

# A New Look at Old Tricks: Fertile Roots of Current Research

Paul B. Kantor  
Professor, LIS/SC&I;  
Computer Science & Operations Research  
Rutgers University

# I1. My perspective

- Our\* problem is to connect two different information systems: the one that we can build, and the one in the user's brain – what is the best “impedance match”.
- How do we do it? we expend great effort on **representing** the information items (“**documents**”); considerably less on representing the user's **need**, propose algorithms to compute a **similarity**, and use computationally tractable **heuristics** to effect the proposed “matching”.
- Wherever possible, we should keep clearly in mind the triad: (representation, similarity, heuristic) rather than bundling all three together under the “trade name” of the software package or the mother company.

Original material © Paul Kantor. 2012

\*Formulation by IJ Good

# I.2. My Apology

- In the ‘good old days of SIGIR’
  - talks were about theory
  - Never burdened by actual results
- in some mature disciplines (chemistry, physics) there is an acceptable division between theory and experimentation
- i used to be a theoretical physicist
- therefore.....*this talk will have no experimental results.* 😊
- *Is Information Retrieval a dogma? A science? An **alchemy**?*

# 13. Outline\*

- 1. The earliest “automated retrieval”
- 2. Vectors and Logarithms: their pragmatic origins
- 3. Probabilistic approaches. A frequentist foundation
- 4. The quest for a theoretical foundation:
- 5. Generative approaches: Language models and topic modeling
- 6. Network approaches
- 7. Binding approaches together

# I.4. Midway between facts and theory

- “This course deals with the facts...[someone else] teaches the theory. My only apology is that next year the facts will be the same, but the theories will be different..”

» *Sam B. Treiman*

- *In IR, a variety of theories, sit behind a factual array of computations -- in many cases the computations are alike, although theories are different*

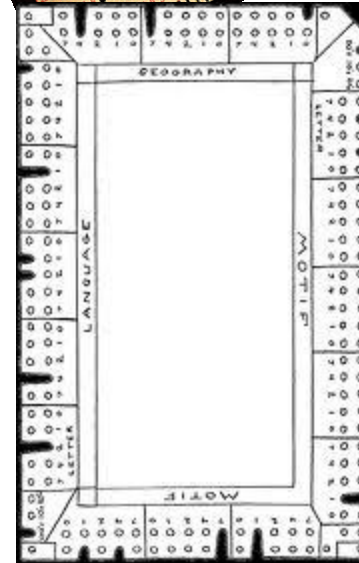
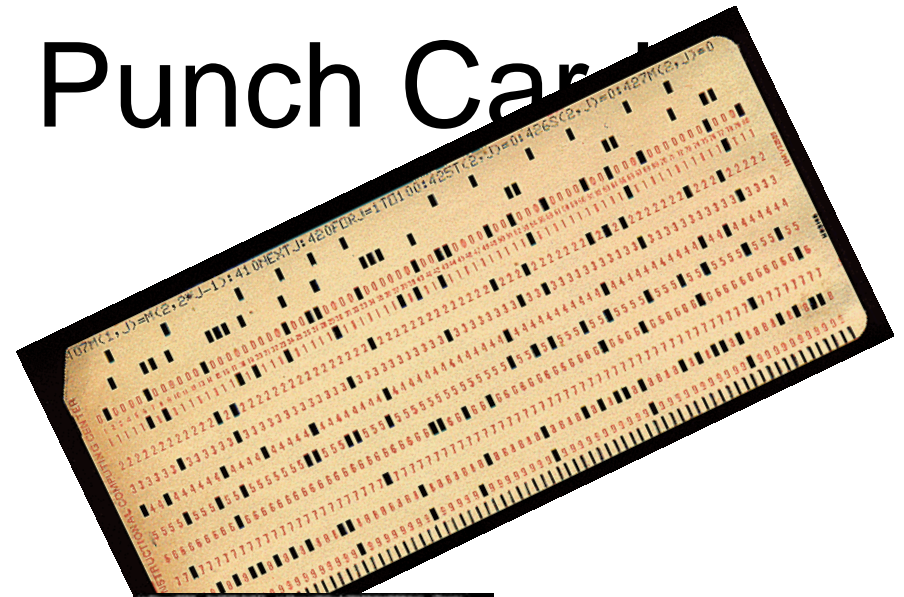
» *His student* Original material © Paul Kantor. 2012

# I.5. Alchemy

- from Medieval Latin: alkimia, from Arabic al-kimiya, from Gk. khemeioa (found c.300 C.E. in a decree of Diocletian against "the old writings of the Egyptians"), all meaning "alchemy." Perhaps from an old name for Egypt (Khemia, lit. "land of black earth," found in Plutarch), or from Gk. khymatos "that which is poured out," from khein "to pour," related to khymos "juice, sap" [Klein, citing W. Muss-Arnolt, calls this folk etymology].
  - » <http://etymonline.com/?term=alchemy>
- "since c.1600 the word has been applied distinctively to the pursuit of the transmutation of baser metals into gold, ..., along with the search for the universal solvent and the panacea\*"
  - » Principe, Lawrence. News report by Reardon, Sara. *The Alchemical Revolution* Science 20 May 2011: Vol. 332 no. 6032 pp. 914-915 DOI: 10.1126/science.332.6032.914

*There are definitely some similarities with the claims made for IR today. And, like the alchemists, our only tools are empirical observation, and the "self-evident" beliefs or axioms handed down to us by our predecessors.*

# A.1. Keysort; Punch Cards



# A.2 The earliest “automated retrieval”

- A. Origins in the American Chemical Society research
  - Encode subject matter// key terms // punch card sorting //
  - A McBee keysort card with N holes can encode  $2^N$  distinct subjects\*; or with 2 rows,  $3^N$
  - Librarians think of subjects as fitting into an outline structure
- B. Kent experiments; Center for Documentation and Communication Research; Western Reserve University (Library School) 1955. Correspondence with Cleverdon – the Cranfield experiments. Limited contact with Salton.
- An Approach to Automated Vocabulary Control in Indexes of Organic Compounds, II; Charles H. Davis
  - J. Chem. Doc., 1969, 9 (4), pp 252–256; DOI: 10.1021/c160035a017; November 1969
- C. Sets and modifications
  - (i) Boolean combinations: A AND (B OR C) AND (D OR E) [CNF]
  - (ii) Quorum type rules; [having more terms is better] ABE > CDE
- D. Key word selection; discriminatory words
- E. *H. P. Luhn; ranking texts*
  - (i) the origins of *tf*
  - (ii) the origins of *idf*
- F. Ranking and retrieval; Luhn 1959



Syn Set

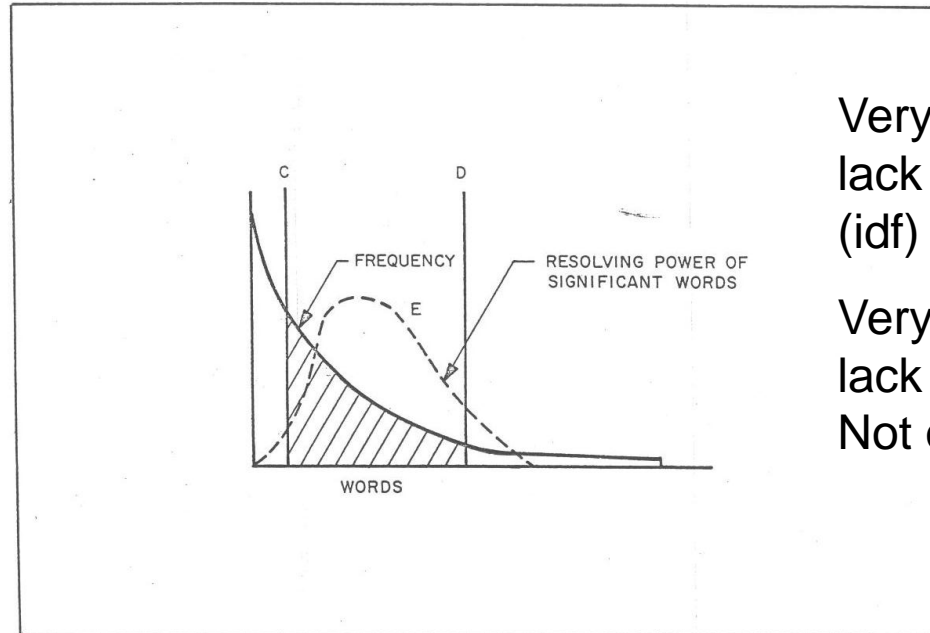
Original material © Paul Kantor. 2012

\* In Jorge Luis Borges' Library of Babel, how long is a book's call number?

# A.3. The first commercial system

- 1967 The online information retrieval system DIALOG is created by Roger Summit.
- 1972 Dialog becomes the world's first commercial online service.
- 1981 Dialog Information Services, Inc. becomes a subsidiary of Lockheed Corp.
- 1988 Dialog Information Services, Inc. is acquired by Knight-Ridder, Inc from Lockheed for \$353 million.
- 1993 DataStar is acquired by Knight-Ridder, Inc. from Radio Suisse for \$36.3 million.

# V.0. tf.idf (From Luhn, some years earlier 1959)



Very common words  
lack resolving power  
(idf)

Very rare words also  
lack it (misspellings?  
Not on topic)

Article: Robert K. Plumb, "Experiments Suggest a New Approach to the Treatment of Heart Attacks," The New York Times, September 22, 1957.

