

# Researchers Beware! The Devil is in the (Mathematical) Models

**Venkatesh Vinayakarao**

[venkateshv@cmi.ac.in](mailto:venkateshv@cmi.ac.in)

<http://vvtesh.co.in>

---

Two Day National Workshop on

**An Insight into Mathematical Modelling in Information and Communication Engineering**

Jointly Organized by Department of CSE and Department of Mathematics

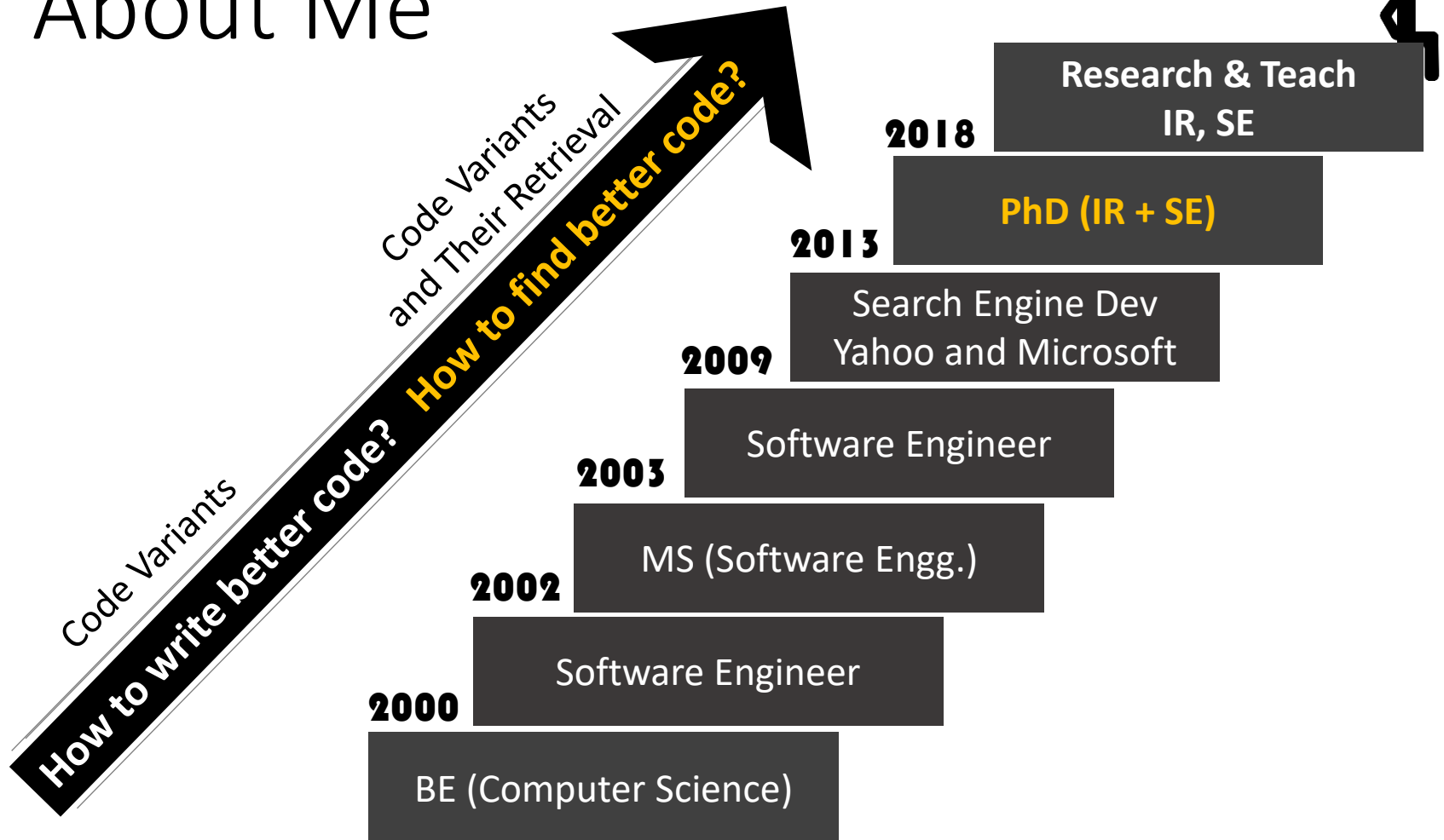
Sri Sivasubramaniya Nadar College of Engineering

---

Mathematics, the science of structure, order, and relation that has evolved from elemental practices of counting, measuring, and describing the shapes of objects.

Britannica, <https://www.britannica.com/science/mathematics>.

