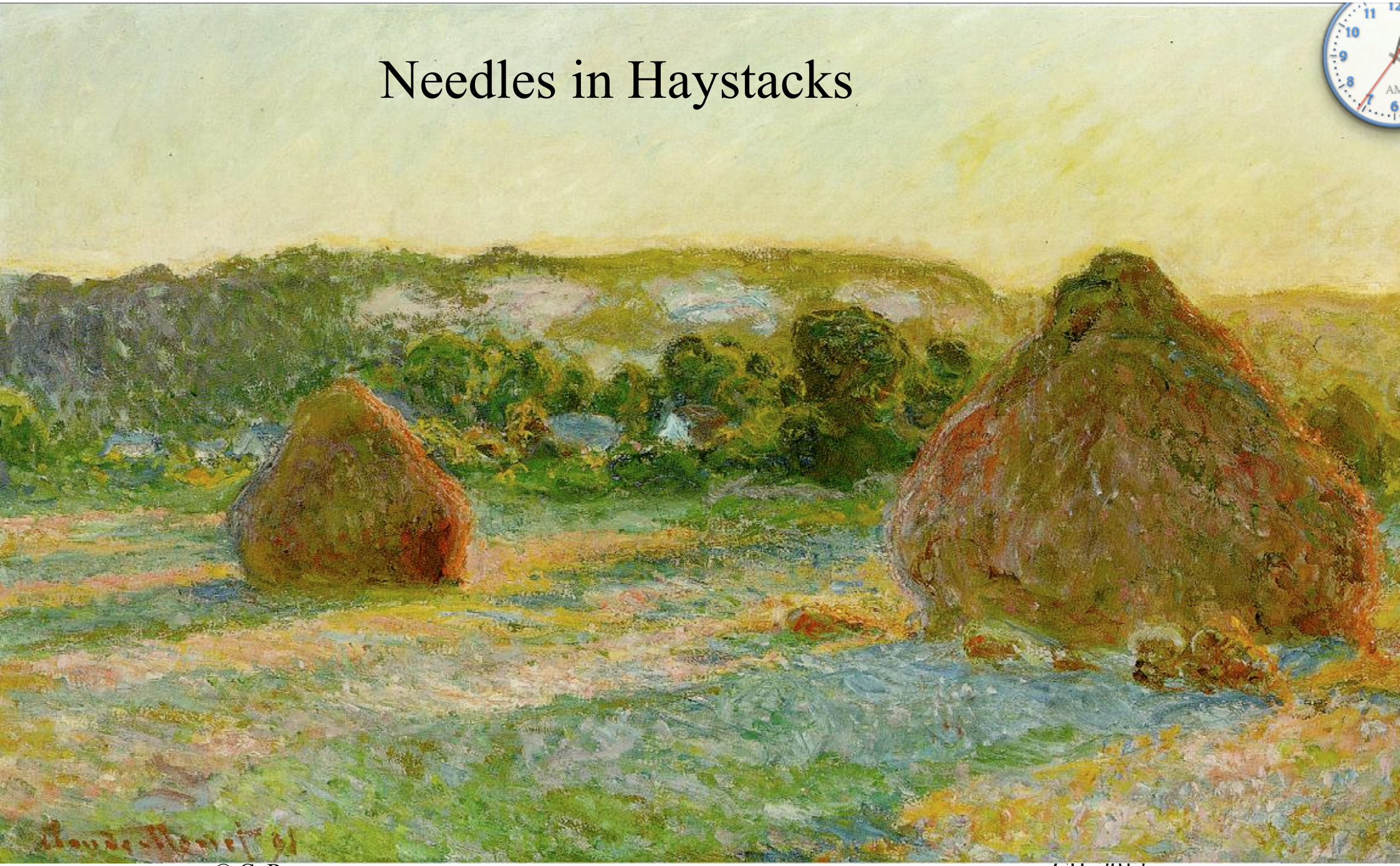
# Needles in Haystacks

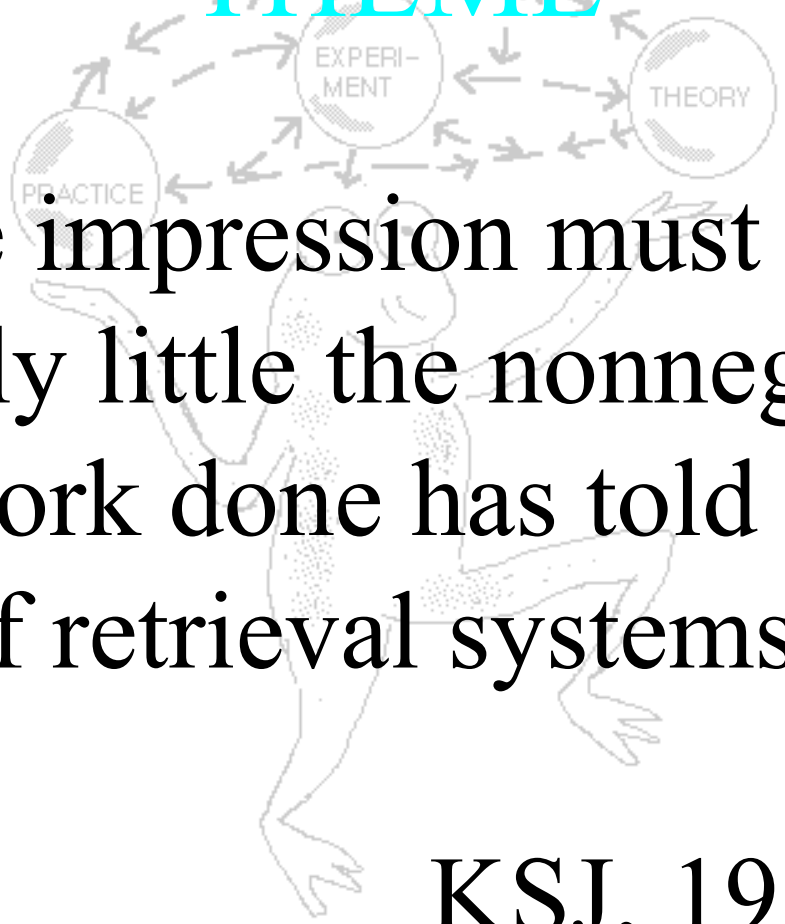



# Past, Present, and Future in Information Retrieval (and a little visualisation)

Keith van Rijsbergen

Zinal, January, 2012

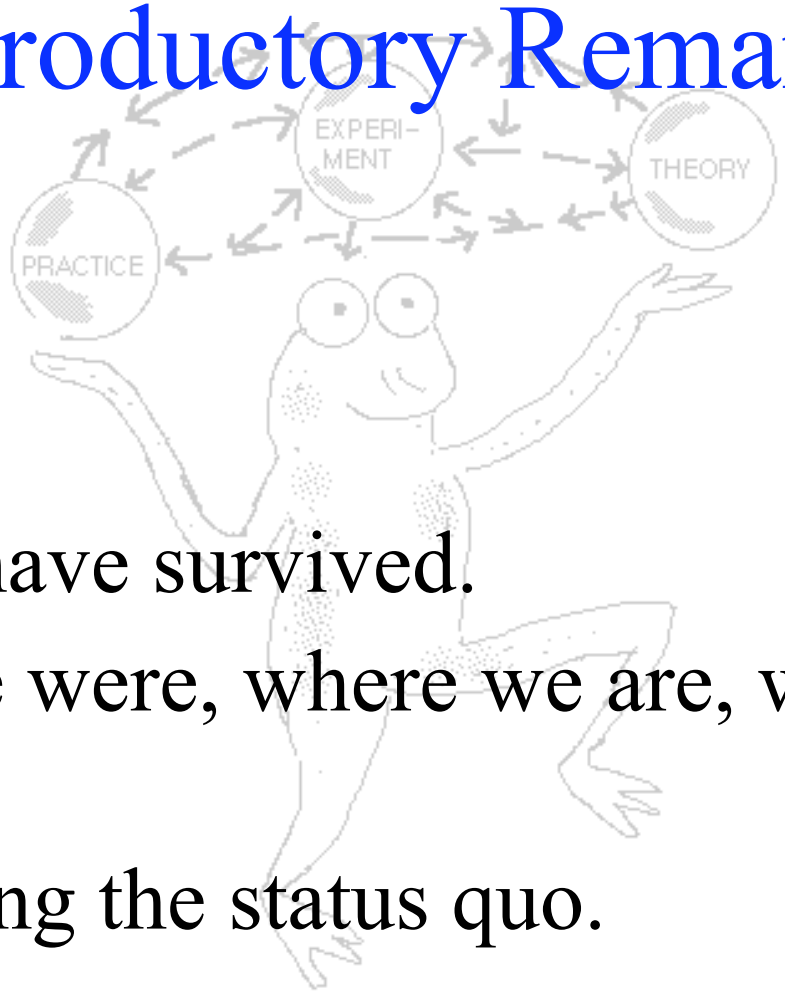
# THEME




"Overall, the impression must be of how comparatively little the nonnegligible amount of work done has told us about the real nature of retrieval systems"

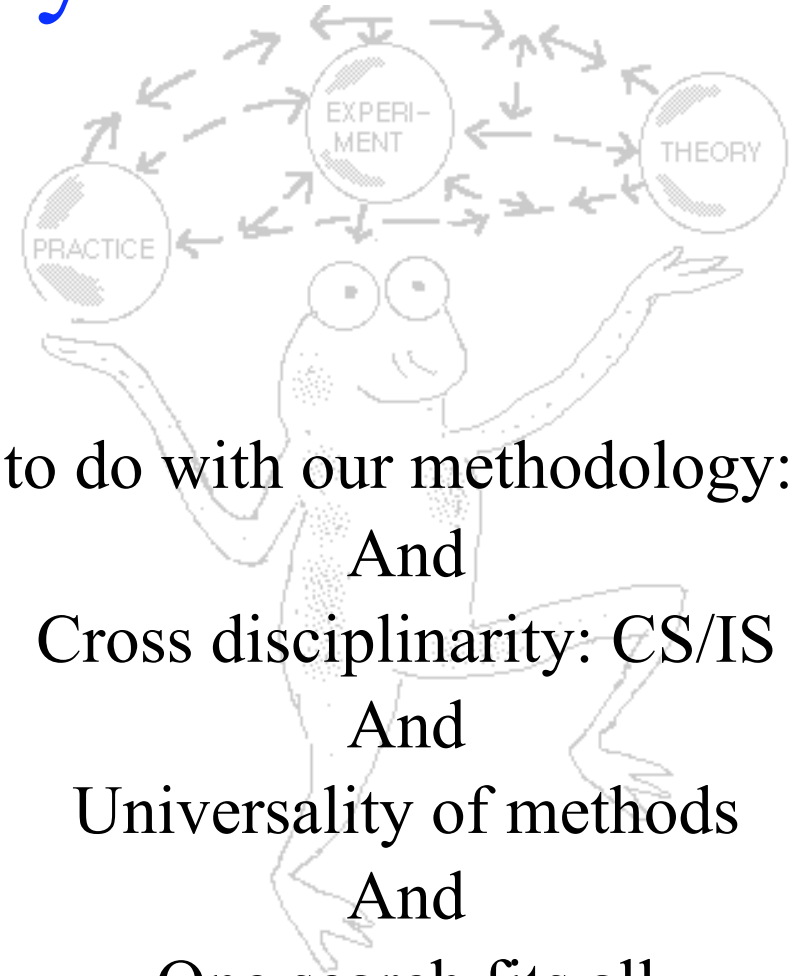
KSJ, 1981.

# Introductory Remarks




- Caveats
- Why we have survived.
- Where we were, where we are, where we are going.
- Challenging the status quo.

# Why are we still here?




Has a lot to do with our methodology: Cranfield

And

Cross disciplinarity: CS/IS

And

Universality of methods

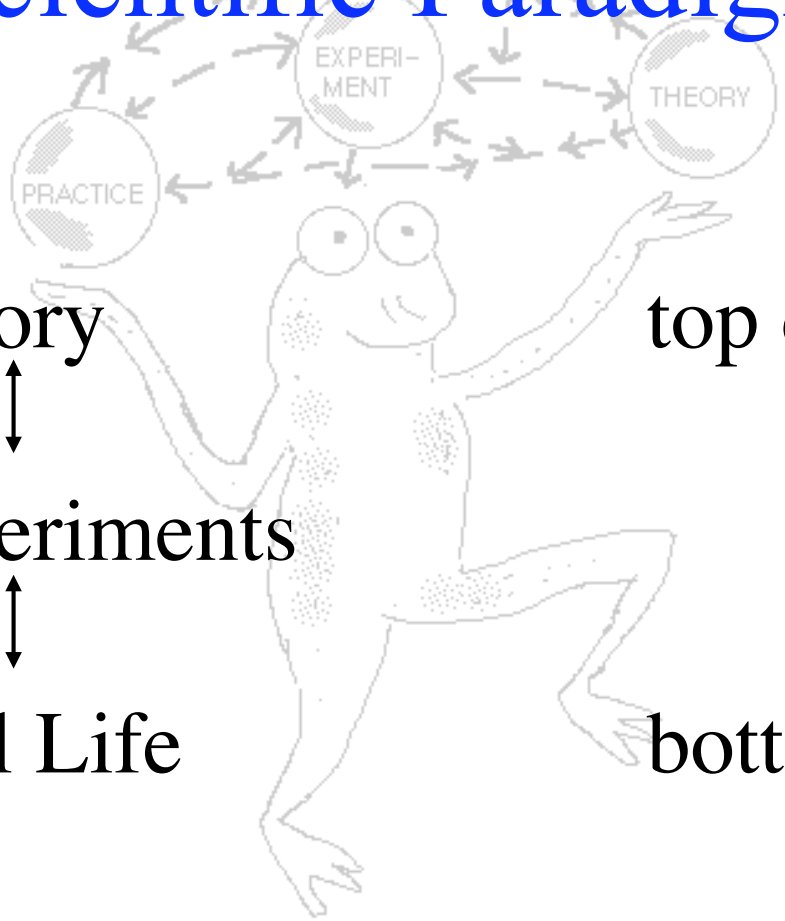
And

One search fits all

# Cranfield Paradigm

- Document collection
- Relevance judgements in advance
- Run strategy A and B
- Evaluate A and B in terms of P & R
- Compare A with B statistically
- State whether A ~ B, A > B, B < A

# Scientific Paradigm




Theory — top down approach

↕

Science — Experiments

↕

Real Life — bottom-up approach

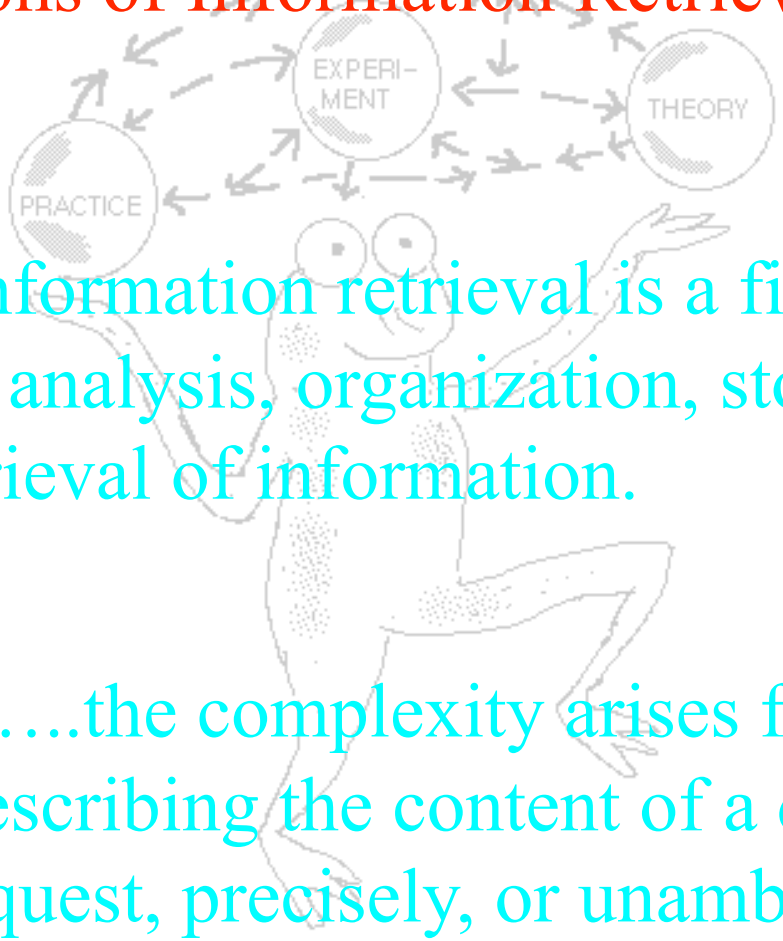
Basic research in IR is concerned with the design of better IR systems.

# What is Information Retrieval?




- General definition
  - Retrieval of unstructured data
  - Most often it is
    - Retrieval of text documents
      - Searching newspaper articles
      - Searching on the Web
  - Other types of retrieval
    - Image retrieval
    - Video retrieval
    - Music retrieval ….

# Definitions of Information Retrieval



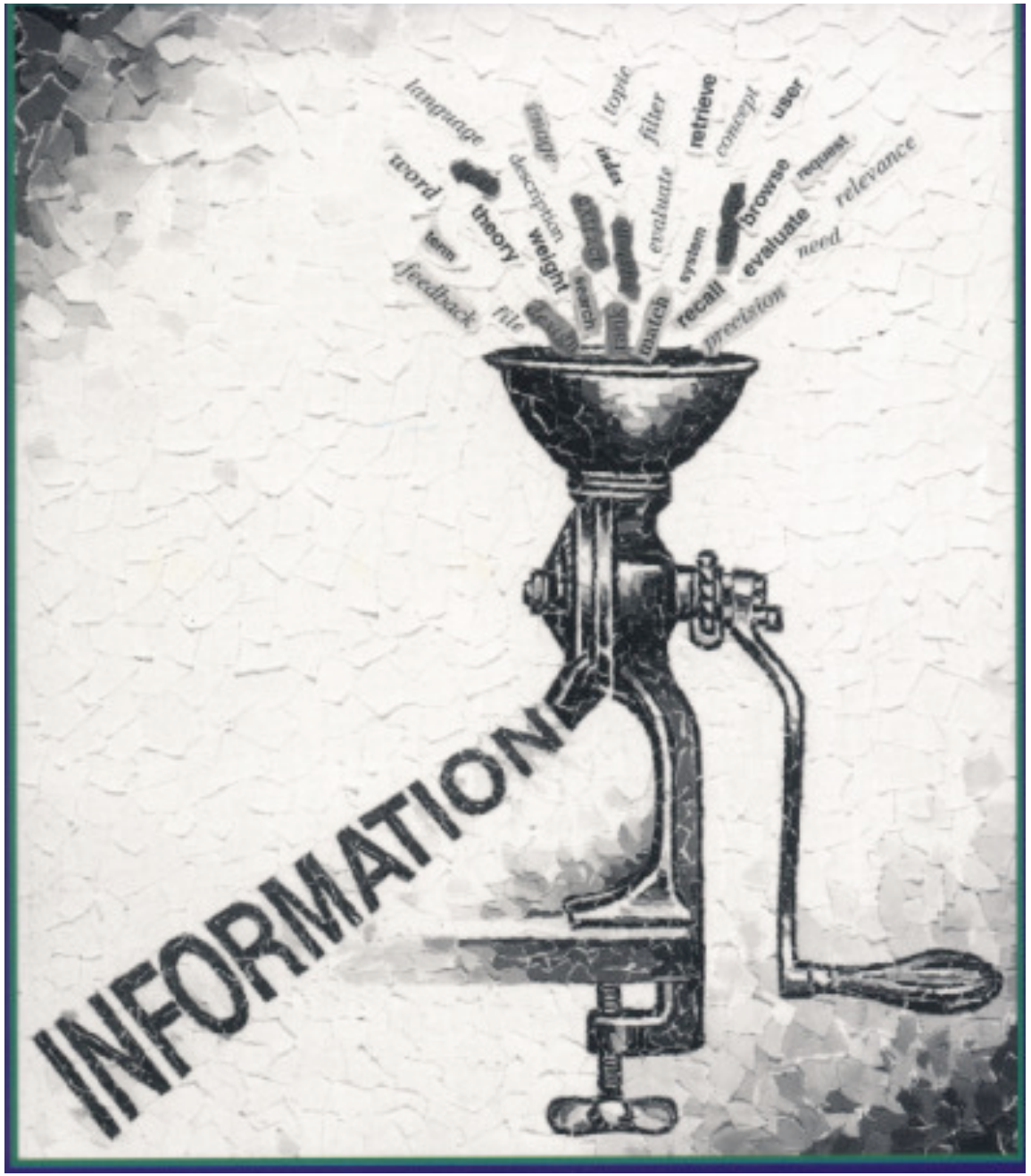

(Salton, 1968) – Information retrieval is a field concerned with the structure, analysis, organization, storage, searching, and retrieval of information.

(Needham, 1977)…..the complexity arises from the Impossibility of describing the content of a document, Or the intent of request, precisely, or unambiguously

# What is it, cont...

- Reference Retrieval?
- What is a document?
- What is a query?
- What is relevance?

The role of Information


INFORMATION
language
topic
filter
retrieve
concept
user
word
image
description
index
relevance
theory
weight
evaluate
request
browse
term
system
evaluate
need
feedback
file
search
match
recall
precision



CH, 2012

# Pre-history

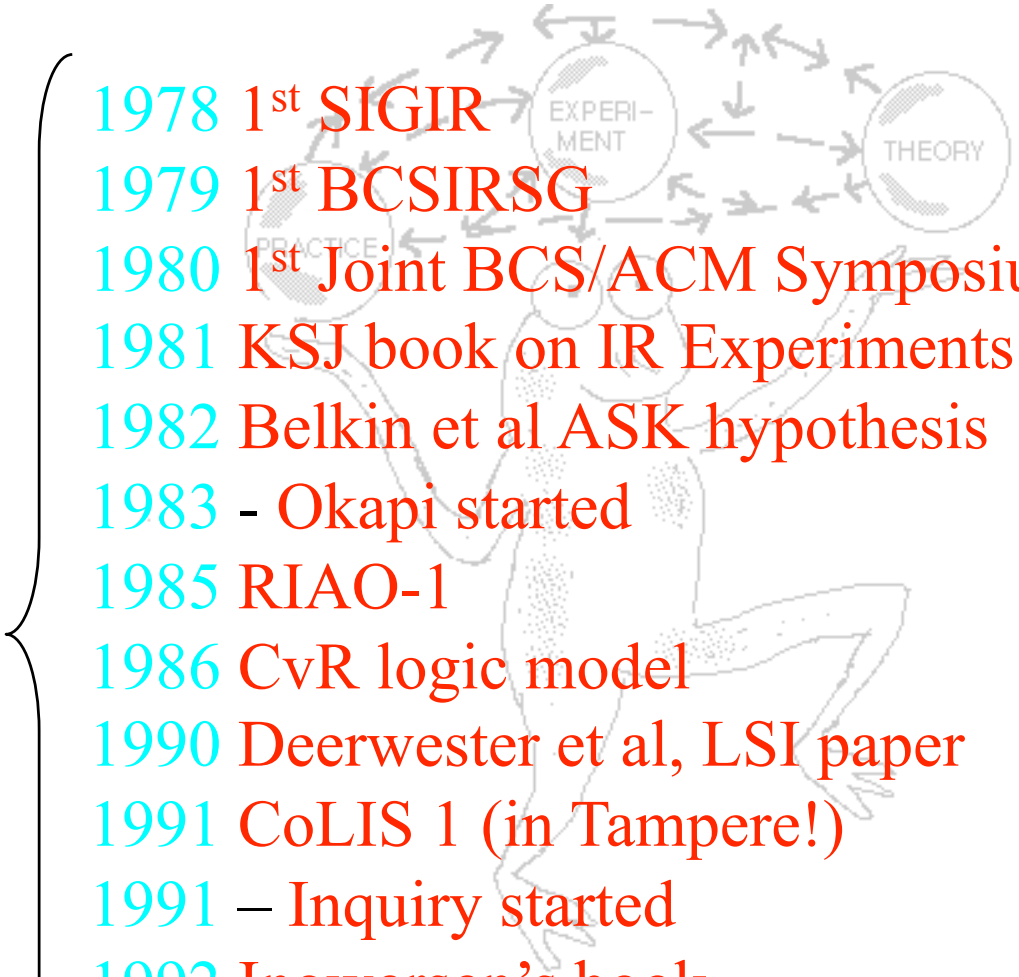

Goldberg (1930) (see Buckland)
Bush (1945)
Bagley (1951) MIT
Fairthorne (1945-52) RAE
Mooers (1952)
Luhn (1958)
Maron, Kuhns, Swanson (1959+)

# Time I (highlights for me)


1952 Mooers coins IR
1958 International Conference on Scientific Information
1960 Cranfield I
1960 Maron and Kuhns paper
1961 Towards IR, RAF
1961 (-1965) Smart built
1964 Washington conference on Association Methods
1966 Cranfield II
1968 Salton's first book

