Attendance code 16th January 2020 is

Intelligent Data Analysis 2020

Prof. Martin Russell

UNIVERSITY OF BIRMINGHAM

Course structure 2020

- 11 x 2 hour lectures - Thursday, 11am – 1pm
  - Text retrieval (1) Intro
    - TF-IDF similarity, vectorization of documents
  - Maths revision 1: Vectors and linear algebra
  - Dimension reduction
    - Principal Components Analysis (PCA)
  - Visualization of high-dimensional data
    - PCA, Topographic maps, t-SNE
  - Clustering and vector quantization
  - Text retrieval (2)
    - Synonym relationships, Latent Semantic Analysis (LSA), Page Rank
  - Classification

UNIVERSITY OF BIRMINGHAM

Assessment

- Assessment
  - 1.5 hour exam in May/June - answer 3 questions from 3
- Assignment for extended module

Exercise sheets

- Weekly exercise sheets on Canvas
- Solutions will be published on Canvas
- Not part of formal assessment

Course Canvas page & C code

- All materials will go on Canvas
- Canvas site for 2020 will contain:
  - Copies of all slides and C code
  - Details of assignment - for IDA (extended)
  - Weekly exercise sheets and solutions
  - Pointers to relevant websites
- C code
  - Simple ANSII C implementations of basic techniques from the course. Compile using the MS Visual Studio .NET command line C compiler.

Office hours

- My office hours:
  - Tuesday 2.30-3.30pm
  - Thursday 2.30-3.30pm (until 13/2/2020 – week 5)
  - Friday 2.30-3.30pm (from 21/2/2020 – week 6)
- My office hours will be in my office in the **Gisbert Kapp building, GK-N429**

Moore’s Law and disk capacity

- Moore’s Law
  - *Technology performance doubles and prices halve every 18 months*
- Implications for data storage
  - Applies to disk capacity
  - We have the potential to record and store online a big proportion of what we do.

How much speech fits on 1TB?

- How much speech data can be stored on a 1TB disk?
- Assume:
  - 16kHz sampling rate (16,000) samples per second (Nyquist) $\max N\,Hz \rightarrow 2N\,Hz$ sampling rate
  - 16 bits per sample
- Then:
  - 1 second of speech requires 32,000 bytes
  - 1TB = 3.125×10^7s = 520,833 mins = 8,681hrs = 362 days

Petabytes

- 1 petabyte of disk space costs
  - $2,000,000 in 2003
  - $25,000 in 2019 (based on Amazon, $100 for 4TB!)
- 1 petabyte = 10^{15} bytes
  - 10^6 – (1MB), 10^9 – (1GB), 10^{12} – (1TB)
  - 10^{15} – zillion
  - 1 zillion used to be synonymous with infinity – an unimaginably large number!

A Petabyte is a lot of data…

- 1PB =
  - 20 million 4-drawer filing cabinets filled with text
  - 13.3 years of HD-TV video
- 1.5PB =
  - Combined size of the 10B photos on Facebook
- 20PB =
  - The amount of data processed by Google per day

(A Google search will find many similar examples)

Accessing data – “aboutness”

- Why store these huge corpora?
- Because **information** in them is potentially **useful**
- But, how can we find the **relevant** items?
  - AV recording of a meeting contains more information than conventional minutes, but only useful with good search functions
- Need to know:
  - What each item in a corpus is **about**
  - Relationships between different corpus items query – corpus
  - Relationships between ‘queries’ and corpus items
- Manual indexing impossible - deal with ‘raw’ data.
- Need to determine **automatically** what a text is **about**