# About Me



# Agenda

## Why should **Researchers** care for **Mathematics**?

### Will Discuss

- ✓ Concepts
- ✓ Illustrations
- ✓ Intuitions
- ✓ Purpose
- ✓ Properties

### Will not Discuss

- ⊗ Details
- ⊗ Definitions
- ⊗ Formalism
- ⊗ Derivations
- ⊗ Proofs

### Three Stories

**Relational Model** in Databases,  
**Vector Space Models** and **Probabilistic Models** in  
Search Engines

# Story 1

Evolution of RDBMS and SQL



Interrelated “Data” as  
a Relation

## The Relational Model

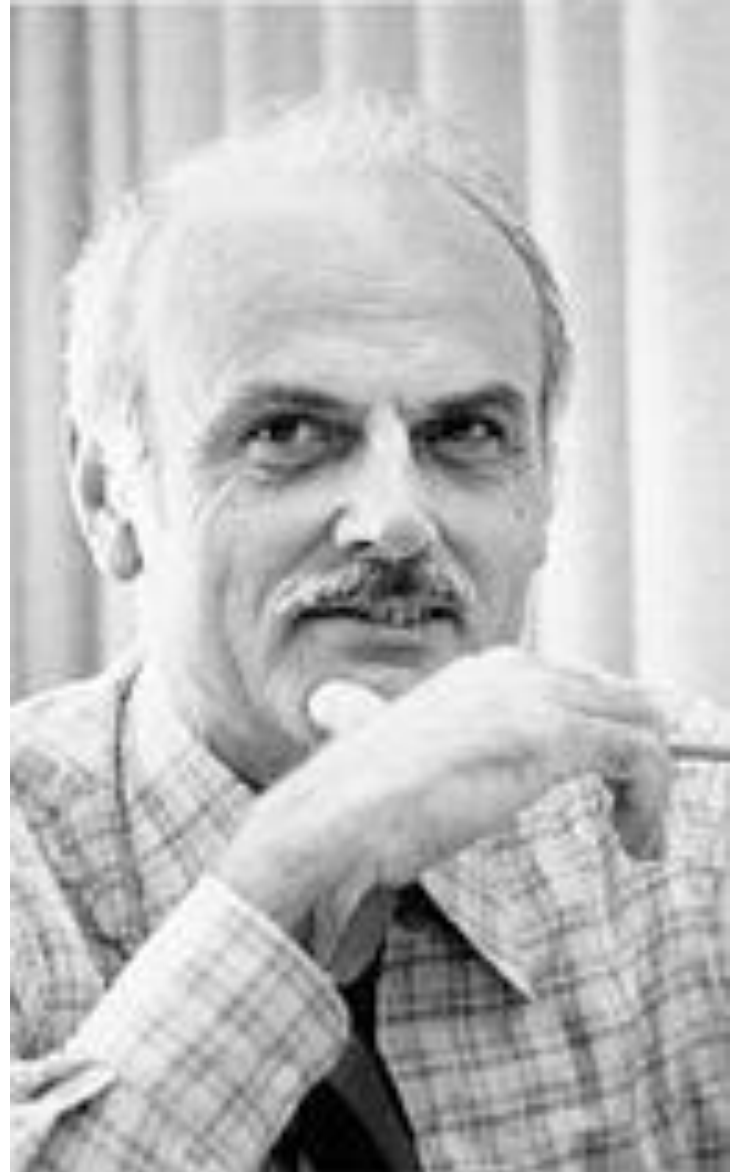
**Edgar F. Codd**

**PhD in Computer Science**

**Winner of the Turing Award**

*“He made other valuable contributions to computer science, but the relational model, a very influential general theory of data management, remains his most mentioned, analyzed and celebrated achievement.”*

— Wikipedia.



# The “Relation” from Math Textbook

- A **relation**  $R$  from a set  $A$  to set  $B$  is a subset of the cartesian product  $A \times B$ .

$$\begin{array}{c} \{\text{blue circle}, \text{black circle}, \text{red circle}\} \\ \text{set A} \end{array} \times \begin{array}{c} \{\text{blue triangle}, \text{red triangle}\} \\ \text{set B} \end{array} = \begin{array}{c} \{ (\text{blue circle}, \text{red triangle}), (\text{blue circle}, \text{blue triangle}), \\ (\text{black circle}, \text{red triangle}), (\text{black circle}, \text{blue triangle}), \\ (\text{red circle}, \text{red triangle}), (\text{red circle}, \text{blue triangle}) \} \\ \text{set of all ordered pairs, } A \times B \\ A \times B = \{ (a, b) \mid a \in A \text{ and } b \in B \} \end{array}$$

---

Let's say a relation exists between the reds

$$\text{Relation RedShapes} = \{ (\text{red circle}, \text{red triangle}) \}$$

# A Relation

- Let the set, **id** = {1,2,3}
- Let the set, **name** = {vv, sd}
- What is **id x name**? \_\_\_\_\_
- We have a **relation** if we assign a sequential id to each name.



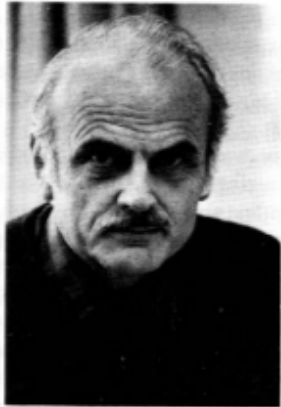
id	name
1	sd
1	vv
2	sd
2	vv
3	sd
3	vv

id	name
1	sd
2	vv



# The 1981 ACM Turing Award Lecture

Delivered at ACM '81, Los Angeles, California, November 9, 1981



The 1981 ACM Turing Award was presented to Edgar F. Codd, an IBM Fellow of the San Jose Research Laboratory, by President Peter Denning on November 9, 1981 at the ACM Annual Conference in Los Angeles, California. It is the Association's foremost award for technical contributions to the computing community.

Codd was selected by the ACM General Technical Achievement Award Committee for his "fundamental and continuing contributions to the theory and practice of database management systems." The originator of the relational model for databases, Codd has made further important contributions in the development of relational algebra, relational calculus, and normalization of relations.

Edgar F. Codd joined IBM in 1949 to prepare programs for the Selective Sequence Electronic Calculator. Since then, his work in computing has encompassed logical design of computers (IBM 701 and Stretch), managing a computer center in Canada, heading the development of one of the first operating systems with a general multiprogramming capability, contributing to the logic of self-reproducing automata, developing high level techniques for software specifica-

tion, creating and extending the relational approach to database management, and developing an English analyzing and synthesizing subsystem for casual users of relational databases. He is also the author of *Cellular Automata*, an early volume in the ACM Monograph Series.

Codd received his B.A. and M.A. in Mathematics from Oxford University in England, and his M.Sc. and Ph.D. in Computer and Communication Sciences from the University of Michigan. He is a Member of the National Academy of Engineering (USA) and a Fellow of the British Computer Society.

The ACM Turing Award is presented each year in commemoration of A. M. Turing, the English mathematician who made major contributions to the computing sciences.

<https://dl.acm.org/doi/pdf/10.1145/1283920.1283937?download=true>

---

## Relational Database: A Practical Foundation for Productivity

E. F. Codd  
IBM San Jose Research Laboratory



# Story So Far...

## What is a Relation?

$$\begin{array}{c} \{ \text{blue circle}, \text{black circle}, \text{red circle} \} \\ \text{set A} \end{array} \times \begin{array}{c} \{ \text{blue triangle}, \text{red triangle} \} \\ \text{set B} \end{array} = \begin{array}{c} \{ (\text{blue circle}, \text{red triangle}), (\text{blue circle}, \text{blue triangle}), \\ (\text{black circle}, \text{red triangle}), (\text{black circle}, \text{blue triangle}), \\ (\text{red circle}, \text{red triangle}), (\text{red circle}, \text{blue triangle}) \} \\ \text{set of all ordered pairs, } A \times B \\ A \times B = \{ (a, b) \mid a \in A \text{ and } b \in B \} \end{array}$$

Let's say a relation exists between the reds:

**Relation R = { (red circle, red triangle) }**

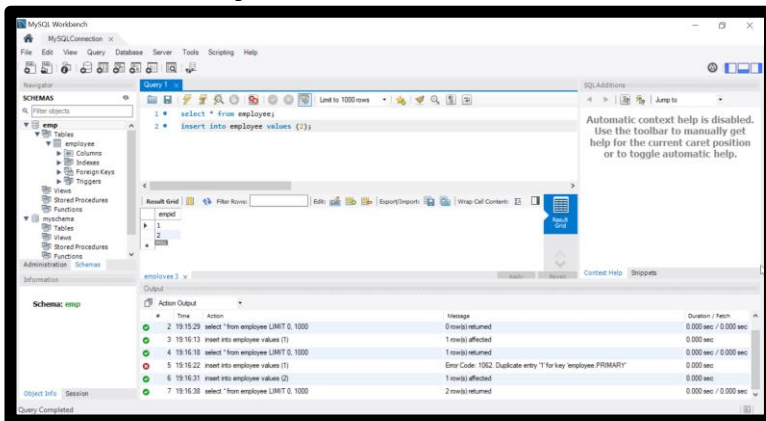
## Relational Data Model

id	name
1	sd
1	vv
2	sd
2	vv
3	sd
3	vv

Vs

id	name
1	sd
2	vv

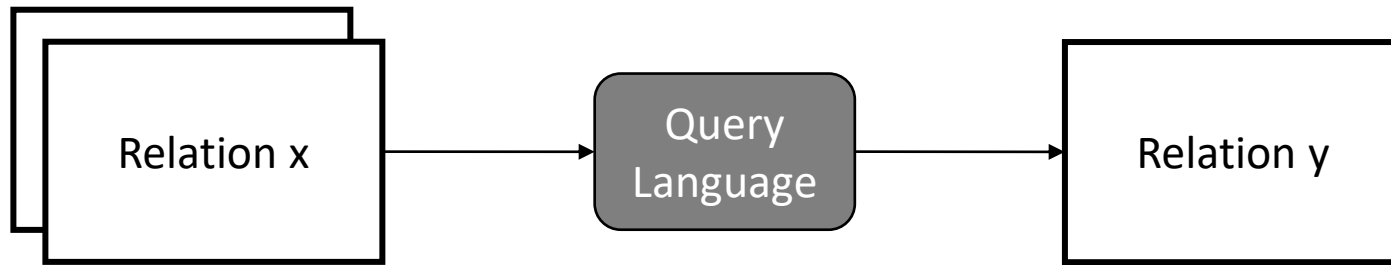
## MySQL – An RDBMS



**Attendance relation**  
**{(1,1), (2,1)}**  
is same as the table

	studentid	sessionid
▶	1	1
	2	1

# Query Languages



**Procedural Language**  
Relational Algebra

**Popular Language**  
SQL

# Select Operation

<i>ID</i>	<i>name</i>	<i>dept_name</i>	<i>salary</i>
22222	Einstein	Physics	95000
12121	Wu	Finance	90000
32343	El Said	History	60000
45565	Katz	Comp. Sci.	75000
98345	Kim	Elec. Eng.	80000
76766	Crick	Biology	72000
10101	Srinivasan	Comp. Sci.	65000
58583	Califieri	History	62000
83821	Brandt	Comp. Sci.	92000
15151	Mozart	Music	40000
33456	Gold	Physics	87000
76543	Singh	Finance	80000

(a) The *instructor* table

- Relational Algebra

$\sigma_{\text{dept\_name}=\text{"Physics"}}(\text{instructor})$

<i>ID</i>	<i>name</i>	<i>dept_name</i>	<i>salary</i>
22222	Einstein	Physics	95000
33456	Gold	Physics	87000

# Project Operation

- Notation:  $\Pi_{A_1, A_2, \dots, A_k}(r)$  where  $A_i$  are attribute names

<i>ID</i>	<i>name</i>	<i>dept_name</i>	<i>salary</i>
22222	Einstein	Physics	95000
12121	Wu	Finance	90000
32343	El Said	History	60000
45565	Katz	Comp. Sci.	75000
98345	Kim	Elec. Eng.	80000
76766	Crick	Biology	72000
10101	Srinivasan	Comp. Sci.	65000
58583	Califieri	History	62000
83821	Brandt	Comp. Sci.	92000
15151	Mozart	Music	40000
33456	Gold	Physics	87000
76543	Singh	Finance	80000

(a) The *instructor* table

- Example of projection:

$\Pi_{ID, name, salary}(instructor)$

# Projection

$\Pi_{ID, name, salary}(instructor)$

<i>ID</i>	<i>name</i>	<i>dept_name</i>	<i>salary</i>
22222	Einstein	Physics	95000
12121	Wu	Finance	90000
32343	El Said	History	60000
45565	Katz	Comp. Sci.	75000
98345	Kim	Elec. Eng.	80000
76766	Crick	Biology	72000
10101	Srinivasan	Comp. Sci.	65000
58583	Califieri	History	62000
83821	Brandt	Comp. Sci.	92000
15151	Mozart	Music	40000
33456	Gold	Physics	87000
76543	Singh	Finance	80000

(a) The *instructor* table



<i>ID</i>	<i>name</i>	<i>salary</i>
10101	Srinivasan	65000
12121	Wu	90000
15151	Mozart	40000
22222	Einstein	95000
32343	El Said	60000
33456	Gold	87000
45565	Katz	75000
58583	Califieri	62000
76543	Singh	80000
76766	Crick	72000
83821	Brandt	92000
98345	Kim	80000

# Union, Selection and Projection

- $\Pi_{course\_id} (\sigma_{semester="Fall" \wedge year=2009} (section)) \cup$   
 $\Pi_{course\_id} (\sigma_{semester="Spring" \wedge year=2010} (section))$

course_id	sec_id	semester	year	building	room_number	time_slot_id
BIO-101	1	Summer	2009	Painter	514	B
BIO-301	1	Summer	2010	Painter	514	A
CS-101	1	Fall	2009	Packard	101	H
CS-101	1	Spring	2010	Packard	101	F
CS-190	1	Spring	2009	Taylor	3128	E
CS-190	2	Spring	2009	Taylor	3128	A
CS-315	1	Spring	2010	Watson	120	D
CS-319	1	Spring	2010	Watson	100	B
CS-319	2	Spring	2010	Taylor	3128	C
CS-347	1	Fall	2009	Taylor	3128	A
EE-181	1	Spring	2009	Taylor	3128	C
FIN-201	1	Spring	2010	Packard	101	B
HIS-351	1	Spring	2010	Painter	514	C
MU-199	1	Spring	2010	Packard	101	D
PHY-101	1	Fall	2009	Watson	100	A



course_id
CS-101
CS-315
CS-319
CS-347
FIN-201
HIS-351
MU-199
PHY-101

Modern day SQL

select ... from ... where ...

Union

select ... from ... where ...