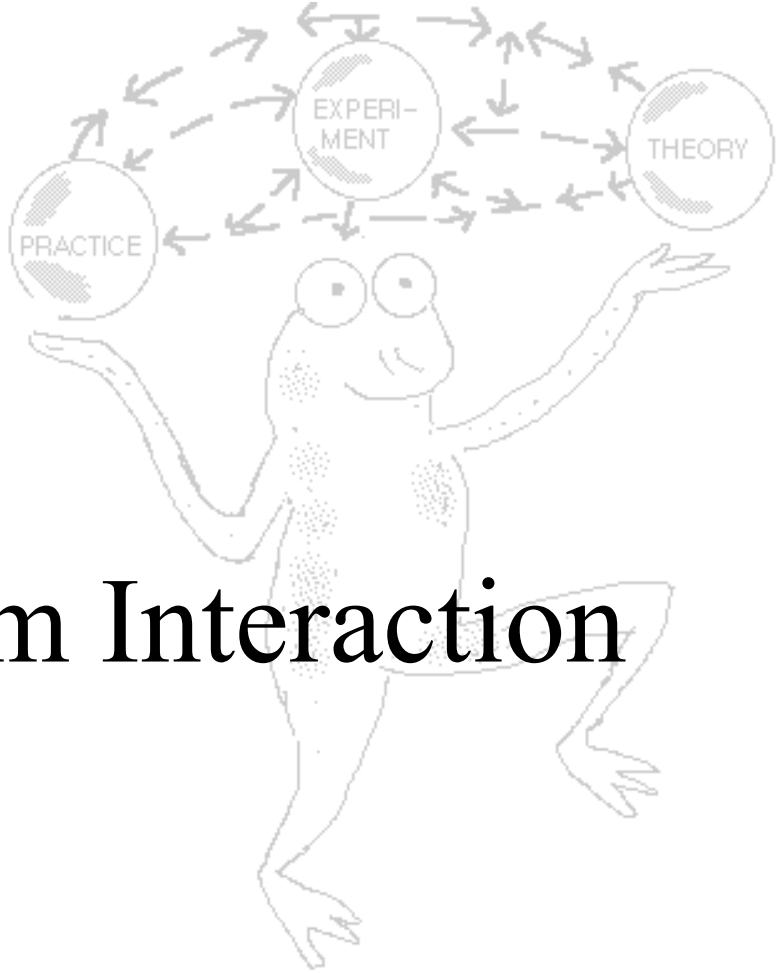
197- Cranfield conferences
1975 CvR's book
1975 Ideal test collection
1976 KSJ/SER JASIS paper .... BM25

# Time II




1978 1st SIGIR
1979 1st BCSIRSG
1980 1st Joint BCS/ACM Symposium (SIGIR 3 )
1981 KSJ book on IR Experiments
1982 Belkin et al ASK hypothesis
1983 - Okapi started
1985 RIAO-1
1986 CvR logic model
1990 Deerwester et al, LSI paper
1991 CoLIS 1 (in Tampere!)
1991 – Inquiry started
1992 Ingwersen's book
1992 TREC-1
1997 Kluwer series on IR
1998 Croft Ponte paper on language models

# Time III



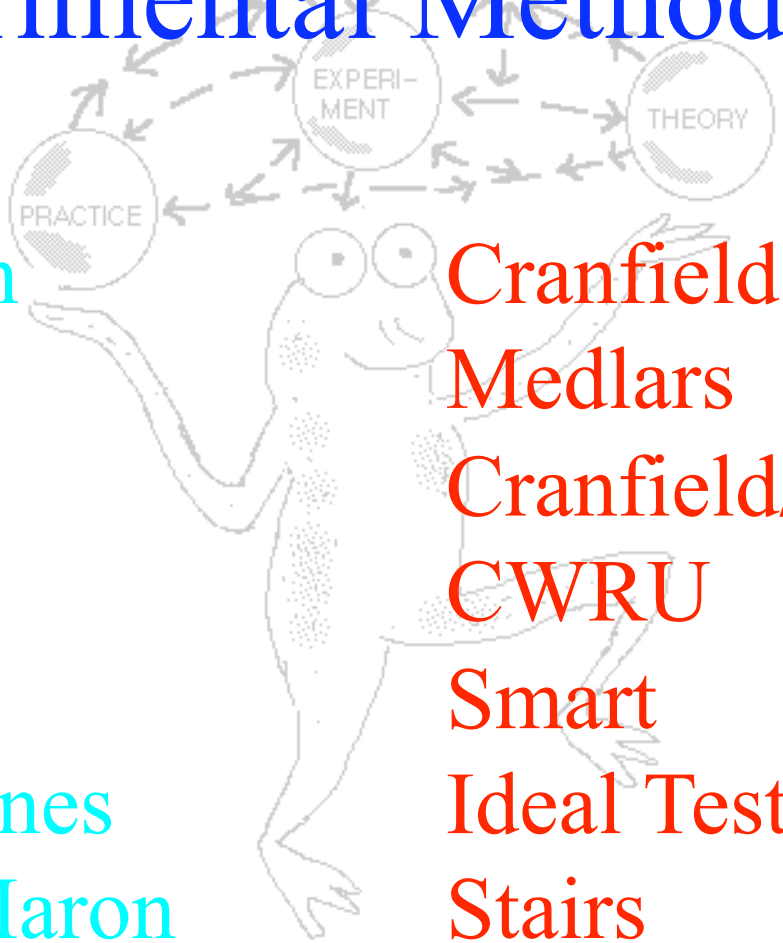

CIKM

Quantum Interaction



CH, 2012

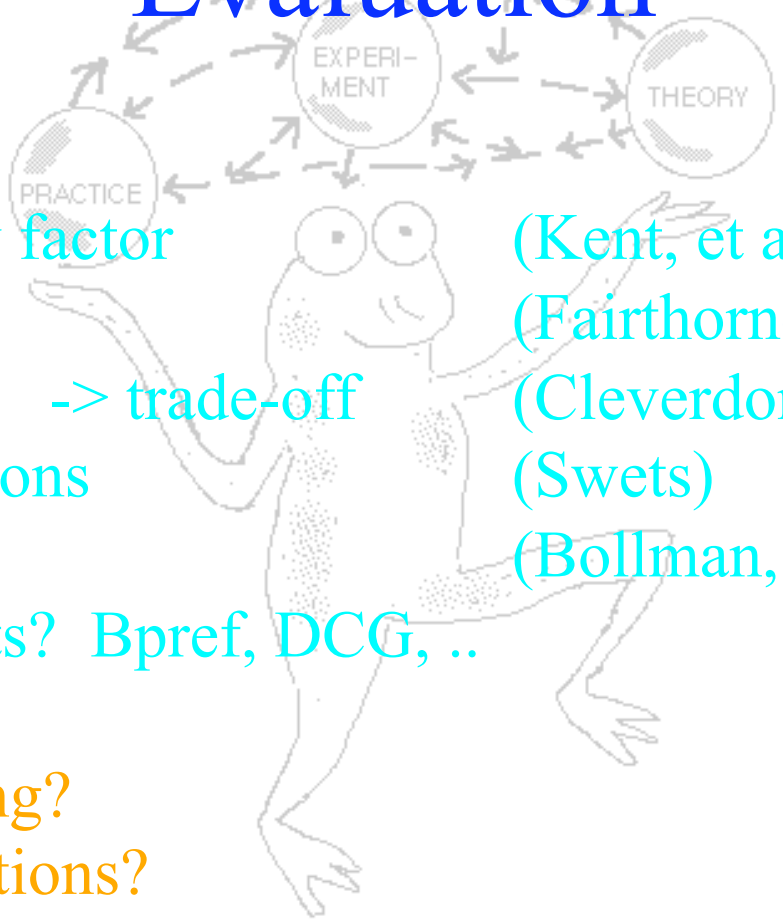
# Experimental Methodology




Cleverdon
Lancaster
Keen
Saracevic
Salton
Sparck Jones
Blair & Maron
Harman/Voorhees

Cranfield
Medlars
Cranfield/Smart
CWRU
Smart
Ideal Test Collection
Stairs
TREC

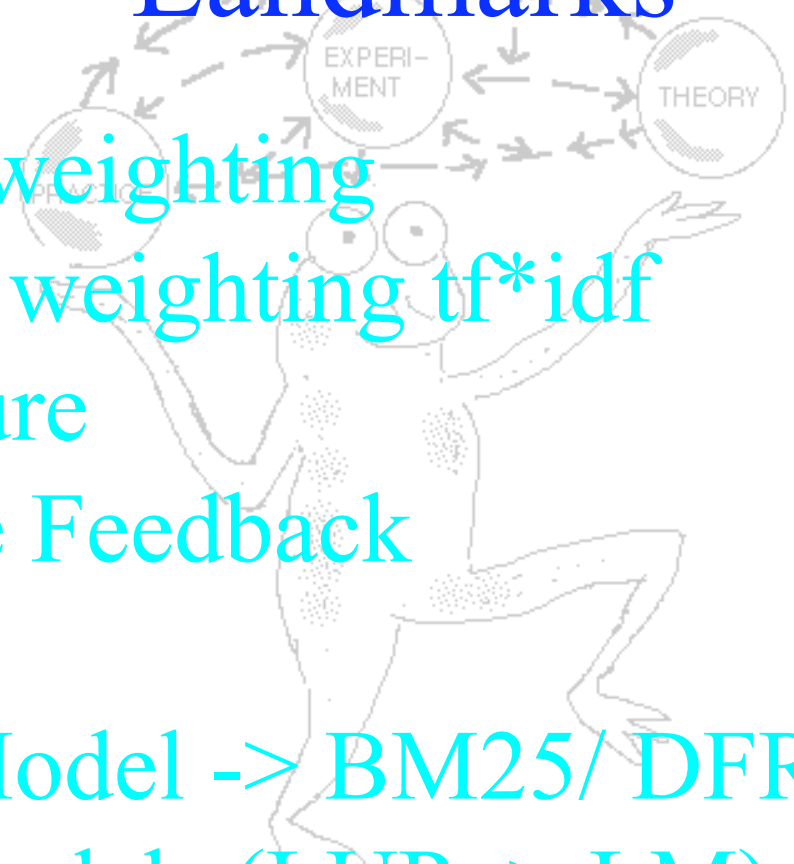
# Evaluation



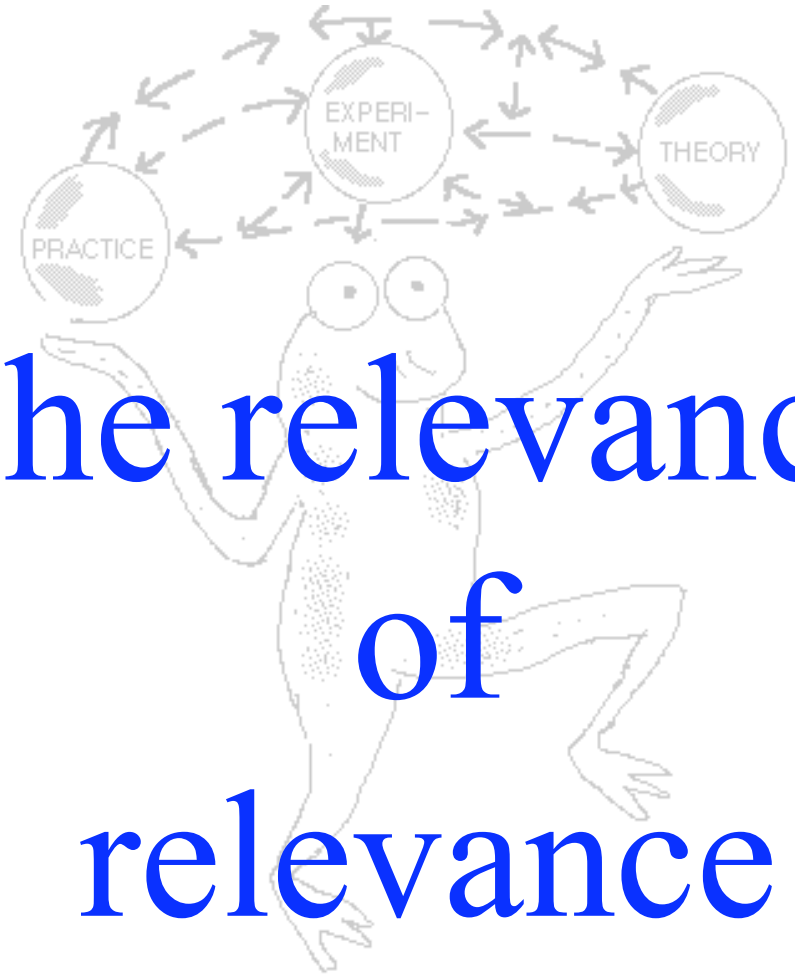

Recall/ pertinency factor (Kent, et al)
ABNO/OBNA (Fairthorne)
Precision, Recall -> trade-off (Cleverdon)
Probabilistic versions (Swets)
Measure-theoretic (Bollman, van Rijsbergen)
New developments? Bpref, DCG, ..

Significance testing?
Sampling distributions?
Incomplete judgments?
Multi-valued relevance?

# Landmarks



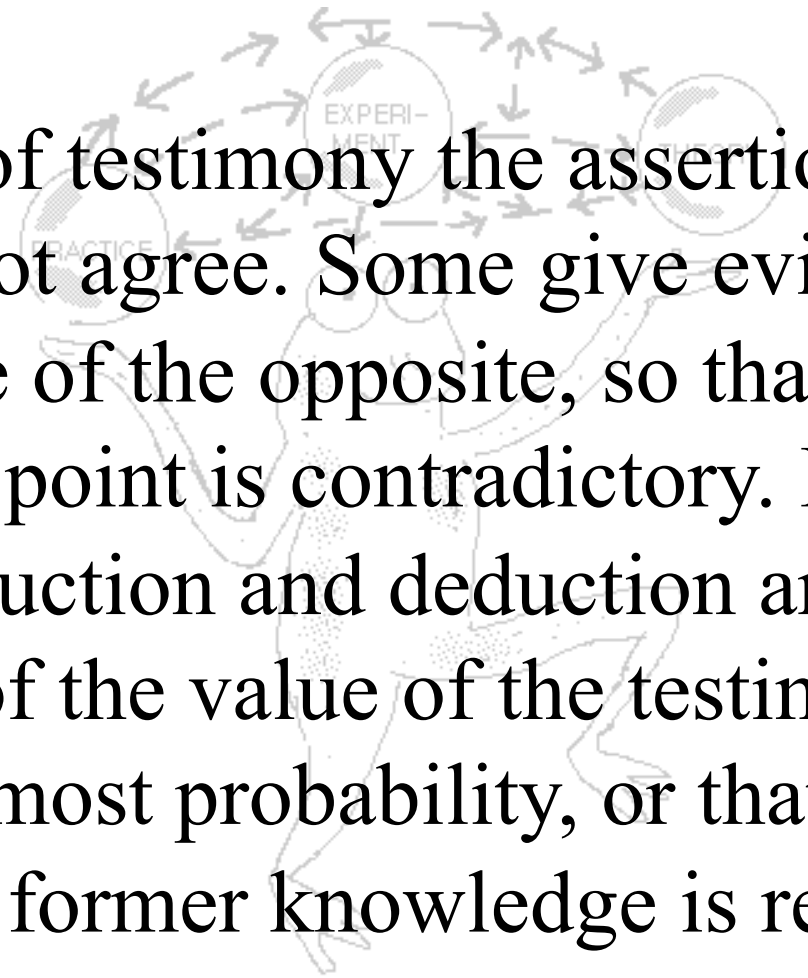

Luhn's tf weighting
Statistical weighting tf*idf
Architecture
Relevance Feedback
Stemming
Poisson Model -> BM25/ DFR
Various models (LUP -> LM)




# The relevance of relevance



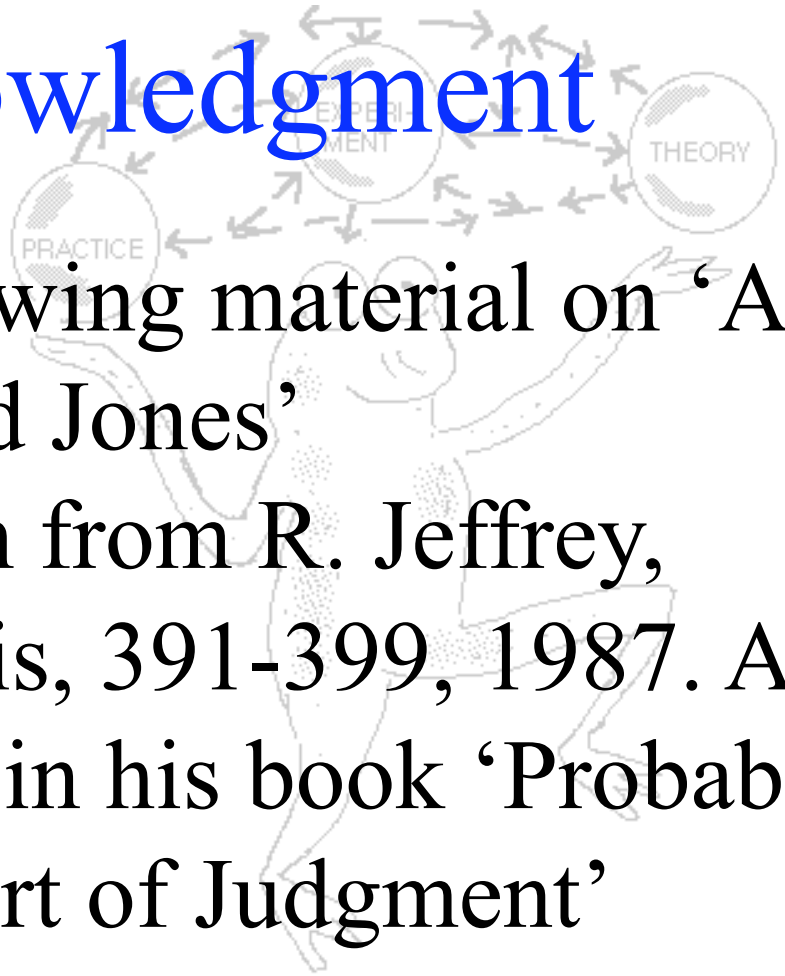
CH, 2012




In most cases of testimony the assertions of all witnesses do not agree. Some give evidence of one kind, some of the opposite, so that the evidence upon the same point is contradictory. In these, cases, the laws of induction and deduction are applied by the jury, to judge of the value of the testimony, and that which affords most probability, or that which most coincides with former knowledge is received.

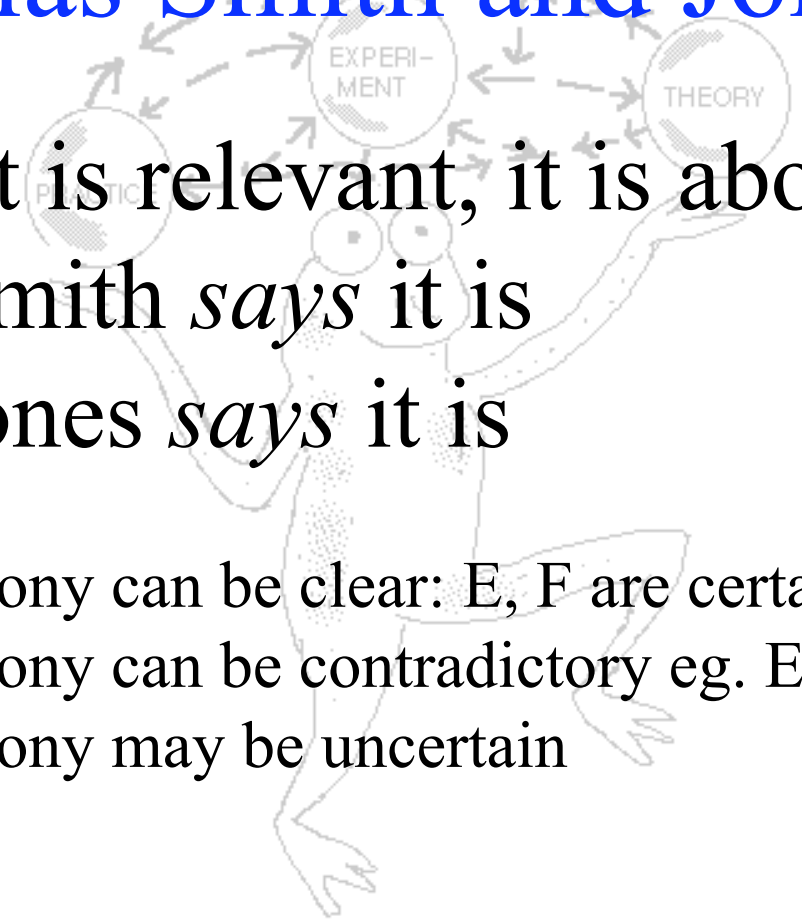
Alfred Smee, 1851

# Acknowledgment




The following material on ‘Alias Smith and Jones’
was taken from R. Jeffrey, Erkenntnis, 391-399, 1987. Also reprinted in his book ‘Probability and the Art of Judgment’

# Alias Smith and Jones




H: it is relevant, it is about.....

E: Smith *says* it is

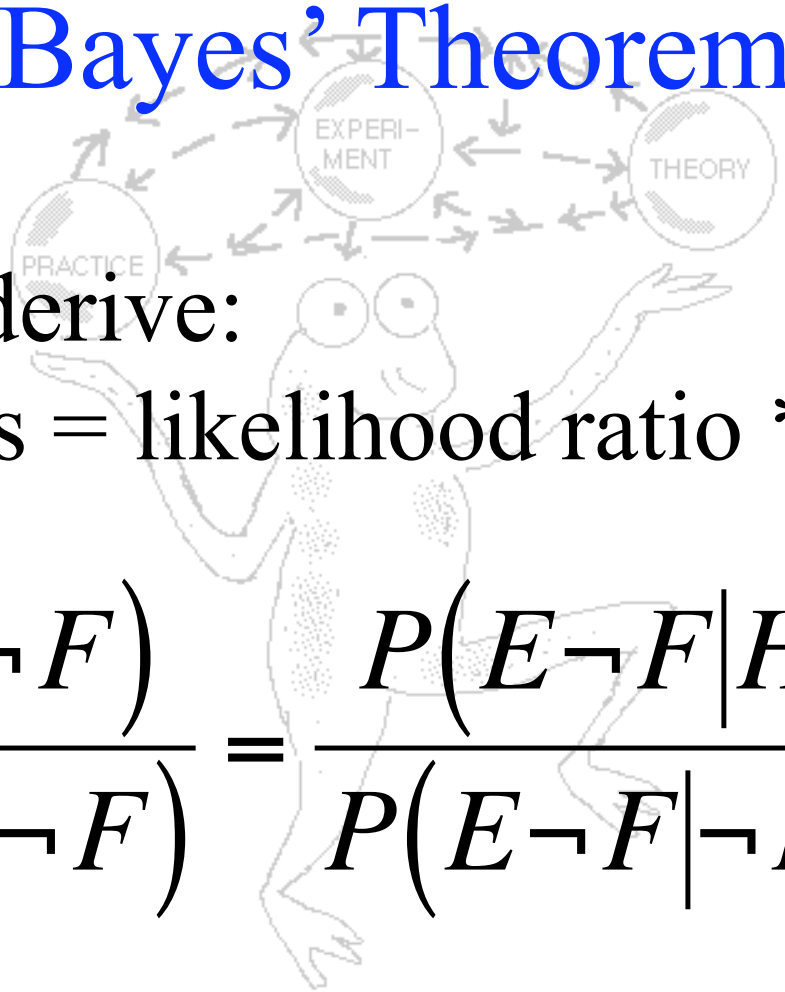
F: Jones *says* it is

- Their testimony can be clear: E, F are certain
- Their testimony can be contradictory eg. E and -F
- Their testimony may be uncertain

Smith and Jones contradict each other

Cancellation?