Auto Abstract: The result is a lead, at least, toward the discovery of compounds that will act like female hormones in lowering the blood cholesterol levels in ailing male heart-attack patients without the feminizing side effects.

# V.1. Vectors and Logarithms: their pragmatic origins

- A. Weighting the Importance of Features
- B. Controlling for document size
  - (i) Salton and the ray space approach
  - (ii) Fox's metrics
  - (iii) "Pivoting and Singhal"
- C. The unreasonable introduction of logarithms
- D. Is there a "vector space of documents and queries"
  - (i) Queries as dual spaces
  - (ii) Non Euclidean metrics
- E. Patterns in spaces
  - (i) Linear subspaces
  - (ii) Manifolds
  - (iii) Relative densities – relation to sensor problems

# V.5.1 Question old assumptions: What kind of a vector space

- The components are real numbers (“over the reals”)
  - should it be over the complex?
    - Van Rijsbergen; case not yet made
- is it a linear vector space?
  - Salton -- by default
  - permits linear combination
  - can be equipped with inner product

## V.5.2. Linear VS - more

- Has a very nice property: the space of linear functions on the space is also a LVS
- so we can ‘represent’ queries and documents in “the same” space\*
- every linear function can be written as an inner product  $(f, d)$  for some (dual) vector  $f$
- so the problem is to map each query into a dual vector.

## V.5.3. Actual computations

- Query  $\rightarrow$  a function  $Q$
- Document  $\rightarrow$  a variable:  $d$
- “Score” *a real number assigned to the doc*

$$Q : D \longrightarrow \mathbf{R}$$

- *or any other ordered set. Not so good if*

$$Q : D \longrightarrow \mathbf{R}^n$$

# V.0.1 Vectors: sets of labeled features

- All the work I know of represents a document in terms of numerical or categorical features
- could be summarized as a set of pairs {label:value}.
- When the values are real numbers, mathematicians call such a set a “vector over the reals”.

# V.0.2 Willful suspension of disbelief

- Usually we deal with sets of vectors that form a 'linear vector space'. This imposes a lot of conditions such as: if  $v, v'$  are vectors, then  $v+3v'$  is also a vector; and so is  $-v$  a vector.
- So we don't really believe that documents naturally form a linear vector space. We agree to pretend.

# V.0.3 Linearity of probability

- Also, we often pretend that the probability function would be linear in  $d$  if we could just find the right representation.
- This means  $p(q, d)$  is linear in  $d$ . But any linear function on  $D$ ,
  - if  $D$  is a linear vector space !
- can be written as an inner product  $(\textit{something}(q), d) = \langle \textit{something} | d \rangle$  (*Dirac*).

## V.0.4. Two notations

- In mathematics we often write  $(x,y)$  for the inner product of the vectors  $x,y$ .
- In quantum mechanics Paul Dirac introduced the notation  $|x\rangle$  for a “state vector  $x$ ” and called it a “ket”. The state space is a Hilbert space, and the dual vector is called a “bra”  $\langle x|$ ; the inner product  $(x,y)$  then becomes the ‘bra(c)ket’  $\langle x|y\rangle$ .

# V.0.5 The dual space

- For finite dimensional (or Hilbert) spaces there is a natural map of the space of *some things representing queries* to the space of vectors such as  $d$ . So we could say that *something*( $q$ ) is the (dual) vector representing the query  $q$ .
- if *something* is also linear in  $q$ , we have the general representation for a linear rule

$$p(q, d) = (Wq, d) = (q, W^T d)$$

## Q.8. Word- or term- based *diagonal* models

- *Consider W. If* the features are indexed by words (stemmed?) and  $W$  is diagonal, we have the general family of functions based on term frequencies: *note indices same:  $t$*

$$p(q, d) = \sum_t \phi(q_t) W_{tt} \eta(d_t)$$

- *truth in advertising. We allow monotone transforms of  $p$  to get actual probability*

# Q.9. Off diagonal terms

- But we know that terms may be “related to” each other,
- And so there are many ingenious schemes
  - (LSI,
  - spreading activation,
  - PCA, ...)
- for computing  $W$  in a way that is not diagonal.

# Q.10. Corpus or bi-corpus

- LSI uses only information about the documents
- If we can assemble enough queries, we might reason about relations among queries, to better infer relations among the documents themselves
  - A. Using only the features
  - B. Using user behavior

# V.0.9. Weakness of linearity



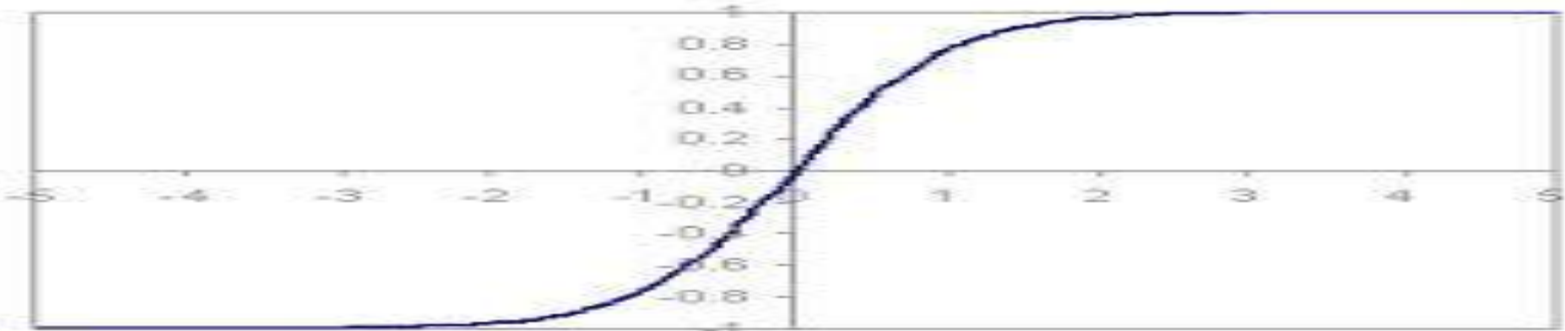
- Linearity has implications
  - For example: linearity means that if
$$d'' \text{ is near } \alpha d + (1 - \alpha)d'$$
    - then the value of  $p$  will be the corresponding linear combinations of values for  $d, d'$
    - this rules out queries whose answers fall into separate (green) islands as shown above
    - Linear mixing of scores cannot solve this problem because linear combinations of linear functions are still linear.

## V.5.4. The static view

- The static view: user unaffected by what he sees
  - There are documents  $D=\{d..\}$  where  $D$  is called the corpus. Might be
    - company files;
    - the web;
    - eavesdropped messages,
    - etc.
- We will explore this first, and then question it.

# V.2. Weighting importance

- Salton: Concepts  $\rightarrow$  words
- Importance  $\rightarrow$  frequency of occurrence; or relative frequency; or “ray space” with all vectors normalized (Euclidean – but why? – simply the form best known to researchers at that time. Maybe another (e.g. Fox) will be better\*.
- Logarithms: Another way of controlling large numbers. Again, simply the slowly growing function best known to researchers at that time. One might, for example, have argued for the hyperbolic tangent, or the logistic function, which are comfortably bounded. What would it mean that a word is “negative in a document”?



\*Beware of a trap for empirical investigations. It is almost surely true that if we take any parameter of a model, and let it vary, *for any given corpus* there will be some value that is better than where we started. It becomes interesting if one value gives improvement for all, or nearly all, corpora

# P.1. Probabilistic approaches. A frequentist foundation

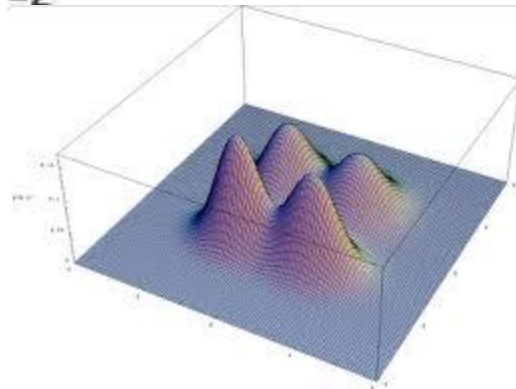
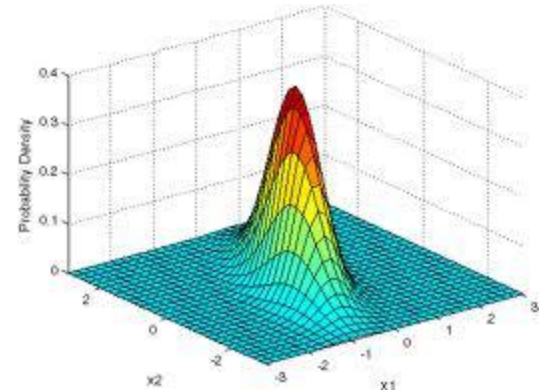
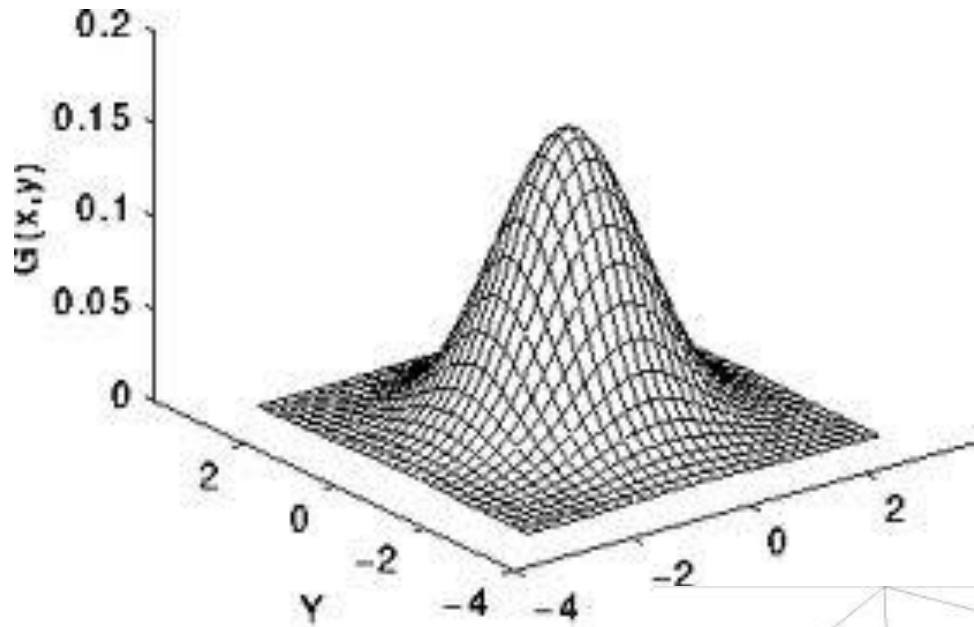
- A. Maron and Kuhns
- B. One user among many – the “system orientation”
  - $\text{Prob}\{d \text{ is good for } q\} \sim N(\text{can use it}|q)/N(\text{all giving } q)$
  - Note: we are pretending that value inheres in a document standing alone. We shall return.
- C. One document and one user – subjective probability
  - Somewhat metaphysical ~ usually treated as above