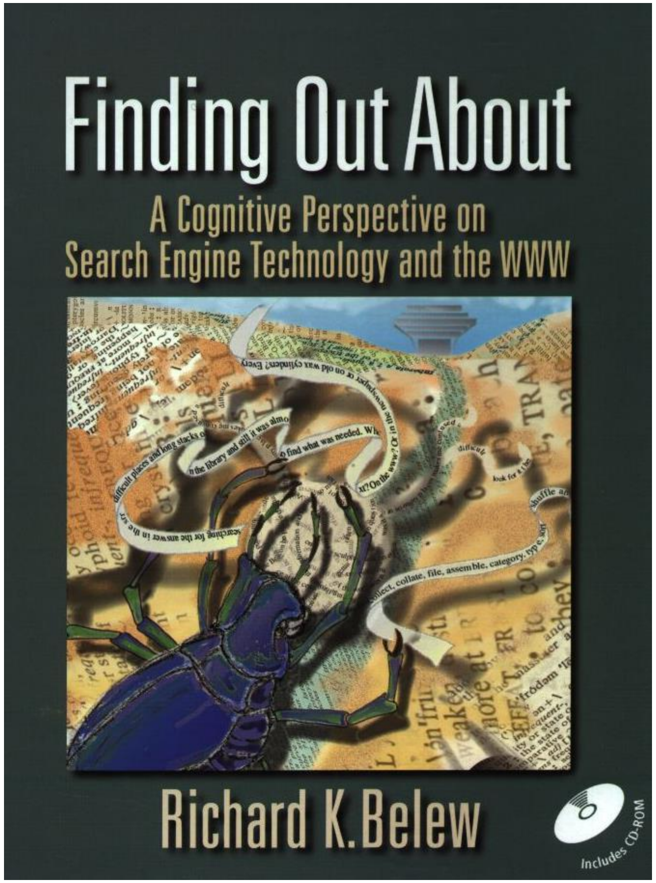
The problem of “Aboutness”

- What is a text, audio signal, or image **about**?
- This is a problem in **semantics**
- This is exactly the type of problem which:
  - Humans are good at, but
  - Computer programmes are particularly bad at!
- For example – “is this image about dogs?”

"Aboutness"

information retrieval

- Richard K Belew
- *Finding Out About: A Cognitive Perspective on Search Engine Technology and the WWW*
- Cambridge University Press, 2001
- Includes CD-ROM & website



When is an image “about” dogs

UNIVERSITY OF BIRMINGHAM

The problem of "aboutness"

- Intuitively, if we focus on text things should be more straightforward
- But even human interpretation of texts may be ambiguous…
- Simple example:

I saw the man on the hill with the telescope

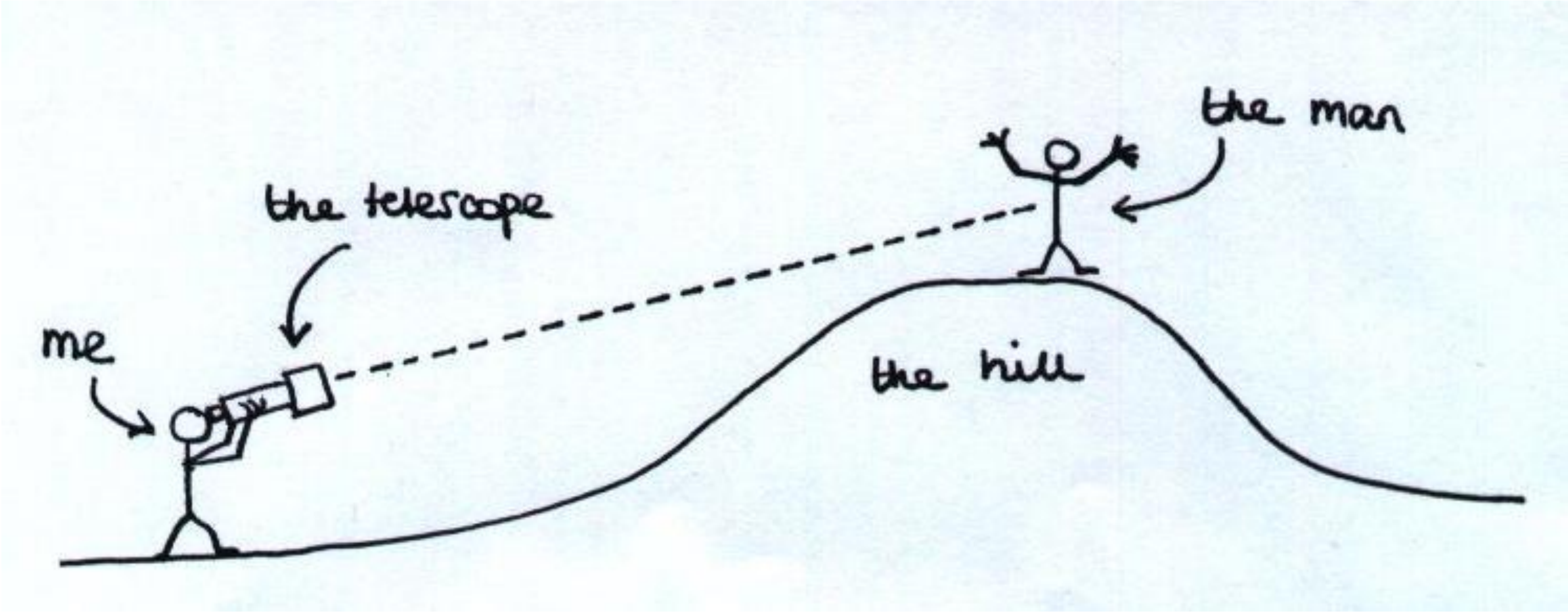
Text Understanding

- How can a machine **understand** what this sentence is about?
- Traditionally this involves:
  - Finding the grammatical role and meaning of each word
  - Analyse – Parsing the word sequence – applying a set of rules to identify the structure of the word sequence relative to a grammar
  - A **grammar** is a model that encodes all of the valid word sequences (sentences) in a language

Text Understanding

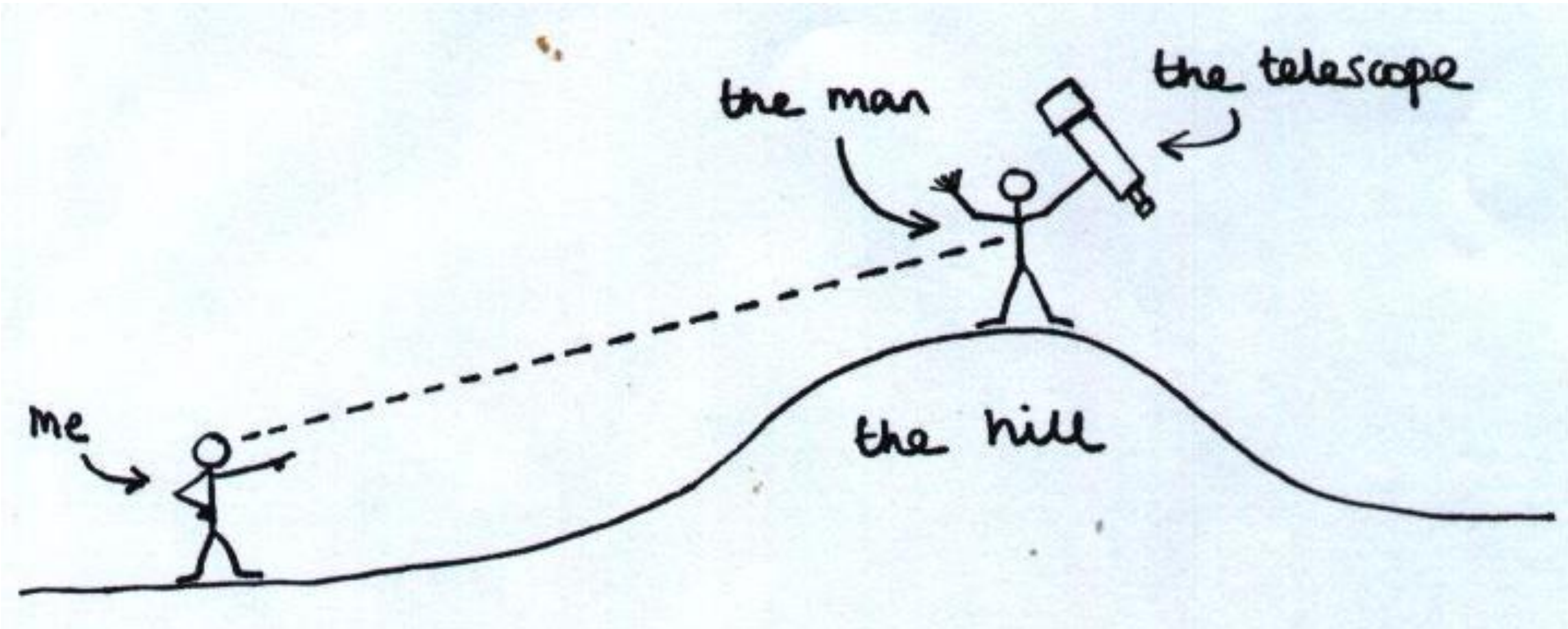
- Words have different meanings and grammatical roles (e.g. “lead” (verb or noun))
- A word sequence may have multiple interpretations relative to the grammar
- A grammatical word sequence may not occur in the given grammar (under generation)
- Conversely, an ungrammatical sentence may be in the grammar (over-generation)