# Bayes' Theorem




you can derive:

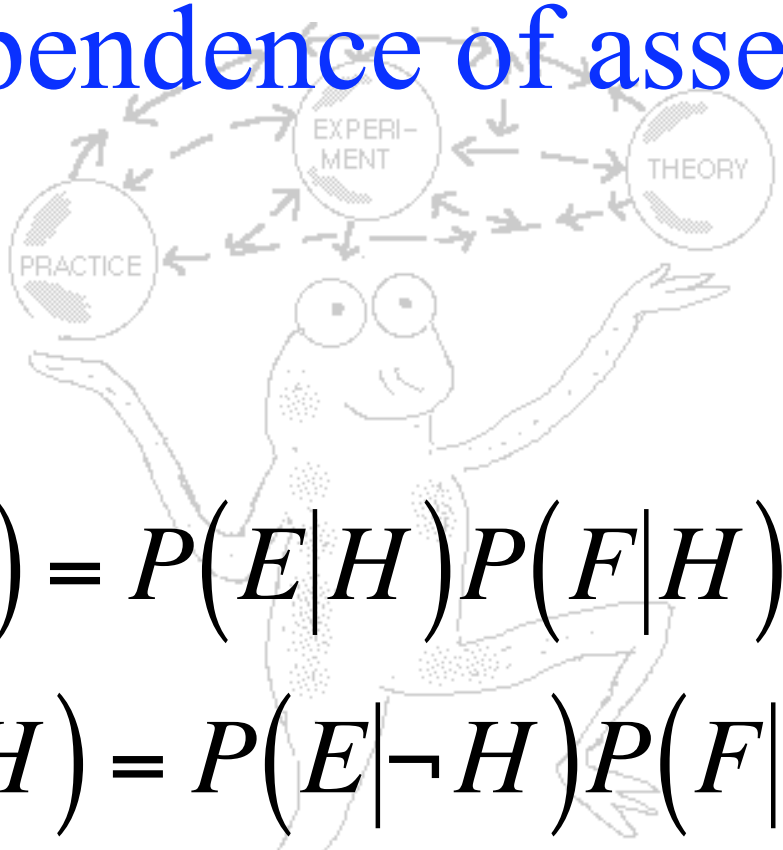
final odds = likelihood ratio * prior odds

$$\frac{P(H|E\neg F)}{P(\neg H|E\neg F)} = \frac{P(E\neg F|H)}{P(E\neg F|\neg H)} \times \frac{P(H)}{P(\neg H)}$$

Cancellation of testimony:

Likelihood ratio = 1

# independence of assessors




$$P(EF|H) = P(E|H)P(F|H)$$

$$P(EF|\neg H) = P(E|\neg H)P(F|\neg H)$$

$$P(EF) = P(E)P(F)?$$

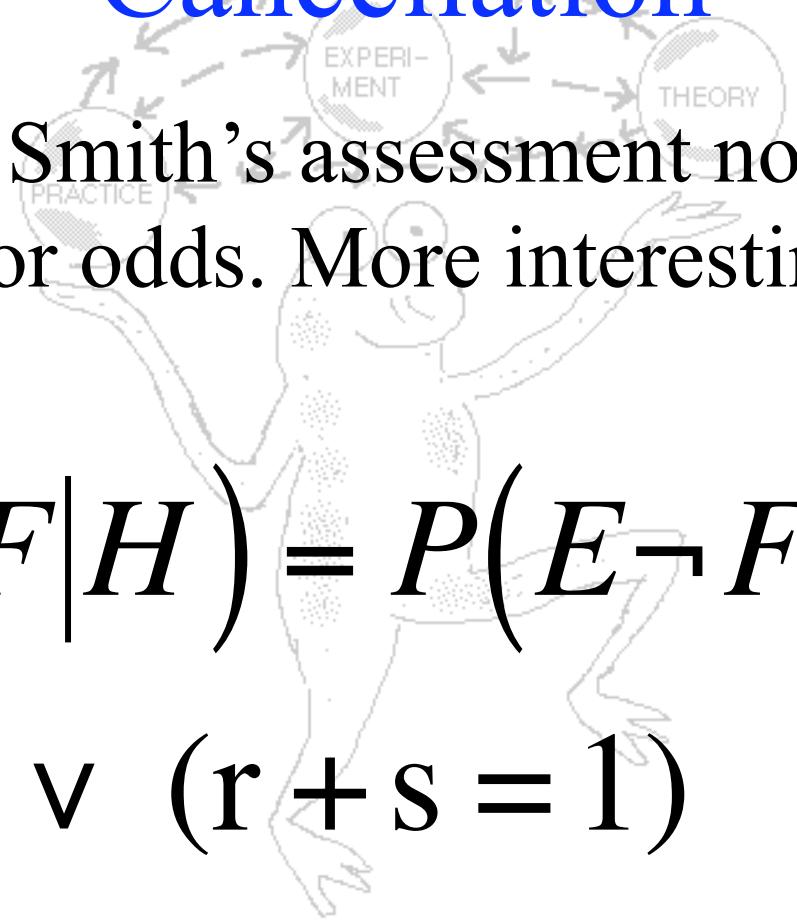
# Conflicting clear testimony of *equally reliable* independent assessors

$$P(E|H) = P(F|H) = r$$

$$P(E|\neg H) = P(F|\neg H) = s$$

$$P(EF|H) = r^2, P(EF|\neg H) = s^2$$

# Cancellation



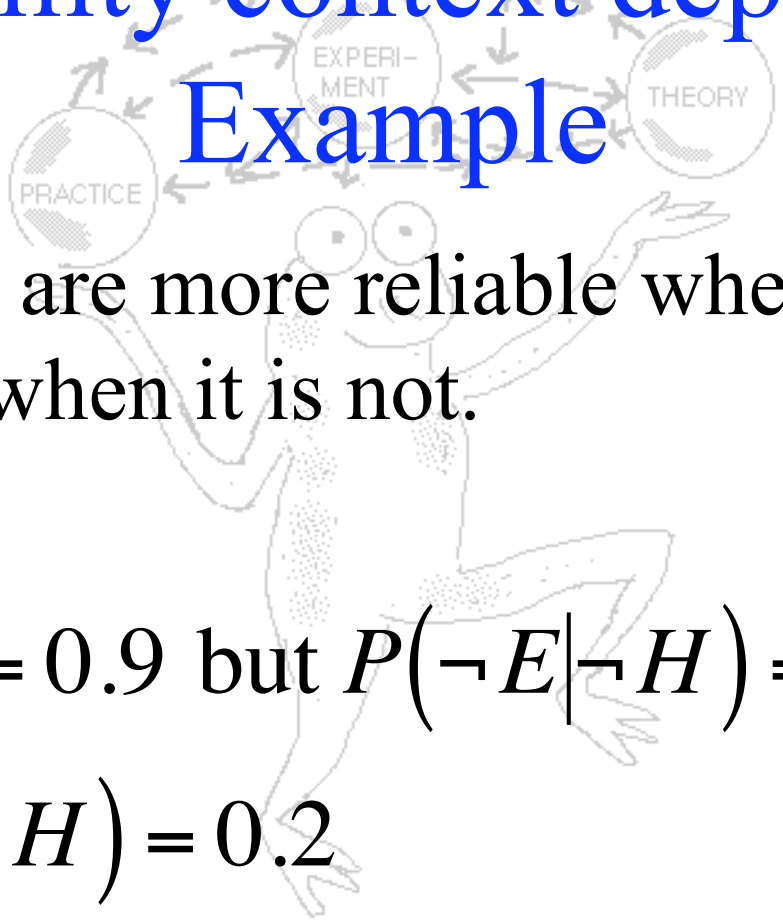

r = s means Smith's assessment nor Jones' changes prior odds. More interesting case:

$$P(E\neg F|H) = P(E\neg F|\neg H) \text{ iff}$$

$$(r = s) \ \vee \ (r + s = 1)$$

Cancellation when : $P(E|H) + P(E|\neg H) = 1$

# Reliability context dependent Example

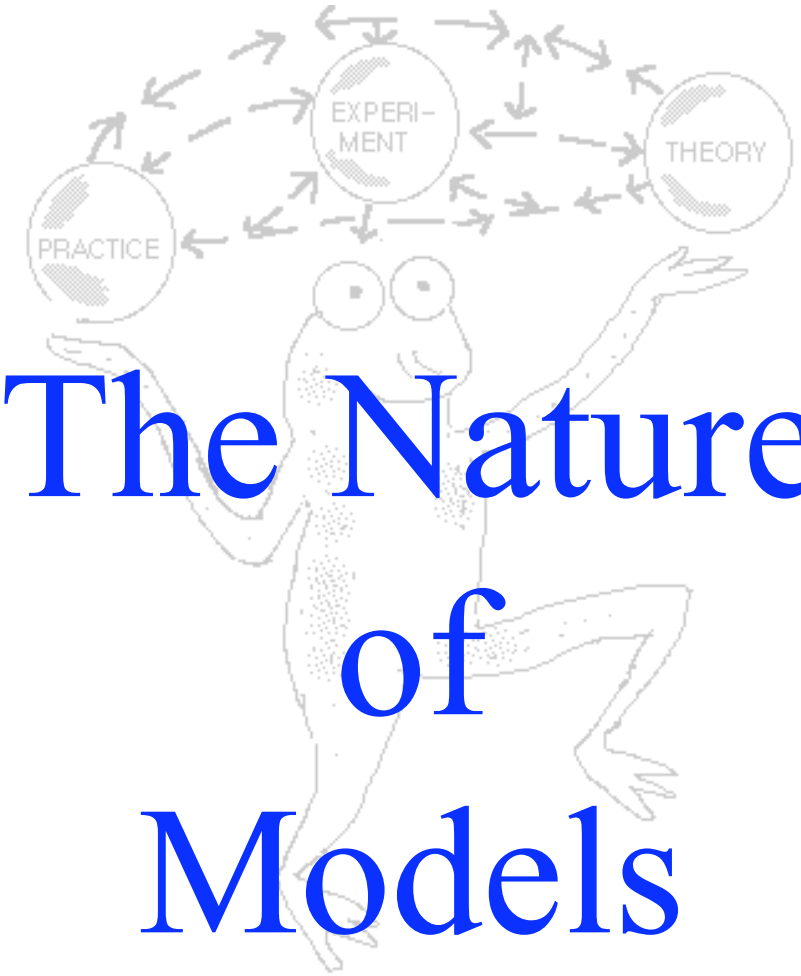



Smith and Jones are more reliable when the document is relevant than when it is not.

$$P(E|H) = 0.9 \text{ but } P(\neg E|\neg H) = 0.8$$

$$\Rightarrow P(E|\neg H) = 0.2$$

$$\therefore P(E|H) + P(E|\neg H) = 1.1$$

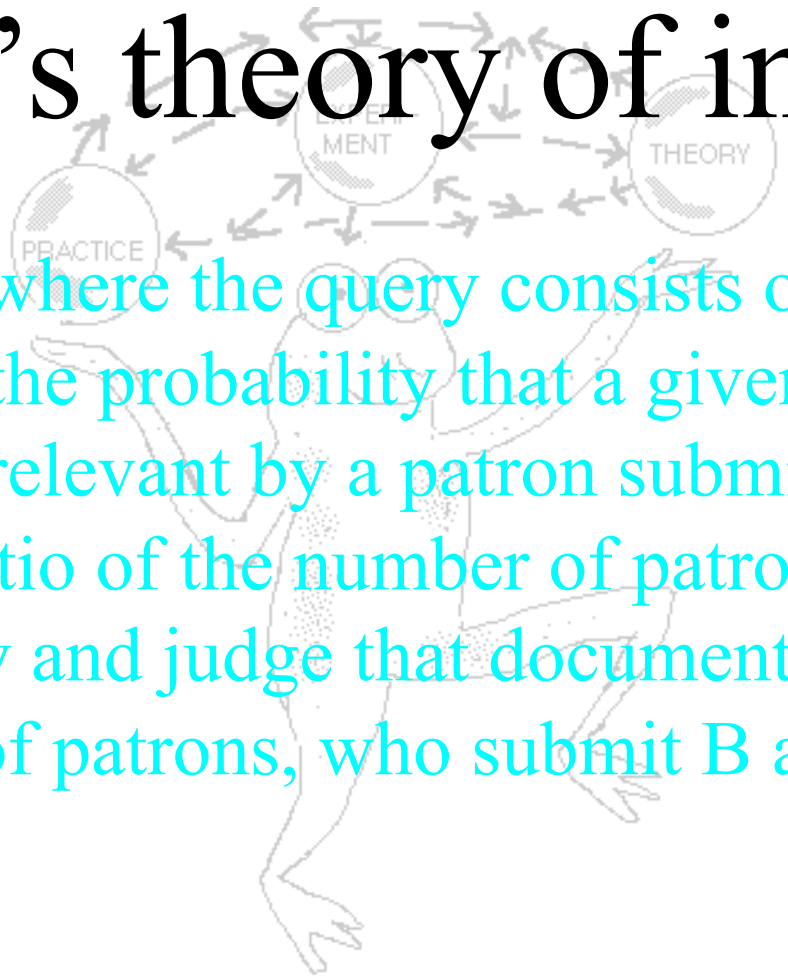



# The Nature of Models



CH, 2012

# Maron's theory of indexing

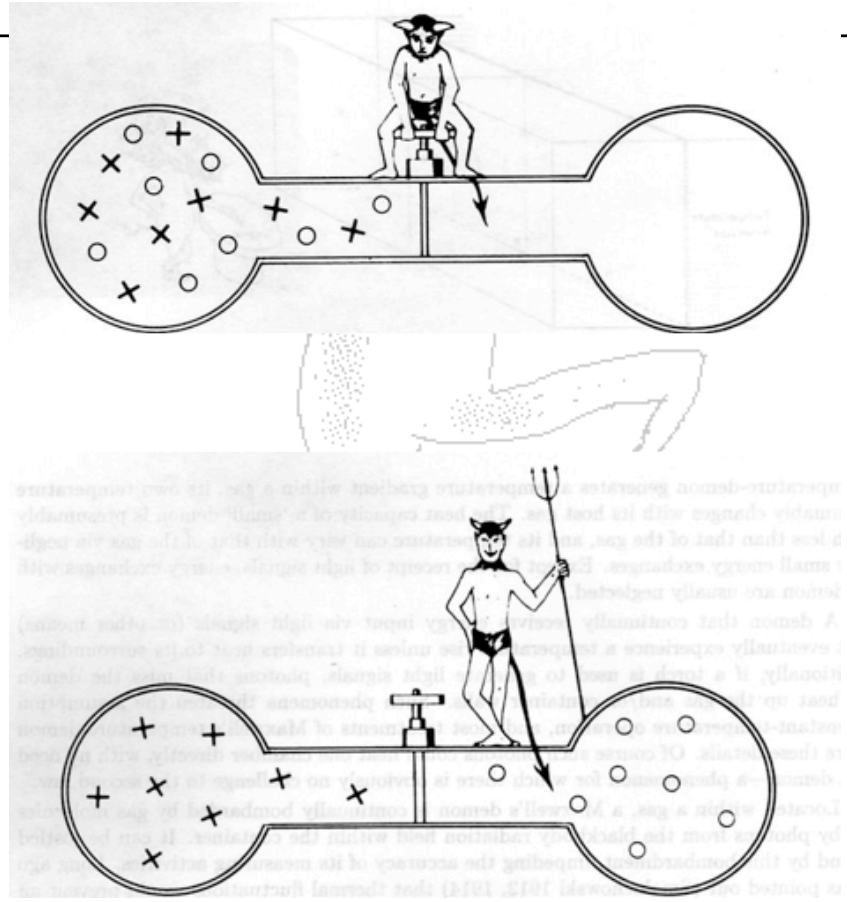



…..in the case where the query consists of single term, call it B, the probability that a given document will be judged relevant by a patron submitting B is simply the ratio of the number of patrons who submit B as their query and judge that document as relevant, to the number of patrons, who submit B as their search query

# What is this? quack

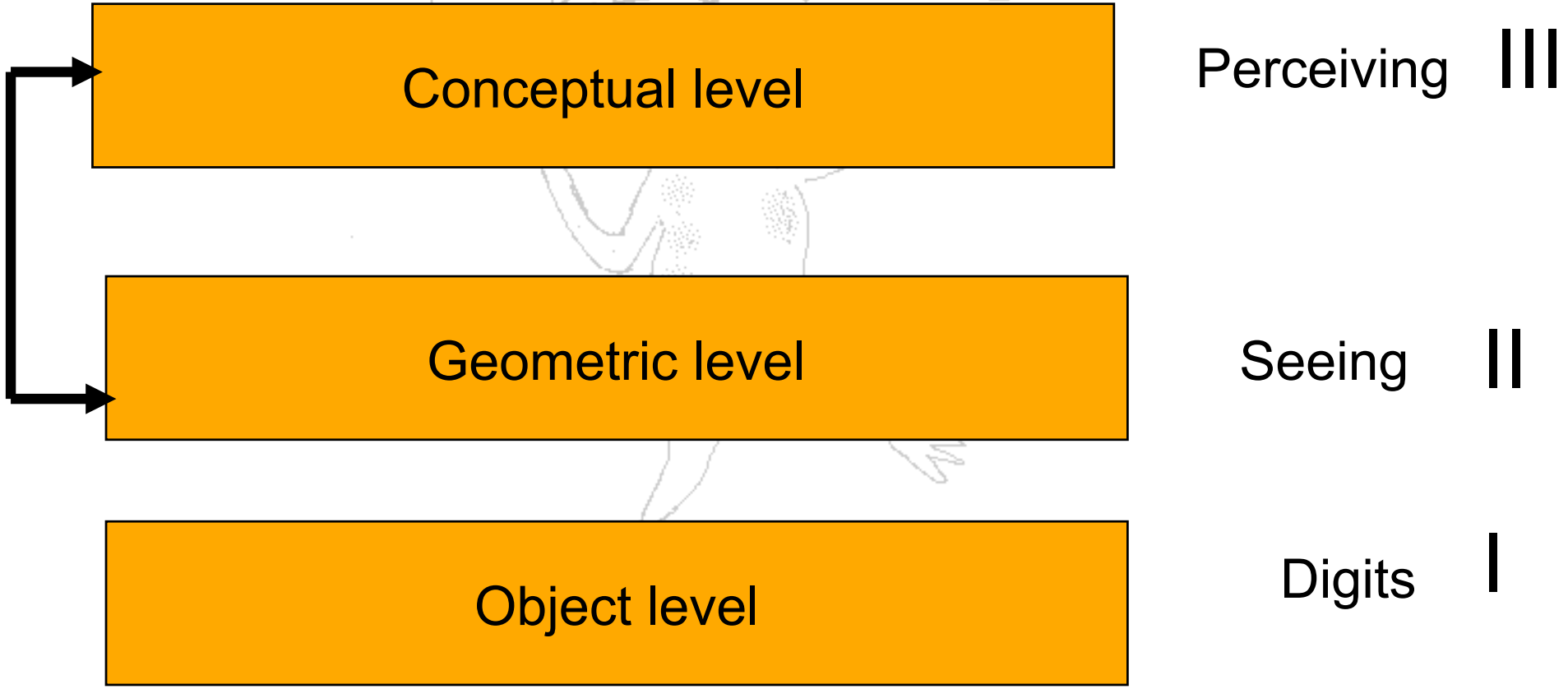
# …..and this?

# IR demon

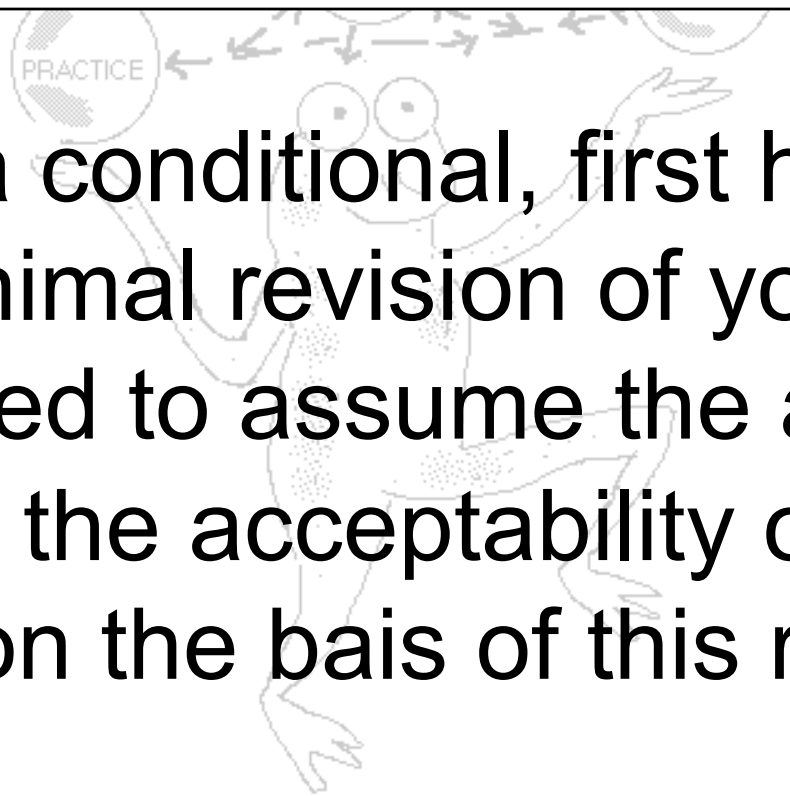


geometric view

# Representation

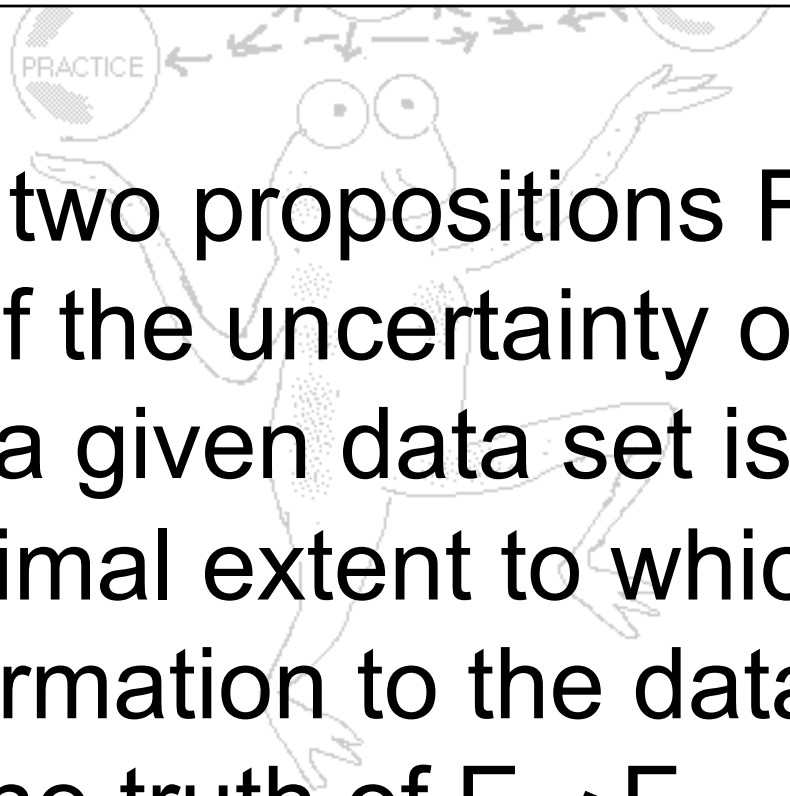


# The logical view



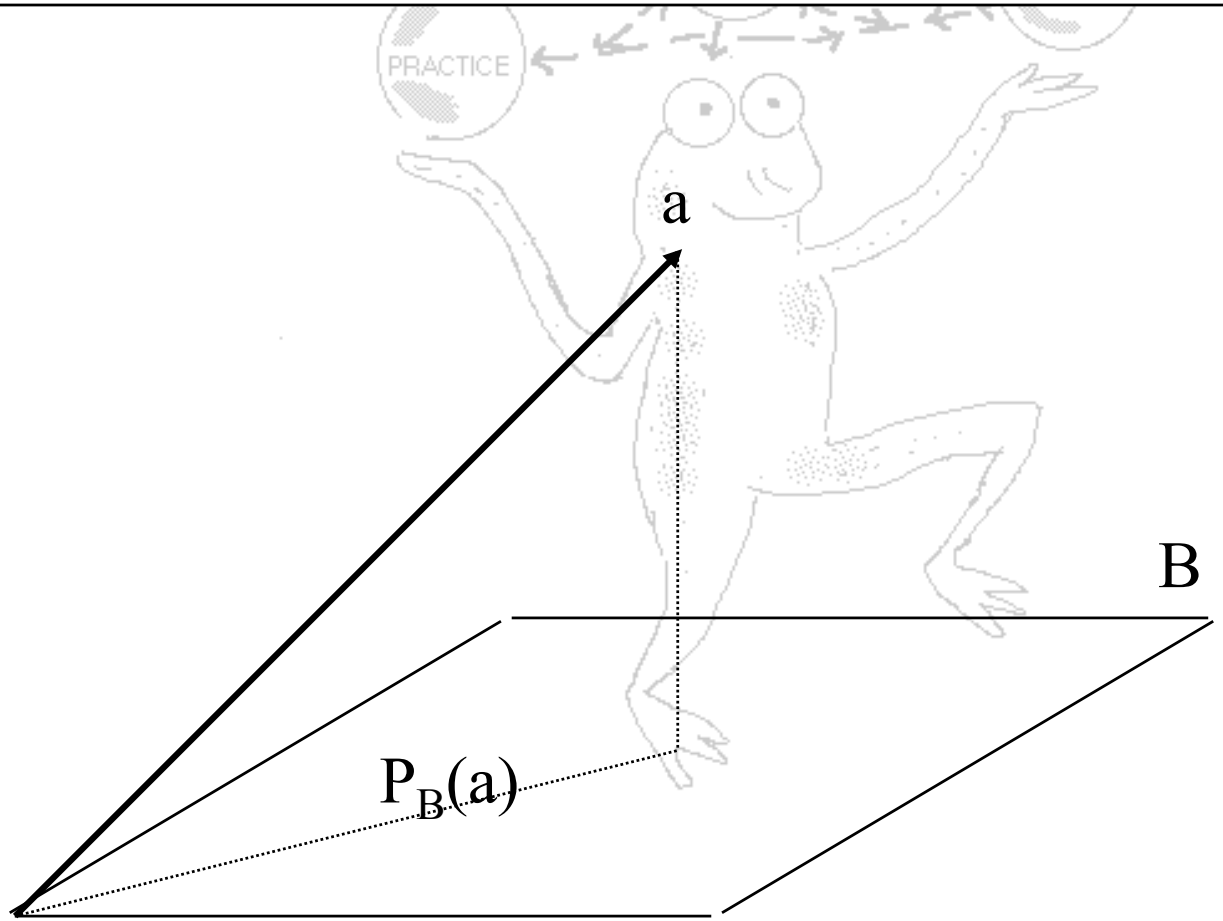

To evaluate a conditional, first hypothetically make the minimal revision of your stock of beliefs required to assume the antecedent. The evaluate the acceptability of the consequent on the bais of this revised body of beliefs.

