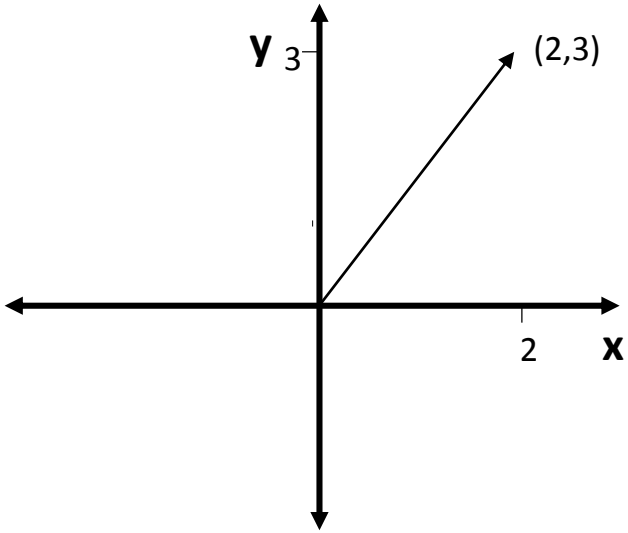
- Geometric entity which has magnitude and direction



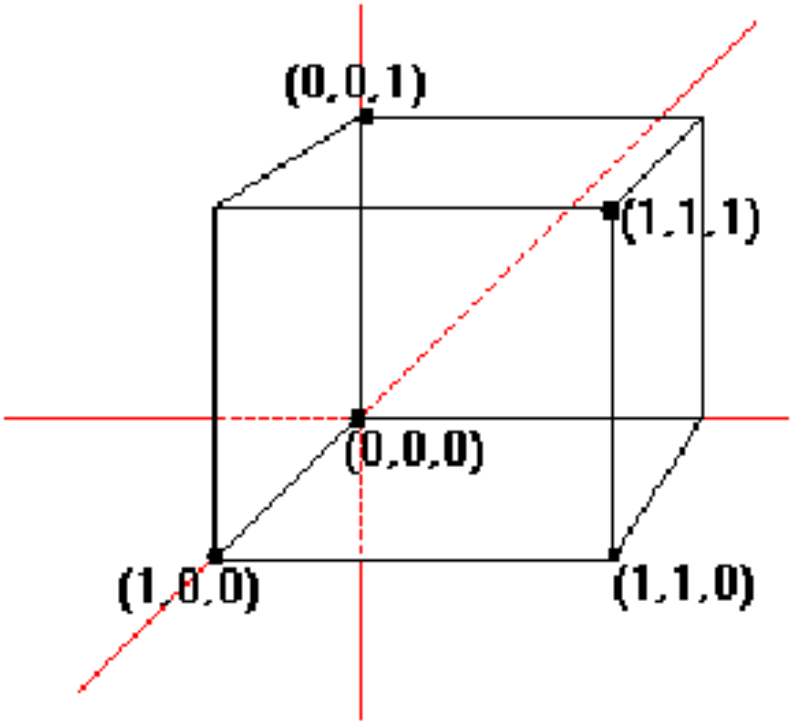
- If (x,y) is our vector of interest, this figure shows $\vec{A}$ vector = (1,1).

How is (2,3) Different?





What is (1,1,1) ?

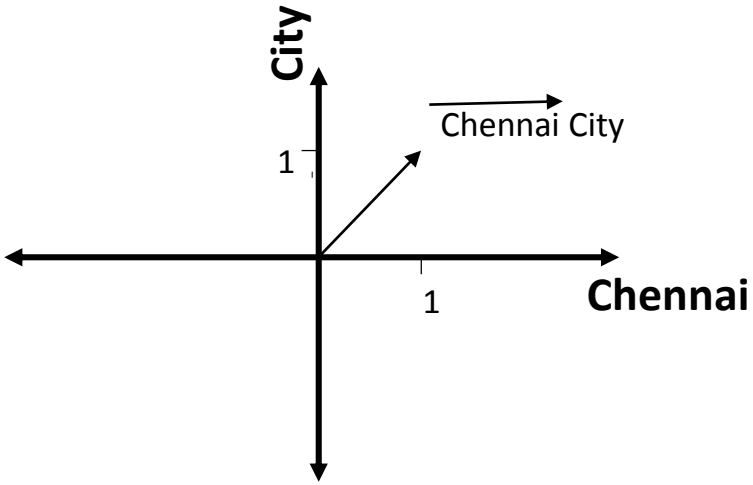


Remember!