I saw the man on the hill with the telescope



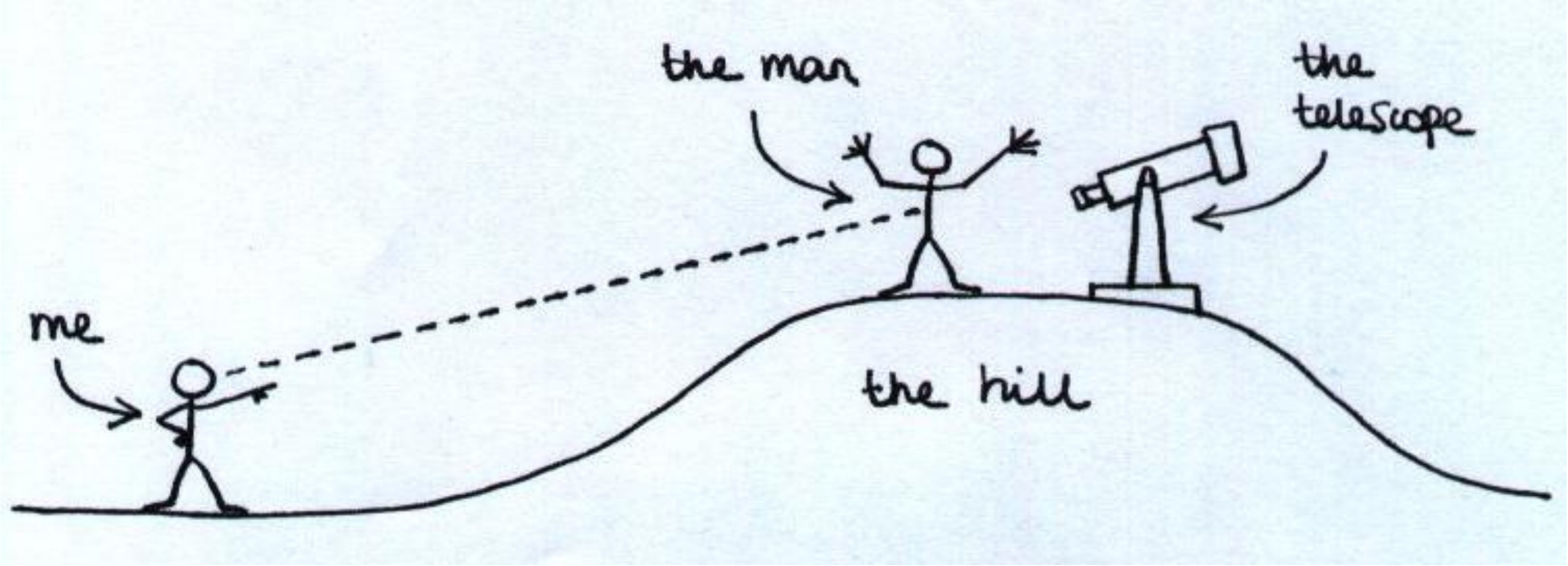
I saw the man on the hill with the telescope