ARTICLE

## A relational model for unstructured documents



**Author:**  [A. Salminen](#) [Authors Info & Affiliations](#)

**Publication:** SIGIR '87: Proceedings of the 10th annual international ACM SIGIR conference on Research and development in information retrieval • November 1987 • Pages 196–207 • <https://doi.org/10.1145/42005.42028>

ARTICLE

## Semantics of extended relational model for managing uncertain information



**Authors:**  [V. S. Alagar](#),  [F. Sadri](#),  [J. N. Said](#) [Authors Info & Affiliations](#)

**Publication:** CIKM '95: Proceedings of the fourth international conference on Information and knowledge management • December 1995 • Pages 234–240 • <https://doi.org/10.1145/221270.221578>

RESEARCH-ARTICLE

## On the Optimization of Recursive Relational Queries: Application to Graph Queries



**Authors:**  [Louis Jachiet](#),  [Pierre Genevès](#),  [Nils Gesbert](#),  [Nabil Layaida](#) [Authors Info & Affiliations](#)

**Publication:** SIGMOD '20: Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data • June 2020 • Pages 681–697 • <https://doi.org/10.1145/3318464.3380567>



# Story 2

Models in Search Engines

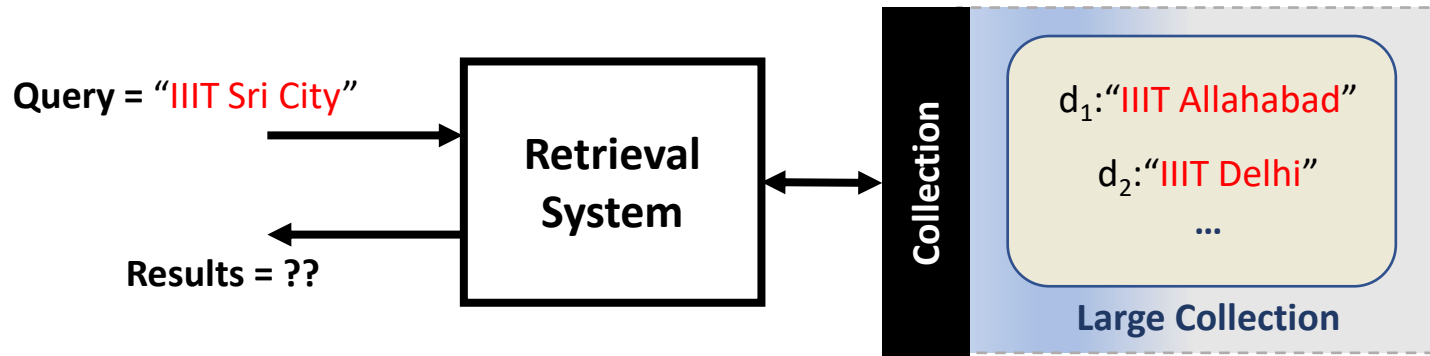


Image Source  
[https://philbradley.typepad.com/phil\\_bradleys\\_weblog/2016/07/100-search-engines-logos-image-for-you-to-use.html](https://philbradley.typepad.com/phil_bradleys_weblog/2016/07/100-search-engines-logos-image-for-you-to-use.html)

# Simple Retrieval Problem

- Say, we have a **collection** with 5 **documents**, each having the following contents
  - d1: IIIT ALLAHABAD
  - d2: IIIT DELHI
  - d3: IIIT GUWAHATI
  - d4: IIIT KANCHIPURAM
  - d5: IIIT SRI CITY
- Assume, the **Query** is
  - IIIT SRI CITY
- Which **document** will you match and why?

# The Problem: How to Build a Retrieval System?



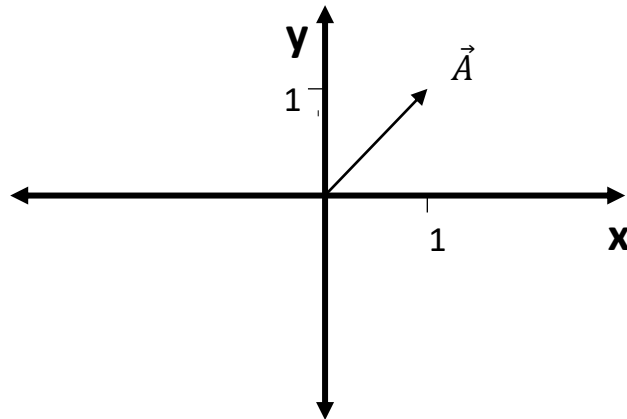
# One (bad) Approach

- First match the **term** IIIT.
  - Filter out documents that contain this term.
- Next match the **term** Sri.
  - Filter out documents that contain this term.
- Next match the **term** City.
  - Filter out documents that contain this term.

**Three iterations!**  
**Quiz: Can we do better?**

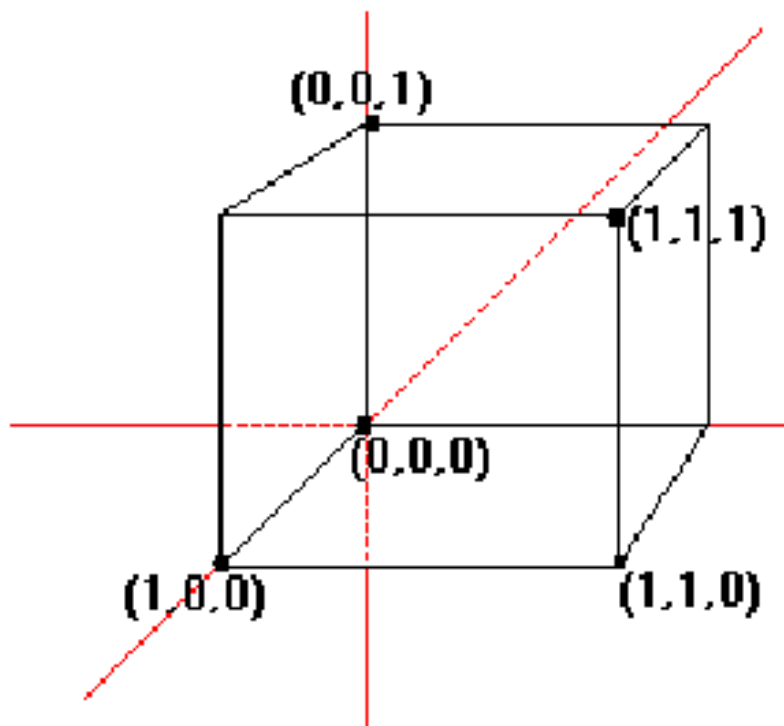
# Vector

- Geometric entity which has magnitude and direction



$\vec{A}$  is fixed at (0,0)

What is  $(1,1,1)$  ?