A number is just a mathematical object. We give meaning to it!

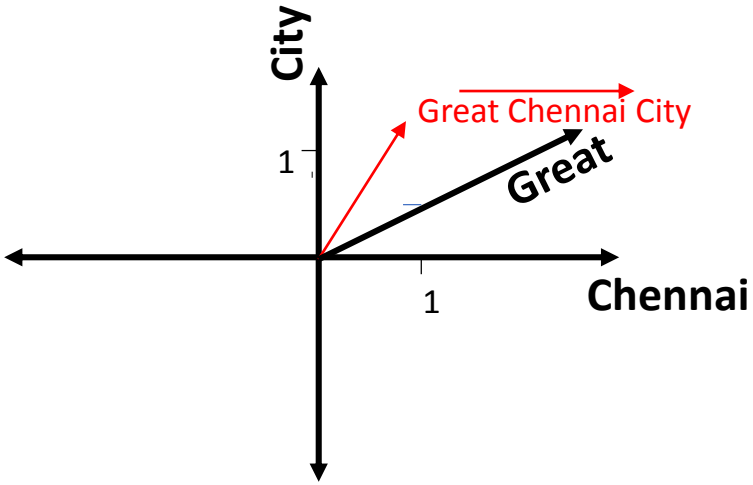
Sentences are Vectors

- "Chennai City" as a vector



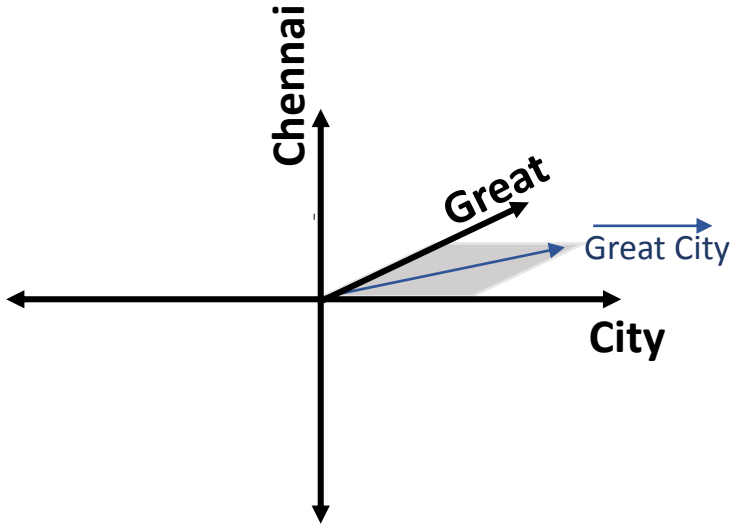
Sentences are Vectors

- "Great Chennai City" is a 3-dimensional vector



Sentences are Vectors

- On this 3D space, "Great City" vector will lie on the x (City) and z (Great) plane. "Great City" is (1,0,1).



Natural Language Phrases as Vectors

Let query q = "IIT Delhi".

Let document, d_1 = "IIT Madras" and d_2 = "IIT Delhi".

	IIT	Delhi	Madras
q	1	1	0
d_1	1	0	1
d_2	1	1	0

q = (1,1,0), d_1= (1,0,1) and d_2 = (1,1,0)

Quiz

- Considering the following vectors:

	IIIT	Sri	City	Delhi
q	1	1	1	0
d_1	1	1	1	0
d_2	1	0	0	1

- What is the Natural Language (NL) equivalent of (0,1,1,0) ?
- What is the NL equivalent of (1,0,0,1) ?
- What is the vector for Delhi?

Similarity Score

- D1 = “Chennai”
- D2 = “Delhi”

- Quiz
  - What is the angle between D1 and D2 vectors?
  - On a scale of 0 – 1, how similar are D1 and D2?

0 – 90 to 1 – 0: How?