# P.2. Machine Learning view: “semantic question”

- D. One user and a feature-based **class** of documents
  - (i) The “cluster hypothesis”
    - Documents that are near each other “share relevance” no matter what the queries.
  - (ii) The “smoothness of relevance”
- E. Research in Pattern Recognition
  - Characterize the points in a space in smooth ways (because “objects of interest have physical extent”)\*
  - (i) Estimating densities empirically / Parzen-Rosenblatt (\*earlier!). Also de la Valle-Poussin. Cubic approximation to the Gaussian

\*Does this have to be true in an abstract space, for a conceptual “object?” What about for a quality? (such as “clarity”, “irony” ...?)

# P.3. Everybody's favorite parametrized distribution: Lagrangean, or Gaussian; Or a mixture.



# P.4. The quest for a theoretical foundation

- A. I.J. Good\* and the weight of evidence. The “*bin*”
  - (i) odds of relevance:
    - $\text{Prob}\{\text{useful} \mid \text{features}\} / \text{Prob}\{\text{not u.} \mid \text{features}\}$
  - (ii) the relevance of log-odds
    - Using logs makes independent pieces of evidence *additive*
- B. Robertson and Sparck-Jones
  - explaining” term weighting
  - estimating parameters from data
- C. Naïve but successful
  - (i) Is there a “deep reason”? Robertson believes not
  - (ii) Robertson and “BM25”.
- D. Hierarchical models .....

Original material © Paul Kantor, 2012

\* We will have more to say about the astounding Dr. Good as the notes progress

# V.\*. What of George Boole?

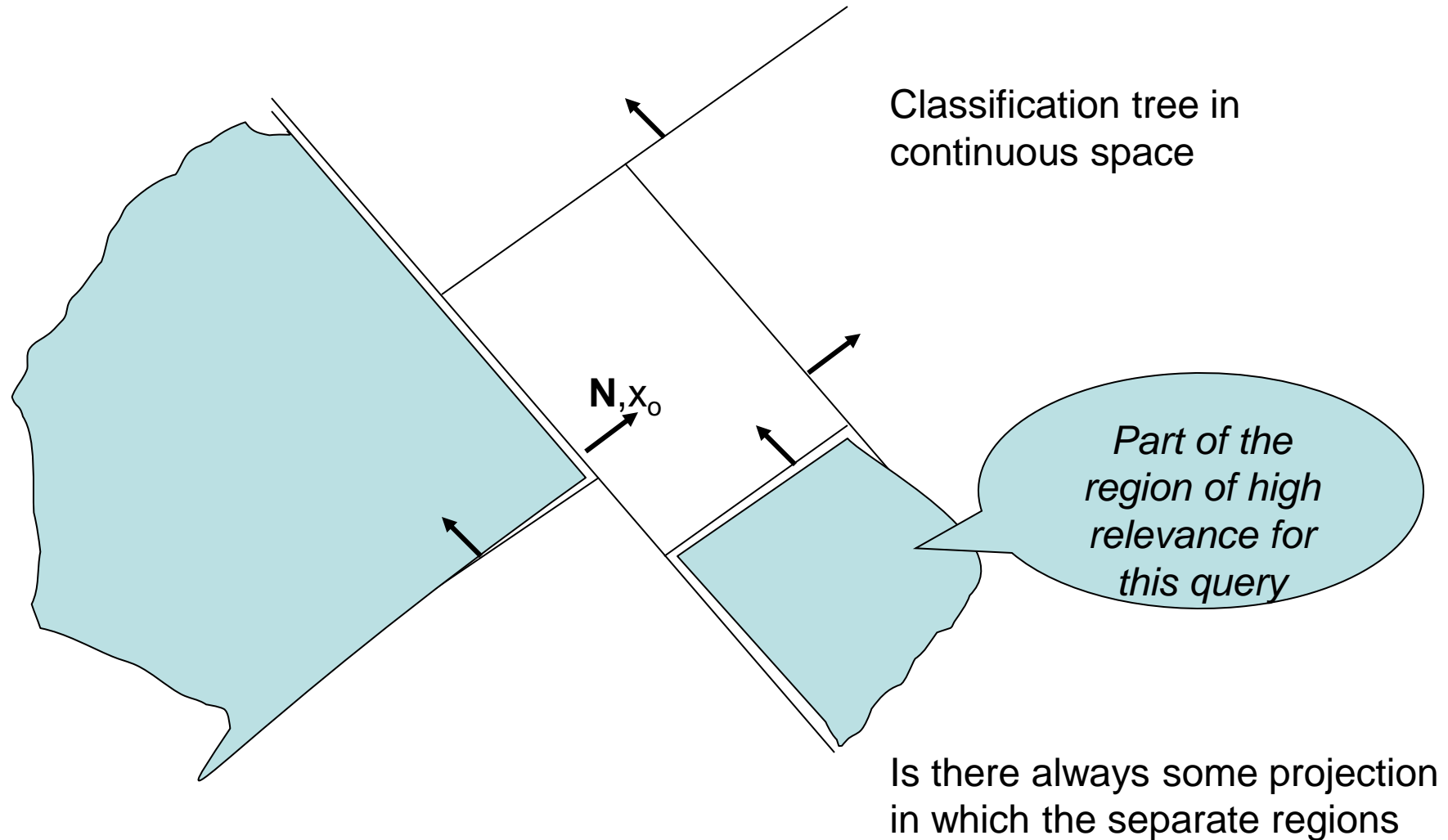
- In the early days (e.g. DIALOG) and still, under the hood at YFSC\* there is an initial Boolean kind of search. And we are permitted to say “not ‘fish’ ”. It seems intuitively clear that this provides a finer lattice for retrieval than can a separating hyperplane.
- Can the vector  $-\mathbf{d}$  be thought to represent negation?

Original material © Paul Kantor, 2012  
\*YFSC = your favorite search company; possible the one that employs you

## V.3. Ways to be not really linear....

- Non-linear features can do it (SVM) – but we must be careful about size of space
- Non –linear combinations rules can do it (Boolean; continuous Boolean; fuzzy Boolean)
- Progress is made, and the web gives us new features (link based)
- And users are ‘satisfied’ .....But *can we do more?*

# V.4. Subdividing space



# P.5. The real curse of dimensionality

- To characterize *regions* we might look at neighbors.
- To see why the problem is tough, ask how many points there have to be in a unit hypercube of dimension  $d$  so that we are guaranteed that every test point is within a distance  $\varepsilon$  of one of them
  - That is, each test point actually has some very near neighbors
- what is the maximum volume that  $K$  hyperballs of radius  $\varepsilon$  can occupy, if they have *no overlap at all*?

## P.6. The geometry is challenging

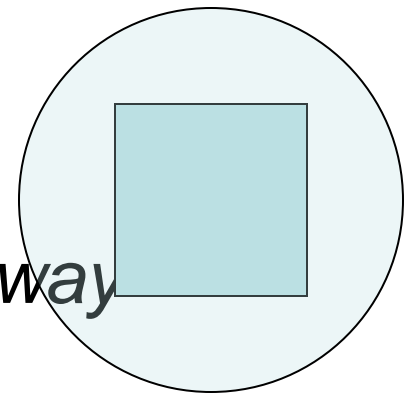
- We know that the answer will be  $O(K\varepsilon^d)$
- We did not anticipate that the constant itself would be important
- It turns out that the constant is inversely related to  $d/2$ -factorial. For  $d$  even this is:

$$V(B(\varepsilon)) = \frac{\pi^{d/2}}{(d/2)!} \varepsilon^d$$

Original material © Paul Kantor. 2012

# P.7. For even moderate dimensions

- $d=100$
- And for  $\varepsilon=1$
- $V(B(\varepsilon))=2.1 \times 10^{-40}$
- *So it would require  $O(5 \times 10^{38})$  of these Balls to fill less than half of the volume.*
- *Not at all like  $d=2$*
- *Therefore, very likely that*
  - *Nearest Euclidean neighbor is far away*



# P.8. Smoothness

- We hope that we can learn about  $p$  by observations. If  $p$  is a completely unsmooth function, this would not be possible. And, as we have just seen, it may not have any labeled points nearby.
- So we assume that it is, in some sense, smooth as a function of position of the point, in the space of dimension  $d$ .