# LUP in Hilbert space

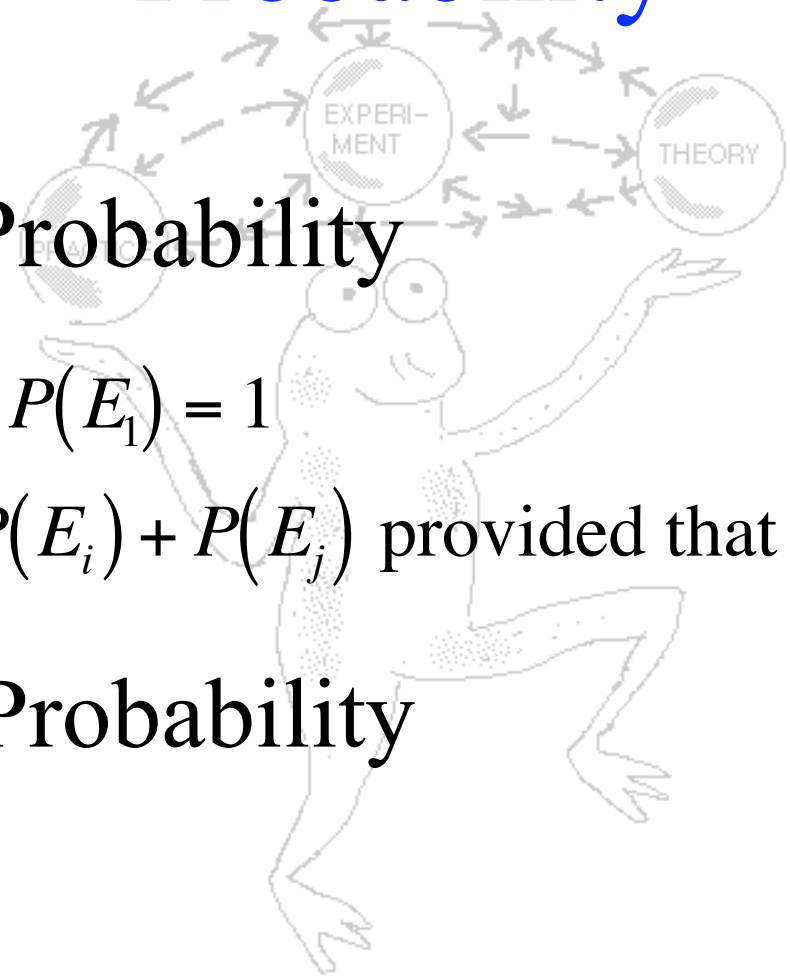



Given any two propositions F and E; a measure of the uncertainty of E→F relative to a given data set is determined by the minimal extent to which we have to add information to the data set, to establish the truth of E→F.

# LUP Example



# Probability




## Classical Probability

$P(E_0) = 0$ and $P(E_1) = 1$

$P(E_i \cup E_j) = P(E_i) + P(E_j)$ provided that $E_i \cap E_j = E_0$

## Quantum Probability

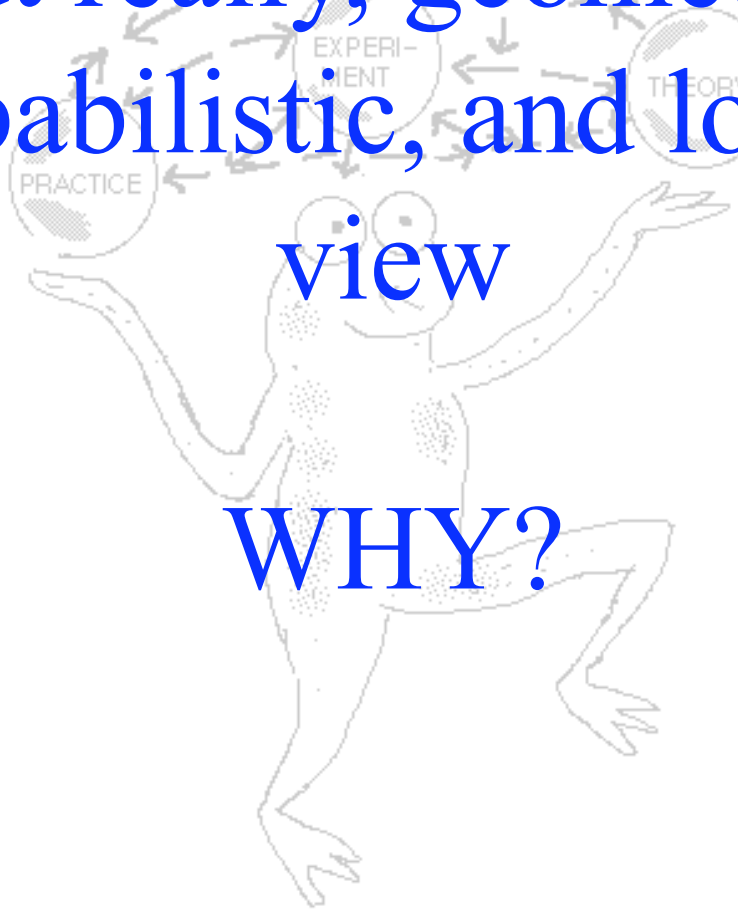
$\mu_\varphi(\Phi) = 0$

$\mu_\varphi(\mathbf{H}) = 1$

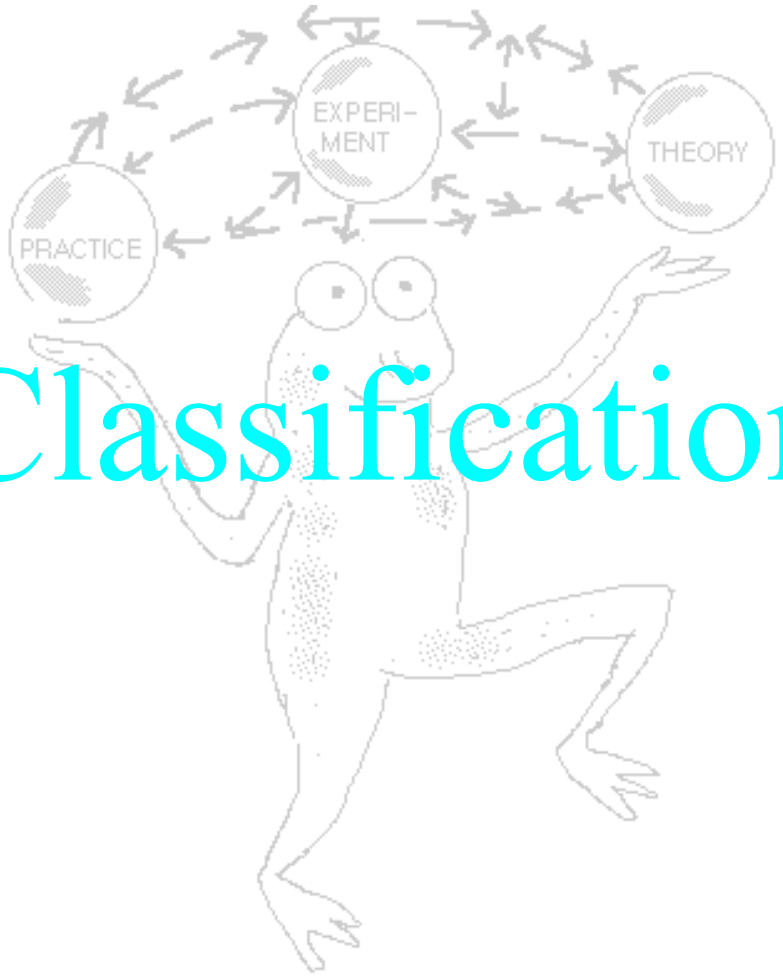
For subspaces $L_i$ and $L_j$, $\mu_\varphi(L_i \oplus L_j) = \mu_\phi(L_i) + \mu_\varphi(L_j)$ provided $L_i \cap L_j = \Phi$

Total Probability $P(A) = P(A|B)P(B) + P(A|B^c)P(B^c) + \_\_$ (interference)

# But really, geometric, probabilistic, and logical view

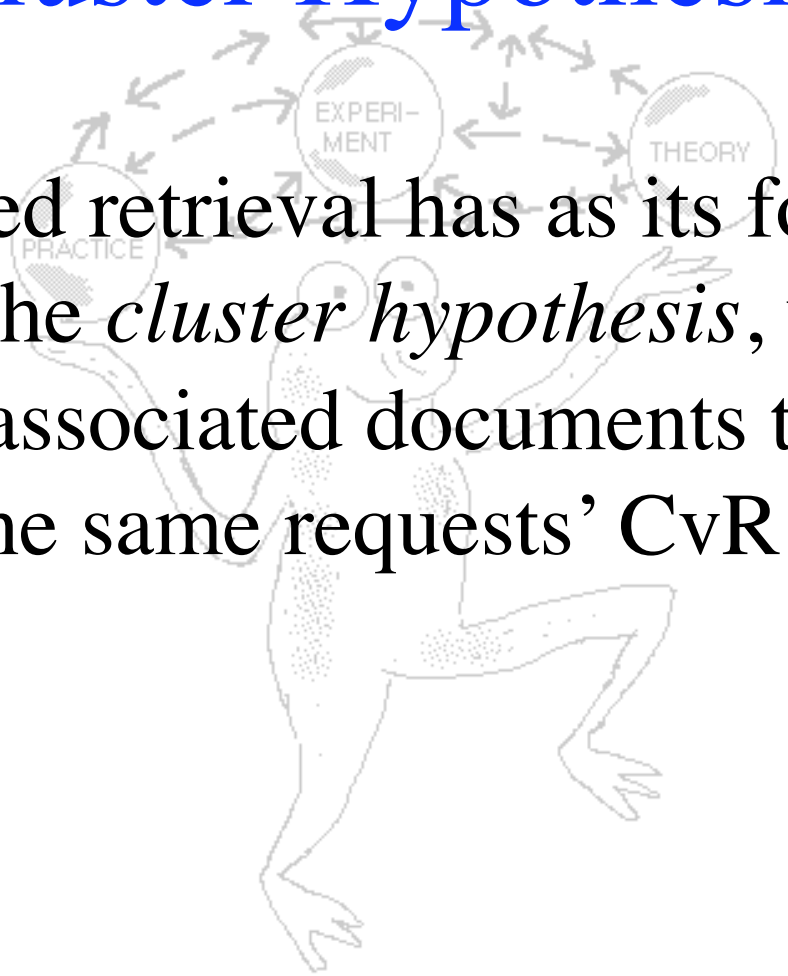



# WHY?



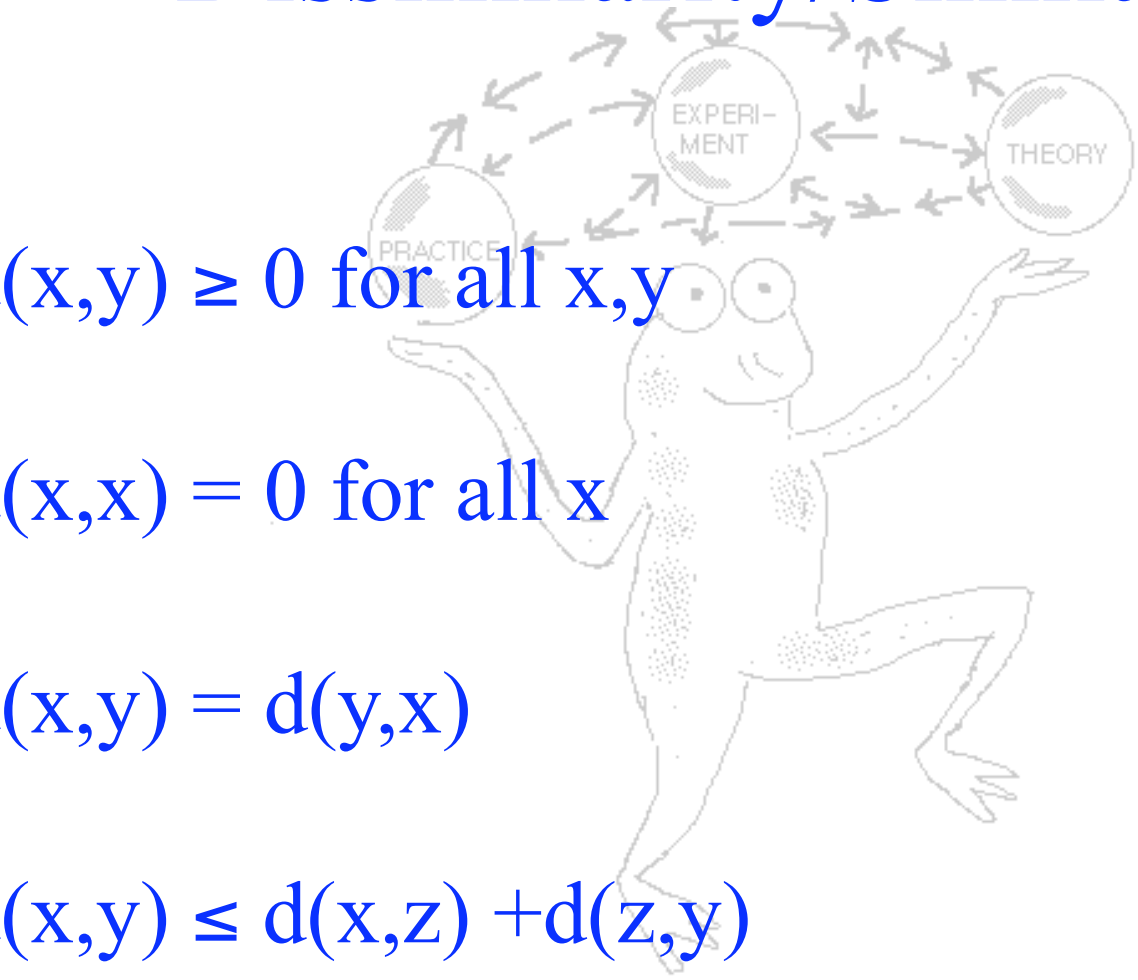

# Classification

# Cluster Hypothesis I




‘Cluster-based retrieval has as its foundation a hypothesis, the *cluster hypothesis*, which states that closely associated documents tend to be relevant to the same requests’ CvR, 1971

# Dissimilarity/Similarity




$d(x,y) \geq 0$ for all x,y

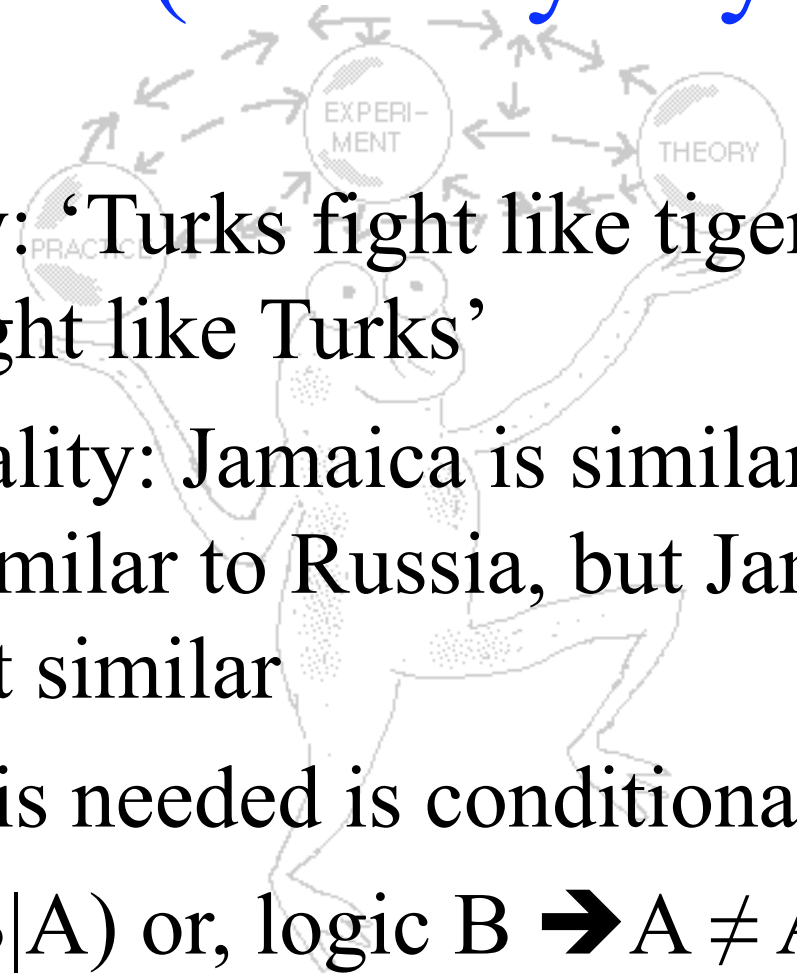
$d(x,x) = 0$ for all x

$d(x,y) = d(y,x)$

$d(x,y) \leq d(x,z) + d(z,y)$

$\{d(x,y) \leq \max [d(x,z), d(z,y)]\}$

# But (Tversky says)

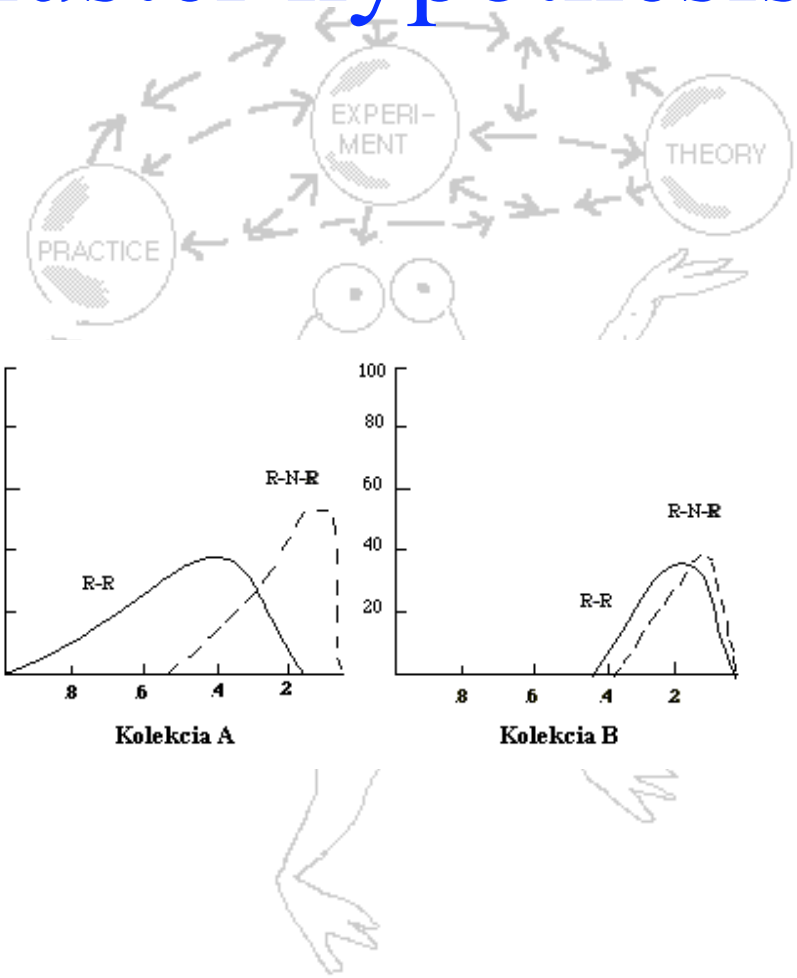



1. Symmetry: 'Turks fight like tigers', not 'Tigers fight like Turks'
2. Δ – inequality: Jamaica is similar to Cuba, Cuba is similar to Russia, but Jamaica and Russia not similar

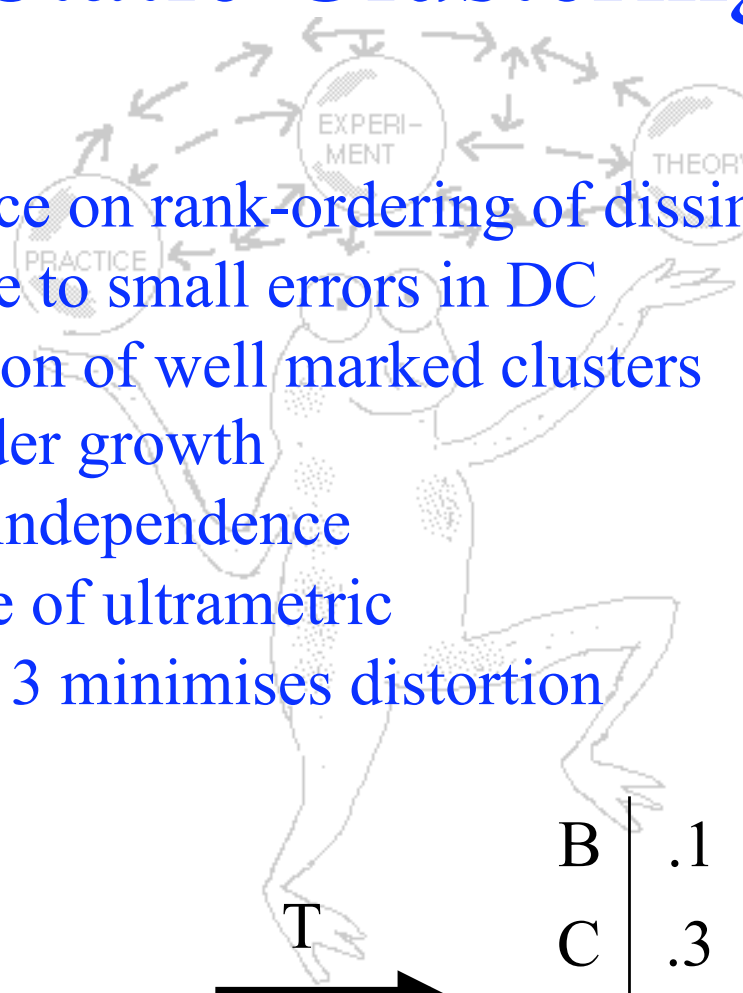
Maybe what is needed is conditional probability

$P(A|B) \neq P(B|A)$ or, logic B ➔ A ≠ A ➔ B

# Cluster hypothesis II



# Static Clustering



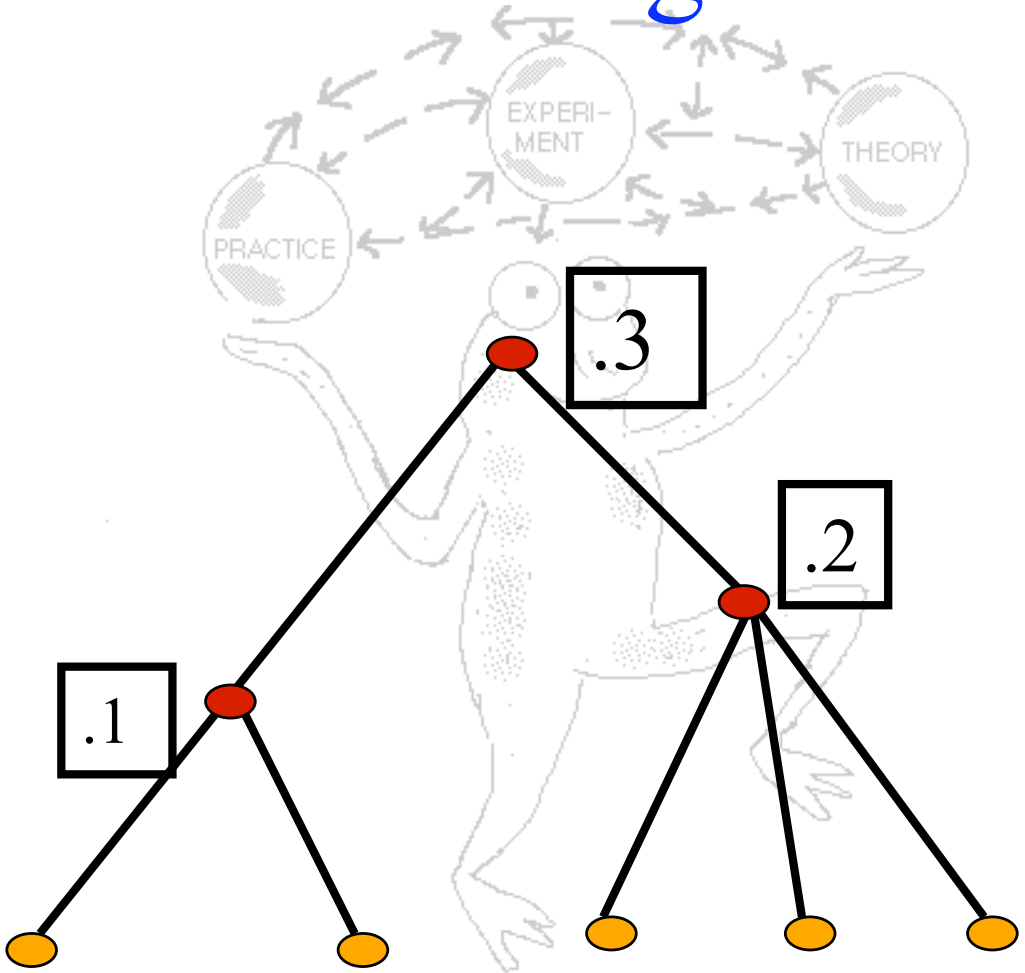

1. dependence on rank-ordering of dissimilarity
2. insensitive to small errors in DC
3. preservation of well marked clusters
4. stable under growth
5. labelling independence
6. invariance of ultrametric
7. subject to 3 minimises distortion

| | | | | |
|---|---|---|---|---|
| B | .1 | | | |
| C | .4 | .2 | | |
| D | .3 | .3 | .3 | |
| E | .4 | .4 | .4 | .1 |
| | A | B | C | D |

T

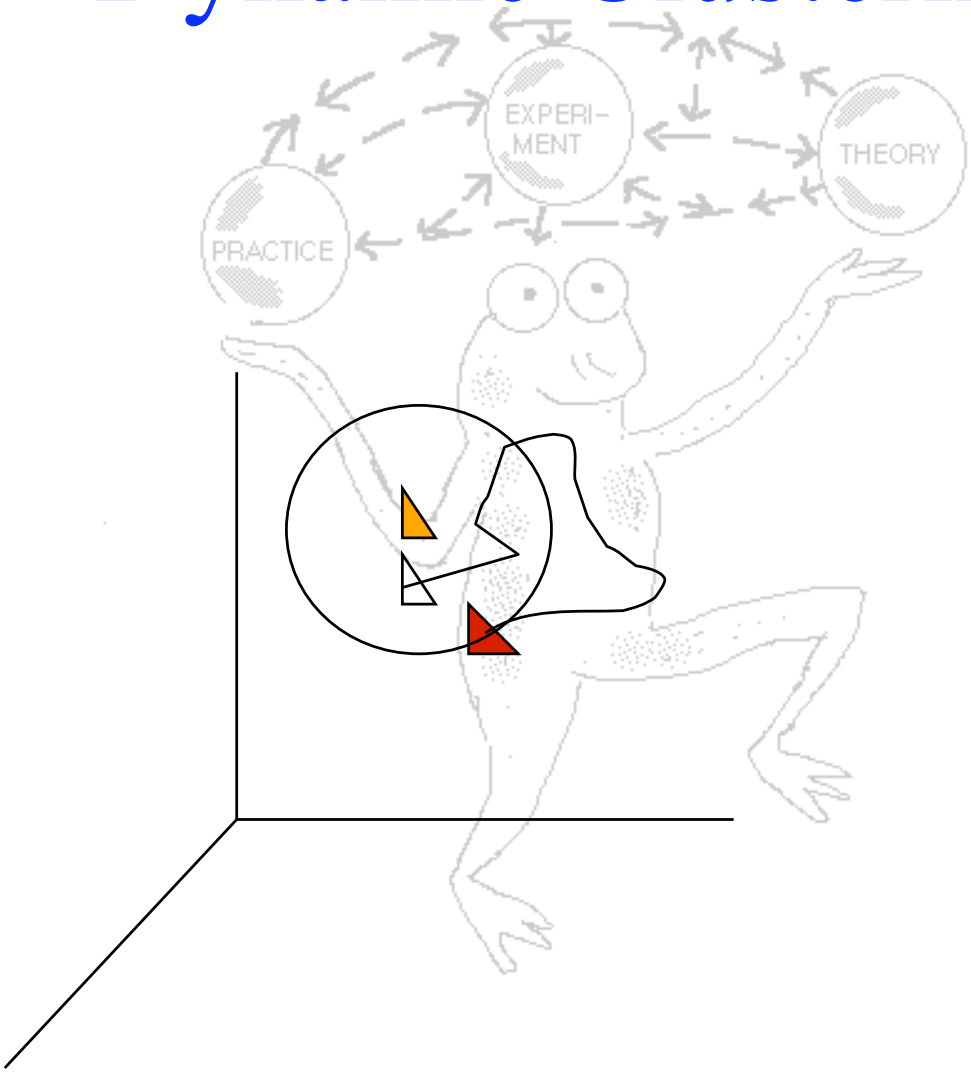
| | | | | |
|---|---|---|---|---|
| B | .1 | | | |
| C | .3 | .3 | | |
| D | .3 | .3 | .2 | |
| E | .3 | .3 | .2 | .2 |
| | A | B | C | D |

# Dendrogram




Spanning tree?

