CS3320 Multimedia Information Retrieval Unit 1: an introduction

George Vogiatzis

Learning outcomes

- After this unit you should
  - Know about the **organisation** of the module
  - Know how the module is **assessed**
  - Have an idea of **what CS3320 is about**

Contact info

- MB211F
- E-mail: g.vogiatzis@aston.ac.uk
- Office hour
  - WASS

Aston University
Birmingham

Module organisation

- Lectures
  - Theory of IR

- Labs
  - Practical illustrations of IR techniques using Python

Assessment

- 80% exam
  - May
  - 1.5 hours
  - Based on lectures and parts of the lab questions
- 20% coursework
  - 10% Coursework 1: Document clustering
    - Spec: week 5
    - Hand-in: week 7
  - 10% Coursework 2: SIFT matching
    - Spec: week 9
    - Hand-in: week 11

How to make most of CS3320

- **Revise** when needed
  - Image Processing (Yr2)
  - Some basic maths
- **Attend** lectures & labs
- **Read** through the directed reading material

Literature

- Lecture notes, lab sheets
- Directed reading material
- Book (optional but v. helpful)
  - **Manning, Raghavan and Schutze: Introduction to Information Retrieval,**
    Cambridge University Press

    will be referred to as (IIR)

Information retrieval

- Information Retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).

IR vs Databases

- Databases = structured data

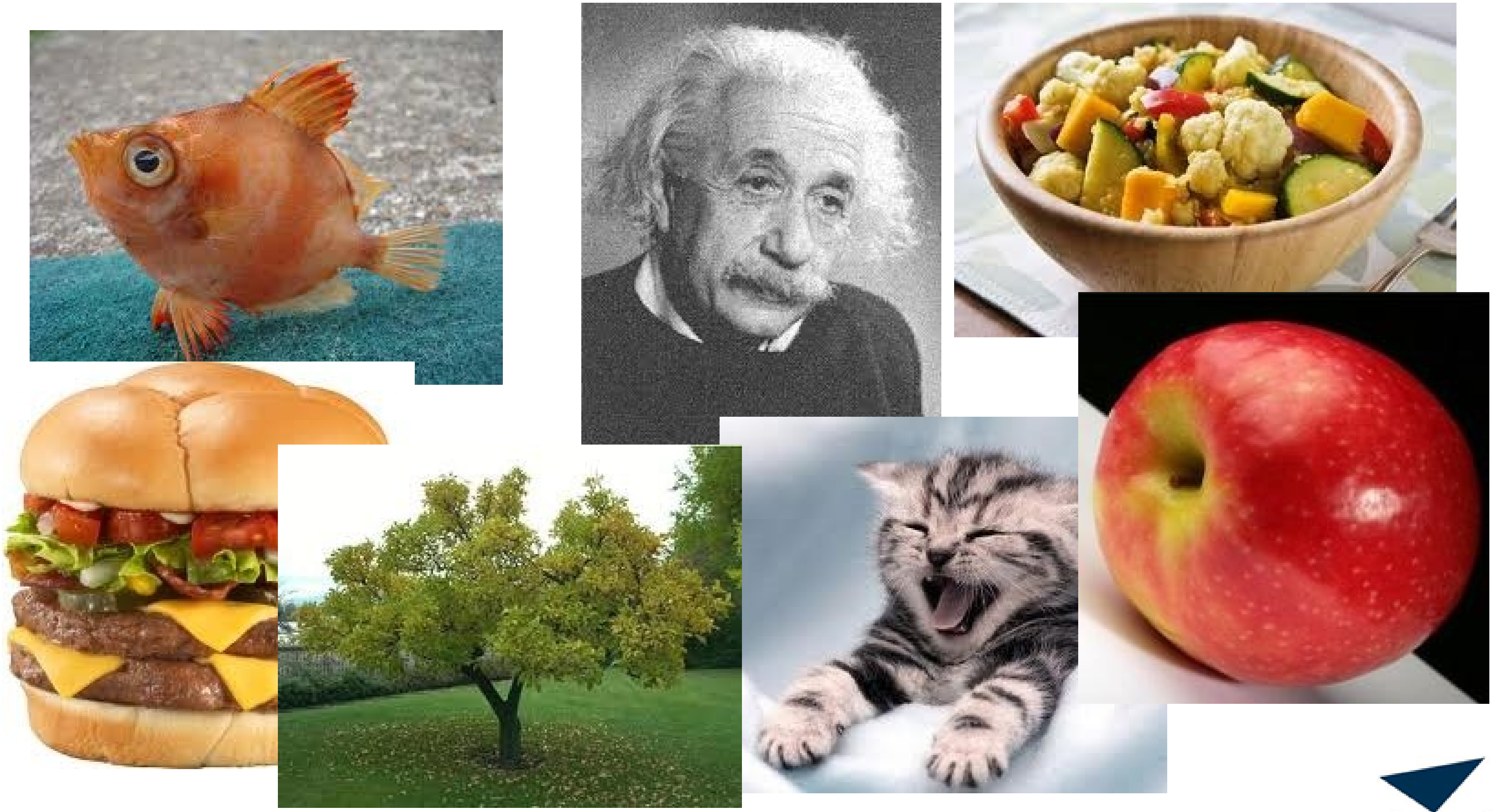
EmployeeID	Name	DateOfBirth	Salary
1	George	11-Jan-1978	£32,000
2	Mary	13-Jun-1975	£35,000
3	Thomas	2-Feb-1986	£25,000

- Searching with SQL querries
  - SELECT * FROM Employee WHERE Salary>£30,000
  - returns:

EmployeeID	Name	DateOfBirth	Salary
1	George	11-Jan-1978	£30,000
2	Mary	13-Jun-1975	£35,000

IR vs Databases

- Unstructured data



IR vs Databases

- Unstructured data

We the people of the United States, in
order to form a more perfect union,
establish justice, insure domestic
tranquility, provide for the common
defense, promote the general
welfare, and secure the blessings of
liberty to ourselves and our posterity,
do orda
Constitu
America
....

We shall fight on the beaches, we shall
fight on the landing grounds, we shall fight
in the fields and in the streets, we shall
fight in the hills; we shall never surrender,
and even if, which I do not for a moment
believe, this island or a large part of it were
subjugated and starving, then our Empire
beyon ded by
the Br e
strugg the new
world, , steps
forth t on of the
old.
....