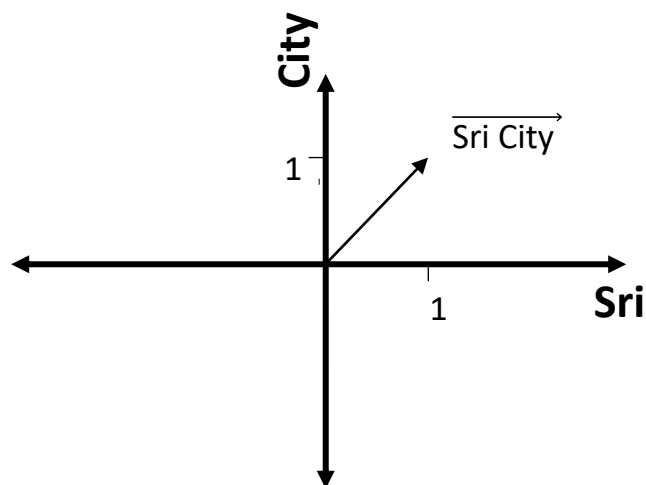
# Remember!

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



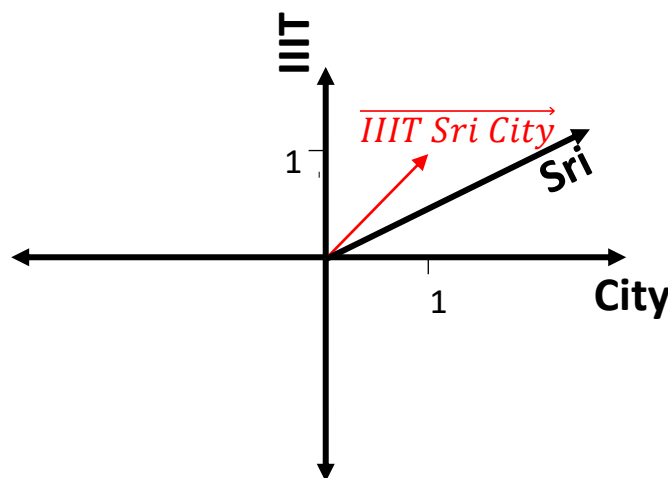
# Sentences are Vectors

- “Sri City” as a vector



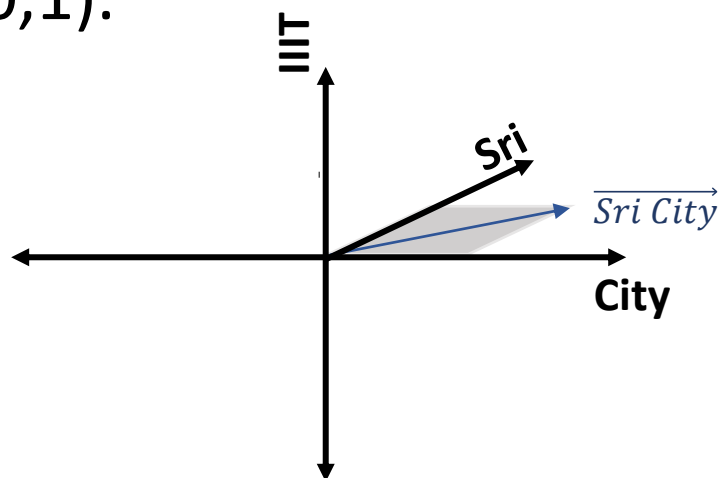
# Sentences are Vectors

- “**IIIT Sri City**” is a 3-dimensional vector



# Sentences are Vectors

- On this 3D space, “Sri City” vector will lie on the x (City) and z (Sri) plane. If  $(x,y,z)$  denotes the vector, “Sri City” is  $(1,0,1)$ .



# Natural Language Phrases as Vectors

Let query  $q$  = “IIIT Sri City”.

Let document,  $d_1$  = “IIIT Sri City” and  $d_2$  = “IIIT Delhi”.

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

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

# Quiz

- Considering the following vectors:

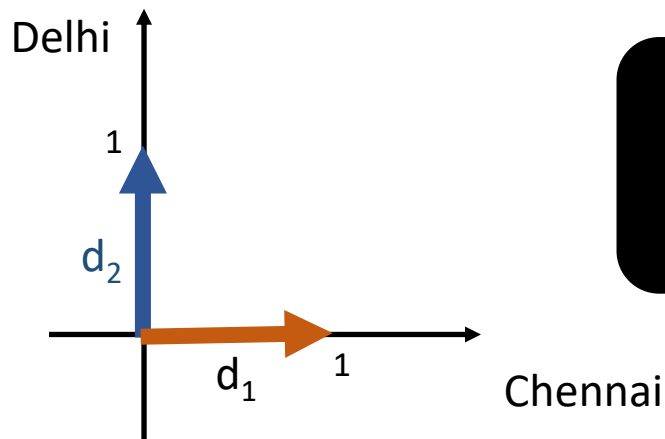
	IIIT	Sri	City	Delhi
q	1	1	1	0
d <sub>1</sub>	1	1	1	0
d <sub>2</sub>	1	0	0	1

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



# Similarity Score

- Assume, we have the following two documents:
  - $d_1$  = “Chennai”
  - $d_2$  = “Delhi”
- On a scale of 0 – 1, how similar are  $d_1$  and  $d_2$ ?
- What is the angle between  $d_1$  and  $d_2$  vectors?



**Can we express  
similarity as a function  
of the angles?**



0 – 90 to 1 – 0: How?

	0°	30°	45°	60°	90°
sin	0	1/2	1/√2	√3/2	1
cos	1	√3/2	1/√2	1/2	0
tan	0	1/√3	1	√3	Not defined

# Back to Trigonometry: Dot Product

- If  $\mathbf{a}$  and  $\mathbf{b}$  are non-unit vectors, what is the cosine of angle between them ( $\cos \theta$ )?

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

(or)

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

# Example

Let query  $q$  = “BITS Pilani”.

Let document,  $d_1$  = “BITS Pilani Goa Campus” and  $d_2$  = “IIT Delhi”.