# P.9. Smoothness requires a metric

- Can we really imagine that documents are of such a nature that one may believe that some pairs are closer than others.
- It's very easy to believe this about documents (although it might well depend on the reader)
- For example:
  - “I like Romeo and Juliet” in English *and*
  - “ 我喜欢罗密欧与朱丽叶 ” in Chinese
- are entirely different for me, but rather close to each other, for many at this workshop. [I hope]

# P.10 We will agree representations admit a metric

- Let's not say that *documents* are close, but lets talk about the *representations* of them.
- So *in some settings* (I hate to be saying this) we can say that *d and d'* are close to each other.
- This is described mathematically by a metric, which we might write as  $\rho(d, d')$ .

# L.1. Generative approaches: Language models and topic modeling

- A. A new layer of formalism
- *B. hidden relations to the older views*
- *C. characterizing a literature, a topic or an interest*
- *D. Conjugate distributions and hyperparameters*
- *E. Mixtures of distributions*
- *F. Random numbers of random variables.*

## L.2. A perspective

- Rather than saying that “topics are characterized by the frequency of characteristic words” say that “documents are created as if by a random model” – except that there is no rationality to the order dependence\*.
- Then the formal machinery of probability theory can be used to decorate the work

Original material © Paul Kantor, 2012

\*For an example that is generative with some order dependence, try Eliza (in emacs): or <http://www.elsewhere.org/pomo/>

# L.3. Bayesian update models

- A mixture model will produce a population that may have different proportions in different members of the same population. But this will only happen if there is a process that separates the population
- If, rather, each item is generated, a word at a time, then the fact that random mixtures of Dirichlet is again Dirichlet with the obvious parameters, reduces the characterization to one involving a matrix decomposition into Positive\* factors.

\*We say  $M$  is positive if all  $M_{ij} \geq 0$ . This is not the same as positive definite, which means  $(x, Mx) \geq 0$  for all  $x$ .

# P.\*. On being positive

- Let  $X = \|x_{dt}\|$  represent the weight (or importance, or some such thing) of document  $d$ , with respect to the axis labeled by “term  $t$ ”.
- When we decompose  $M = X^T X$  we normally allow negative elements. In fact, the eigenvectors of the decomposition are mutually orthogonal, so there can be only one (non-zero) eigenvector with all elements non-negative
- This leads to the quest for non-negative factorizations, which are also called Dirichlet models. The hope is that by being “at every step meaningful” the resulting decompositions will be more effective at matching computer results to the human quest for meaning.

# V.\*. Can there be negative documents?

- Rather as communication scholars assert that “you can not not communicate”, there does not seem to be any real meaning to the notion of the document “-d”.
- Therefore, while we may be content to obtain useful results from a formalism whose individual steps lack a sensible interpretation, there is pressure to work in the “world of positive numbers”

# N.1. Network approaches

- 0. Graph theory and IR
- A. Origins in citation studies
  - (i) Impact Factor
  - (ii) Network instabilities
  - (iii) Co-citation analysis
- B. Warren; Garfield; Small; Narin;
- C. Kleinberg. Directed graphs. Hubs and authorities
- D. Brin & Page: Page Rank
- E. Ask.com and the “media war”
- *F.* Networks social and other for information finding
  - (i) Folksonomies and Tagging
  - (ii) Association of persons
  - (iii) Privacy and social stability

# N.2. Impedance\*

- “...a more interesting model involving oriented linear graphs is one in which the nodes represent something more general than just fields of knowledge. In fact, imagine that we have a node for every field of knowledge, document, proportion, phrase, word, or customer of the library. Each pair of nodes, A and B, is joined by a path with an arrow on it and an associated measure of the "impedance" for going from A to B, or the "relevance" of B to A. The relevance of B to A is not necessarily equal to that of A to B, so that we must have two paths, one in each direction. Denote the impedance from A to B by  $Z_{AB}$ . This may be a vector, since there are various kinds of relevance; but if it is a number,  $Z_{AB}^{-1}$  is a measure of the relevance of B to A. Now the \$64,000 question is how to define  $Z_{AB}$ ?”
- Graph distance (shortest path; undirected)
- Conditional probability of use (directed)
- Reference (directed; binary); Citation (inverse graph)
- **Topical:** “Interest relevance. If customer A is known to be interested in a certain field of knowledge ... a document, B, highly relevant to this field .... is relevant to A "by interest" (and conversely).”
- **Association.** “The relevance of one word, i, to another one, j, may be measured by the "association factor"  $p_{ij} / (p_i p_j)$ , where  $p_i$  is the probability of i, and  $p_{ij}$  that of i and j in the same context.”\*

# E.0.5 Binding approaches together

- A. No one method works
- B. The kinds of synthesis
  - (i) Feature expansion
  - (ii) Rule combination
- C. The scope of synthesis
  - (i) Global
  - (ii) User-dependent
  - (iii) Task-dependent **the problem is to recognize task classes other than by the empirical effectiveness of methods, for them\*.**
- D. Usage as a meta-feature – “collaborative indexing”
  - (i) Covert collection
  - (ii) Overt cooperation

\*E.g., currently name searches on the Web fail because returns are dominated by social sites, which happens not to be what I want.

# Q.1. There are needs: quests

- There is some set of ‘information needs’ (“quests”) that we hope to serve or support  
 $Q = \{q \dots\}$
- One might say that the combined set  $(Q, D)$  is a special kind of entity (a “bicorpus”) that defines the static problem that we face.

# Q.2. A System sees only the query

- system sees not the Quest, but a denatured indication of it, the query, hereafter, *q*.
  - set of words (most web queries); or a
  - set of words with weights (E.g. Inquiry) or a
  - set of ordered word strings (“like this”) or a
  - Boolean structure (these AND (those OR others) BUT NOT undesirables) or
  - windows: this(5w)that; this(5n)that. DIALOG®

## Q.3. Representing the Query: less is more (popular)

- This was a surprise to many who are expert in the more esoteric forms,
- The most simple form of query, {words, more, words} has become hugely popular\*.

Original material © Paul Kantor. 2012  
\* See Bingle (Harry Lillis "Bing" Crosby (May 3, 1903 – October 14, 1977) )

## Q.4. System returns lists

- at each stage in a search, the system presents to the user an ordered (ranked) list of ***offered*** documents.
  - Offer\_1
  - Offer\_2
  - Offer\_3
  - Offer\_4
  - Offer\_5

# Q.5. What *order* is best?

- The goal is to put this in “the best” order.
- “best” may mean:
  - “maximum total expected clicks [for system revenue]  
OR
  - “minimum clicks until satisfaction” [for customer satisfaction and thus loyalty].
  - In practice ‘largest number of reasonably satisfied users’. (*Not a formula; may be litigated*)
  - But also: what order will make the systems’ next response most valuable?<sup>(1)</sup> Make the whole stream of future responses most valuable? <sup>(2)</sup>

\* (1) Callan; Zhai; (2) Frazier Original material © Paul Kantor. 2012

# Q.6. Offers Out of order are less good

- Consider the short list:
  - Offer\_4
  - Offer\_5
- Suppose ‘satisfaction’ is some kind of discounted gain. Then if Offer\_5 would completely satisfy more people than Offer\_4, we have them in the wrong order here.

# Q.7. The case where a single item satisfies

- To provide a formalism, we will focus on the case of ‘finding any one item that satisfies’ and
- consider the function  $p(q,d)$ , *defined as*
- the probability that a user whose quest has been currently represented by  $q$  will be completely satisfied by finding document  $d$

# C.0. Promoting diversity

- We need ways to make them different.
- (a) find centroid (first); look at the  $u-c$  then most remote; then most remote from those two,...
- (b) portfolio --- look at each vector (centered) and divided by its norm (so the inner product is a “correlation”) and “minimize the variance of the selected set” → find negatively ‘correlated’ documents
- anti-clustering – use any similarity measure – form an inverse; then do ordinary (hierarchical) clustering on the resulting measure.