# Dynamic Clustering

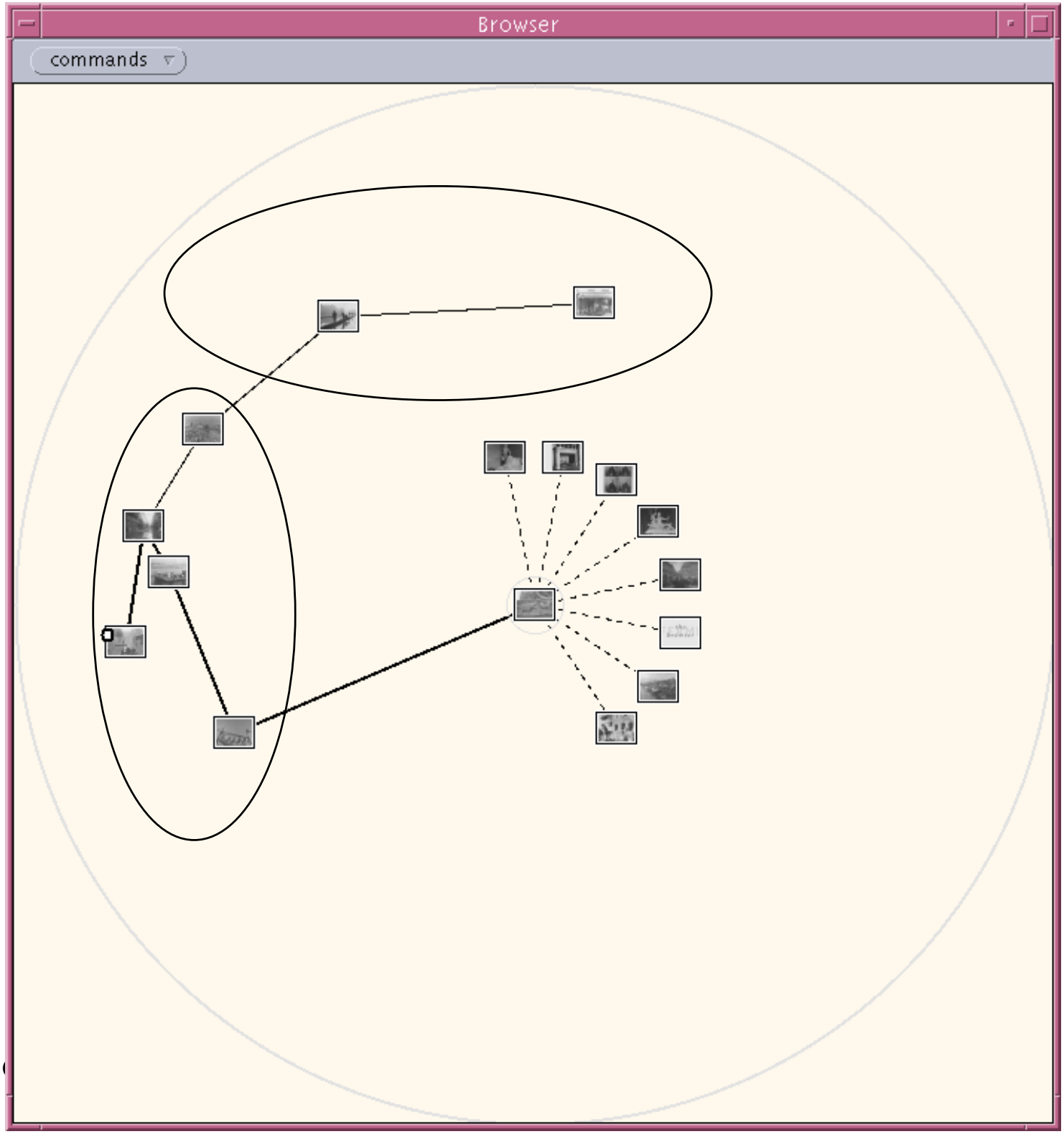


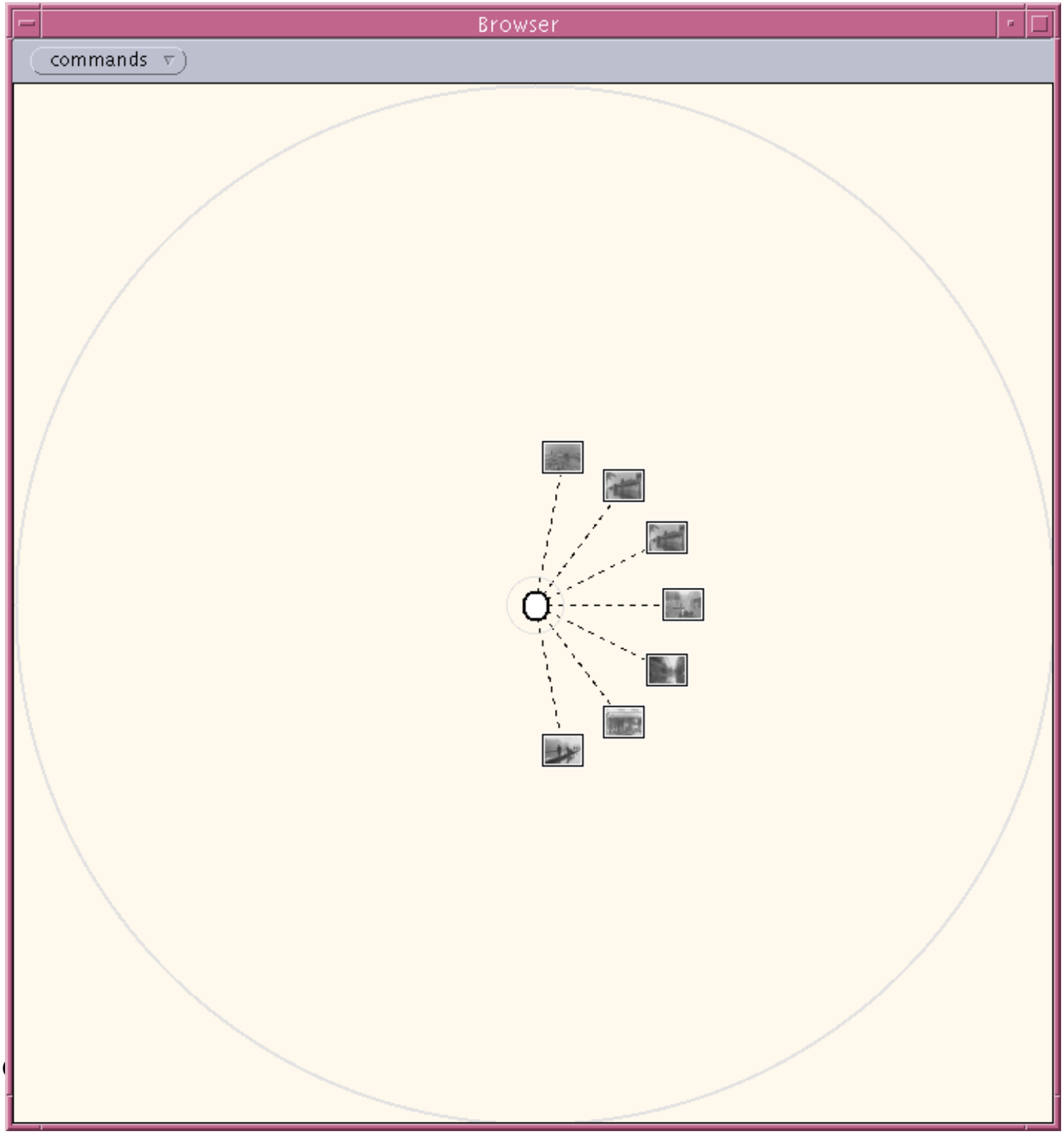

Hilbert-Schmidt: (A,B) = trace(A'B)

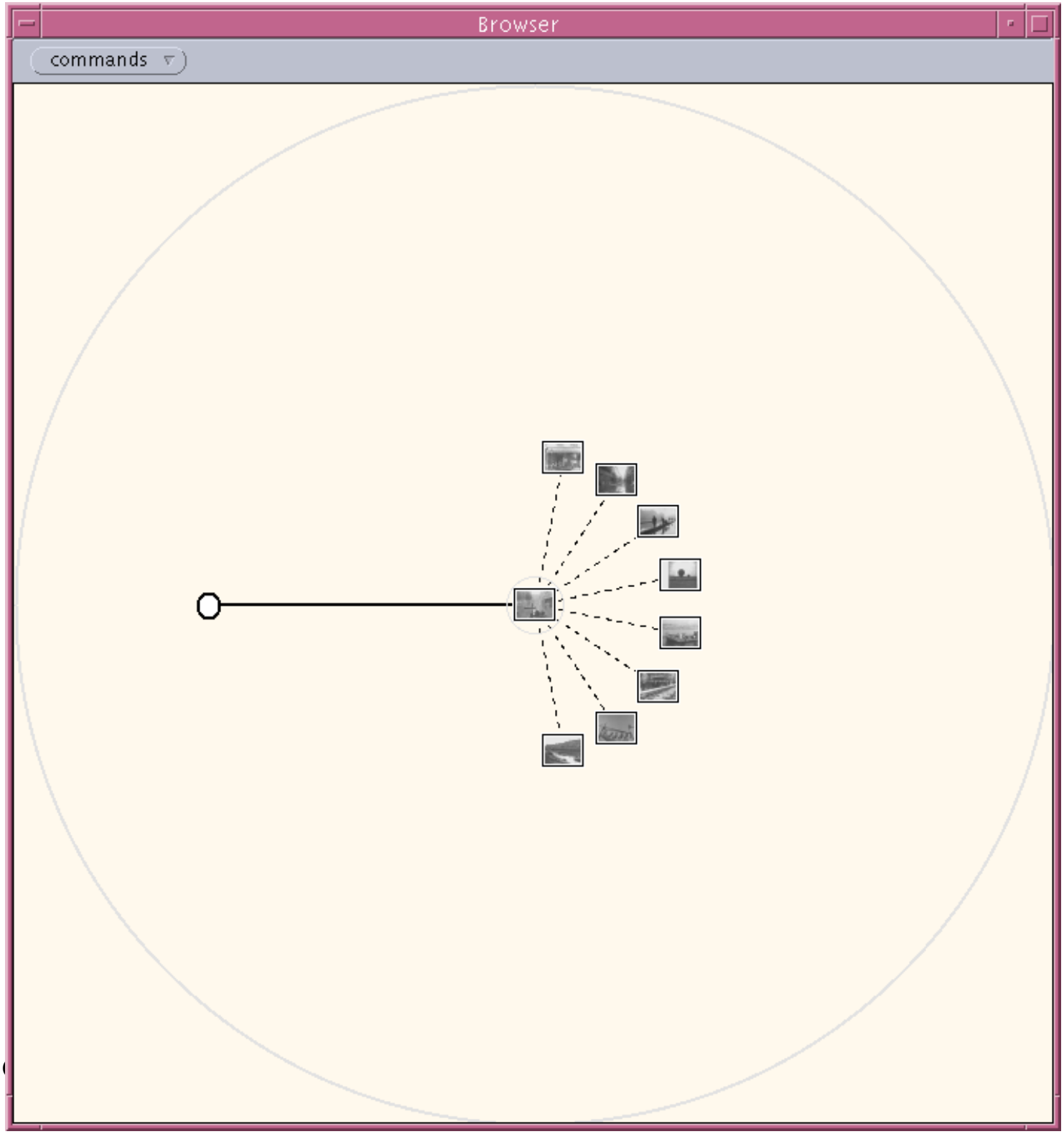
# Pseudo Relevance Feedback

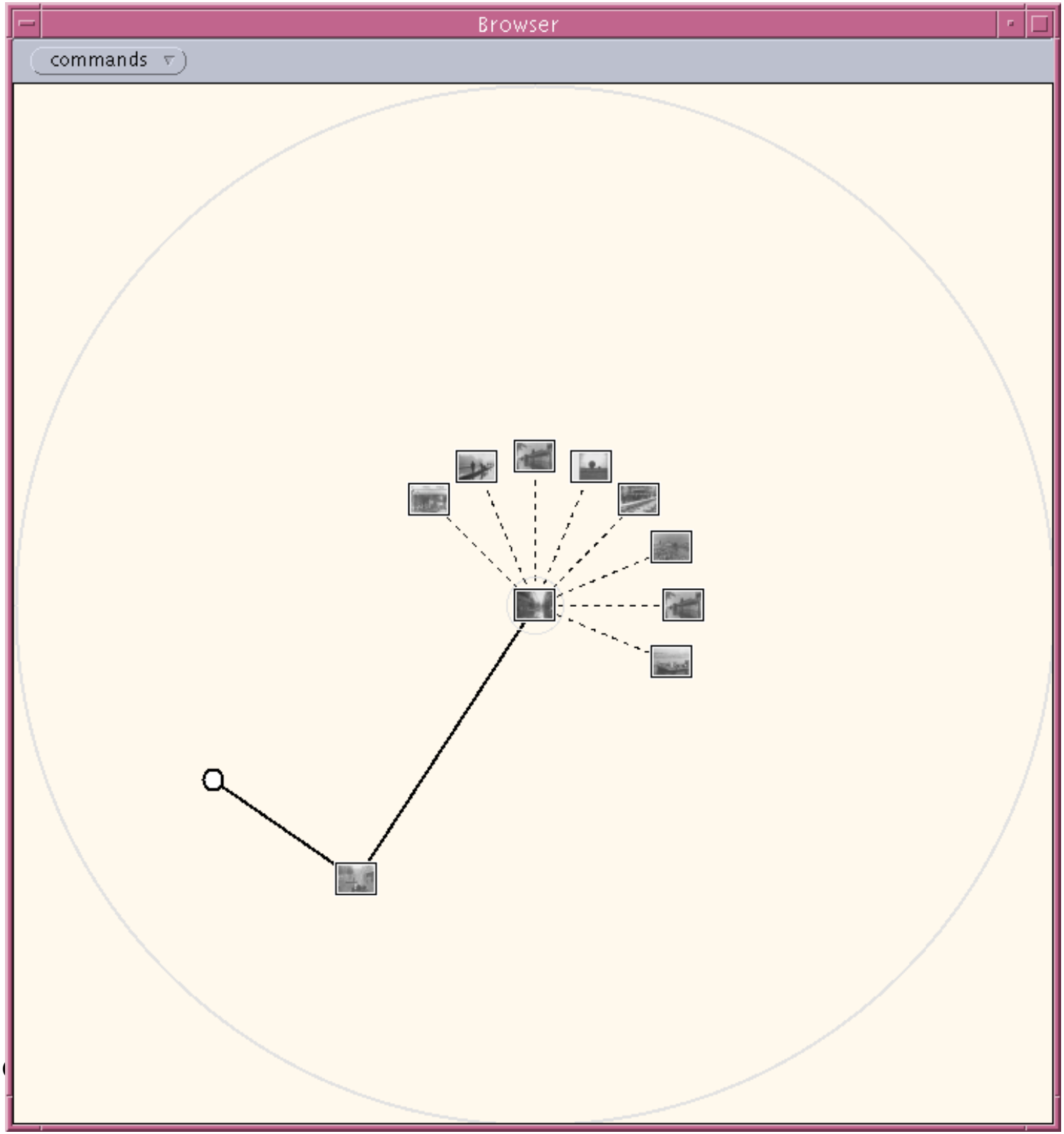

How does it differ from Relevance Feedback?