	BITS	Pilani	Goa	Campus	IIT	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  $d_2 = (0,0,0,0,1,1)$

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

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

# An Assumption

**More frequent appearance of a query term  
implies higher document relevance.**

# 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}~~



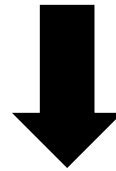
Vadivelu, a popular  
tamil actor

# 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”  $\rightarrow \{\text{IIIT}, \text{Sri}, \text{City}\}$
- “IIIT Sri City, Sri City”  $\rightarrow \{\text{IIIT}, \text{Sri}, \text{City}\}$   
(Assuming, we ignore the punctuations)

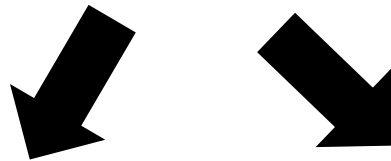


	IIIT	Sri	City
q	1	1	1



# Bag of Words Representation

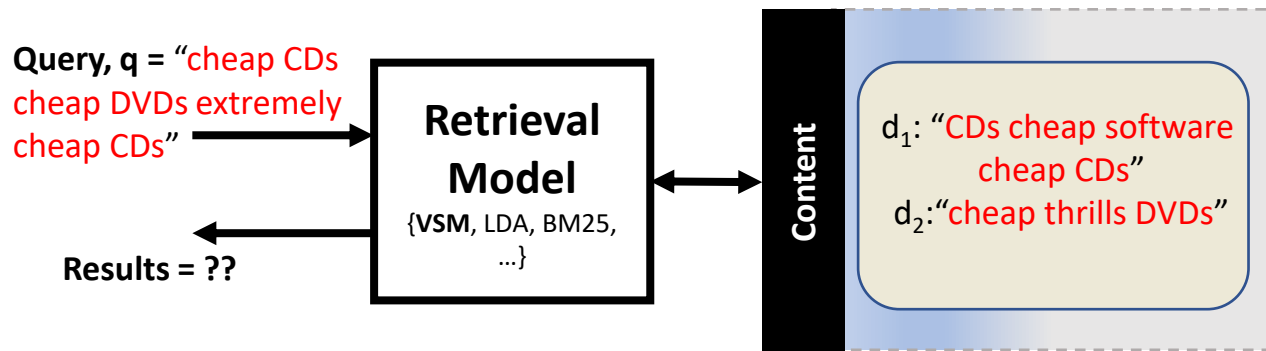
- “IIIT Sri City”  $\rightarrow \{\text{IIIT}, \text{Sri}, \text{City}\}$
- “IIIT Sri City, Sri City”  $\rightarrow [\text{IIIT}, \text{Sri}, \text{Sri}, \text{City}, \text{City}]$



IIIT Sri City				IIIT Sri City, Sri City			
	IIIT	Sri	City		IIIT	Sri	City
q	1	1	1	q	1	2	2

Leads to different vectors

# 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

$\text{sim}(q, d_1) = 0.86$

$\text{sim}(q, d_2) = 0.59$

ARTICLE

## The basic ideas in neural networks



**Authors:** [David E. Rumelhart](#), [Bernard Widrow](#), [Michael A. Lehr](#) [Authors Info & Affiliations](#)

**Publication:** Communications of the ACM • March 1994 • <https://doi.org/10.1145/175247.175256>

RESEARCH-ARTICLE

## word2vec, node2vec, graph2vec, X2vec: Towards a Theory of Vector Embeddings of Structured Data



**Author:** [Martin Grohe](#) [Authors Info & Affiliations](#)

**Publication:** PODS'20: Proceedings of the 39th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems • June 2020 • Pages 1–16 • <https://doi.org/10.1145/3375395.3387641>

RESEARCH-ARTICLE

## CC2Vec: distributed representations of code changes



**Authors:** [Thong Hoang](#), [Hong Jin Kang](#), [David Lo](#), [Julia Lawall](#) [Authors Info & Affiliations](#)

**Publication:** ICSE '20: Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering • June 2020 • Pages 518–529 • <https://doi.org/10.1145/3377811.3380361>



# Story 3

---

A Probabilistic Model



# The Law – Robert M. Coates

From the book, “The World of Mathematics – Volume IV”.

**Triborough Bridge, NY, USA.**  
(aka Robert F. Kennedy Bridge)



Late 1940s, NY: No other bridge or main highway was affected, and though the two preceding nights had been equally balmy and moonlit, on both of these the bridge traffic had run close to **normal**.

And then, one day...



It just looked as if **everybody** in Manhattan who owned a car had decided to drive out to Long Island that evening.

# No Reason!



**Sergeant:** “I kept askin’ them” he said, “Is there night football somewhere that we don’t know about? Is it the races you’re goin’ to?”

But the funny thing was half the time they’d be askin’ me. “What’s the crowd for, Mac?” they would say. And I’d just look at them.

If normal things stop happening, if  
we lose regularities in life, our  
planet could become unlivable!

# Time for Action

- At this juncture, it was inevitable that Congress should be called on for action.



- Senator said, “*You can control it*”. Re-education and reforms were decided upon. He said, (we need to lead people back to) “*the basic regularities, the homely **averageness** of the American way of life*”.



# The Law of Large Numbers

Known as the Fundamental theorem of Probability

**The average of the results obtained from a large number of trials should be close to the expected value, and will tend to become closer as more trials are performed.**

# Probabilistic Retrieval

- Information Need: Taj Mahal
- Let a query  $q$  be “Taj”
- Let the results be:
  - d1: Taj
  - d2: Taj Mahal
  - d3: Taj Tea
- Two judges were asked to provide relevance judgments:

Document	Judge 1	Judge 2
Taj	R	N
Taj Mahal	R	R
Taj Tea	N	N

# Probability of Relevance

- Documents can have probability of being relevant and of being non-relevant at the same time.
- Example:
  - Documents in our collection :

Document	$P(R=0 d,q)$	$P(R=1 d,q)$
Taj	0.5	0.5
Taj Mahal	?	?
Taj Tea	?	?