I saw the man on the hill with the telescope



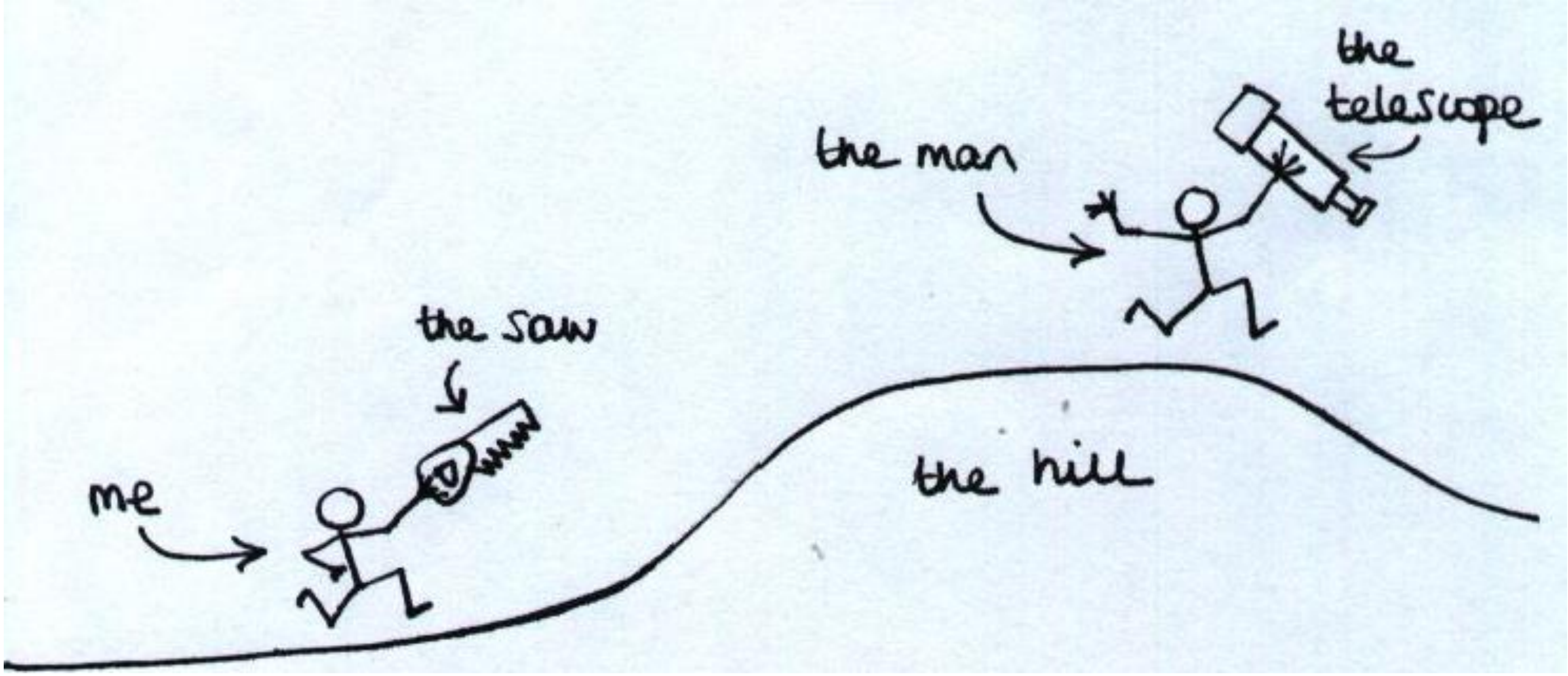
I saw the man on the hill with the telescope