# C.1. Why should we cluster?

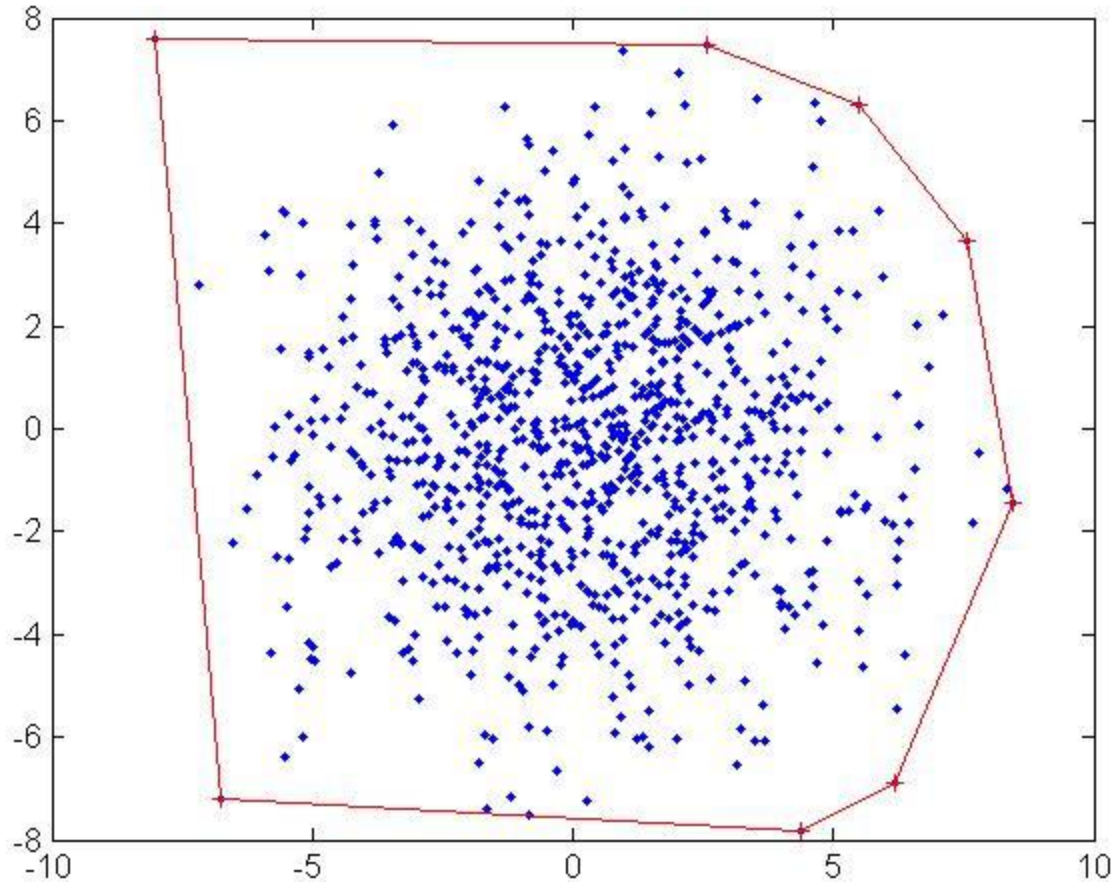
- We need it for the multi-lithic case
  - If the same query can arise from several quests that are quite different
  - This is the multilithic case.
- Then, if components of the answer form single islands,
  - we believe the set  $U$  consists of components that are 'clusterable'.
  - cluster  $U$  by your favorite means
  - my favorite means: lower order projections and convex hull – not yet based on evidence.

## C.2. A suggestion

- Think of the entities as points in a vector space.
  - This set has a convex hull
  - Points on the hull are as different from each other as they can be
  - Restrict the search for exemplars to that hull
  - There are fast convex hull algorithms,  $O(N)$ , rather than
  - Suppose there are  $H$  clusters to be found

# C.3. So it looks like this

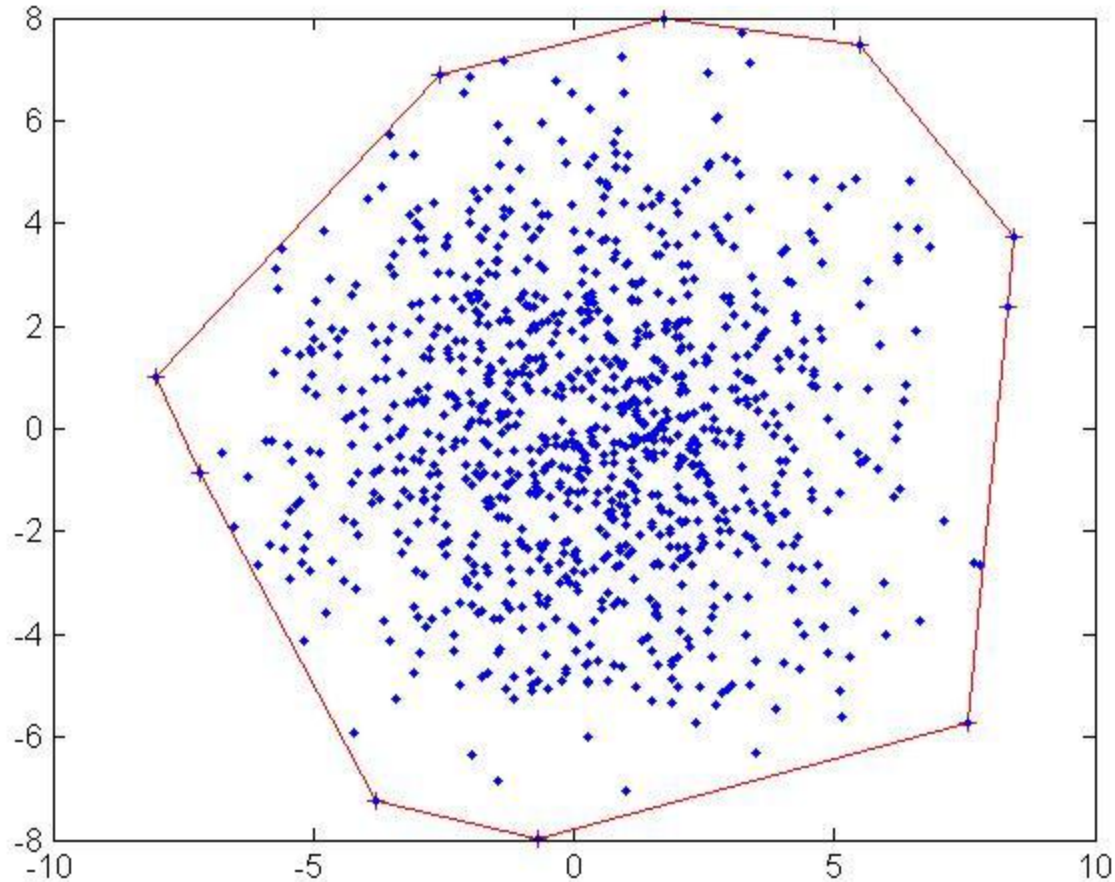
$|H|=8$ ;  $N=1,000$



# C.4. Random Projections

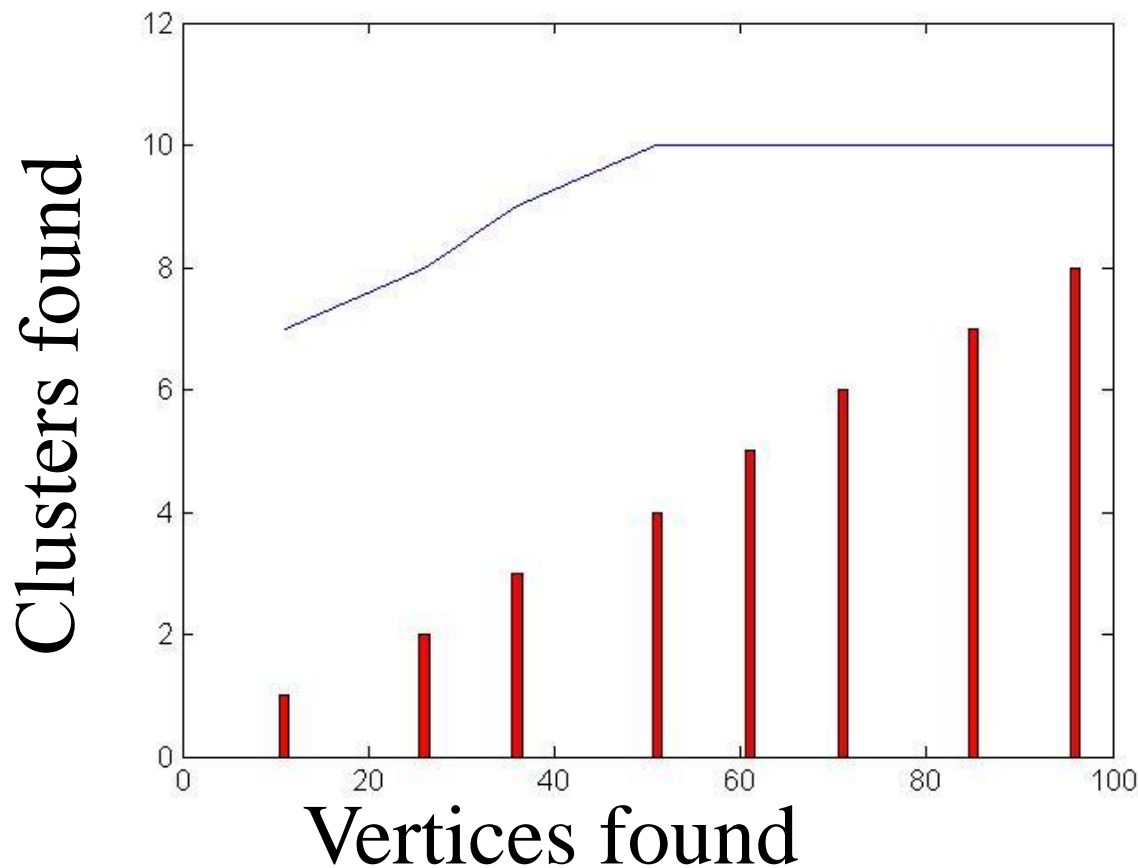
- Rather than work in the high dimensional space,
- (1) We could project onto a smaller space ( $d=10$ , or  $15$ )
- (2) But we are even lazier, and simply consider the pairs of axes
- This gives us very easy 2-D hull problems
- We solve them over and over, and keep track of the vertices (entities) that appear