$R = 0 \rightarrow$  Non-Relevant

$R = 1 \rightarrow$  Relevant

# Probability of Relevance

- Documents can have probability of being relevant and of being non-relevant at the same time.
- Example:
  - Documents in our collection :

Document	$P(R=0 d,q)$	$P(R=1 d,q)$
Taj	0.5	0.5
Taj Mahal	0	1
Taj Tea	1	0

$R = 0 \rightarrow$  Non-Relevant

$R = 1 \rightarrow$  Relevant

# Probability Ranking Principle

**Rank documents by the  
probability of relevance,  
 $P(R=1 | q, d)$   $R \in \{0, 1\}$**

# Probability Ranking Principle

**Rank documents by the  
probability of relevance,  
 $P(R=1 | q, d)$   $R \in \{0, 1\}$**

Document	$P(R=0   d, q)$	$P(R=1   d, q)$
Taj	0.5	0.5
Taj Mahal	0	1
Taj Tea	1	0

$R = 0 \rightarrow$  Non-Relevant

$R = 1 \rightarrow$  Relevant

**Search Result:**

1. Taj Mahal
2. Taj
3. Taj Tea

# Bayes Optimal Decision Rule

d is relevant if  
 $P(R=1 | d, q) > P(R=0 | d, q)$


Document	$P(R=0   d, q)$	$P(R=1   d, q)$
Taj	0.5	0.5
Taj Mahal	0	1
Taj Tea	1	0

**Search Result:**

1. Taj Mahal

# Predicting Relevance

Document	$P(R=0   d, q)$	$P(R=1   d, q)$
Taj	0.5	0.5
Taj Mahal	0	1
Taj Tea	1	0



This is user given relevance.

Can we  
estimate/predict  
relevance based  
on term  
occurrence ?



# Predicting Relevance

- You may use labeled set from judges (or mined from clicklogs)
- You may assume query and document as set of words.

Query	Document	Relevance
q1 = (x1,x2,...)	d1 = (..xi, xj,...)	1
q1	d2	1
q1	d3	0
q2	d1	0
q2	d2	0
q2	d3	1



This is user given relevance.

Can we estimate/predict relevance ?

# Binary Independence Model (BIM)

- Each document is a binary vector of terms.
- Occurrence of terms is mutually independent.

$$P(R = 1|d, q) = \frac{P(d|R = 1, q)P(R = 1|q)}{P(d|q)}$$

Bayes Rule

# Example

$$P(R = 1|d, q) = \frac{P(d|R = 1, q)P(R = 1|q)}{P(d|q)}$$

Bayes Rule

- $P(d=\text{Taj} | R=1, q) = ?$
- $P(R=1 | q) = ?$
- $P(d | q) = ?$

Document	$P(R=0   d, q)$	$P(R=1   d, q)$
Taj	0.5	0.5
Taj Mahal	0	1
Taj Tea	1	0

# Quiz

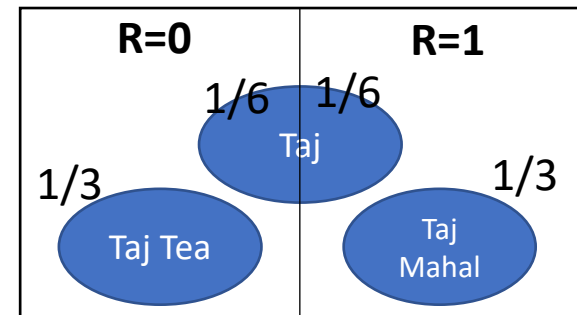
$$P(R = 1|d, q) = \frac{P(d|R = 1, q)P(R = 1|q)}{P(d|q)}$$

Bayes Rule

- $P(d=\text{Taj} | R=1, q) = 1/3$
- $P(R=1 | q) = 1/2$
- $P(d=\text{Taj} | q) = 1/3$

$$P(R=1 | d=\text{Taj}, q) = (1/3)(1/2)/(1/3) = 1/2$$

Document	$P(R=0   d, q)$	$P(R=1   d, q)$
Taj	0.5	0.5
Taj Mahal	0	1
Taj Tea	1	0



Euler Diagram

# Predicting Relevance

- Odds of Relevance is easier to calculate

$$O(R|\vec{x}, \vec{q}) = \frac{P(R=1|\vec{x}, \vec{q})}{P(R=0|\vec{x}, \vec{q})} = \frac{\frac{P(R=1|\vec{q})P(\vec{x}|R=1, \vec{q})}{P(\vec{x}|\vec{q})}}{\frac{P(R=0|\vec{q})P(\vec{x}|R=0, \vec{q})}{P(\vec{x}|\vec{q})}} = \frac{P(R=1|\vec{q})}{P(R=0|\vec{q})} \cdot \frac{P(\vec{x}|R=1, \vec{q})}{P(\vec{x}|R=0, \vec{q})}$$

Constant term.

- In BIM, we assume that the term occurrence is mutually independent

$$\begin{aligned} \frac{P(\vec{x}|R=1, \vec{q})}{P(\vec{x}|R=0, \vec{q})} &= \prod_{t=1}^M \frac{P(x_t|R=1, \vec{q})}{P(x_t|R=0, \vec{q})} = \prod_{t:x_t=1} \frac{P(x_t=1|R=1, \vec{q})}{P(x_t=1|R=0, \vec{q})} \cdot \prod_{t:x_t=0} \frac{P(x_t=0|R=1, \vec{q})}{P(x_t=0|R=0, \vec{q})} \\ &= \prod_{t:x_t=q_t=1} \frac{p_t}{u_t} \cdot \prod_{t:x_t=0, q_t=1} \frac{1-p_t}{1-u_t} \end{aligned}$$

# Retrieval Status Value

Not document specific.  
Constant for a query.



$$\prod_{t: x_t = q_t = 1} \frac{p_t(1 - u_t)}{u_t(1 - p_t)} \cdot \prod_{t: q_t = 1} \frac{1 - p_t}{1 - u_t}$$

$$\text{RSV} = \prod_{t: x_t = q_t = 1} \frac{p_t(1 - u_t)}{u_t(1 - p_t)}$$

**RSV is used for ranking documents.**

ARTICLE

## Probabilistic model for contextual retrieval



**Authors:** [Ji-Rong Wen](#), [Ni Lao](#), [Wei-Ying Ma](#) [Authors Info & Affiliations](#)

**Publication:** SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval • July 2004 • Pages 57–63 • <https://doi.org/10.1145/1008992.1009005>

## Using Term Location Information to Enhance Probabilistic Information Retrieval



**Authors:** [Baiyan Liu](#), [Xiangdong An](#), [Jimmy Xiangji Huang](#) [Authors Info & Affiliations](#)

**Publication:** SIGIR '15: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval • August 2015 • Pages 883–886 • <https://doi.org/10.1145/2766462.2767827>

RESEARCH-ARTICLE

## Probabilistic Topic Models for Text Data Retrieval and Analysis



**Author:** [ChengXiang Zhai](#) [Authors Info & Affiliations](#)

**Publication:** SIGIR '17: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval • August 2017 • Pages 1399–1401 • <https://doi.org/10.1145/3077136.3082067>