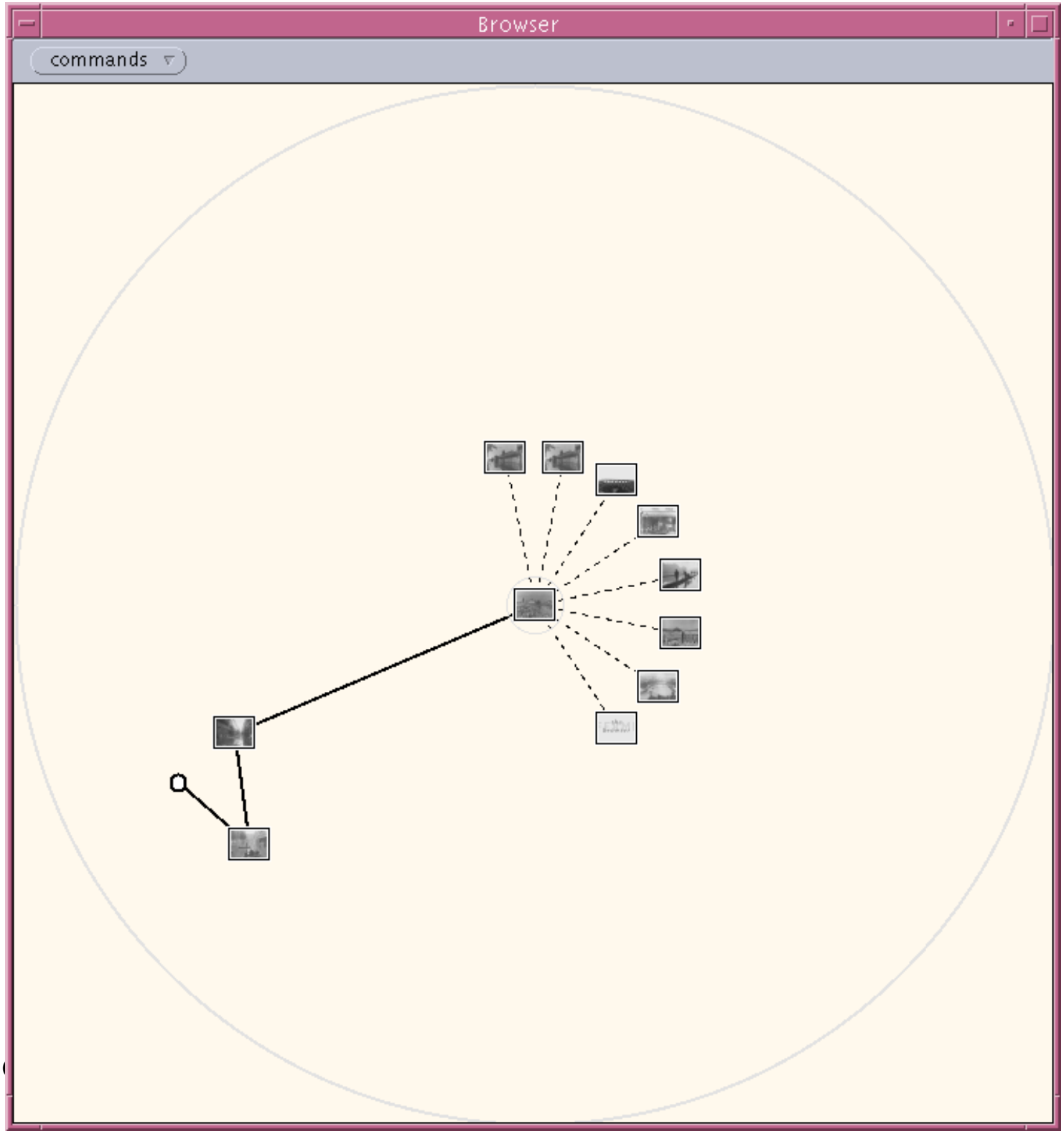
Also, think of Context

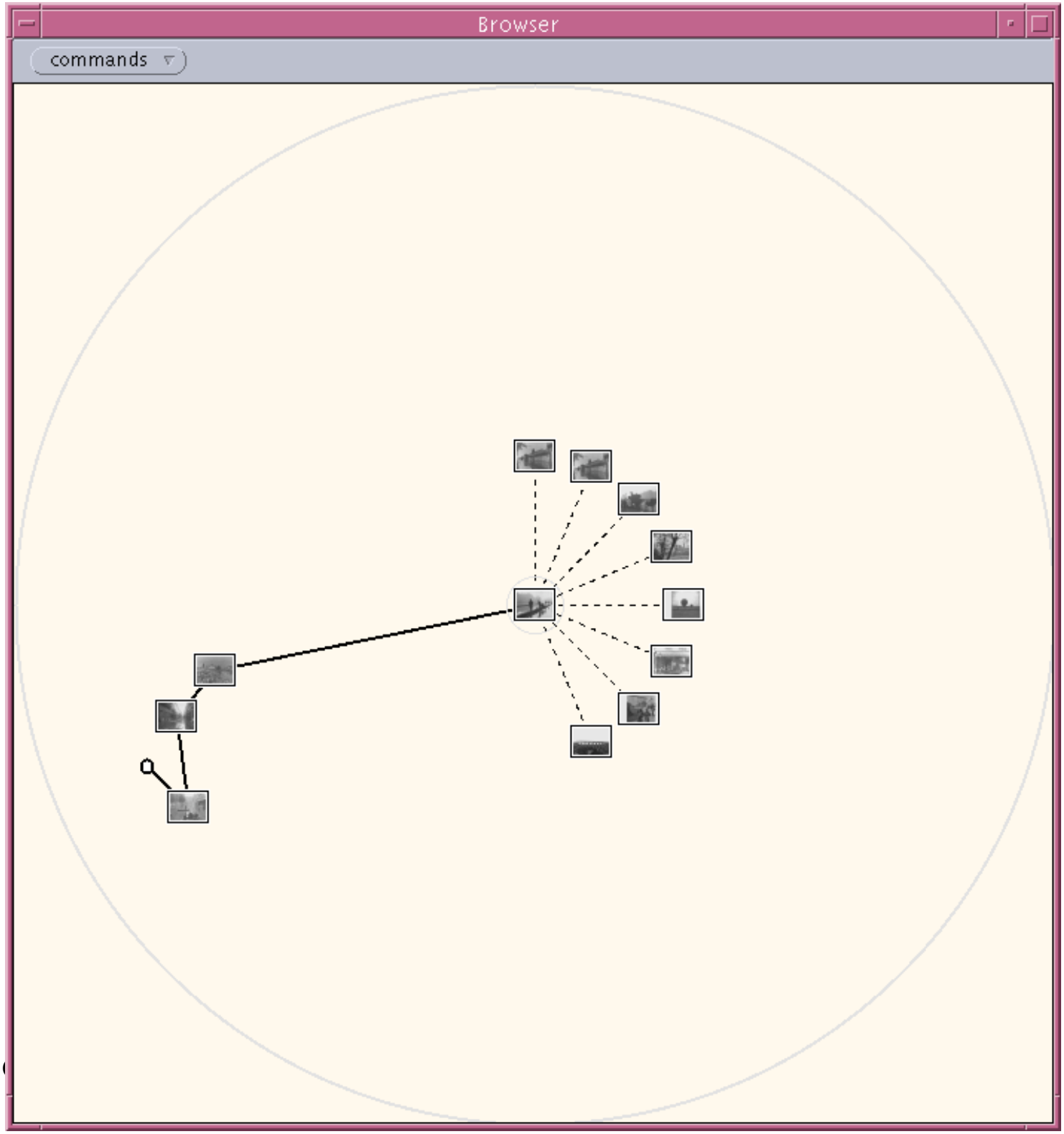

Browser
commands

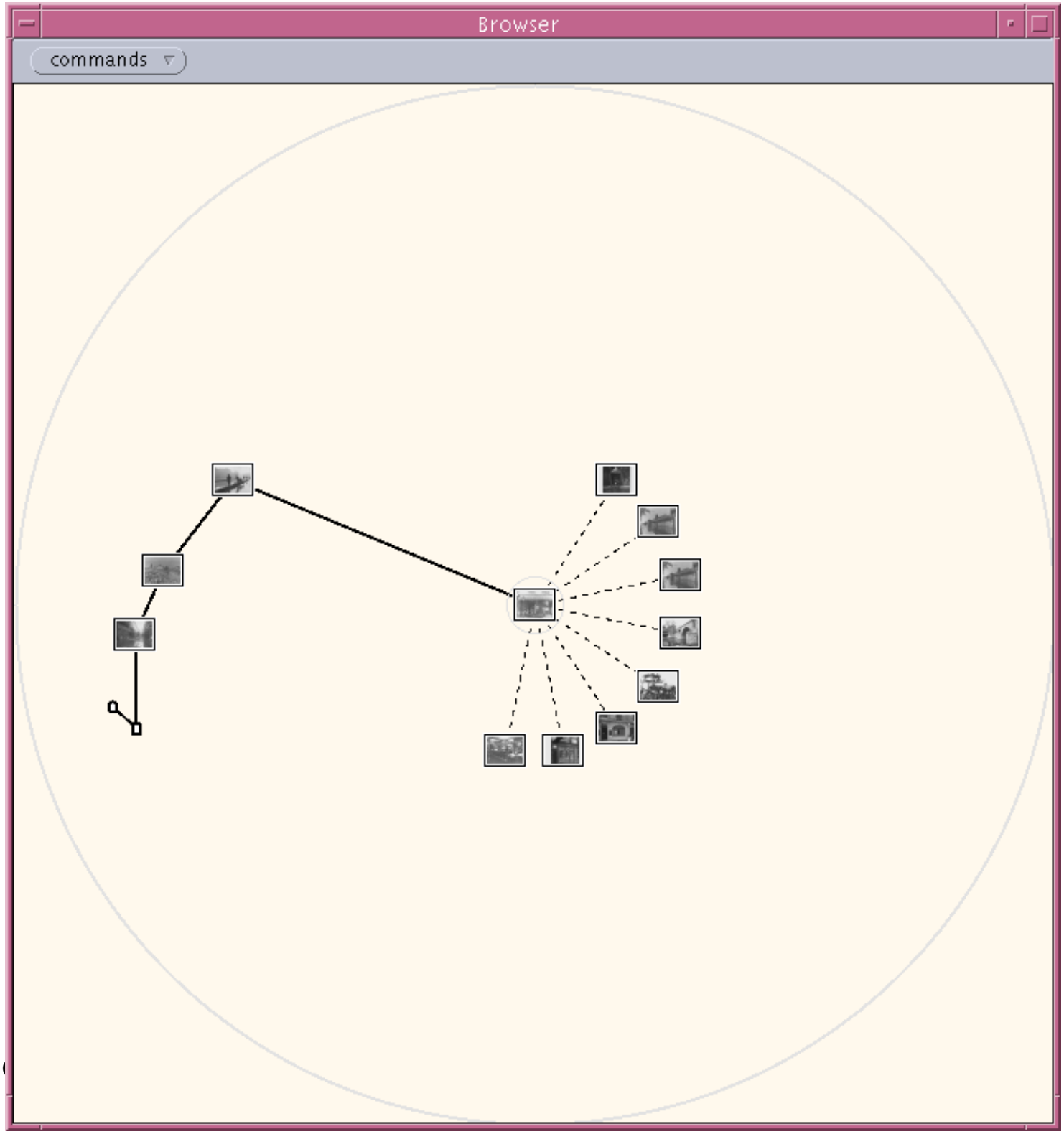

Browser
commands

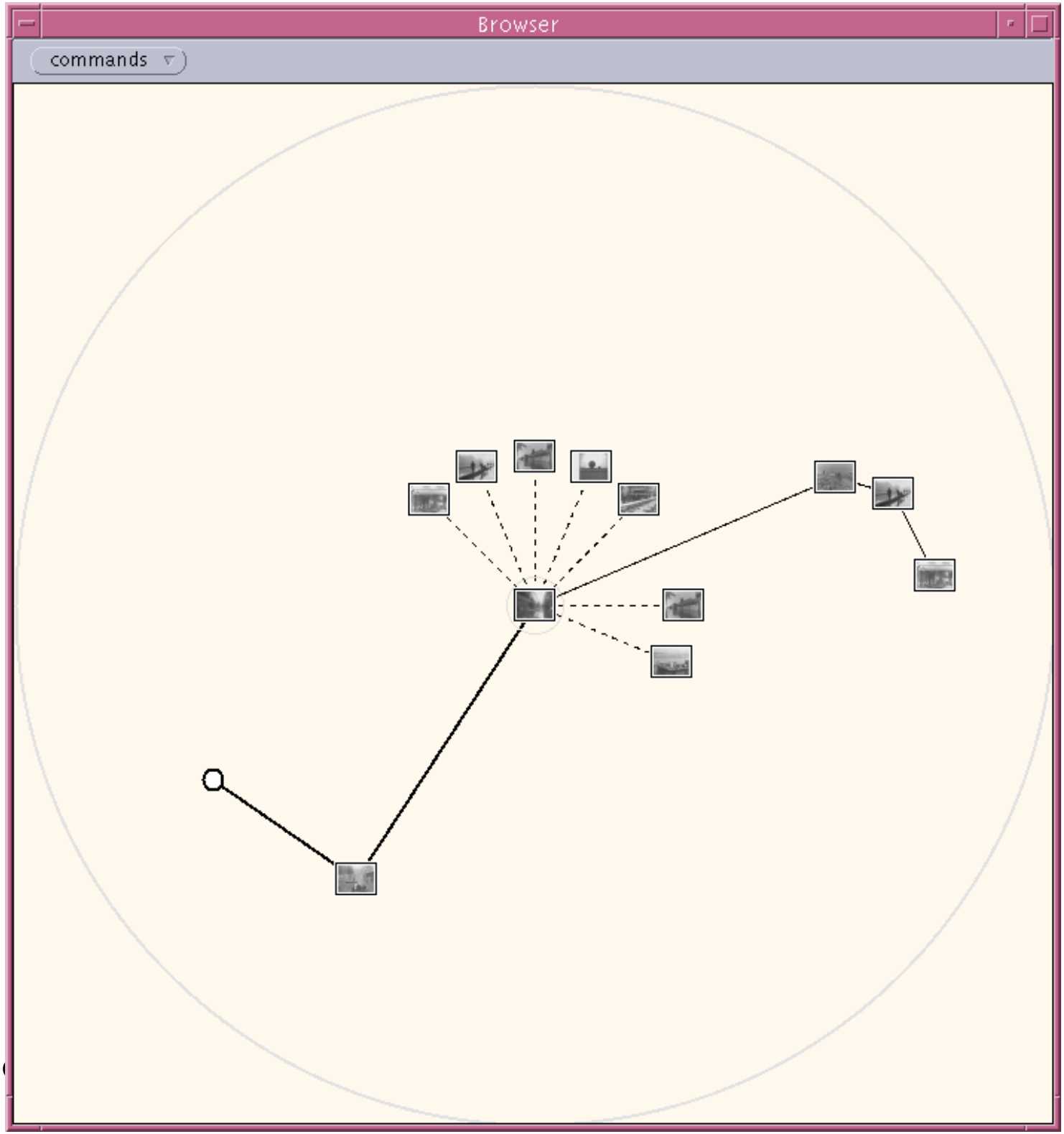

Browser
commands

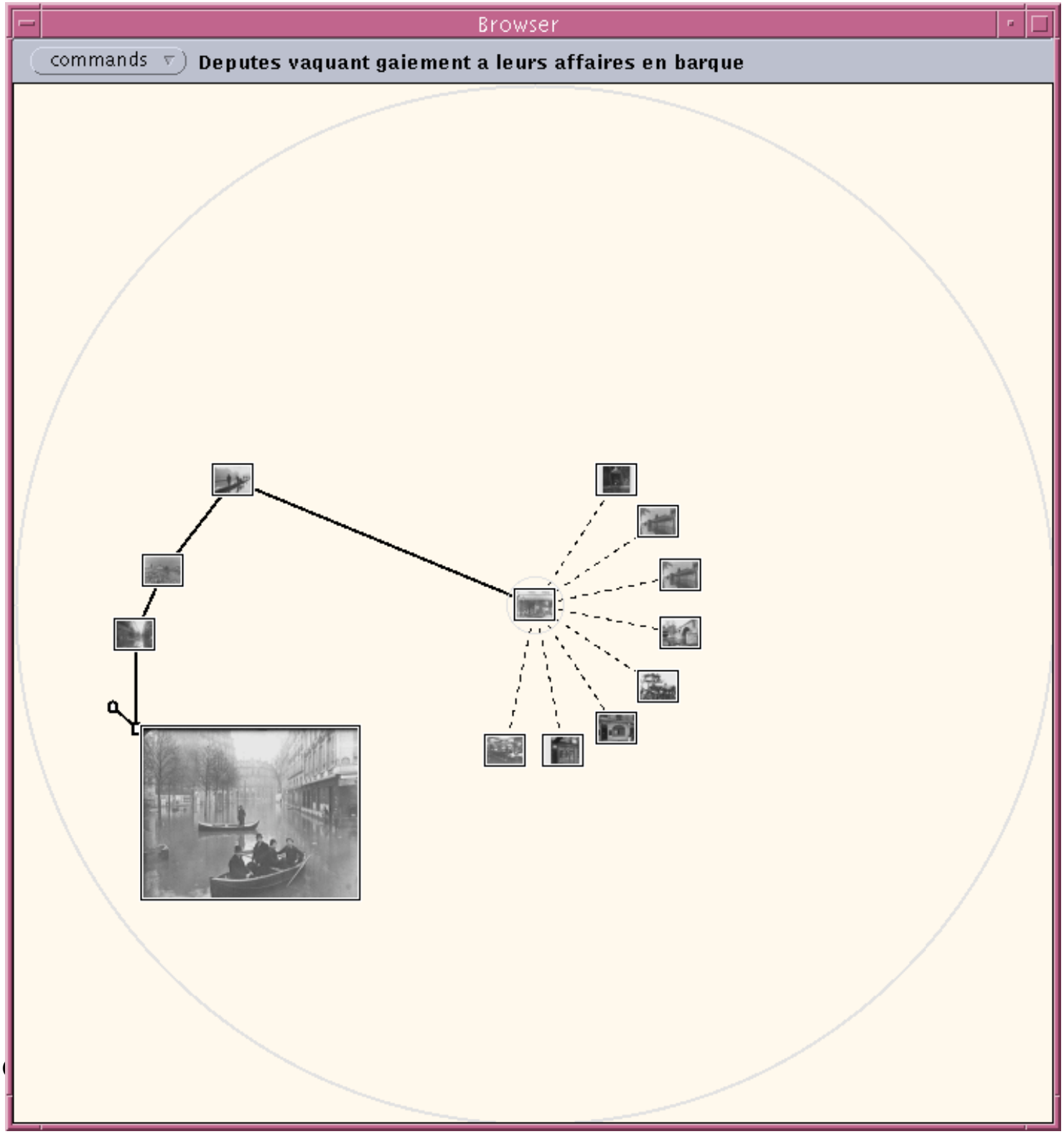

Browser
commands

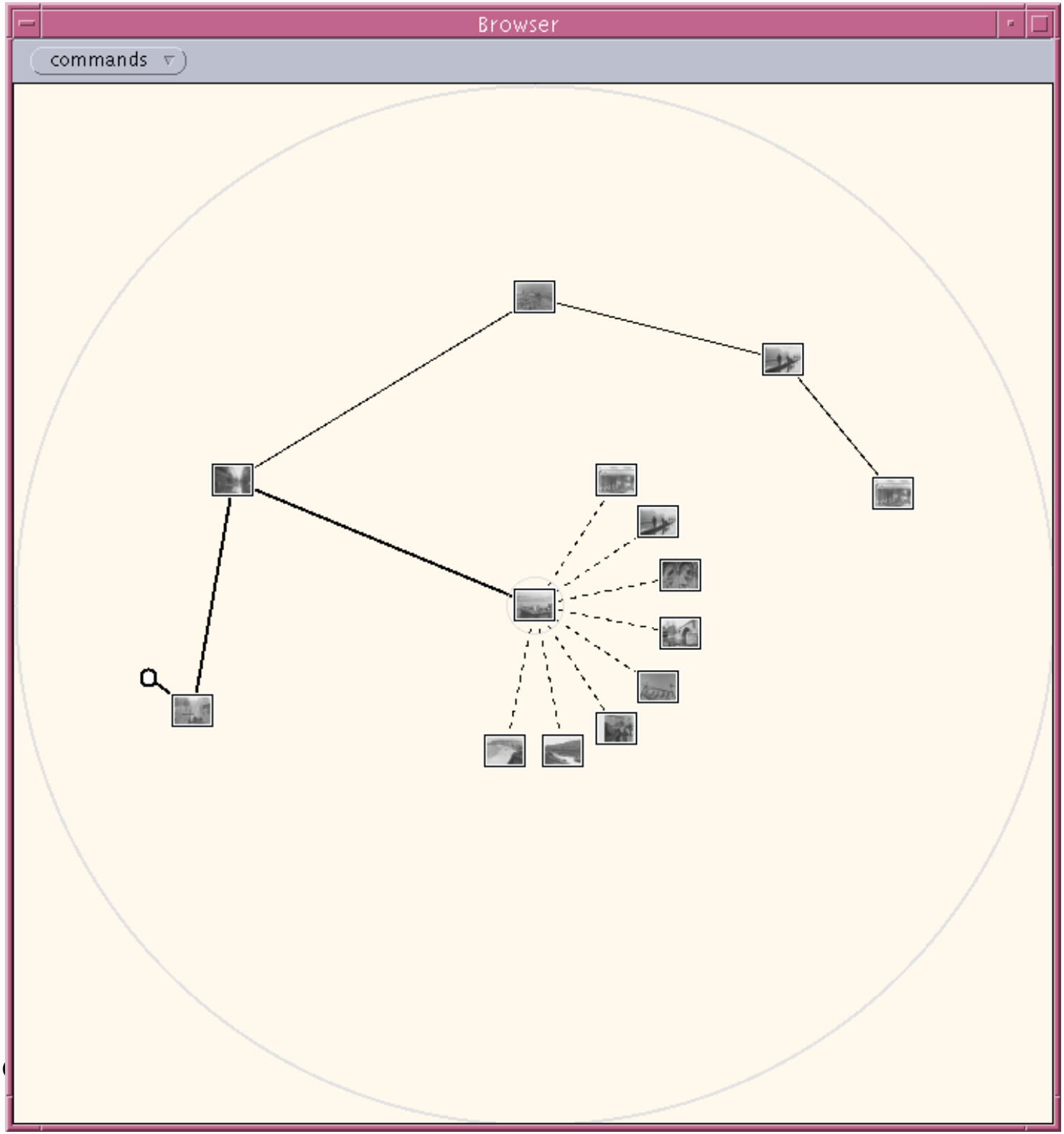

Browser
commands

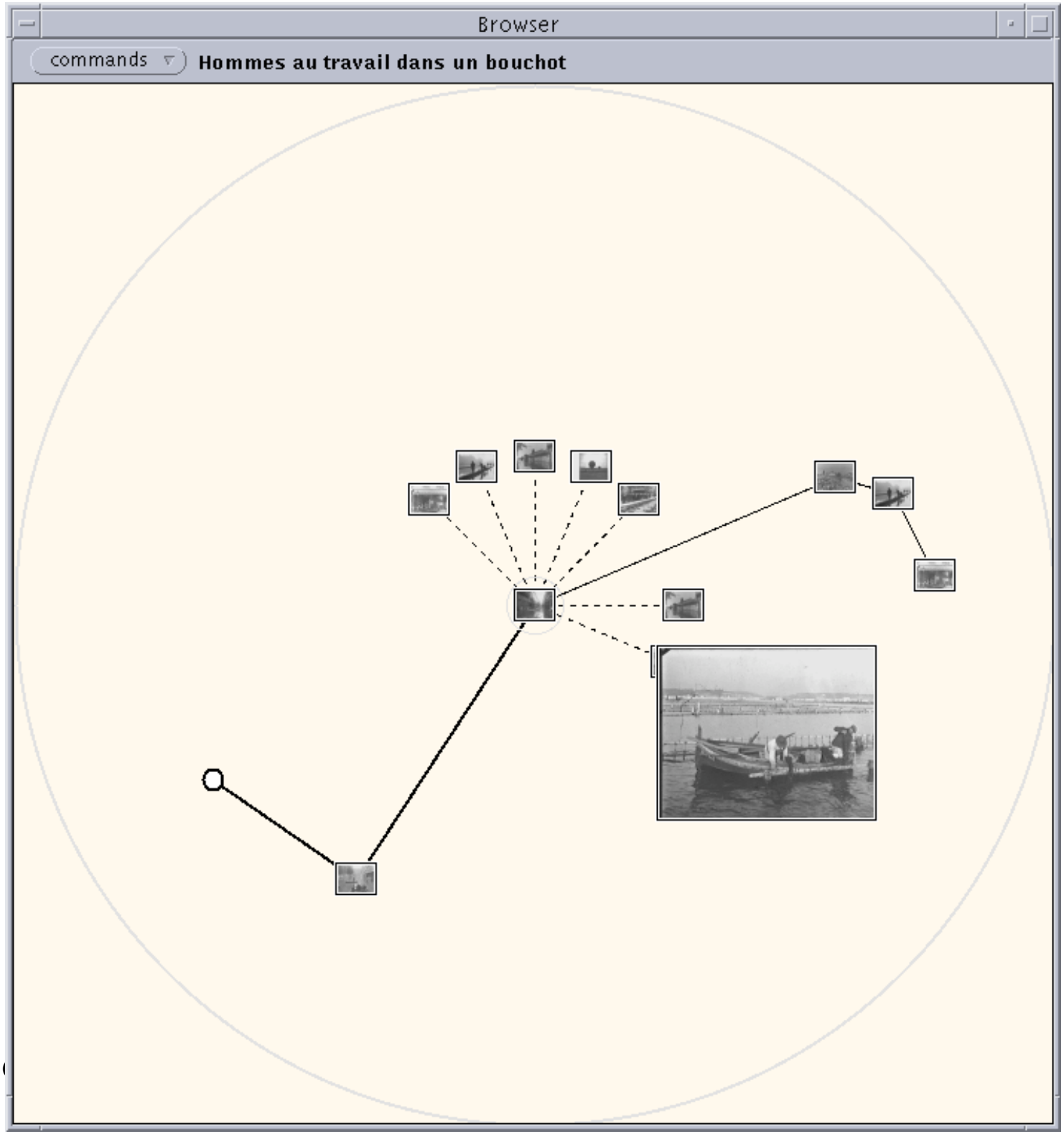

Browser
commands

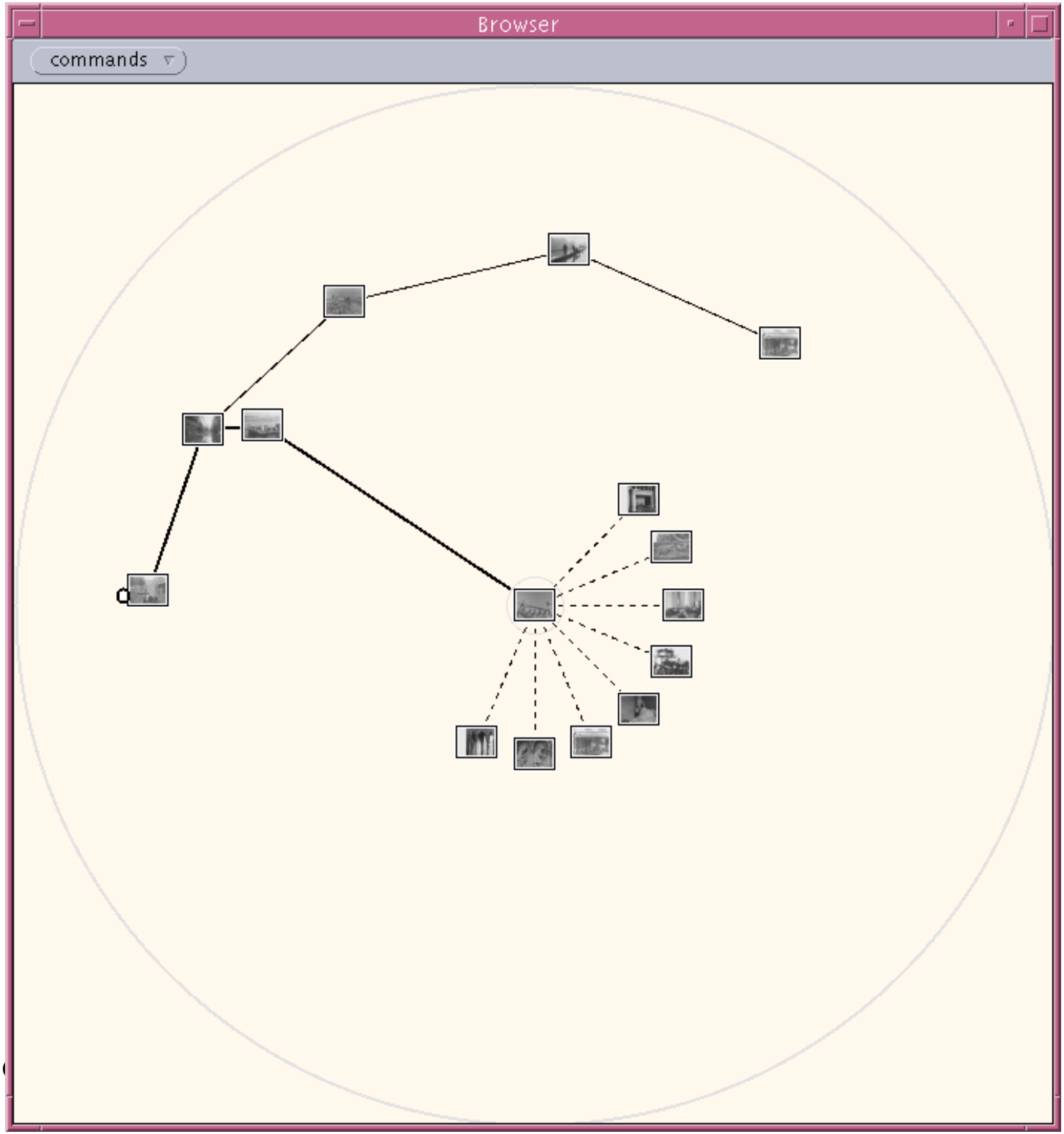

Browser
commands

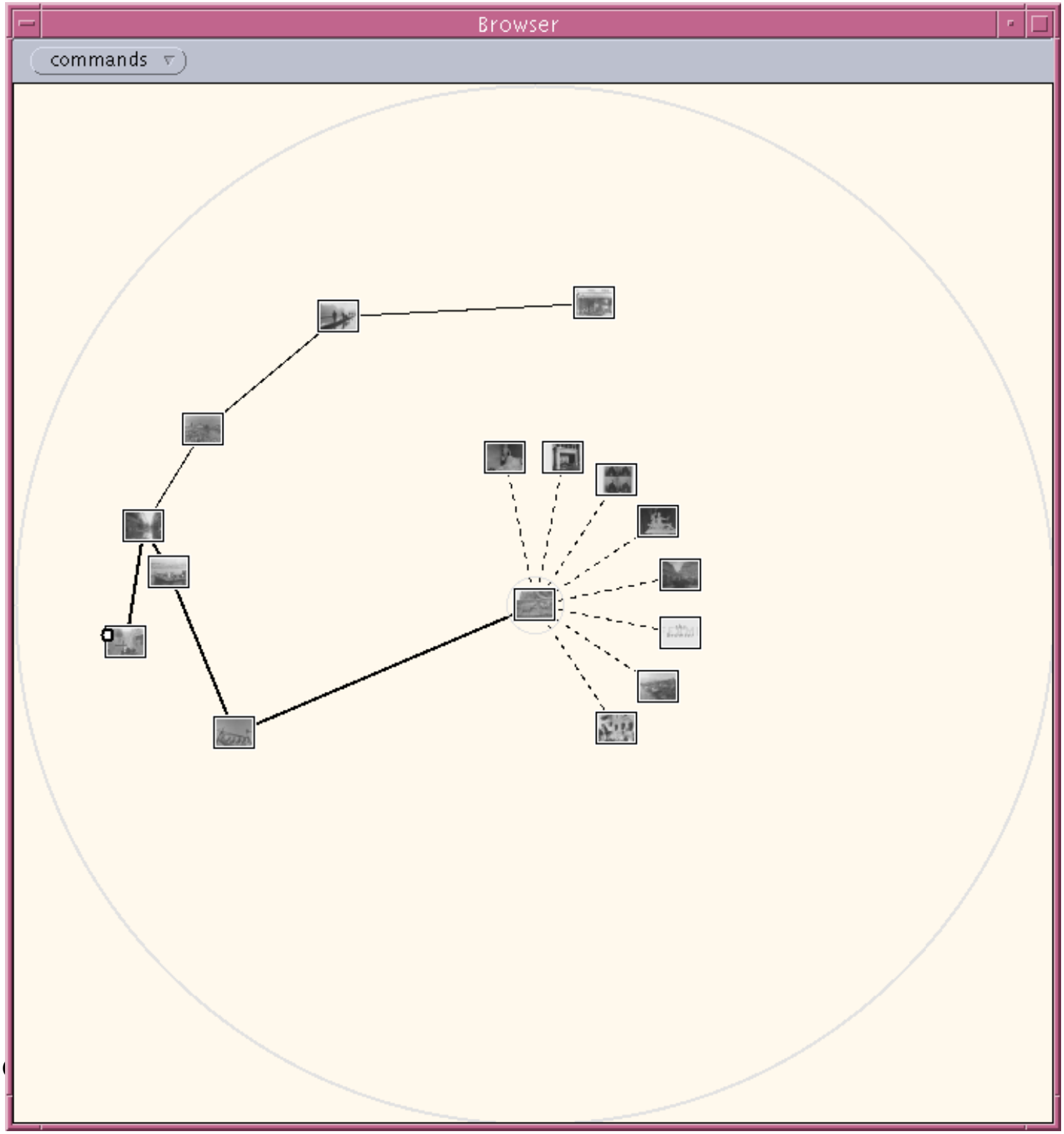

Browser
commands

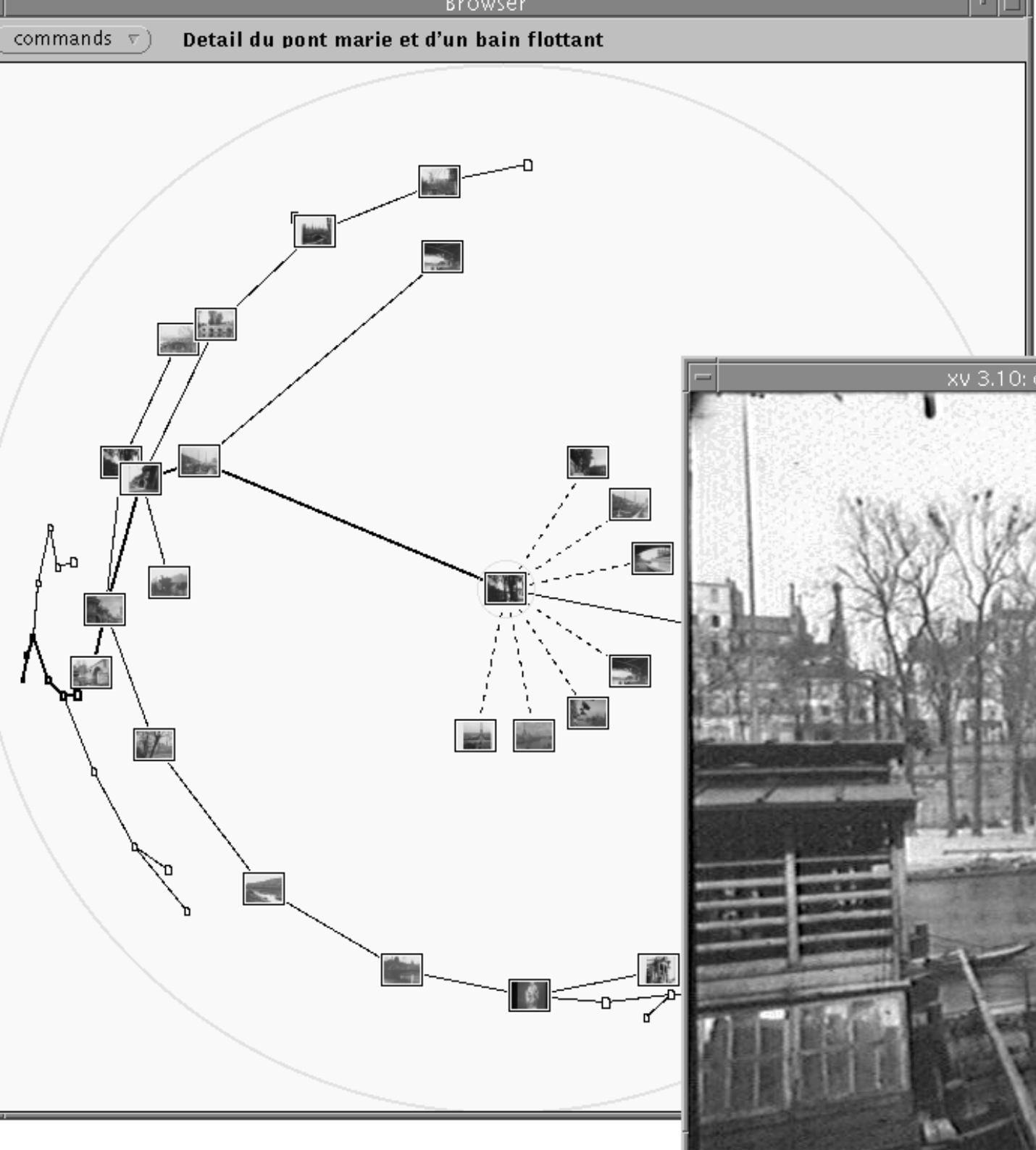

Browser
commands
Deputes vaquant gaiement a leurs affaires en barque

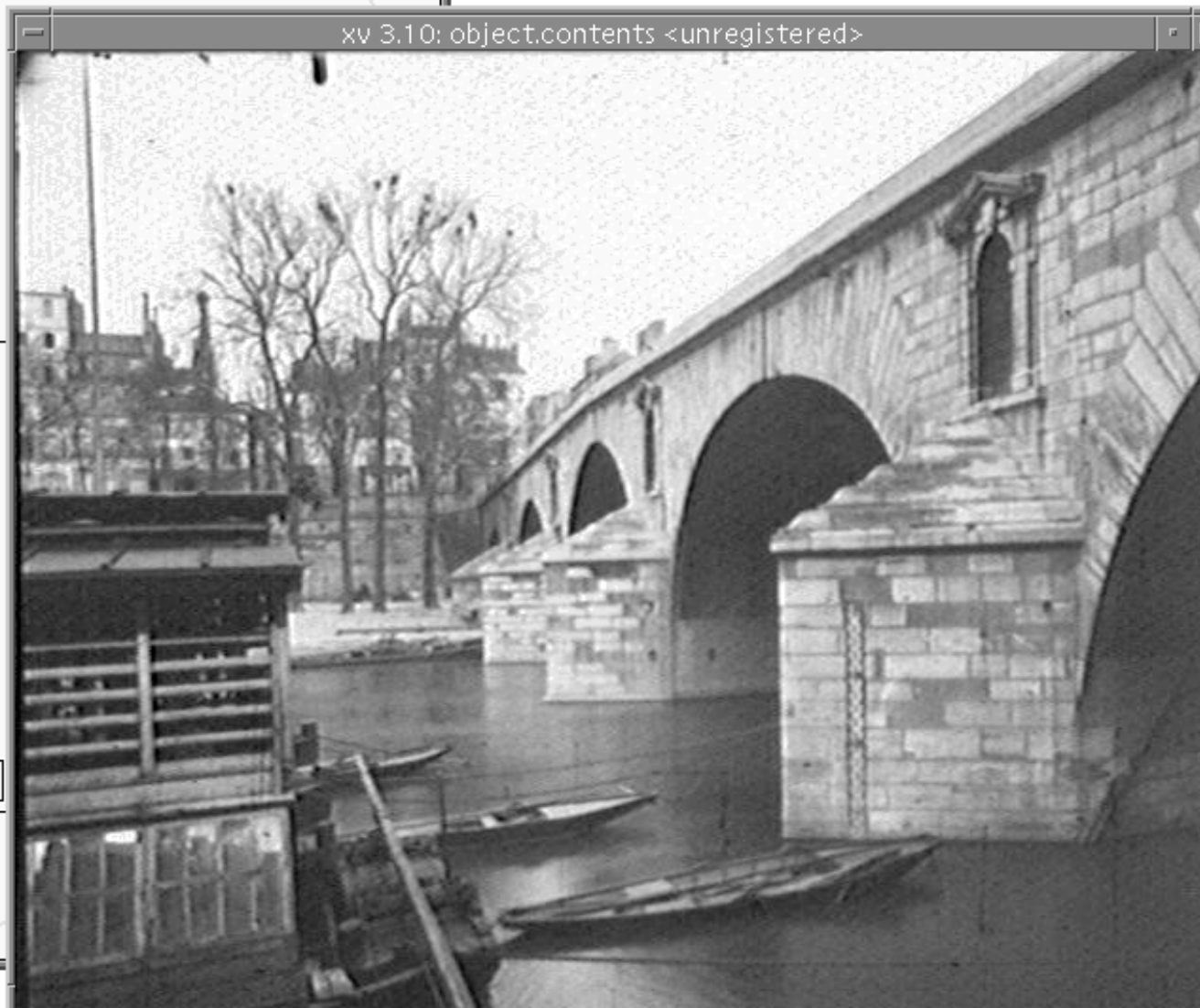

Browser
commands


Browser
commands
Hommes au travail dans un bouchot

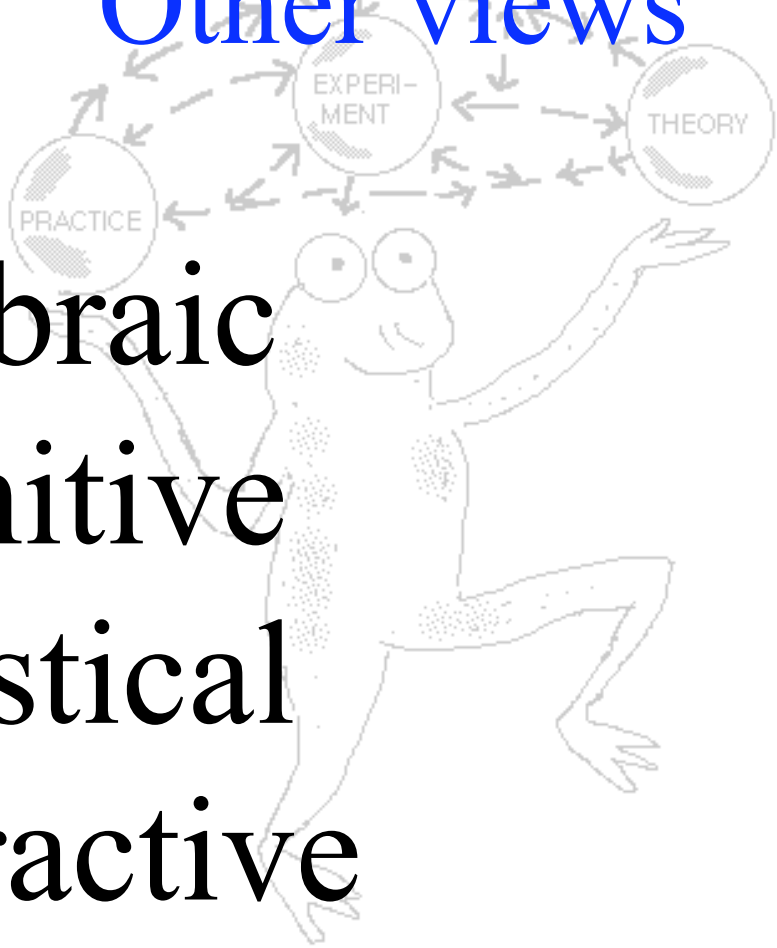

Browser
commands


Browser
commands


Browser
commands
Detail du pont marie et d'un bain flottant

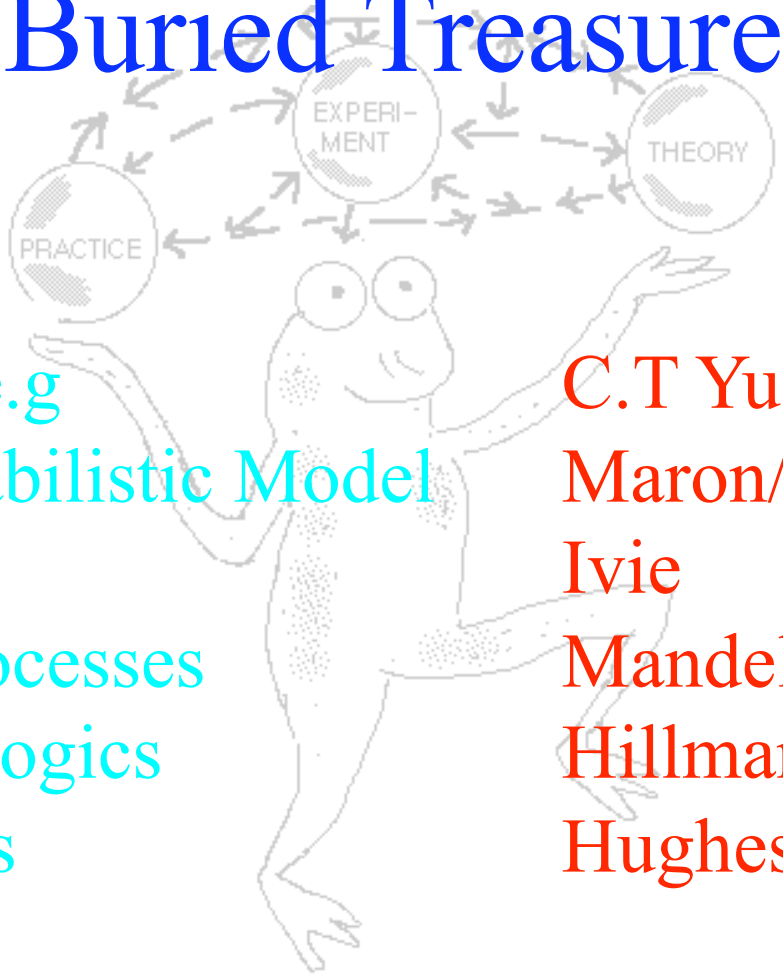

xv 3.10: object.contents <unregistered>

# Other views



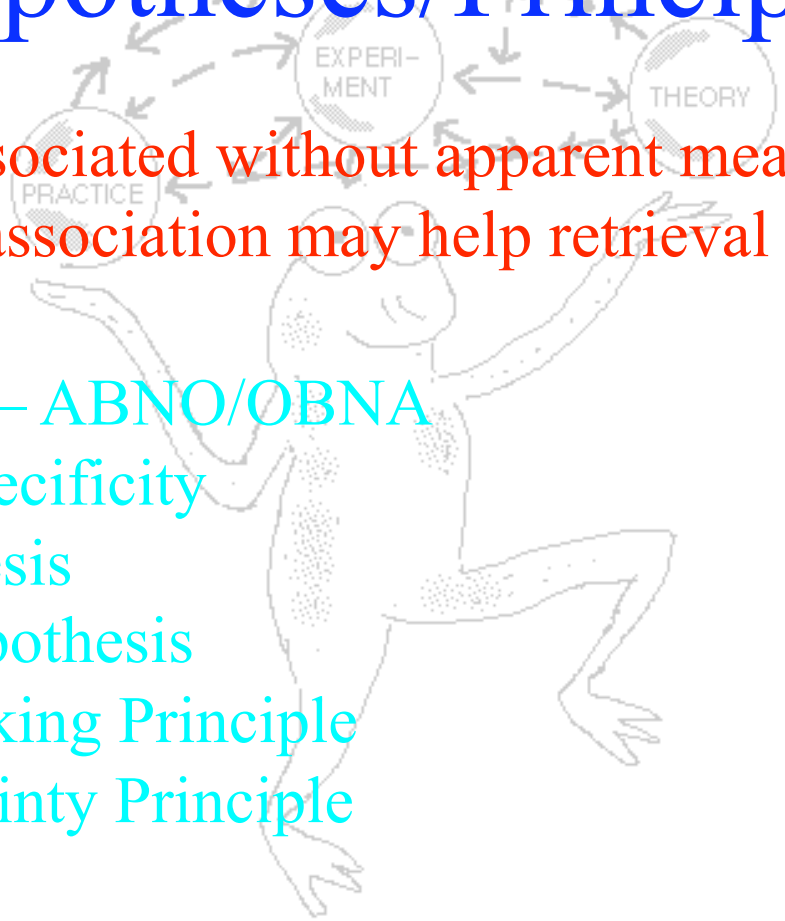