# C.5. Simulation Studies: A projection



# C.6. Number of projection pairs

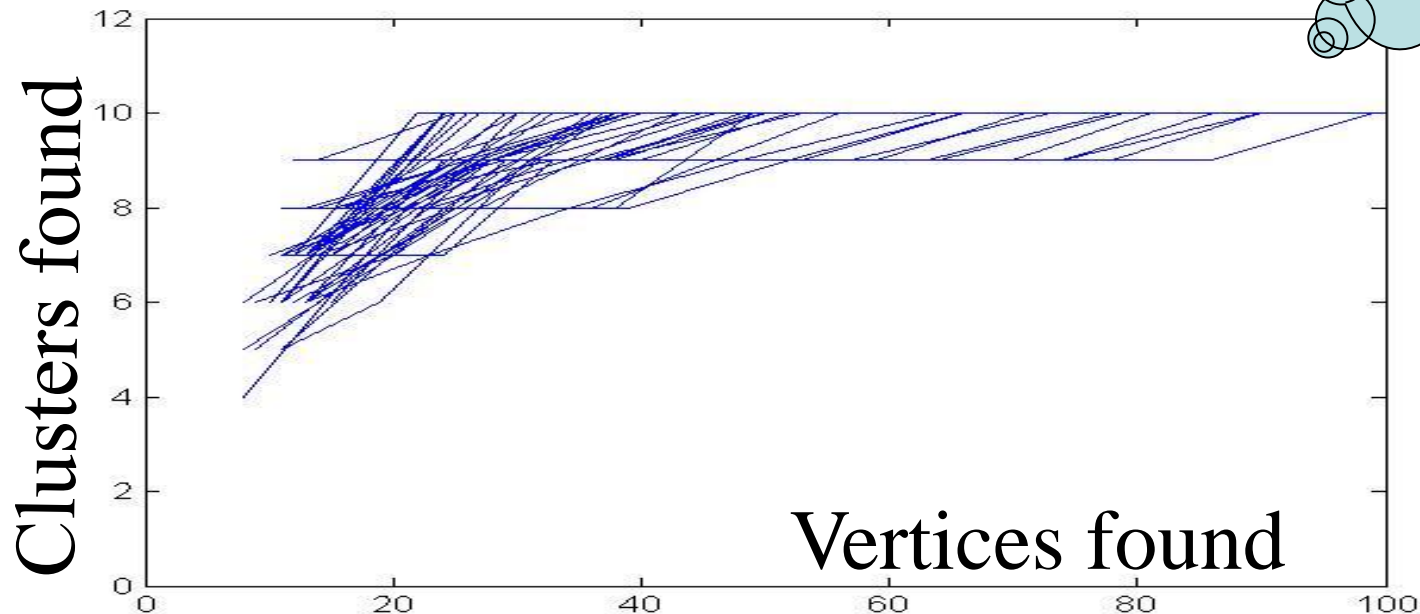
Each red bar marks another pair



# C.7. How fast does this work?

- We ask if we have picked up all ten clusters, after saving  $N$  different vertices and their projection hulls. Here are quite a few

Optical illusion



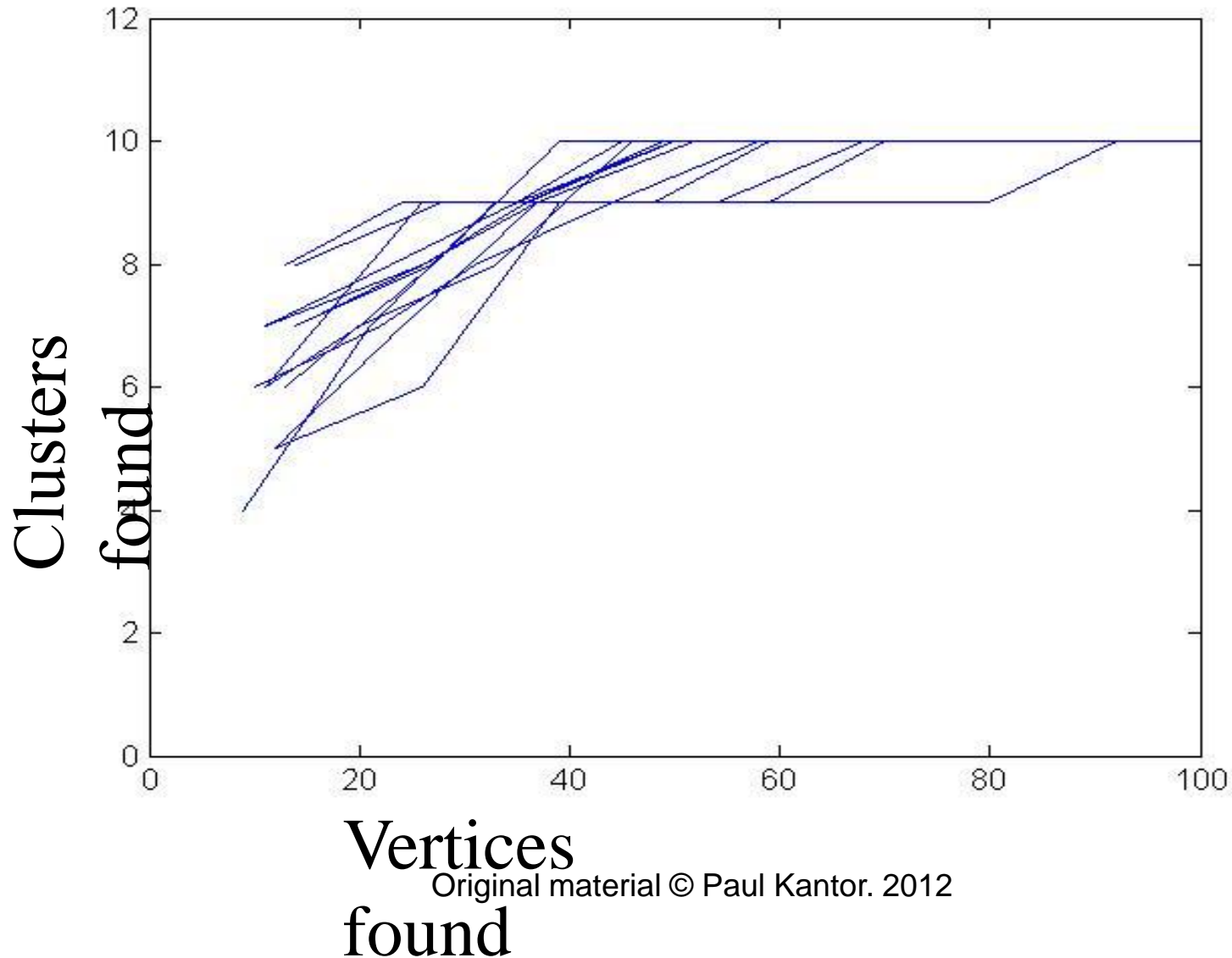
# C.8. How well does it work?

- The result is fundamentally stochastic:
  - $\text{Prob}\{\text{still missing some clusters} \mid v \text{ vertices}\}$
  - Decreases; details not worked out yet
  - Question: if we color the dots in one cluster, do they seem attracted to their vertices
    - would require that the vertices from one cluster be near each other along the convex hull
    - label them to see
      - **if not, must cluster real documents to find the centers.**

# C.9. But

- Using the 2-D projections plus
- Convex Hull trick

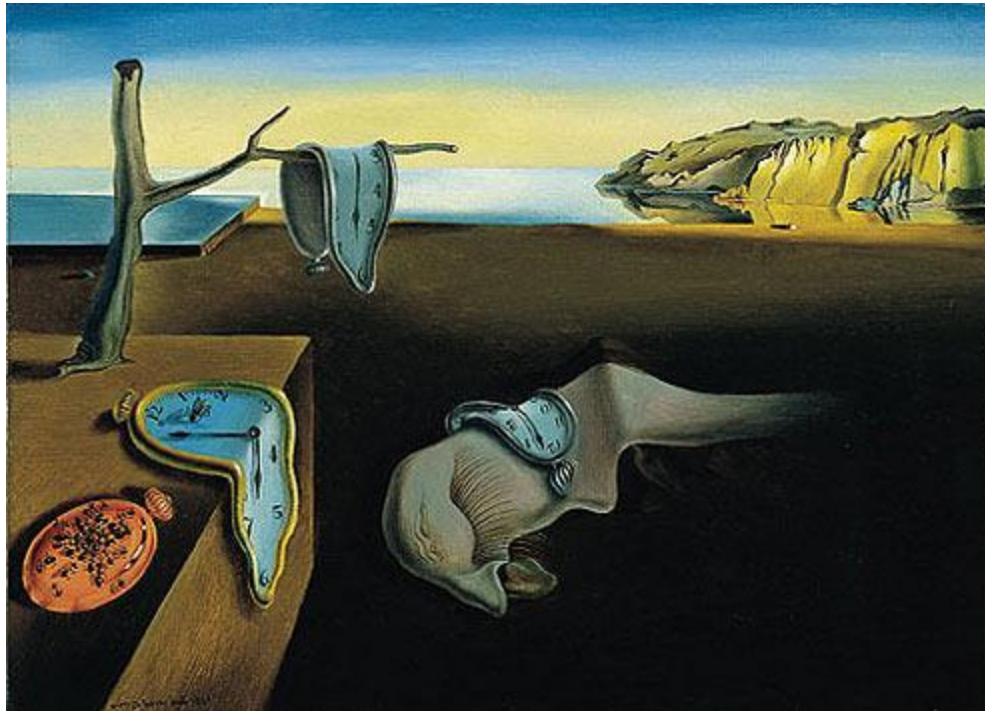
# C.10. Even a very small cluster (size $n=2$ ) can't hide





## M.3. Real Users are Affected by what they see: The persistence of memory

- Can we represent the fact that a user does not want to see the same thing over and over?