# Summary: Relational Model

## What is a Relation?

$$\{\text{blue circle, black circle, red circle}\} \times \{\text{blue triangle, red triangle}\} = \begin{matrix} (\text{blue circle}, \text{blue triangle}), (\text{blue circle}, \text{red triangle}), \\ (\text{black circle}, \text{blue triangle}), (\text{black circle}, \text{red triangle}), \\ (\text{red circle}, \text{blue triangle}), (\text{red circle}, \text{red triangle}) \end{matrix}$$

set of all ordered pairs,  $A \times B$   
 $A \times B = \{(a, b) \mid a \in A \text{ and } b \in B\}$

Let's say a relation exists between the reds:

**Relation R =  $\{(\text{red circle}, \text{red triangle})\}$**

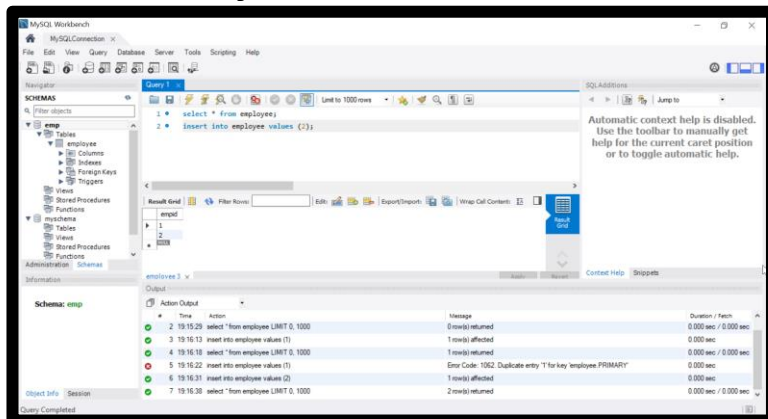
## Relational Data Model

id	name
1	sd
1	vv
2	sd
2	vv
3	sd
3	vv

Vs

id	name
1	sd
2	vv

## MySQL – An RDBMS



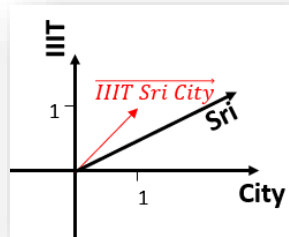
**Attendance relation**  
 $\{(1,1), (2,1)\}$   
is same as the table

	studentid	sessionid
▶	1	1
	2	1



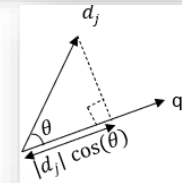
# Summary: Vector Space Model

## Model Documents as Vectors

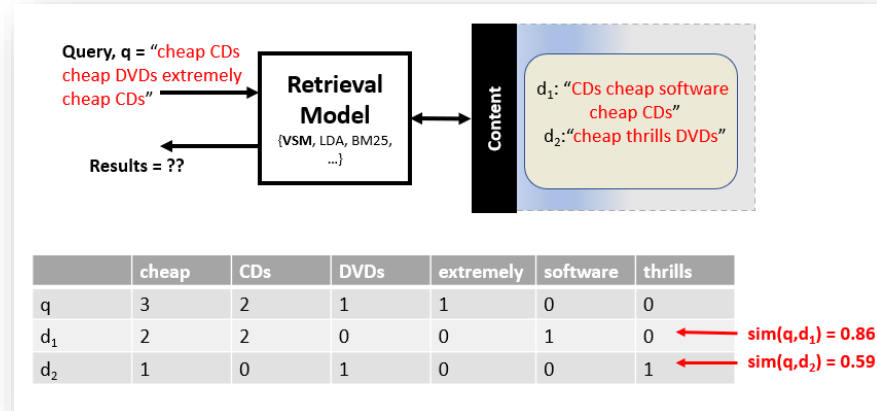


## Compute Similarity

$$\cos(\theta) = \frac{d_j \cdot q}{||d_j|| ||q||}$$

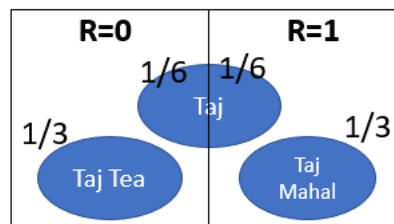


## A Worked-Out Example



# Summary: Probabilistic Model

Document	$P(R=0   d, q)$	$P(R=1   d, q)$
Taj	0.5	0.5
Taj Mahal	0	1
Taj Tea	1	0



$$RSV = \prod_{t: x_t = q_t = 1} \frac{p_t(1 - u_t)}{u_t(1 - p_t)}$$

RSV is used for ranking documents.

# References

- An Introduction to Information Retrieval.  
Christopher D. Manning, Prabhakar Raghavan,  
Hinrich Schutze.
- Search Engines: Information Retrieval in Practice.  
Bruce Croft, Donald Metzler, Trevor Strohman

## VENKATESH VINAYAKARAO

Lecturer  
Chennai Mathematical Institute  
venkateshv@cmi.ac.in  
[CV](#)

### Research



My research interest is in building **Search Engines**. How does Google work? How to search trillions of documents within microseconds? How to evaluate if Google is better or Bing? Broadly, **Information Retrieval** is the field of study that investigates these questions. My current focus is on investigating the techniques to search for source code. You will see me discussing Programming Languages, Program Analysis and Software Engineering. More about my research is [here](#).

If you are looking for an answer to an even more fundamental and important question: Why to study information retrieval?, enjoy the [video](#) from my students of the 2018 IR offering at IIITS - Mounika, Parkhi and Pragna.

**Publications:** [DBLP](#) [Google Scholar](#)

### Teaching

Term@CMI: Feb - Mar 2021: [RDBMS, SQL and Visualization](#)  
Term@CMI: Dec - Jan 2020: Advanced Information Retrieval  
Term@CMI: Aug - Nov 2020: [Information Retrieval](#)  
Term@CMI: Jan - Apr 2020: Big Data and Hadoop  
Term@CMI: Jan - Apr 2020: Applied Program Analysis  
Term@CMI: Oct - Nov 2019: RDBMS, SQL and Visualization  
Term@CMI: Aug - Sep 2019: Information Retrieval  
Term@CMI: Mar - Apr 2019: Program Analysis  
Term@IIITS: Aug - Dec 2018: Information Retrieval  
Term@IIITS: Aug - Dec 2018: Computer Programming

### Talks

— Tweets by @venkvinn

# Thank You

[venkateshv@cmi.ac.in](mailto:venkateshv@cmi.ac.in)

This slide deck is available at <http://vvtesh.co.in/>.