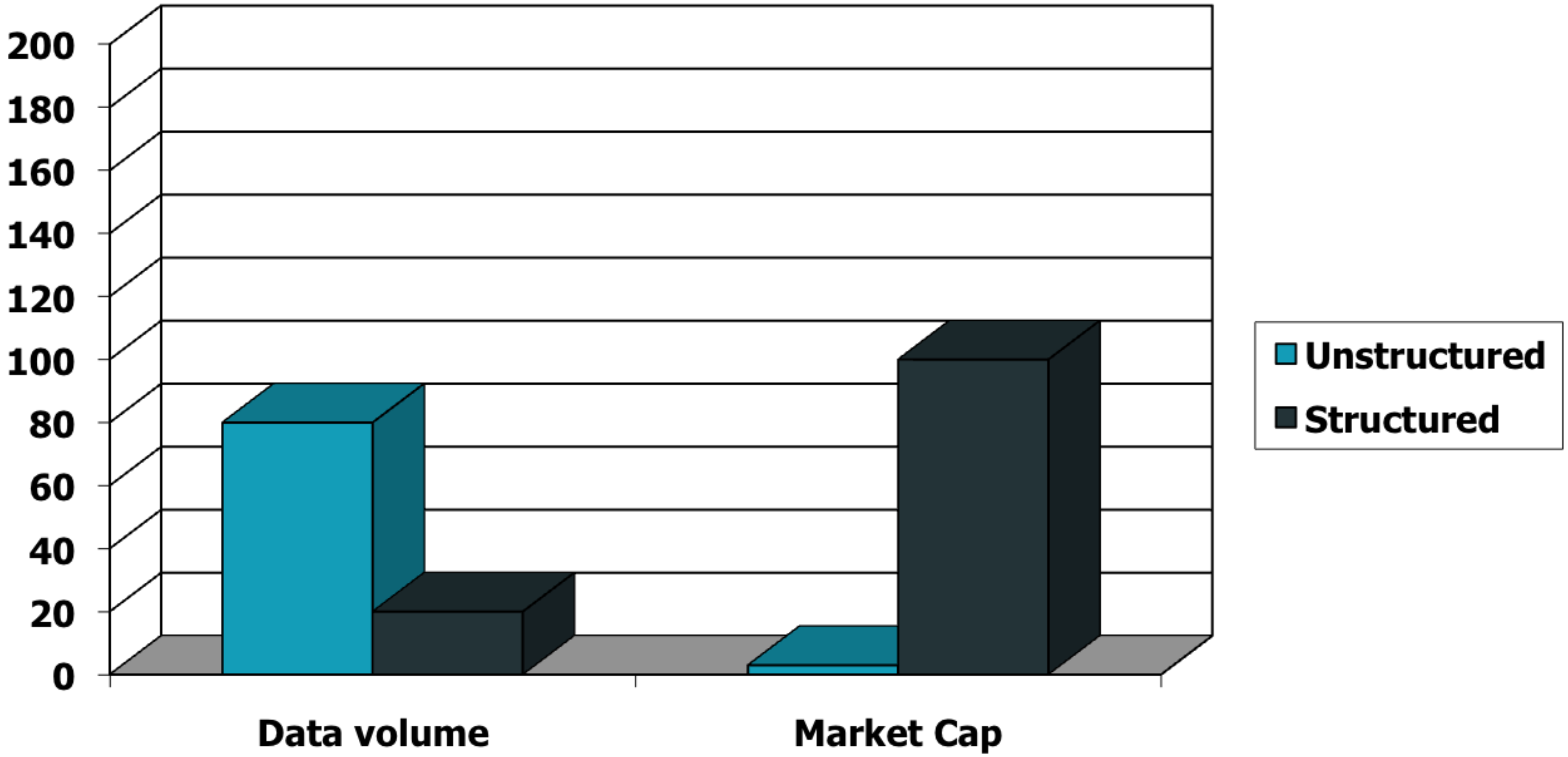
To be, or not to be: that is the
Whether 'tis nobler in the mind
The slings and arrows of outra
fortune,
Or to take arms against a sea
And by opposing end them? To die: to
sleep;
No more; and by a sleep to say we en
The heart-ache and the thousand natu
shocks
That flesh is heir to, 'tis a consummat
Devoutly to be wish'd. To die, to sleep
....

Imagine there's no
Heaven
It's easy if you try
No hell below us
Above us only sky
Imagine all the people
Living for today

Imagine there's no
countries
It isn't hard to do
Nothing to kill or die for
And no religion too
Imagine all the people
Living life in peace
...

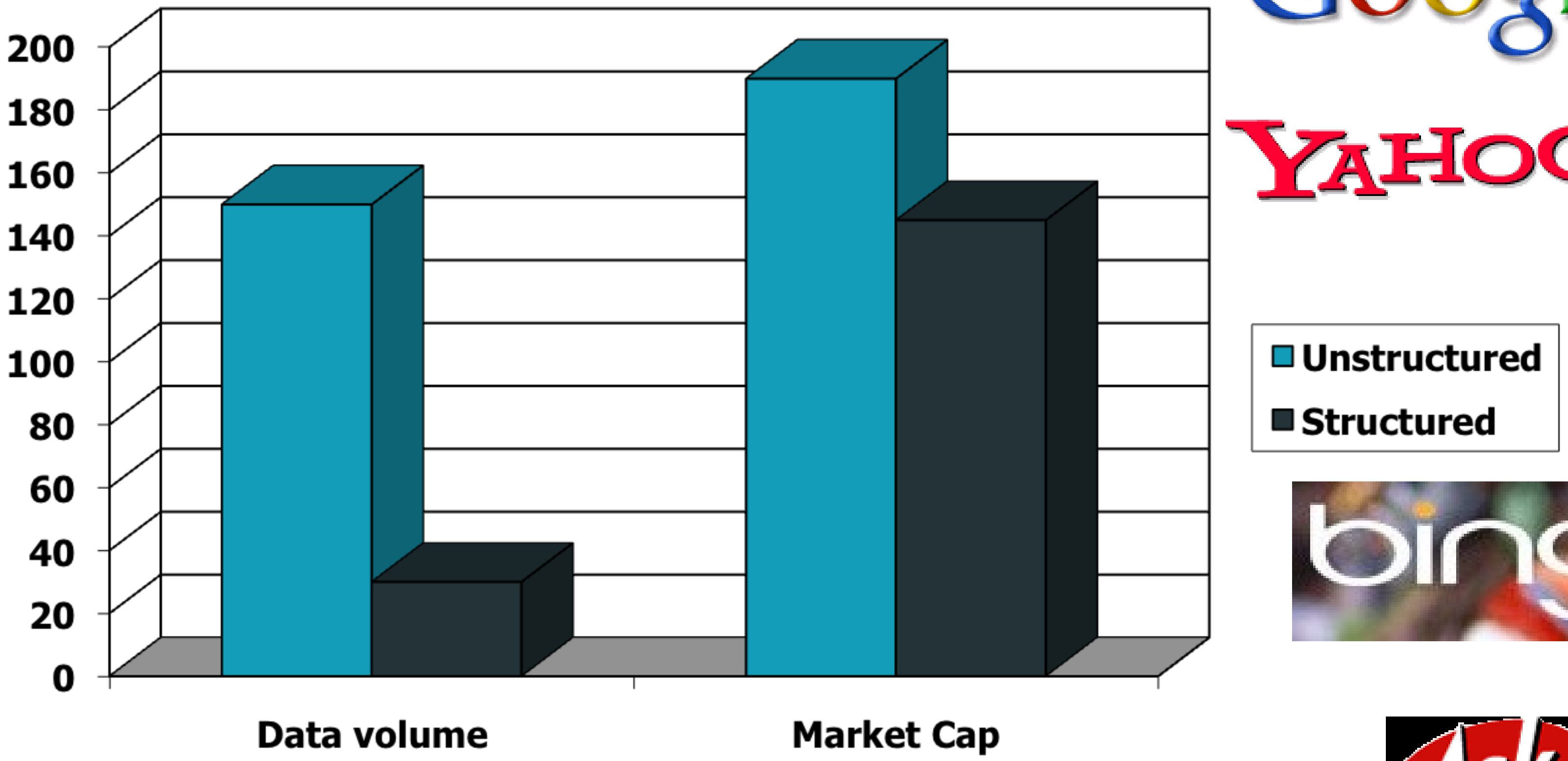
Unstructured vs Structured

- Text vs databases in 1996



Unstructured vs Structured

- Text vs databases in 2009



Google™

YAHOO!®

bing

Ask.com™

Topics that will be covered

- Part I: document retrieval
  - Boolean retrieval
  - Ranked retrieval
    - Term weighting
    - Vector space model
  - Classification
  - Clustering
- Part II: Content based Image retrieval
  - “How to turn an image into 1000 words”

Boolean retrieval

- Which plays of Shakespeare contain the words ***Brutus*** *AND* ***Caesar*** but *NOT* ***Calpurnia***?
- The Boolean retrieval model is being able to ask a query that is a Boolean expression:
  - Boolean Queries are queries using *AND*, *OR* and *NOT* to join query terms
    - Views each document as a set of words
    - Is precise: document matches condition or not.
  - Perhaps the simplest model to build an IR system
- Primary commercial retrieval tool for 3 deca
- Many search systems you still use are Boole
  - Email, library catalog, Mac OS X Spotlight
  - Commercial legal/medical search services

Ranked retrieval

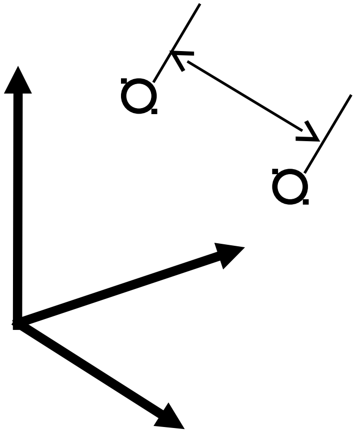
- Boolean queries give inclusion or exclusion of docs.
- Often we want to rank/group results
  - Need to measure proximity from query to each doc.
  - Need to decide whether docs presented to user are singletons, or a group of docs covering various aspects of the query.