	0°	30°	45°	60°	90°
sin θ	0	$\frac{1}{2}$	$\frac{1}{\sqrt{2}}$	$\frac{\sqrt{3}}{2}$	1
cos θ	1	$\frac{\sqrt{3}}{2}$	$\frac{1}{\sqrt{2}}$	$\frac{1}{2}$	0
tan θ	0	$\frac{1}{\sqrt{3}}$	1	$\sqrt{3}$	Not defined

Back to Trigonometry: Dot Product

- If x and y are non-unit vectors, what is the cosine of angle between them (cos Θ)?

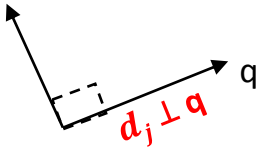
$$\mathbf{a} \cdot \mathbf{b} = \|\mathbf{a}\| \, \|\mathbf{b}\| \cos(\theta)$$

$$Cosine\ Similarity = \cos(\theta) = \frac{a.b}{||a||\,||b||}$$

Matching Documents to Queries

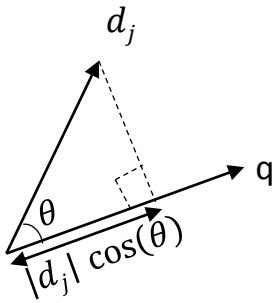
- Document as a vector of term-occurrence

$$d_j = (w_{1j}, w_{2j}, \ldots w_{nj})$$

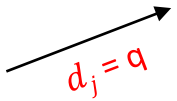


- Query as a vector of term-occurrence

$$q = (w_{1q}, w_{2q}, \ldots wmq)$$



- Similarity between these vectors can be represented as



$$Cosine\ Similarity = \cos(\theta) = \frac{d_j . q}{||d_j||\ ||q||}$$

Example

Let query q = "BITS Pilani".

Let document, d_1 = "BITS Pilani Goa Campus" and d_2 = "IIIT Delhi".

	BITS	Pilani	Goa	Campus	IIIT	Delhi
q	1	1	0	0	0	0
d_1	1	1	1	1	0	0
d_2	0	0	0	0	1	1

In our VSM, q = (1,1,0,0,0,0), d_1= (1,1,1,1,0,0) and d_2 = (0,0,0,0,1,1)

similarity(d_1, q) = $$\frac{d_1.q}{||d_1||\,||q||} = \frac{1.1 + 1.1}{\sqrt{1^2+1^2+1^2+1^2}\,\sqrt{1^2+1^2}} = 0.71.$$

similarity(d_2, q) = $$\frac{d_2.q}{||d_2||\,||q||} = 0.$$

Which of the Following are Sets?

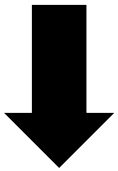
- {1, 2, 3, 4, 5, 6, 5, 7, 8, 9, 10, 11, 12, 13}
- {A, B, C, D, E, F, G, H, I, I, J, K, L, M, N, O}
- {apple, banana, orange, apple, banana, orange}

Bag

- {1, 2, 3, 4, 5, 6, 5, 7, 8, 9, 10, 11, 12, 13}
- {A, B, C, D, E, F, G, H, I, I, J, K, L, M, N, O}
- {apple, banana, orange, apple, banana, orange}

Set of Words Representation

- "IIIT Sri City" → {IIIT, Sri, City}
- "IIIT Sri City, Sri City" → {IIIT, Sri, City}

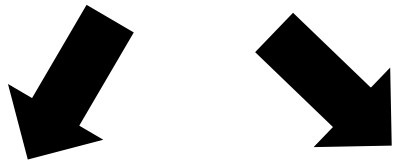


	IIIT	Sri	City
q	1	1	1

Leads to same term-document matrix

Bag of Words Representation

- "IIIT Sri City" → {IIIT, Sri, City}
- "IIIT Sri City, Sri City" → [IIIT, Sri, Sri, City, City]