I saw the man on the hill with the telescope



I saw the man on the hill with the telescope

I saw the man on the hill with the telescope



I saw the man on the hill with the telescope

Analysis

- Example illustrates two different problems
  – Different grammatical parses of same word sequence

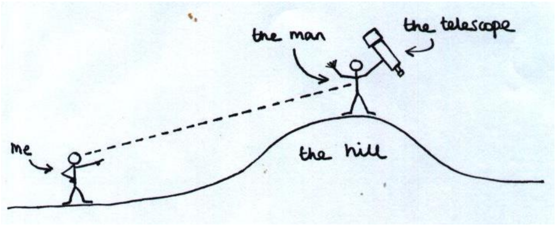
I saw the man on the hill with the telescope

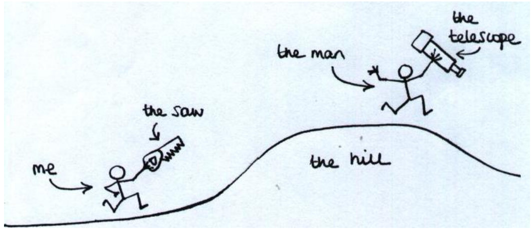
vs

I saw the man on the hill with the telescope

  – Identical parses but different interpretations of words

I saw the man on the hill with the telescope





- Move towards **Machine Learning**

What is Data Mining?

- Mining
  - *Digging deep into the earth, to find hidden, valuable materials*
- Data Mining
  - Analysis of large data corpora: biomedical, acoustic, video, text,… to discover **structure**, **patterns** and **relationships**
  - Corpora too large for human inspection
  - Patterns and structure may be hidden

Related "hot" topics

- "Big Data"
- Pattern recognition/processing
  - As a prerequisite for Data Mining (e.g. ASR for spoken data retrieval)
  - As a consequence of Data Mining
- Machine learning
  - (Deep) Neural Networks, "Deep Learning"
- Data Visualization
  - Dimension reduction

Some example data

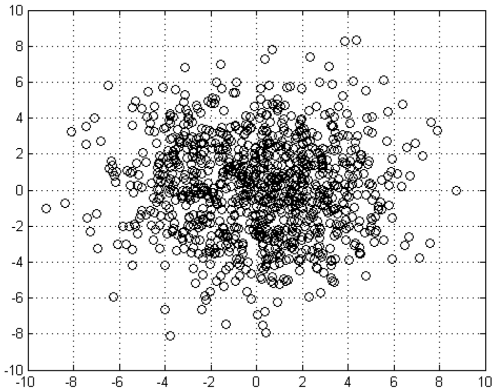


Fig 1: Single, spherical cluster centred at origin.

Correlation X

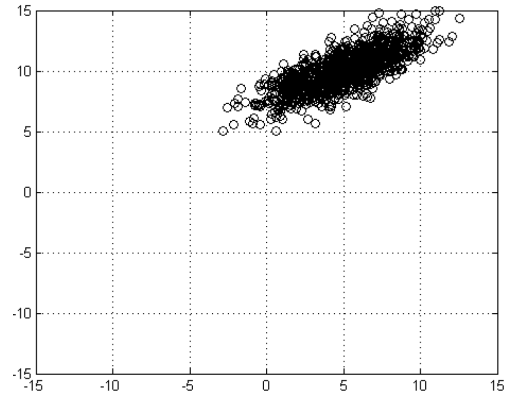


Fig 2: Single, arbitrary elliptical cluster

Correlation ✓

diff. variance

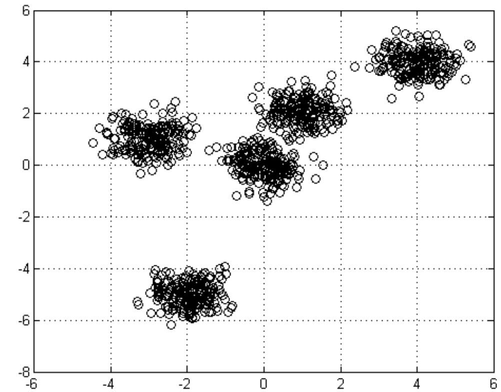


Fig 3: Multiple, arbitrary elliptical clusters

What is Information Retrieval (IR)?