  - Assign score to each term that is matched
    - If a document includes the term *many times* => high score
    - If that term is *rare* => even higher score!

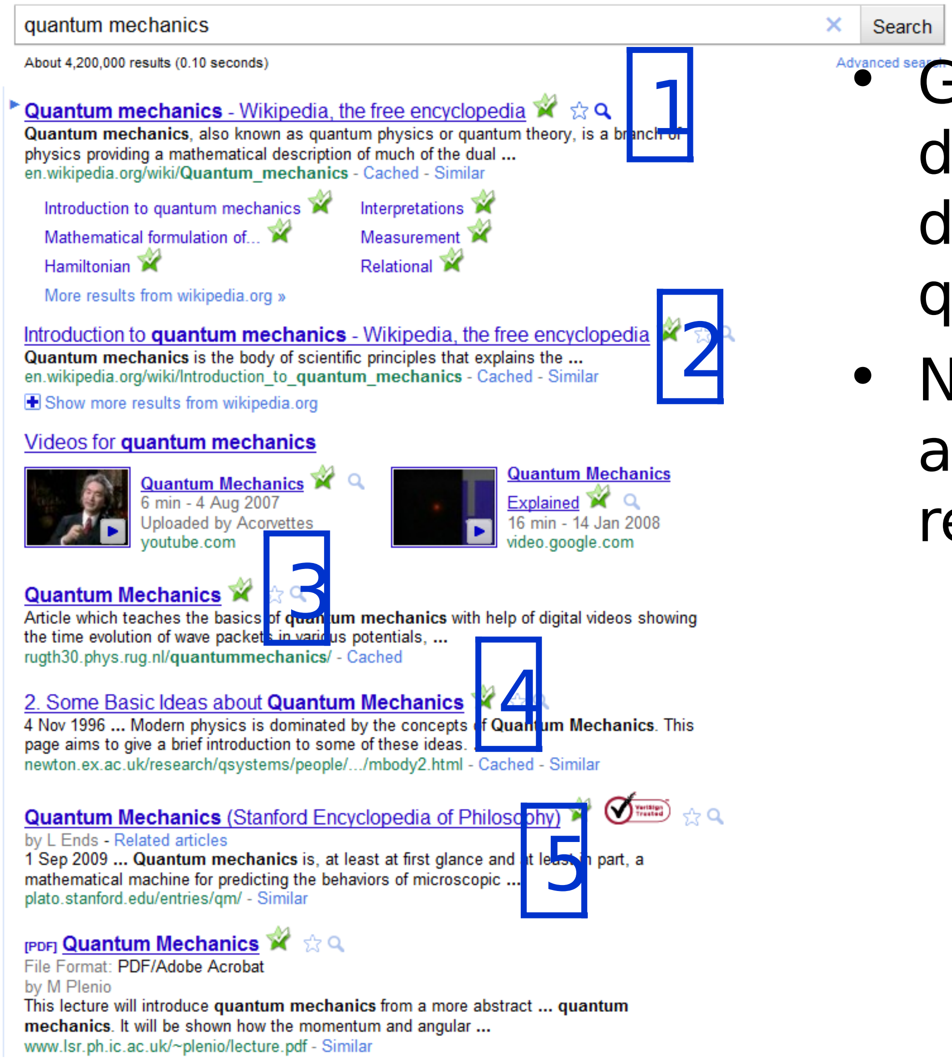
Vector space model

- Ranked retrieval needs a concept of distance between queries and documents

- Solution: convert each document/query into a point in N-dimensional space
- Distance between documents/queries is geometric distance between two points



Vector space model



quantum mechanics

About 4,200,000 results (0.10 seconds)

Quantum mechanics - Wikipedia, the free encyclopedia
Quantum mechanics, also known as quantum physics or quantum theory, is a branch of physics providing a mathematical description of much of the dual ...
en.wikipedia.org/wiki/**Quantum_mechanics** - Cached - Similar

Introduction to quantum mechanics
Mathematical formulation of...
Hamiltonian
Interpretations
Measurement
Relational
More results from wikipedia.org »

Introduction to **quantum mechanics** - Wikipedia, the free encyclopedia
Quantum mechanics is the body of scientific principles that explains the ...
en.wikipedia.org/wiki/Introduction_to_**quantum_mechanics** - Cached - Similar
Show more results from wikipedia.org

Videos for **quantum mechanics**

Quantum Mechanics
6 min - 4 Aug 2007
Uploaded by Acorvettes
youtube.com

Quantum Mechanics Explained
16 min - 14 Jan 2008
video.google.com

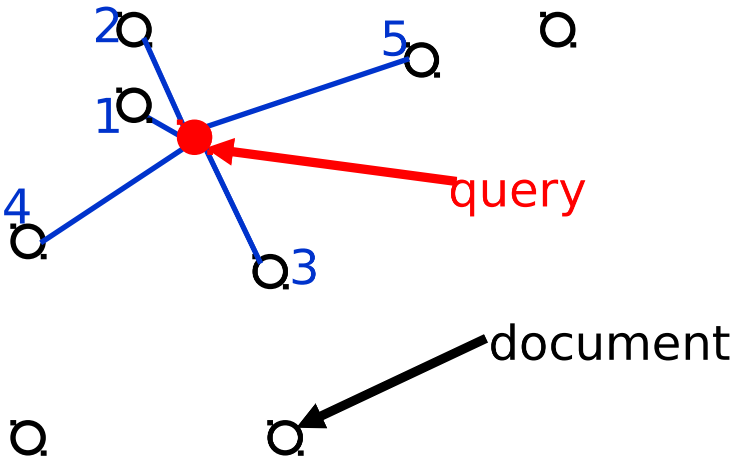
Quantum Mechanics
Article which teaches the basics of **quantum mechanics** with help of digital videos showing the time evolution of wave packets in various potentials, ...
rugth30.phys.rug.nl/**quantummechanics**/ - Cached

2. Some Basic Ideas about Quantum Mechanics
4 Nov 1996 ... Modern physics is dominated by the concepts of **Quantum Mechanics**. This page aims to give a brief introduction to some of these ideas. ...
newton.ex.ac.uk/research/qsystems/people/.../mbody2.html - Cached - Similar

Quantum Mechanics (Stanford Encyclopedia of Philosophy)
by L Ends - Related articles
1 Sep 2009 ... **Quantum mechanics** is, at least at first glance and at least in part, a mathematical machine for predicting the behaviors of microscopic ...
plato.stanford.edu/entries/qm/ - Similar