IIIT Sri City

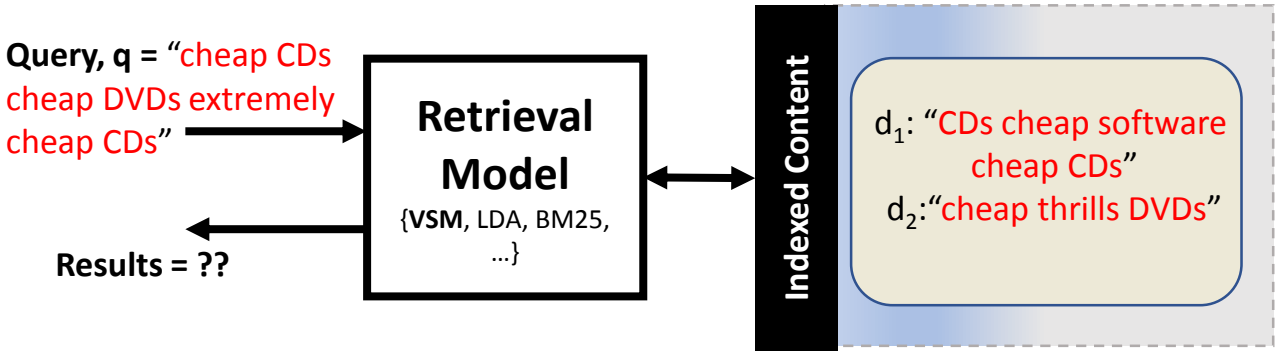
	IIIT	Sri	City
q	1	1	1

IIIT Sri City, Sri City

	IIIT	Sri	City
q	1	2	2

Leads to different term-document matrix

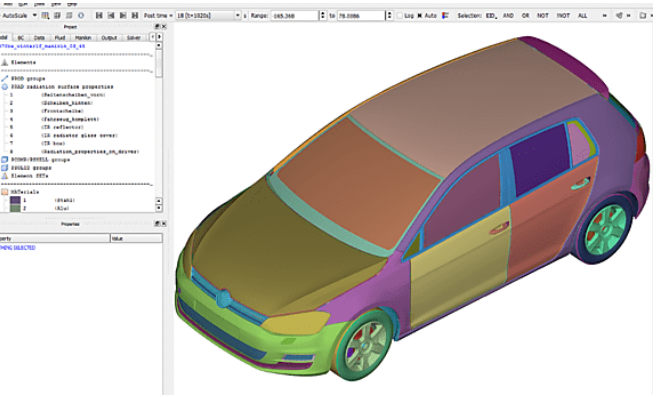
Which Document to Retrieve?

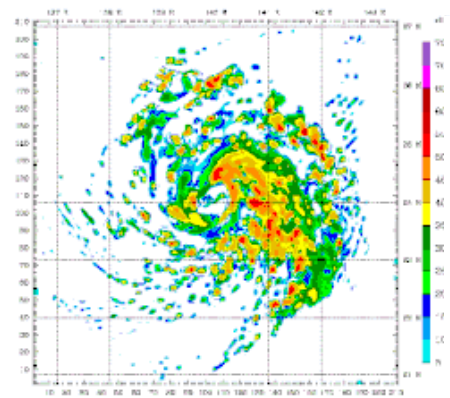


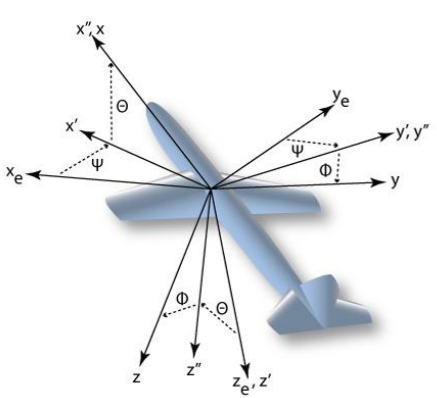
	cheap	CDs	DVDs	extremely	software	thrills
q	3	2	1	1	0	0
d_1	2	2	0	0	1	0
d_2	1	0	1	0	0	1

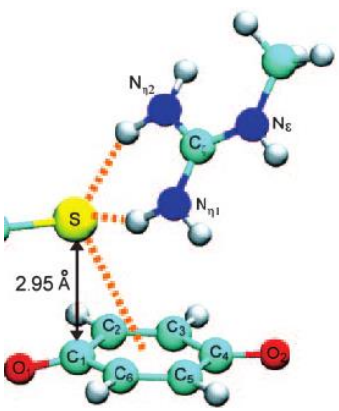
$sim(q,d_1) = 0.86$

$sim(q,d_2) = 0.59$









"Abstraction is one of the greatest visionary tools ever invented by human beings to imagine, decipher, and depict the world."

Jerry Saltz