- algebraic
- cognitive
- statistical
- interactive

# Buried Treasure




| | |
|---|---|
| Dependence e.g | C.T Yu |
| Unified Probabilistic Model | Maron/Cooper/SER |
| Co-relevance | Ivie |
| Stochastic Processes | Mandelbrot/Herdan |
| Brouwerian Logics | Hillman |
| Error Analysis | Hughes/Cover/Duda |

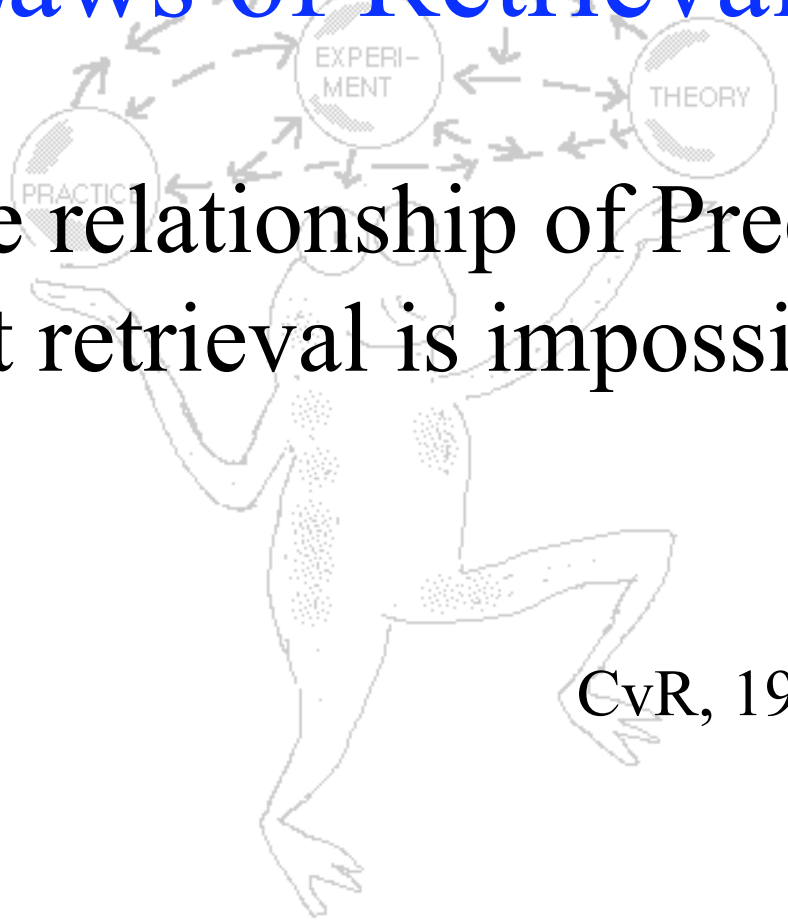
# Hypotheses/Principles




Items may be associated without apparent meaning but exploiting their association may help retrieval

P & R trade-off – ABNO/OBNA
Exhaustivity/Specificity
Cluster Hypothesis
Association Hypothesis
Probability Ranking Principle
Logical Uncertainty Principle
ASK
Polyrepresentation

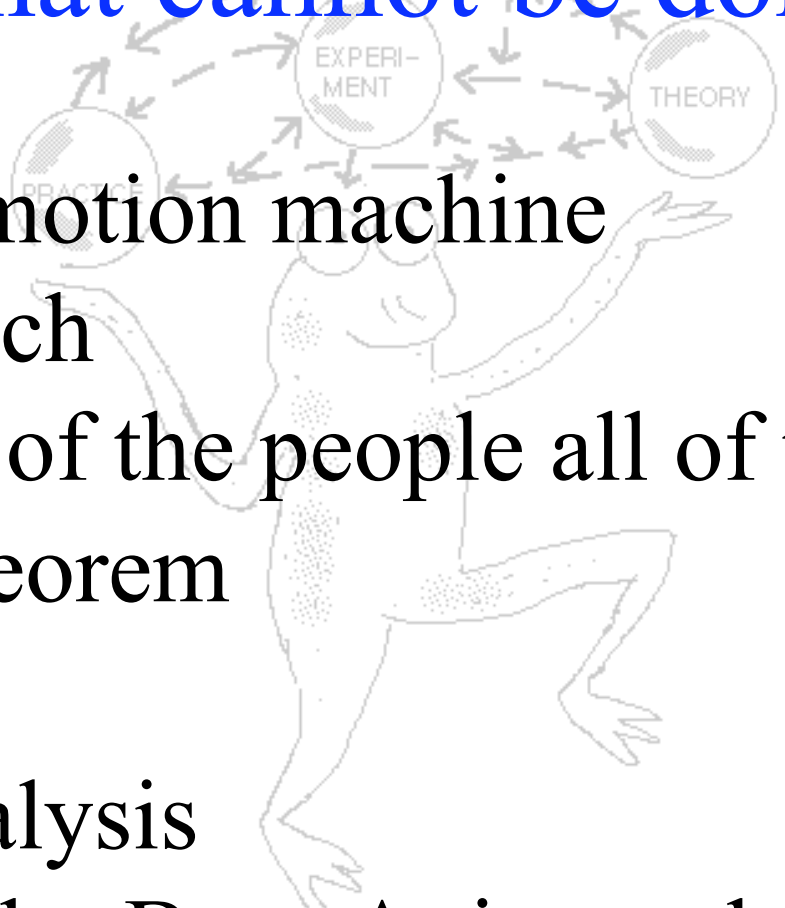
# Laws of Retrieval?




1. Inverse relationship of Precision/Recall
2. Perfect retrieval is impossible
3. ....

CvR, 1979

# What cannot be done!

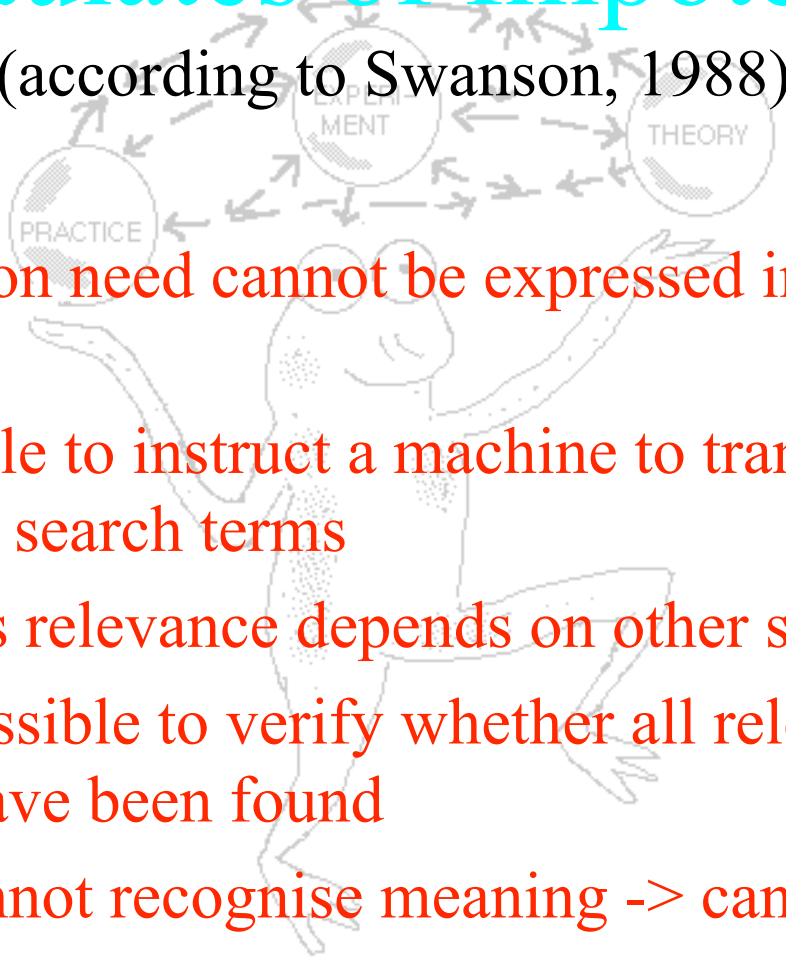



perpetual motion machine
no free lunch
fooling all of the people all of the time
Godel's theorem

Failure Analysis
Examine why B << A, instead of, why of why A >>B.