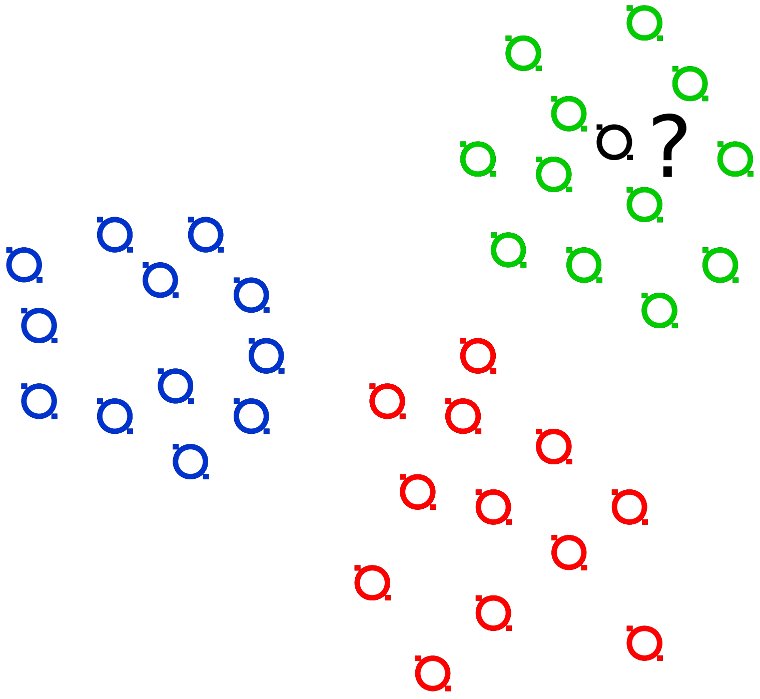
[PDF] **Quantum Mechanics**
File Format: PDF/Adobe Acrobat
by M Plenio
This lecture will introduce **quantum mechanics** from a more abstract ... **quantum mechanics**. It will be shown how the momentum and angular ...
www.lsr.ph.ic.ac.uk/~plenio/lecture.pdf - Similar

- Given a query, we rank all documents according to distance away from the query
- Nearest documents are assumed to be more relevant



Document classification

- Automatically determining the category/class of a document
- E.g. I have three types of email in my inbox, work-related, personal, and spam.



- I manually label some emails and
- Expect the system to automatically label new ones

Document clustering

- Without any manual supervision, determine what categories of document exist in our collection.

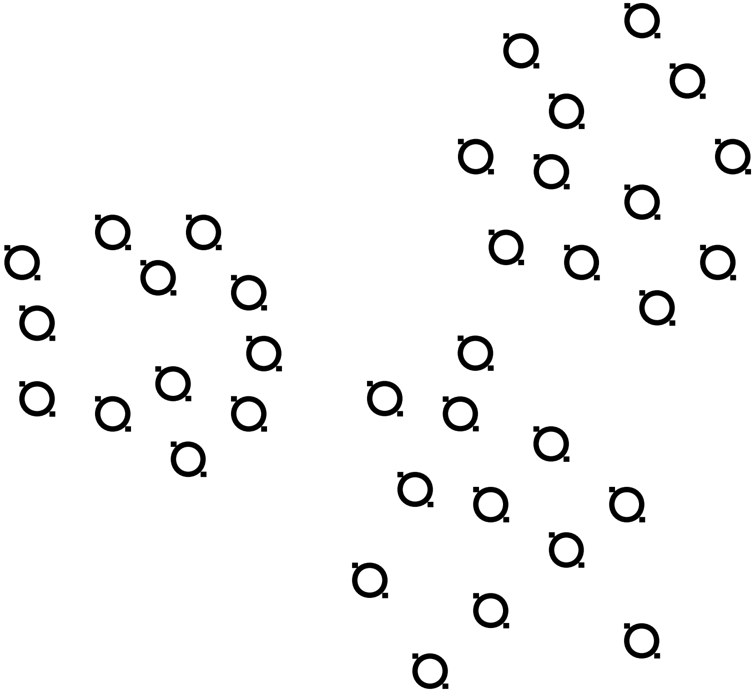


Image retrieval

- Entering tags manually is time-consuming process
- Solution: Content Based Image Retrieval
- “Find me images similar to this”

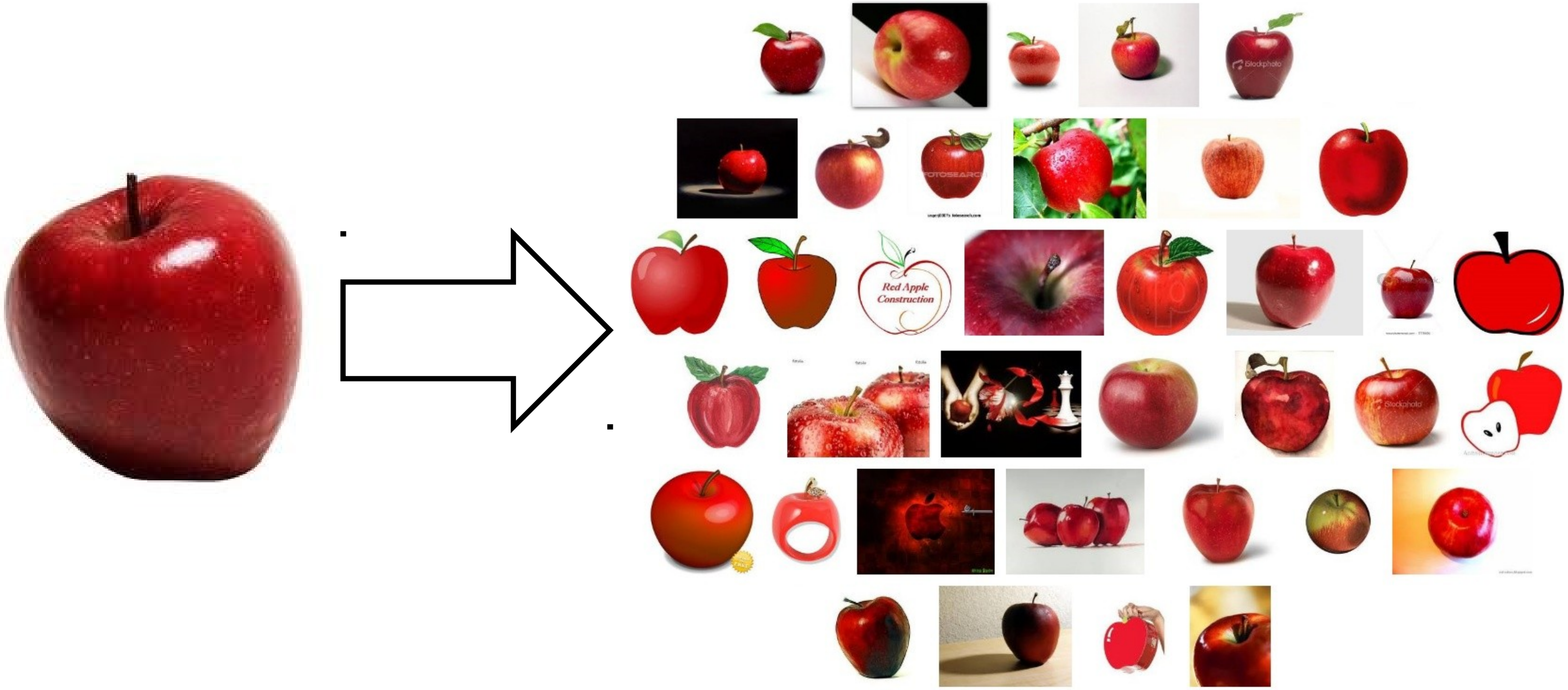
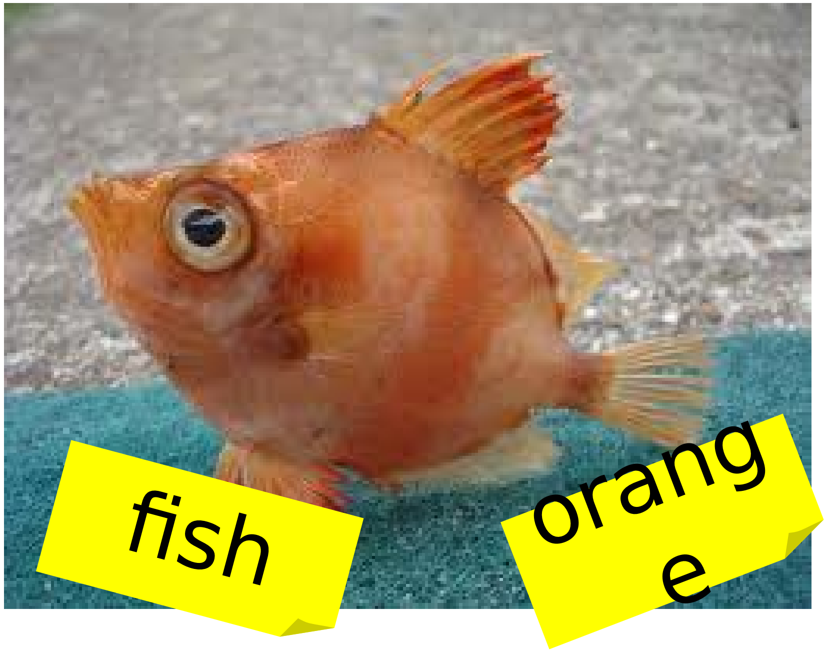