Copy of a  
Painting  
by  
Salvador  
Dali

# M.4. The Density Matrix\*

- In Quantum Mechanics, the D.M. this is used to represent states of physical systems.
- Let's speculate that one may use a Density matrix to represent states of the user:
- The density matrix can include both quantum effects and classical. I am using it only for classical mixtures, but I am borrowing the quantum idea of the 'effect of measurement'.

\* To be fair, these ideas are not "old" in the literature of IR. But they are old "in the real world". Birkhoff and von Neumann (1936) and von Neumann and Wigner (1928) have drawn on this literature, with a quite different perspective.

# M.5. Projections

- If  $|k\rangle$  is a particular state then the operator written as  $|k\rangle\langle k|$  (**note that they are in the reverse order**) has the effect of projecting any ket  $|x\rangle$  into its component along the ket  $k$ .
- $x \rightarrow |k\rangle\langle k|x\rangle$  corresponds to  $x \rightarrow (k,x)k$  in ordinary  $(,)$  notation.

# M.6. Sum of projections on pure states

*We can form a weighted sum of projections*

$$s^q = \sum_k p_k^q |k\rangle\langle k|$$

*The effect of observation. Perhaps it is true that*

*if the user sees document  $m$ , the user's state  $s^q$  is changed*

$$s \rightarrow (1 - |m\rangle\langle m|)s(1 - |m\rangle\langle m|)$$

# M.7. The users' state

“collapses”

- Presumably, when  $|m\rangle$  is seen, some value “accumulator” is increased by an amount related to  $p$
- The transformation of  $s$  means that the state  $|m\rangle$  is not worth seeing again
- This is an irreversible “collapse of the state”.

# M.8. Details are not complete

- “Steal this idea” 😊

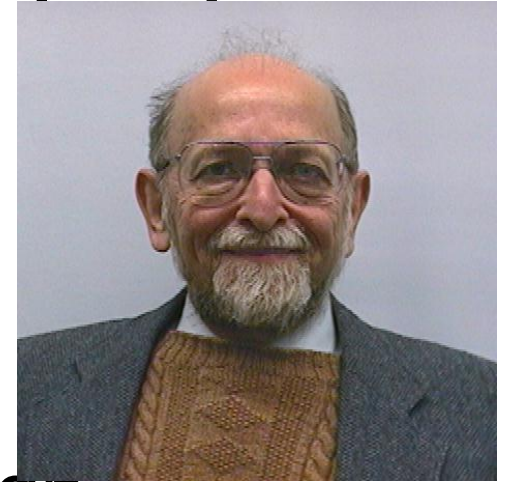
# S.1. Can we move Beyond Ranking?

- All of this research is done in the context of presenting a ranked list at any step in the search
- And, indeed, this is surely a hard enough problem
- but might we serve the users better if we made our problem even harder?.
- Consider the modern food court

## S.2. The food court

- When we go to the food court, there are many options
- I prefer Chinese food for lunch
- why isn't the Chinese food shop "first" (like a ranked list)
  - **when I arrive?**
- But, of course, that would not make sense for my friend who likes Indian food

# S.3. How do people's work?



- Philip Golden (1926-2006)
- An innovator in cafeteria design
  - We do not have a photo
- \* my mother's brother. We don't have a photo, but I resemble him a great deal

# S.4. Benefits of innovation

- Cafeteria design
- Used to be linear – slide trays on rails; then pay
- He designed
  - free flow layout
  - customers can move to any station (in the forward direction)
  - computable increase in throughput, freshness of food;
  - decrease in unsold food; less customer irritation

# S.5. A talk should have pictures

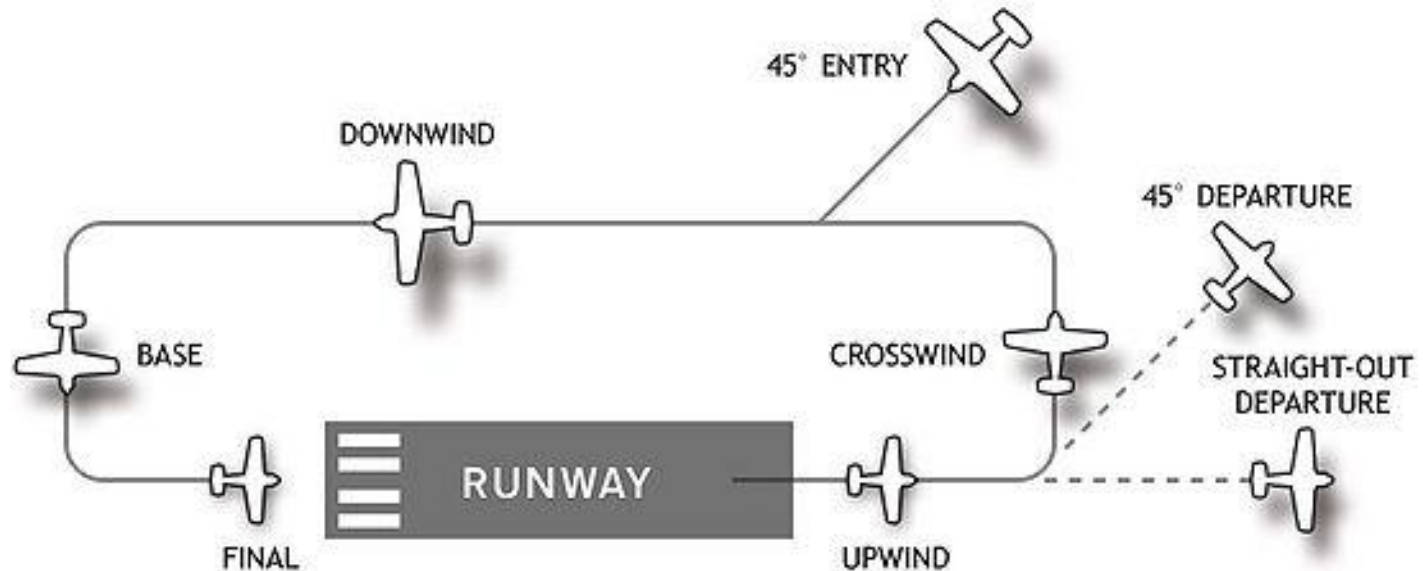
- 'and what is the use of a book,' thought Alice, 'without pictures or conversations?'
- *“Alice in Wonderland”, Lewis Carroll,*
- I have searched for relevant images of the free flow concept online.....
- Search not yet perfect



# S.6. Query: Cafeteria free flow



# S.7. Query: Cafeteria Flow Pattern



(This is actually the standard or 'left' pattern for landing an airplane)

# S.8. Query: Cafeteria traffic pattern (via Bing)

**Young first-time voters  
Following Obama for change  
they can believe in**



Not the very first  
retrieval, happily.

## S.9 Cafeteria Free Flow Failed Completely

- It took several years to learn that:
  - The customers were the problem
  - They stayed in the line and slid their trays along
  - They “knew” that jumping ahead in the line is impolite!!
- It took several years to engineer the solution

# S.10 New Layout

- Break the tray slide rails
  - Open the layout
  - Make it confusing!!
  - And thus, Improve service
- 
- Is it possible to do this for search engines?

# B1. Final Challenge

- Can we, working in Search and Retrieval find a way to imitate that success, and make it possible to present the *diverse aspects* of a “response” in a non-linear fashion.
- To do this means that we might take advantage of the human ability to observe in a non-linear way...

# B2. The organization of the Cerebral cortex: Human visual ability

- We know that the human brain is still more complex than even the fastest supercomputer
  - The K computer, named after the Japanese word "kei", which stands for 10 quadrillion,[1] is a supercomputer being produced by **Fujitsu and located at the RIKEN Advanced Institute for Computational Science** campus in Kobe, Japan.
  - On the 20th of June 2011, TOP500 Project Committee announced that K topped the LINPACK benchmark with the performance of 8.162 petaflops, or  $8.162 \times 10^{15}$  calculations per second, with a computing efficiency ratio of 93.0%, making it the fastest supercomputer in the world. The **previous record holder was the Chinese computer NUDT Tianhe-1A**, which performed at 2.507 petaflops [Wikipedia: <http://en.wikipedia.org/wiki/TOP500>]
    - The K uses a special six-dimensional torus network interconnect called Tofu, and a Tofu-optimized Message Passing Interface based on the open-source Open MPI library.[8][9] Users can create application programs adapted to either a one-, two-, or three-dimensional torus network.
  - The system adopts a two-level local/global file system with parallel/distributed functions, and provides users an automatic staging function for moving files between global and local file systems [Wikipedia: [http://en.wikipedia.org/wiki/K\\_computer](http://en.wikipedia.org/wiki/K_computer)]
- **The human makes use of (we believe) highly parallel and a mix of special purpose** “computational circuits” – although future advances in brain science may eventually replace the “computational circuit” metaphor with some other, more accurate, concept.
- As an example, **can you readily spot a “familiar figure”** in the next three slides? [Raise your hand when you spot it]. Two hands if you see two. Wave them if you recognize three.