- Underlying principles of Search Engine technology
- Finding out About… [Belew]
- Retrieving Information from text sources
- Retrieving Information from other sources
  - Spoken Data Retrieval
  - Bio-informatics
- In IDA we will focus on **text retrieval**

IR vs Database Retrieval

- IR is <u>not</u> 'database retrieval'
- Databases are characterised by:
  - Strong prior assumptions about
    - Salient properties of data
    - Format
    - Logical relations between data items
    - Likely user queries
  - Formal, restrictive query syntax
  - Need for dedicated maintenance to keep it up-to-date
  - Gives specific replies to specific queries


Expedia.co.uk Travel > Book cheap flights, travel, hotels & package holidays - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.expedia.co.uk/Default.asp?CCheck=1
Go Links
Expedia.co.uk
Expedia Sale! Family trips from £405 per family
home deals flights & charters hotels cars holidays cruises guides maps insurance customer support
SALE
Site Map | My Trips | My Profile
Welcome to Expedia.
Sign in - Sign up - It's Free!
London Hotels from £52
SALE
Fun Breaks from £121
TRAVELLER TOOLS
Airport Guides
Arrivals/Departures
Flight Timetables
Currency Converter
World Guide
Weather
BUILD YOUR PERFECT TRIP
Flight only
Hotel only
Car only
Flight + Hotel
Flight + Hotel + Car
Flight + Car
Book together and save!
Tell me more
Departing from:
Birmingham
Depart:
5/2/2004
Evening
Going to:
Edinburgh
Return:
8/2/2004
Midday
Adults: (12-64)
Seniors: (65+)
Children: (2-11)
Infants: (under 2)
2
0
0
0
Search
More flight search options: Additional airports, multiple destinations...
SALE
Bag your perfect trip...
City Breaks
Flight + Hotel from £99
Including: New York, Prague, Rome...
Luxury Breaks
Flight + Hotel from £231
Including: Los Angeles, Miami, St. Lucia...
Fun Breaks
WELCOME TO Fabulous LAS VEGAS NEVADA
Flight + Hotel from £121
Including: Amsterdam, Barcelona, Las Vegas...
Family Trips
Relaxing Getaways
Ski & Snow
HOTEL DEALS from
Las Vegas £13
London £52
New York £49
Orlando £38
Paris £37
Rome £51
Amsterdam £44
UK & Ireland £34
More hotel deals...
FLIGHT+HOTEL from
Shopping Breaks £111
Caribbean Deals £456
Last Minute Deals £97
Regional Departures £96
Fly-Drive Deals £124
Florida Holidays £283
More flight+hotel deals
FLIGHT DEALS from
Barcelona £81
New York £184
Orlando £224
More flight deals...
CAR DEALS
Price per day from
Portugal £10
Florida £17
Internet
Start
C:\My Documents\Publica...
Expedia.co.uk Travel ...
Microsoft PowerPoint - [O...
EN
12:52

IR vs Database Retrieval

- IR (Finding Out About)
  - No prior assumptions about:
    - Salient properties of data
    - Format of data
    - Logical relations between data items
  - Less specific ‘natural language’ queries
  - Source information remains up-to-date
  - Much less focussed replies

Relevant topics in mathematics

- Vectors and matrices
  - Data that we analyse is generally vector data
  - A single data point may comprise multiple measurements
  - Words or documents typically represented as vectors
  - A basic understanding of the mathematics of vectors (linear algebra) is crucial for intelligent data analysis and text retrieval
- Probability
- Next lecture – linear algebra revision

Summary

- Introduction to course components
  - Background mathematics
  - Data visualization and data mining
  - Information retrieval
- Motivation
  - Availability of huge corpora of raw data
- Problems
  - Aboutness