Image retrieval

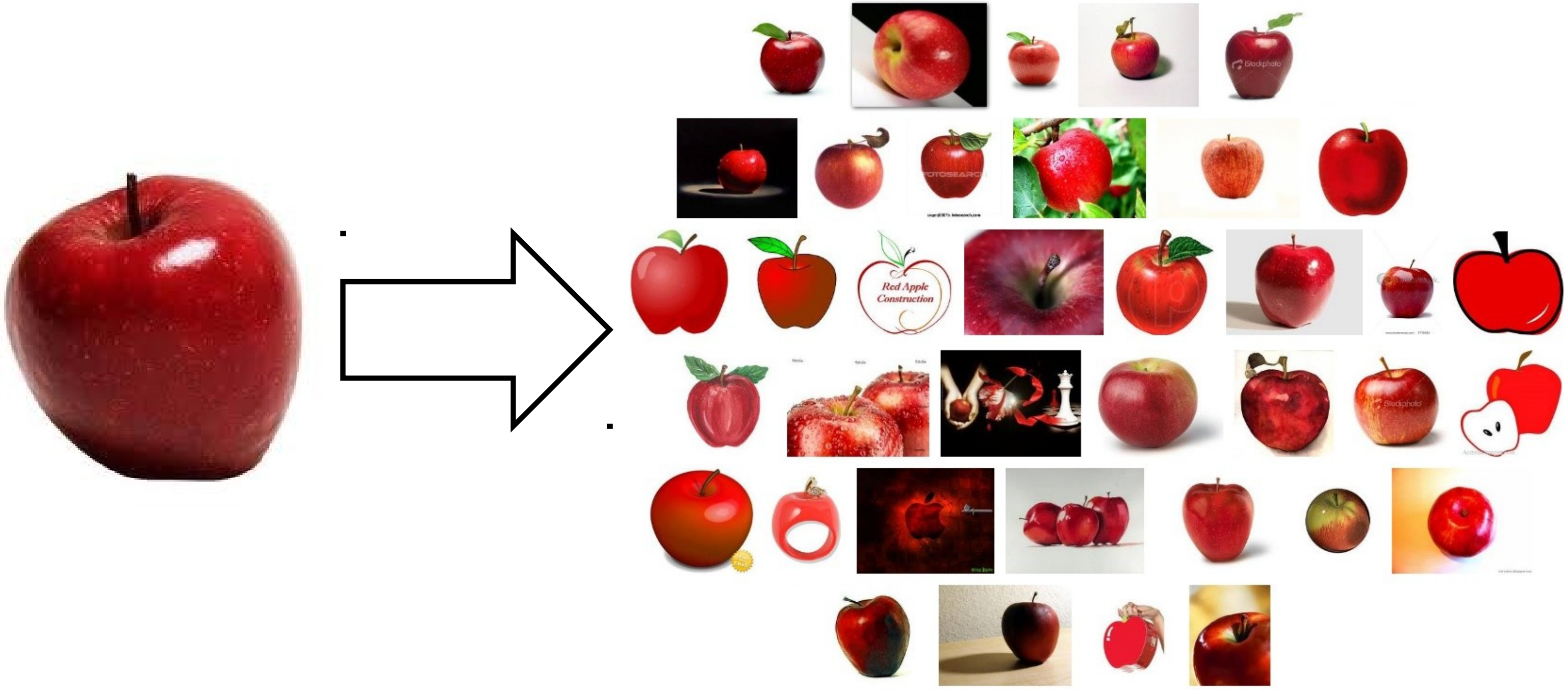
- How do we query and retrieve images?
- Simple solution:
  - Tag images with keywords
  - Search collection using tags => document retrieval



Aston University
Birmingham

Image retrieval

- Entering tags manually is time-consuming process
- Solution: Content Based Image Retrieval
- “Find me images similar to this”



Content-based image retrieval

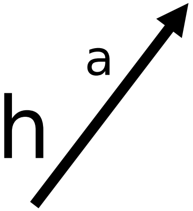
- Searching large image databases by reducing image to a small collection of features e.g.
  - Colour histograms
  - Image shapes
  - Visual words
    - Detect interesting image locations
    - Describe those locations
    - Descriptions become “words” that characterise the image as if it was a document

Vectors

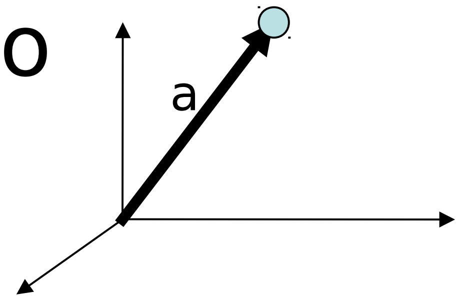
- “1d-array of numbers”
  - co-ordinates

$$a = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}$$

- direction + length



- notation: *a, b, c*,...
- equality: same direction + same length
- vectors can be used to define points.



Vector operations

- Scalar multiplication

$$\lambda a = \lambda \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} \lambda a_1 \\ \lambda a_2 \\ \lambda a_3 \end{bmatrix}$$

- E.g.

$$5 \begin{bmatrix} 2 \\ 3 \\ 4 \end{bmatrix} = \begin{bmatrix} 10 \\ 15 \\ 20 \end{bmatrix}$$

Vector operations (2)

- Vector addition

$$a + b = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} a_1 + b_1 \\ a_2 + b_2 \\ a_3 + b_3 \end{bmatrix}$$

- E.g.

$$\begin{bmatrix} 2 \\ 3 \\ 4 \end{bmatrix} + \begin{bmatrix} -3 \\ 4 \\ 2 \end{bmatrix} = \begin{bmatrix} -1 \\ 7 \\ 6 \end{bmatrix}$$

Vector operations (3)

- Dot product

$$a \cdot b = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} \cdot \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = a_1 b_1 + a_2 b_2 + a_3 b_3$$

- E.g.

$$\begin{bmatrix} 2 \\ 1 \\ 3 \end{bmatrix} \cdot \begin{bmatrix} 2 \\ 5 \\ -3 \end{bmatrix} = 2 \times 2 + 1 \times 5 + 3 \times (-3) = 0$$

Aston University
Birmingham

Vector operations (4)