# B3. Do you recognize anyone here?



# B4. Do you recognize anyone here?



B5. Do you recognize anyone here?



## E.1. Syntax and Semantics Have to Matter: Seeking a cure for the common cold

- “This paper discusses a cure for warts, and explains why it is not possible to cure the common cold.”
- “This paper discusses a cure for the common cold, and explains why it is not possible to cure warts.”
  - Syntax; Modality; not captured by term-based features.

# E.2. Summary: A list of hot new ideas\*

Classification and indexing Trees - ???

Botanical classification – Long tails

Lattice theory - Ontology

The algebra of classes, and Boolean algebra (the algebra of logic) - Everywhere

Syntax – NLP [e.g. History of Mathematics vs. Mathematics of History]

Partial ordering – classifications replaced with vector rankings

Oriented linear graphs – citation; reference

Oriented linear graphs with impedance -- co-occurrence; co-citation;

Random branching theory

Random motion on a network - - PageRank

Clumps - Clustering

Semantics – “the meaning of meaning” --- collaborative methods?

Mechanical translation - Statistical

Information flow & information theory – still not used?

Cerebral cortex & artificial intelligence – as Ghandi said of “Western civilization”  
‡

Theories of evolution – Genetic Algorithms?

Rational behavior – A cost matrix?

Hierarchies of memories in a computer --Everywhere

\*SPECULATIONS CONCERNING INFORMATION RETRIEVAL: IJ Good. IBM Research Report RC-78 December 10, 1958; ‡. “I think it would be a good idea.”  
Original material © Paul Kantor. 2012

## E.3. “Jack” Good

- Irving John ("I.J."; "Jack") Good (9 December 1916 – 5 April 2009)
- He was born **Isadore Jacob Gudak** to a Polish-Jewish family in London. He later anglicized his name to Irving John Good and signed his publications "**I. J. Good.**"
- An appreciation by Prof. Doron Zeilberger, of Rutgers.  
<http://www.math.rutgers.edu/~zeilberg/mamarim/mamarimPDF/jack.pdf>

# E.4 Thank you

- My thanks to the organizers, especially Yi Zhang, for their patience and support.
- To my thesis advisor, Sam B. Treiman, whose influence was greater than either of us imagined
- To Rutgers University office of the Vice President for Research, Michael Pazzani, for support of the projection clustering work
- Many excellent papers at this year's SIGIR, which I wish I could have linked to these slides

# R.1. Some useful readings

- I have put some readings on a Sakai site at Rutgers. When you register for the tutorial, you will be invited to join that site, and to set up a password that will enable you to access them.
- `Kantor_tutorial_2011@sakai.rutgers.edu`

# R1. More Readings

- [1] R.O. Duda, P.E. Hart, and D.G. Stork. Pattern classification. 2001.
- [2] I.J. Good. Probability and the Weighing of Evidence. Griffin, 1950
- [3] I.J. Good. Rational decisions. Journal of the Royal Statistical Society. Series B (Methodological), 14(1):107{114, 1952.
- [4] I.J. Good. Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables. The Annals of Mathematical Statistics, 34(3):911{934, 1963.
- [5] IJ Good and RA Gaskins. Nonparametric roughness penalties for probability densities. Biometrika, 58(2):255{277, 1971.
- [6] T. Hastie, R. Tibshirani, and J.H. Friedman. The elements of statistical learning: data mining, inference, and prediction. Springer Verlag, 2009.
- [7] H.P. Luhn. A statistical approach to mechanized encoding and searching of literary information. IBM Journal of research and development, 1(4):309{317, 1957.
- [8] H.P. Luhn. A business intelligence system. IBM Journal of Research and Development, 2(4):314{319, 1958.
- [9] H.P. Luhn. The automatic creation of literature abstracts. IBM Journal of research and development, 2(2):159{165, 1958.
- [10] P. Massart. The Tight Constant in the Dvoretzky-Kiefer-Wolfowitz Inequality Ann. Probab. Volume 18, Number 3 (1990), 1269-1283.
- [11] Ricci, F.; Rokach, L.; Shapira, B.; Kantor, P.B. (Eds.) Recommender Systems Handbook, Springer 1st Edition., 2011, XXIX, 842 p. 20 illus.

# R.2. Readings on Smoothing Functions

- V. A. Epanechnikov (1969). Theory Probab. Appl. 14, pp. 153-158 (6 pages) Non-Parametric Estimation of a Multivariate Probability Density
- Also Sacks and Ylvisaker: asymptotically optimal among non-negative kernels for  $D_2$  densities\*
- Cline: Mean Integrated Squared Error. Call a kernel admissible if and only if the only kernel that can do better, for a given value of  $n$  (the number of sample points) is the same kernel
- Cline show that this is related to the Fourier transform of the kernel. The best kernels are not non-negative! (remember this point).
- Cline, DBH (1988) Admissible Kernel Estimators of Multivariate Density. Annals of Statistics v16(4):1421-1427 (*JSTOR*)
- For a current view, see the Dagstuhl collection:
- <http://www.dagstuhl.de/Materials/AbstractListing/Files/09181-abstracts-collection.pdf>

# R3. Readings on Clusters and Exploration of User needs; Quantum Mechanics

- Yisong Yue, Thorsten Joachims: Interactively optimizing information retrieval systems as a dueling bandits problem. ICML 2009: 151
- David M. Blei, Peter Frazier: Distance dependent Chinese restaurant processes. ICML 2010: 87-94
- Sinead Williamson, Chong Wang, Katherine A. Heller, David M. Blei: The IBP Compound Dirichlet Process and its Application to Focused Topic Modeling. ICML 2010: 1151-1158
- Pierre Hansen, Brigitte Jaumard: Cluster analysis and mathematical programming. Math. Program. 79: 191-215 (1997) *and succeeding; most recent*
- Daniel Aloise, Pierre Hansen: Evaluating a branch-and-bound RLT-based algorithm for minimum sum-of-squares clustering. J. Global Optimization 49(3): 449-465 (2011)
- Garrett Birkhoff and John Von Neumann. (1936) The Logic of Quantum Mechanics. The Annals of Mathematics. Second Series, 37(4):823-843. Article Stable URL: <http://www.jstor.org/stable/1968621>

# R3. Some readings on n-gram representation

- Strong claims:
  - Damashek, M. (1995). Gauging similarity with n-grams: Language-independent categorization of text. *Science*, 267(5199), 843
  - These are character *n*-grams
  - *Motivated by a “center of gravity concept”*
  - *Controversial (denounced by Salton)*
  - *Effective*
- Also shown effective for “corrupted text” (produced by simulated OCR errors):
  - Kantor, P. B., & Voorhees, E. M. (2000). The TREC-5 confusion track: Comparing retrieval methods for scanned text. *Information Retrieval*, 2(2/3), 165-176.
- And of course the idea of a “centroid” is present somehow in all efforts to “cluster” documents.

# R5. A note on limitations

- The limitations of the static assumption, and the “point-based” approach to value have been known since long before web searching. See, for example:

- » Kantor, P. B. (1982). Evaluation and Feedback in Information Storage and Retrieval Systems. In Annual Review of Information Science and Technology Vol. 17. (pp. 99-120). White Plains, NY: Knowledge Industry Publications, Inc

# App.\*. Moment of inertia

- The moment of inertia tensor is closely related to the matrix that lies at the heart of linear dimensionality reduction, SVD.

# App.\*. Taking vector space “physically”

- Operations on vectors:

$$\exists 0 : v + 0 = v$$

*Distributivity*

$$c(\mathbf{x} + \mathbf{y}) = c\mathbf{x} + c\mathbf{y}$$

*so*

*$-1\mathbf{x}$  is an additive inverse*

- Therefore we should be able to do a lot of “physical” things like “defining the center of gravity” of a concept; or the “moment of inertia”

# App.P.12. Cline excerpt – admissible smoothing functions

135.] Thus an example of an integrable and admissible kernel with the same property is the kernel with transform

$$\psi(\mathbf{t}) = \begin{cases} 1, & \text{if } \|\mathbf{t}\| \leq \alpha - 1, \\ \alpha - \|\mathbf{t}\|, & \text{if } \alpha - 1 < \|\mathbf{t}\| \leq \alpha, \\ 0, & \text{if } \alpha < \|\mathbf{t}\|. \end{cases}$$

In one dimension, this kernel is the difference of two Bartlett kernels,

$$\kappa(x) = \frac{1 - \cos \alpha x}{\pi x^2} - \frac{1 - \cos(\alpha - 1)x}{\pi x^2}.$$

Some familiar kernels are not admissible. For example, the parabolic kernel [Epanechnikov (1969)] and its extension for multivariate density estimators,

$$\kappa(\mathbf{x}) = \frac{\Gamma(2 + d/2)}{\pi^{d/2}} \max(0, 1 - \|\mathbf{x}\|^2),$$

given by Sacks and Ylvisacker (1981), clearly are not admissible. Of course, these were chosen to optimize the *asymptotic* MISE under the restrictions that the kernel be nonnegative and that the density be twice continuously differentiable. If using a nonnegative kernel is of primary concern, then the Epanechnikov kernel has the asymptotic edge and no other nonnegative kernel is uniformly better for all finite sample sizes. On the other hand, there are often good reasons for using kernels which take negative values. For example, one can achieve a convergence rate of  $n^{-5/6}$ , even without assuming a continuous second derivative, by using a kernel of order 3 or greater [Cline and Hart (1986) and Cline