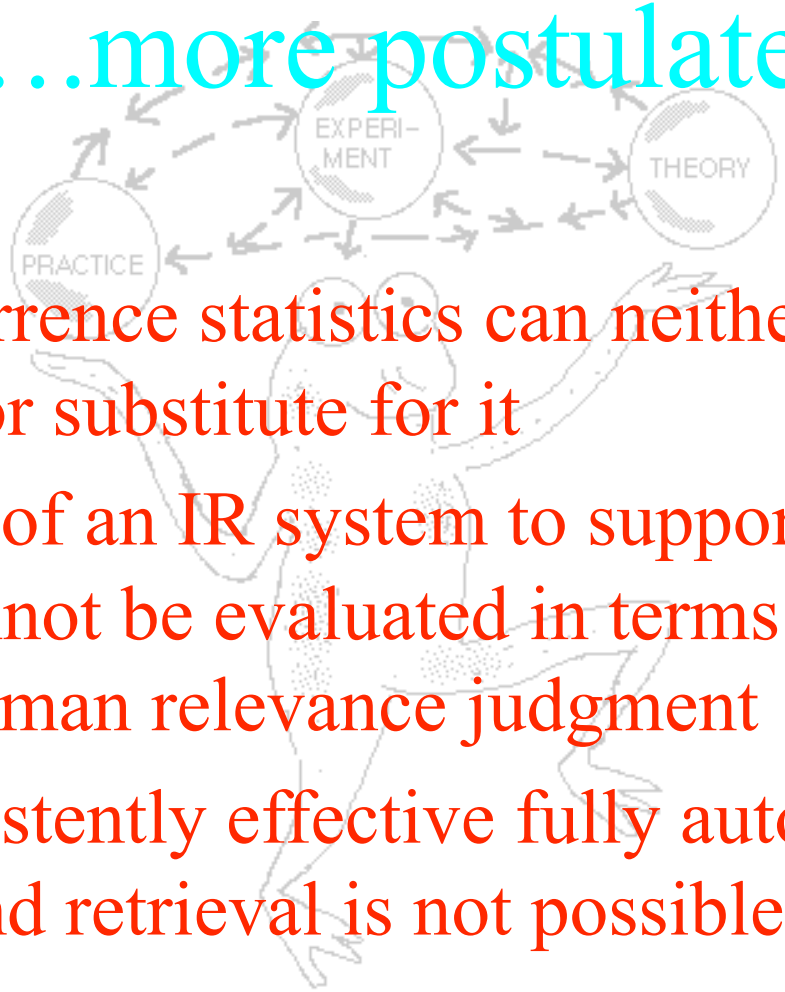
# Postulates of Impotence

(according to Swanson, 1988)



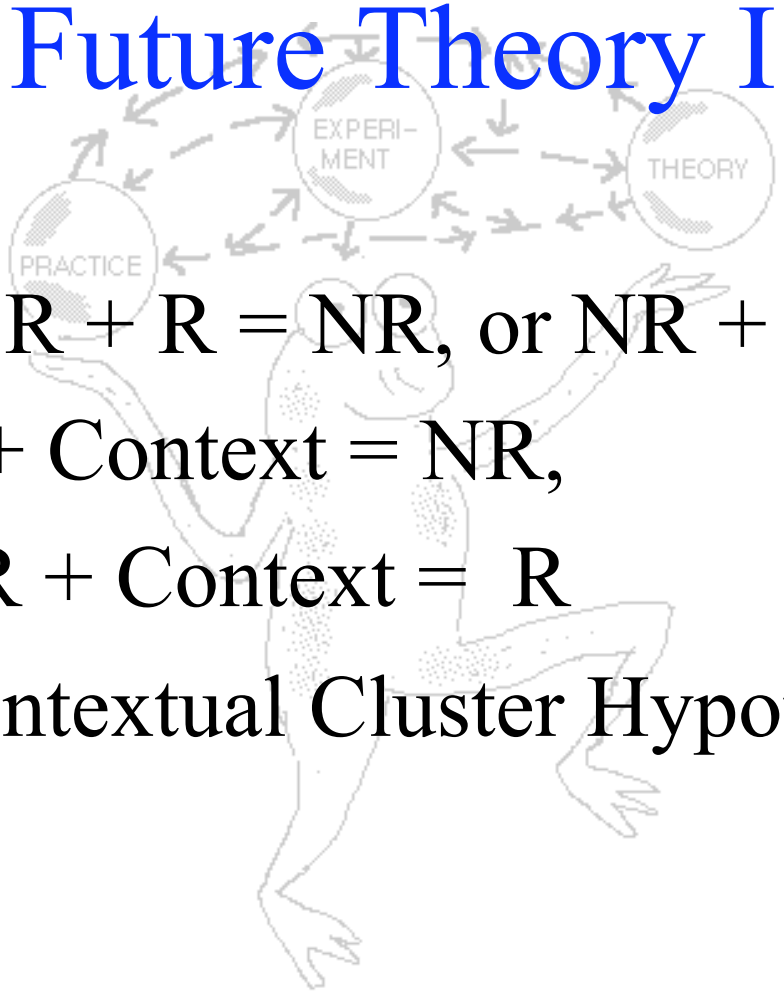

- An information need cannot be expressed independent of context
- It is impossible to instruct a machine to translate a request into adequate search terms
- A document's relevance depends on other seen documents
- It is never possible to verify whether all relevant documents have been found
- Machines cannot recognise meaning -> can't beat human indexing etc

# ….more postulates




- Word-occurrence statistics can neither represent meaning nor substitute for it
- The ability of an IR system to support an iterative process cannot be evaluated in terms of single-iteration human relevance judgment
- Thus, consistently effective fully automatic indexing and retrieval is not possible

# Future Theory I



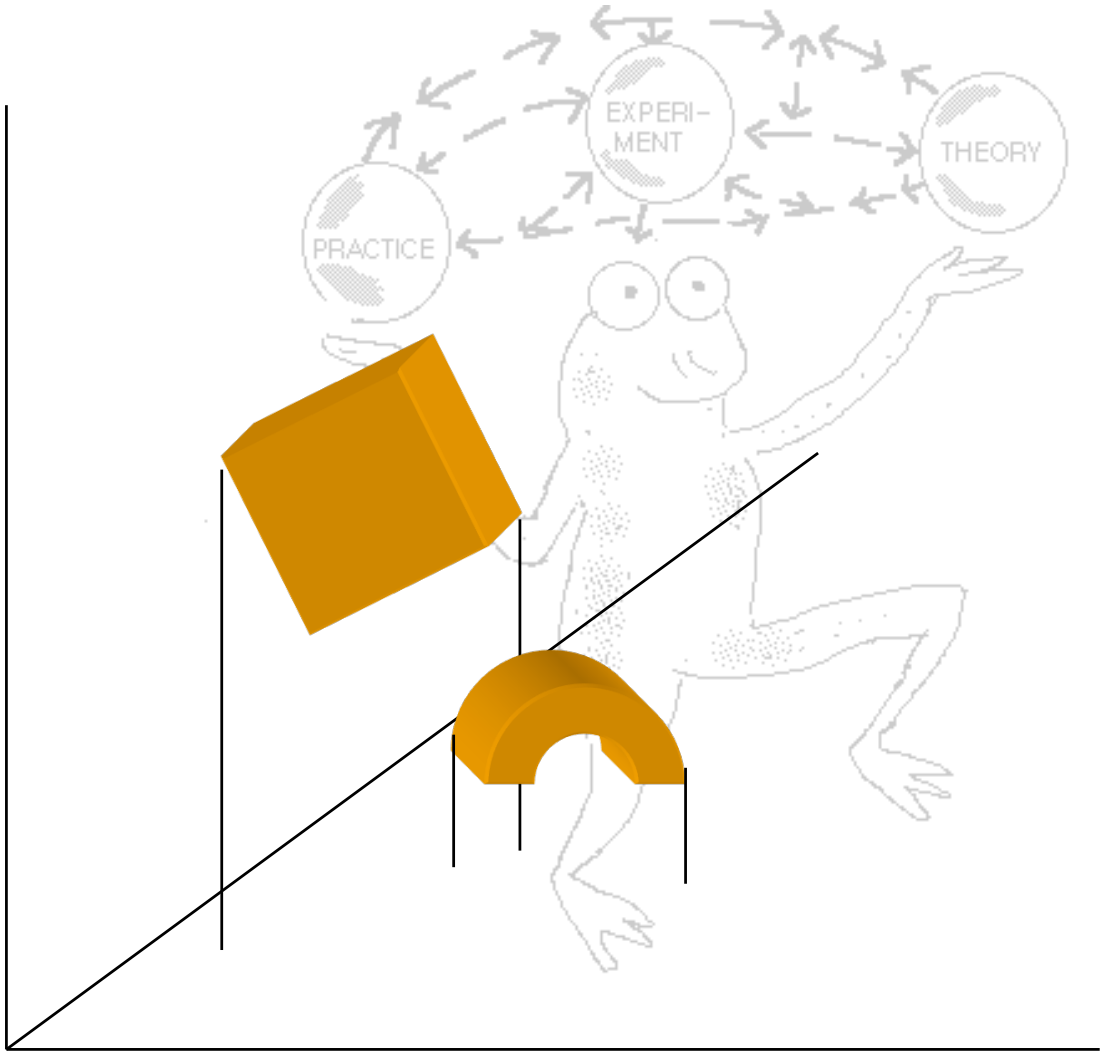

Relevance: R + R = NR, or NR + NR = R

Context: R + Context = NR,

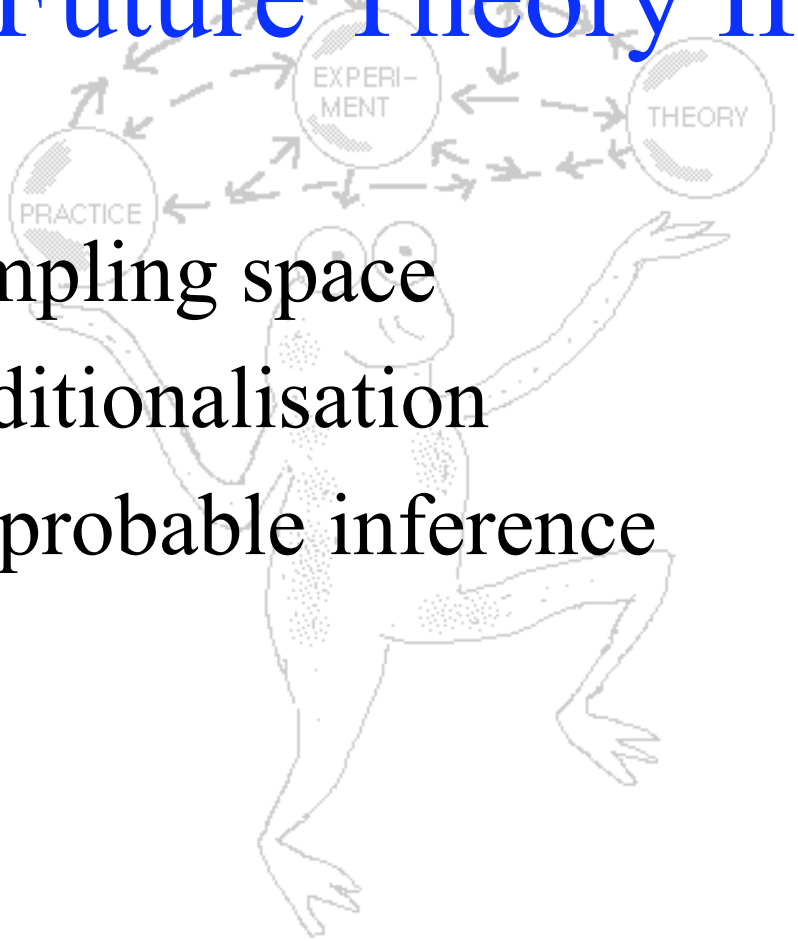
NR + Context = R

Contextual Cluster Hypothesis

# CCH



# Future Theory II

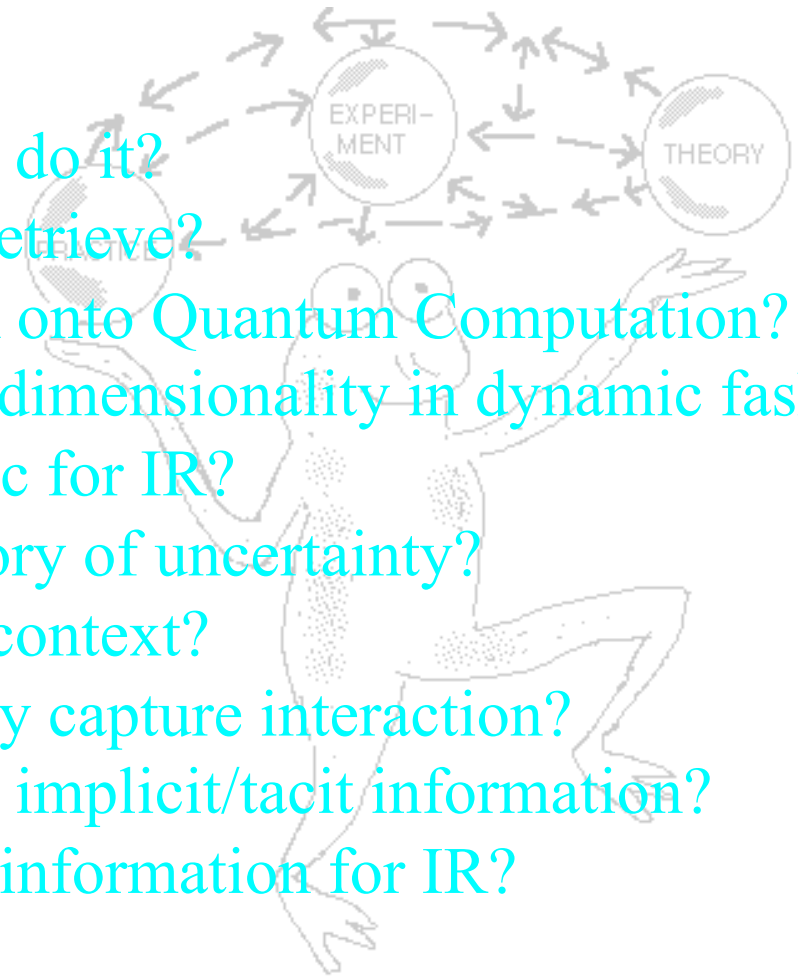



P(X|Y) - sampling space

$P_Y(X)$ - conditionalisation

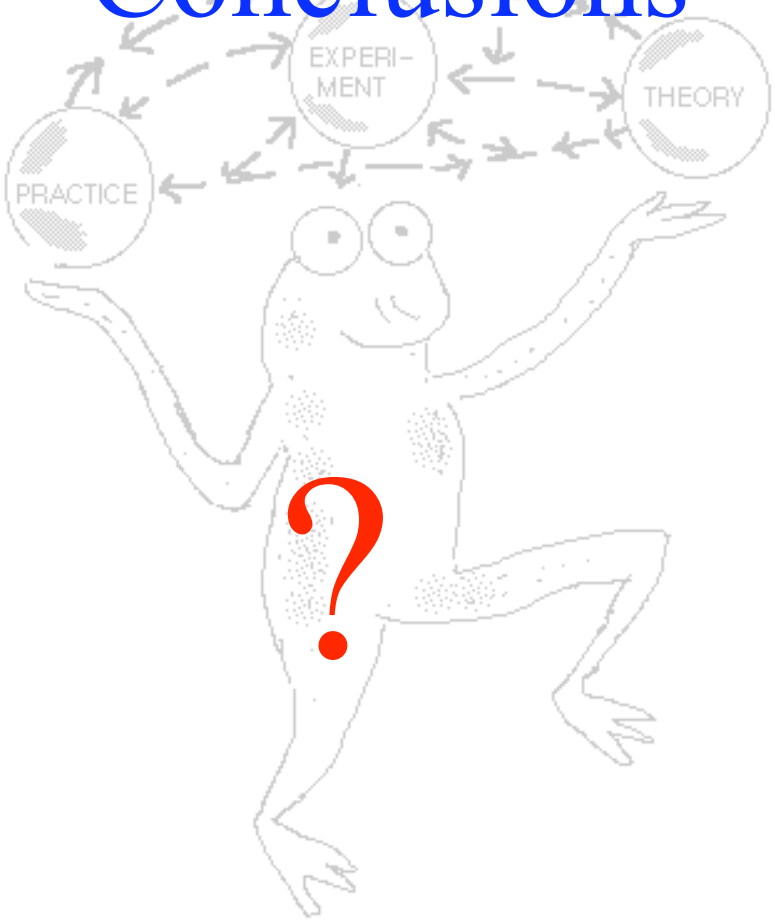
P(Y -> X) - probable inference

# Areas of Research




- How does the brain do it? (neuroscience)
- How do we see to retrieve? (computer vision)
- How do we map IR onto Quantum Computation? (QM)
- How do we reduce dimensionality in dynamic fashion? (Statistics)
- What is a good logic for IR? (mathematical logic)
- What is a good theory of uncertainty? (frequency/geometry)
- How do we model context? (HCI)
- How do we formally capture interaction?
- How do we capture implicit/tacit information?
- Is there a theory of information for IR?

# Conclusions



# Readings I

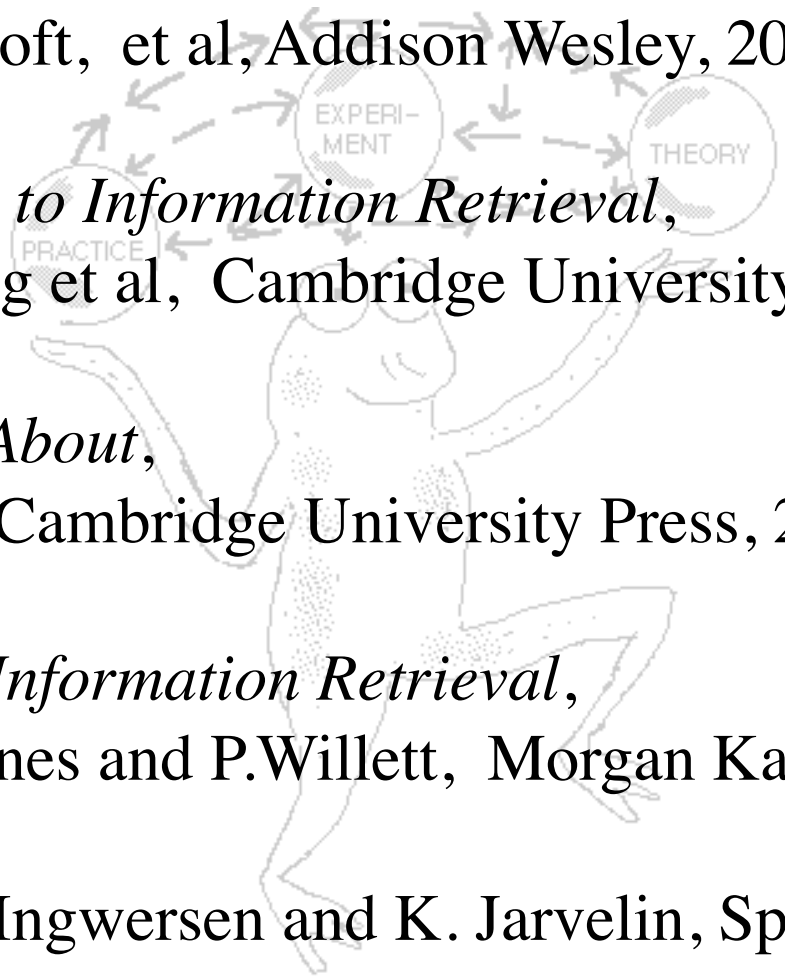
1. *Search Engines: Information Retrieval in Practice*,
   W. Bruce Croft, et al, Addison Wesley, 2010.

2. *Introduction to Information Retrieval*,
   C.D. Manning et al, Cambridge University Press, 2008.

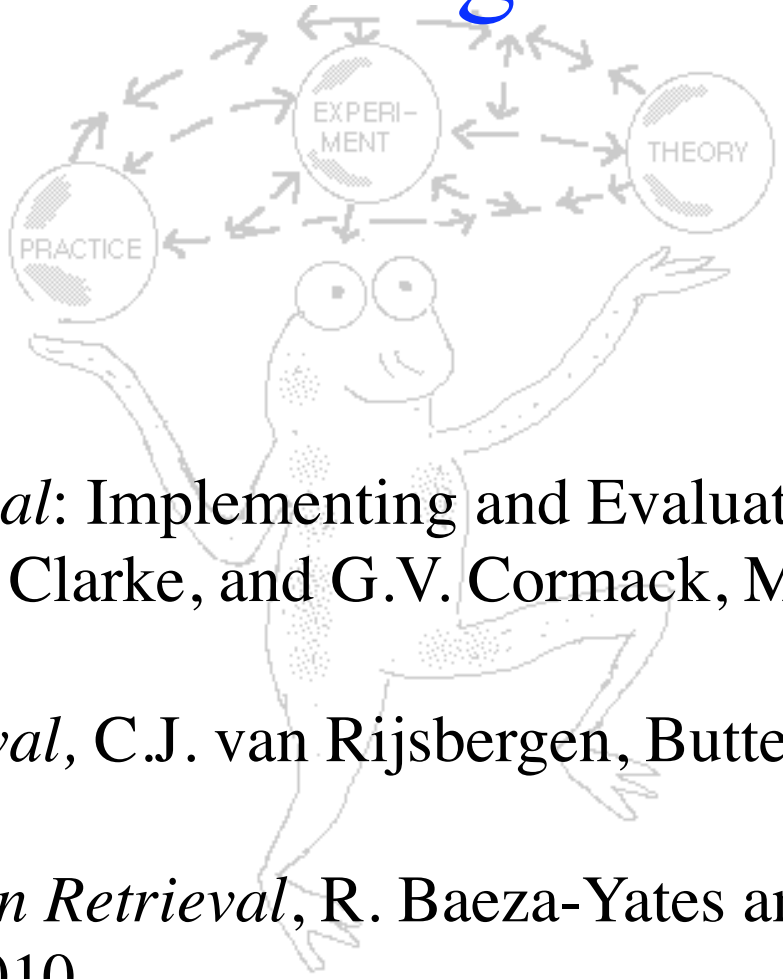
3. *Finding out About*,
   R.K. Belew, Cambridge University Press, 2000.

4. *Readings in Information Retrieval*,
   K. Sparck Jones and P.Willett, Morgan Kauffmann, 1997.

5. *The Turn*, P. Ingwersen and K. Jarvelin, Springer, 2005.



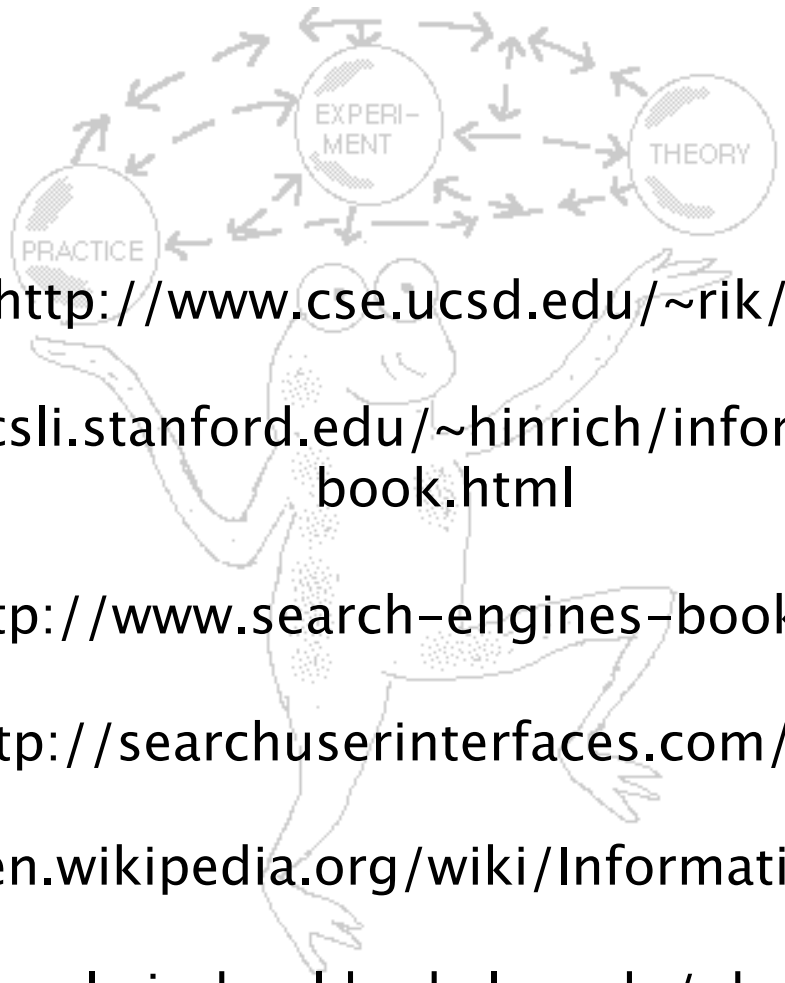
# Readings II




6. *Information Retrieval*: Implementing and Evaluating Search Engines
   S. Buttcher, C.L.A Clarke, and G.V. Cormack, MIT, 2010

7. *Information Retrieval*, C.J. van Rijsbergen, Butterworths, 1979.

8, *Modern Information Retrieval*, R. Baeza-Yates and B. Ribero-Neto
   Addison Wesley, 2010




http://www.cse.ucsd.edu/~rik/foa/

http://www-csli.stanford.edu/~hinrich/information-retrieval-book.html

http://www.search-engines-book.com/

http://searchuserinterfaces.com/book/

http://en.wikipedia.org/wiki/Information_retrieval

http://people.ischool.berkeley.edu/~hearst/irbook/



CH, 2012