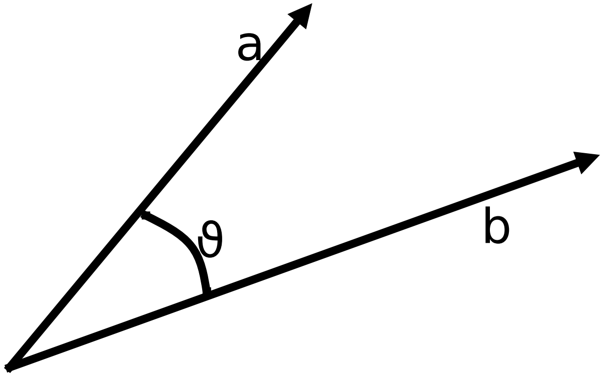
- Length

$$||a|| = \sqrt{a \cdot a} = \sqrt{a_1^2 + a_2^2 + a_3^2}$$

- Unit length vector $\frac{a}{||a||}$

- Angle between two vectors

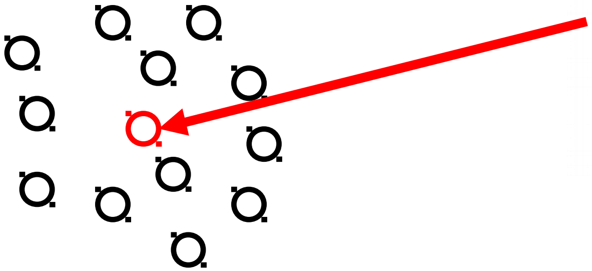
$$\theta = \cos^{-1}\left(\frac{a \cdot b}{||a||\ ||b||}\right)$$



- a·b>0 →θ<90° a·b=0 →θ=90° a·b<0 →θ>90°

Vector centroid

- Calculates the “mean point” for a set of point



$$c = \frac{1}{N} \sum_{i=1}^{N} x_i$$

Vector centroid

- Example N=4

$$x_1=\begin{bmatrix}2\\-3\\1\end{bmatrix},\quad x_2=\begin{bmatrix}1\\2\\0\end{bmatrix}\quad x_3=\begin{bmatrix}-2\\4\\5\end{bmatrix}\quad x_4=\begin{bmatrix}0\\1\\-2\end{bmatrix}$$

$$c=\frac{1}{4}\left(\begin{bmatrix}2\\-3\\1\end{bmatrix}+\begin{bmatrix}1\\2\\0\end{bmatrix}+\begin{bmatrix}-2\\4\\5\end{bmatrix}+\begin{bmatrix}0\\1\\-2\end{bmatrix}\right)=\frac{1}{4}\begin{bmatrix}1\\4\\4\end{bmatrix}=\begin{bmatrix}1/4\\1\\1\end{bmatrix}$$

Aston University
Birmingham

Matrices

- Matrices
  - array of numbers: n x m matrix
    - n rows
    - m columns
  - notation: A,B,C

$A = [a_{ij}]$ e.g. $A = \left[\begin{array}{ccc} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{array} \right]$

  - Vector is special case of matrix
    - nx1 (column vector), 1xn (row vector)

Matrix operations (1)

- Scalar multiplication $\lambda A = \lambda \left[a_{ij} \right] = \left[\lambda a_{ij} \right]$
  e.g.

$$\lambda \left[\begin{array}{ccc} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{array} \right] = \left[\begin{array}{ccc} \lambda a_{11} & \lambda a_{12} & \lambda a_{13} \\ \lambda a_{21} & \lambda a_{22} & \lambda a_{23} \end{array} \right]$$

$$5 \left[\begin{array}{ccc} 1 & 2 & 3 \\ -1 & 5 & 1 \end{array} \right] = \left[\begin{array}{ccc} 5 & 10 & 15 \\ -5 & 25 & 5 \end{array} \right]$$

Aston University
Birmingham

Matrix operations (2)

- Matrix additio $A+B=[a_{ij}]+[b_{ij}]=[a_{ij}+b_{ij}]$ e.g.

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} + \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{bmatrix} = \begin{bmatrix} a_{11}+b_{11} & a_{12}+b_{12} \\ a_{21}+b_{21} & a_{22}+b_{22} \\ a_{31}+b_{31} & a_{32}+b_{32} \end{bmatrix}$$

$$\begin{bmatrix} 1 & 2 & 3 \\ -1 & 5 & 1 \end{bmatrix} + \begin{bmatrix} -1 & 2 & 1 \\ -1 & 3 & 3 \end{bmatrix} = \begin{bmatrix} 0 & 4 & 4 \\ -2 & 8 & 4 \end{bmatrix}$$

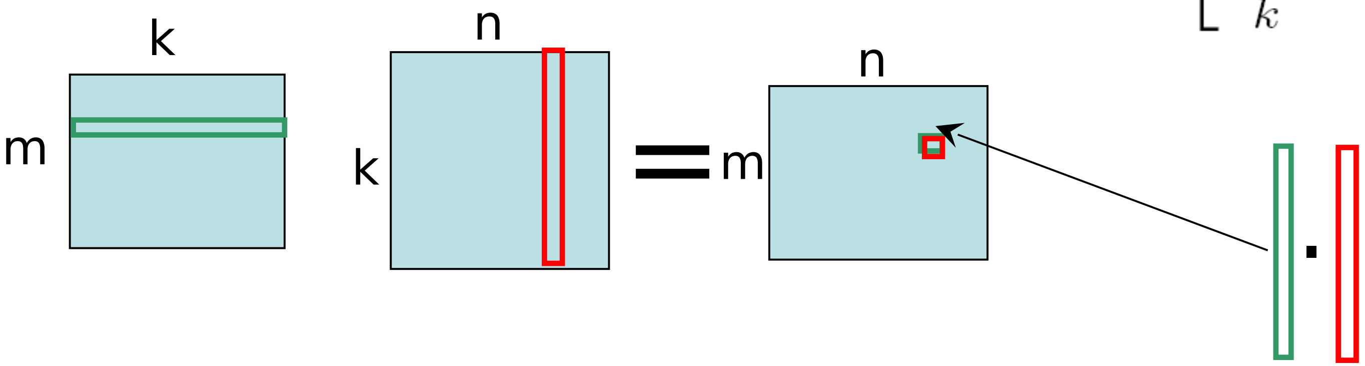
Aston University
Birmingham

Matrix operations (3)

- Matrix multiplication

$$AB = [a_{ij}]\,[b_{ij}] = \left[\sum_k a_{ik} b_{kj}\right]$$

- e.g.



$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{bmatrix} = \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31} & a_{11}b_{12} + a_{12}b_{22} + a_{13}b_{32} \\ a_{21}b_{11} + a_{22}b_{21} + a_{23}b_{31} & a_{21}b_{12} + a_{22}b_{22} + a_{23}b_{32} \end{bmatrix}$$

$$\begin{bmatrix} 1 & -1 \\ -1 & -1 \end{bmatrix} \begin{bmatrix} 0.5 & -0.5 \\ -0.5 & -0.5 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Matrix operations (4)

- Matrix transpose $A^T = [a_{ij}]^T = [a_{ji}]$
  - Swap rows and columns e.g.

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix}^T = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix}$$

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}^T = \begin{bmatrix} 1 & 3 \\ 2 & 4 \end{bmatrix}$$

Matrix operations (5)

- Matrix inverse $AA^{-1} = I$
- I is identity matrix such th $AI = IA = A$
- E.g. $I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

$$\begin{bmatrix} 1 & -1 \\ -1 & -1 \end{bmatrix}^{-1} = \begin{bmatrix} 0.5 & -0.5 \\ -0.5 & -0.5 \end{bmatrix}$$

Note: inverse defined ONLY for square matrices

Note(2): some square matrices do not have inverse e $\begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$

Aston University

Birmingham

What is a good a search engine?

- How fast does it index?
  - E.g. Bytes per hour
- How fast does it search?
  - e.g., latency as a function of queries per second
- What is the cost per query?
  - In £££

What is a good a search engine?

- All the preceding criteria are measurable: we can quantify speed/size/money
- However the key measure is user happiness
- What is user happiness?
- Factors include:
  - Speed of response
  - Size of index
  - Uncluttered UI
  - Most important: relevance
  - (actually, maybe even more important: it’s free)
- None of these is sufficient: could be super-fast but useless answers won’t make a user happy.
- How do we quantify user happiness?

Who is the user?

- Who is the user we are trying to make happy?
- Web search engine: searcher. Success: Searcher finds what she was looking for.
- Web search engine: advertiser. Success: Searcher clicks on ad.
- Ecommerce website: buyer. Success: Buyer buys something.
- Ecommerce website : seller. Success: Seller sells something.
- Enterprise CEO: Success: Employees are more productive (because of effective search).

Who is the user?

- Who is the user we are trying to make happy?
- Web search engine: searcher. Success: Searcher finds what she was looking for. Measure: rate of return to this search engine
- Web search engine: advertiser. Success: Searcher clicks on ad. Measure: click-through rate
- Ecommerce website: buyer. Success: Buyer buys something. Measures: time to purchase, fraction of “conversions” of searchers to buyers
- Ecommerce website : seller. Success: Seller sells something. Measure: profit per item sold
- Enterprise CEO: Success: Employees are more productive (because of effective search). Measure: profit of the company

Relevance

- User happiness is equated with the relevance of search results to the query
- But how do you measure relevance?
- Standard methodology in IR consists of three elements
  - A benchmark document collection
  - A benchmark suite of queries
  - An assessment of the relevance of each query-document pair

Relevance: query vs. Information need

- Relevance to what?
- First take: relevance to the query
- “Relevance to the query” is very problematic.

- Information need i : “I am looking for information on whether drinking red wine is more effective at reducing your risk of heart attacks than white wine.”
  - This is an information need, not a query.

- Query q: [red wine white wine heart attack]
- Consider document d′: *At heart of his speech was an attack on the wine industry lobby for downplaying the role of red and white wine in drunk driving.*
  - d′ is an excellent match for query q . . .
  - d′ is not relevant to the information need i .

Aston University
Birmingham

Relevance: query vs. Information need

- User happiness can only be measured by relevance to an information need, not by relevance to queries

Precision and recall

- Precision (P) is the fraction of retrieved documents that are relevant

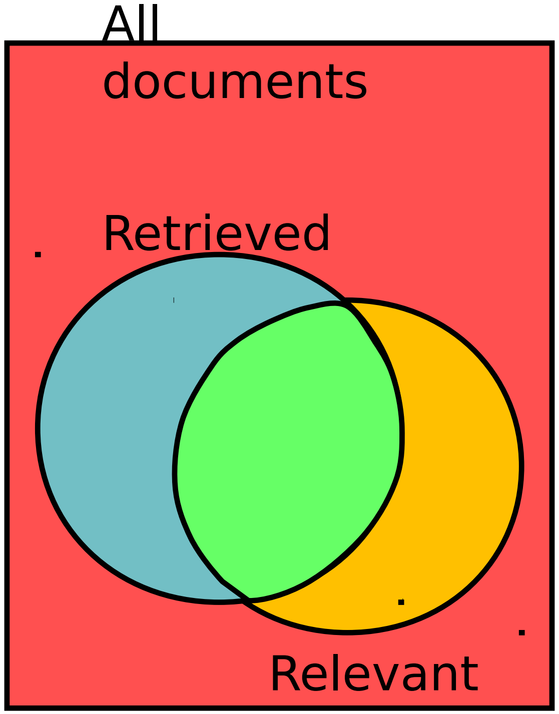
$$\text{Precision} = \frac{\left(\text{relevant documents retrieved}\right)}{\left(\text{retrieved items}\right)} = P\left(relevant \,|\, retrieved\right)¿$$

- Recall (R) is the fraction of relevant documents that are retrieved

$$\text{Recall} = \frac{\left(\text{relevant documents retrieved}\right)}{\left(\text{relevant items}\right)} = P\left(retrieved \,|\, relevant\right)$$

Precision and recall

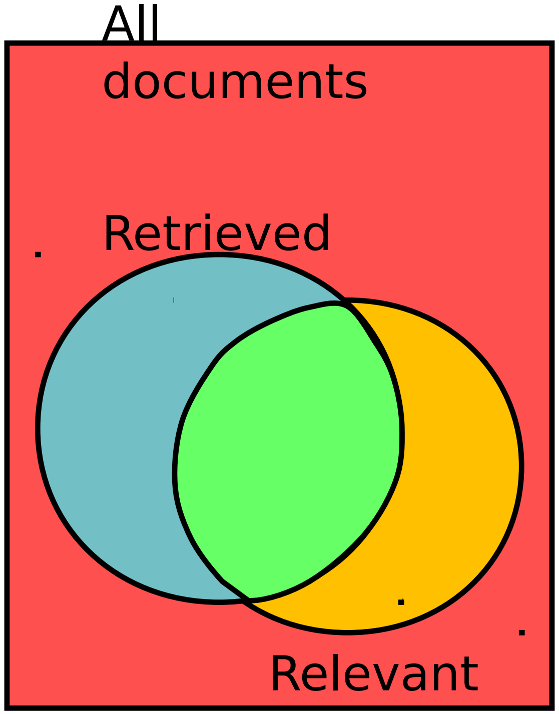
	Relevant	Nonrelevant
Retrieved	tp	fp
Not Retrieved	fn	tn



- Precision P = tp/(tp + fp)
- Recall R = tp/(tp + fn)

Example

	Relevant	Nonrelevant
Retrieved	7	13
Not Retrieved	3	100000



- Precision P = 7/(7 + 13)
- Recall R = 7/(7 + 3)

Precision/recall tradeoff

- You can increase recall by returning more docs.
- Recall is a non-decreasing function of the number of docs retrieved.
- A system that returns all docs has 100% recall!
- The converse is also true (usually): It’s easy to get high precision for very low recall.
- Suppose the document with the largest score is relevant. How can we maximize precision?

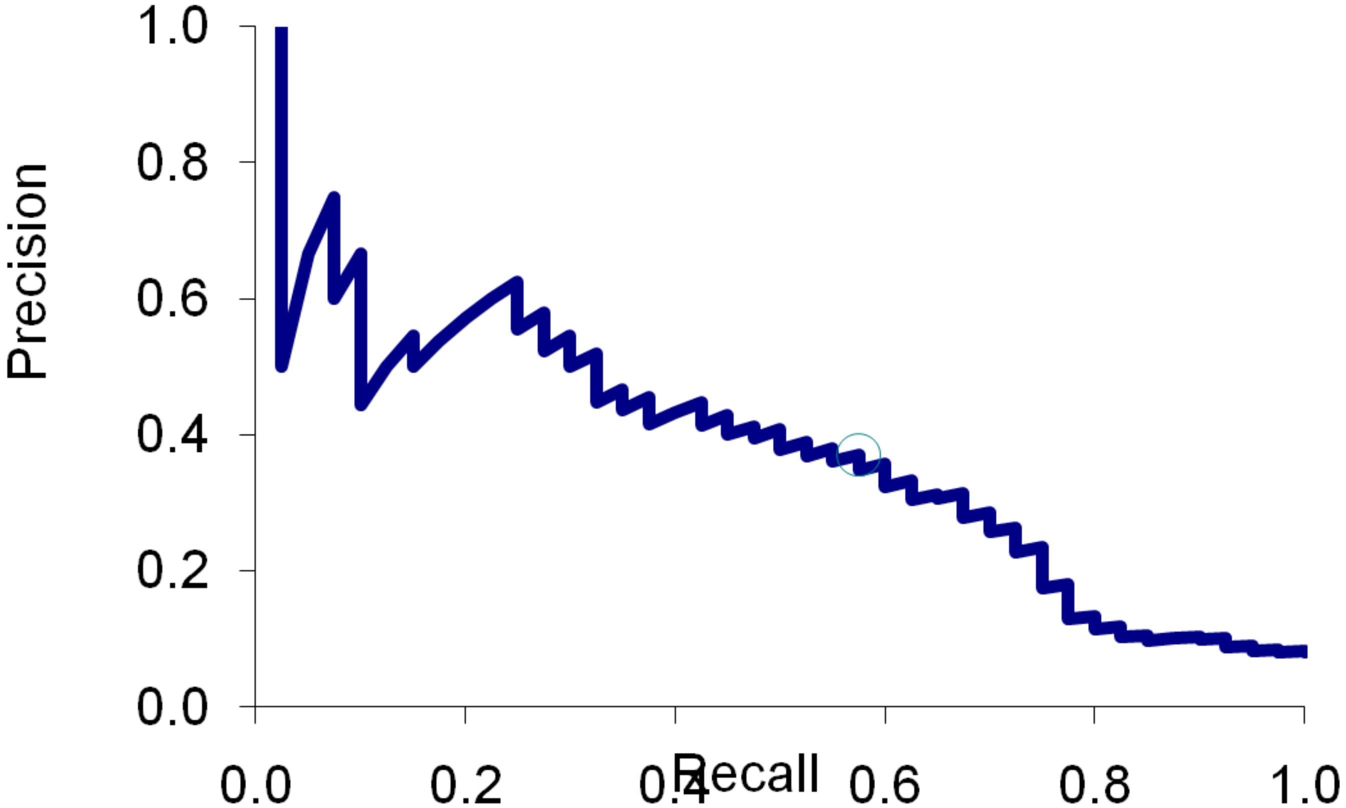
Excercise 1

- An IR system returns eight relevant documents and ten non-relevant documents. There are a total of twenty relevant documents in the collection. What is the precision of the system on this search, and what is its recall?

Evaluating ranked results

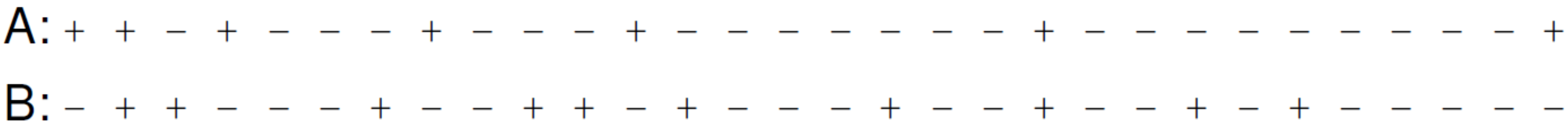
- Evaluation of ranked results:
  - The system can return any number of results
  - By taking various numbers of the top returned documents (levels of recall), the evaluator can produce a *precision-recall curve*

Precision-Recall curve



Exercise 2

- Two retrieval engines A and B index the same document collection. The same query returns the top-30 documents for each retrieval engine. The ranked list is ordered by the relevance to the query. The following listings denote relevant documents with a '+' and non-relevant ones with

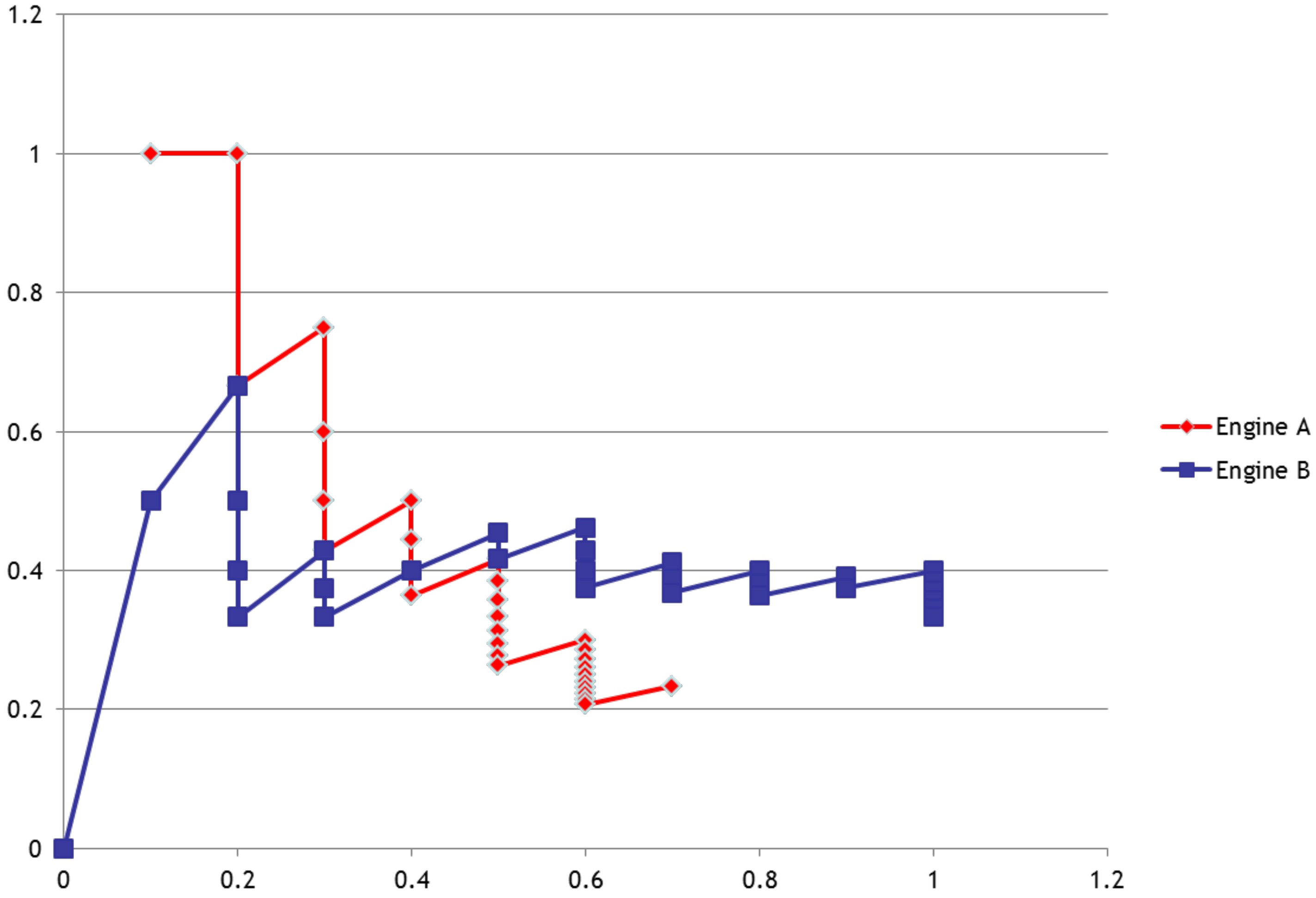


A: + + − + − − − + − − − + − − − − − − − + − − − − − − − − − +

B: − + + − − − + − − + + − + − − − + − − + − − + − + − − − − −

- If the total number of relevant documents is 10, calculate the first 5 points on the precision-recall curve for each of the two search engines.
- Which one is better based on this data?

Precision-recall curve



Aston University
Birmingham

Summary

- Module admin
- How it will be assessed
- Sneak preview of some of the topics
  - Information Retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).
  - What makes a good search engine?
  - Precision/recall
- Next week:
  - Boolean retrieval & some Python