

CS124

Introduction to Lab 3: Information Retrieval

Catalogs are old!

Two famous lost catalogs from 100-200 BC

- Pinakes ("tables") of Callimachus describing the Library of Alexandria!
- Han Dynasty China Liu Xin's *Qilüe* (七略; "Seven Surveys")

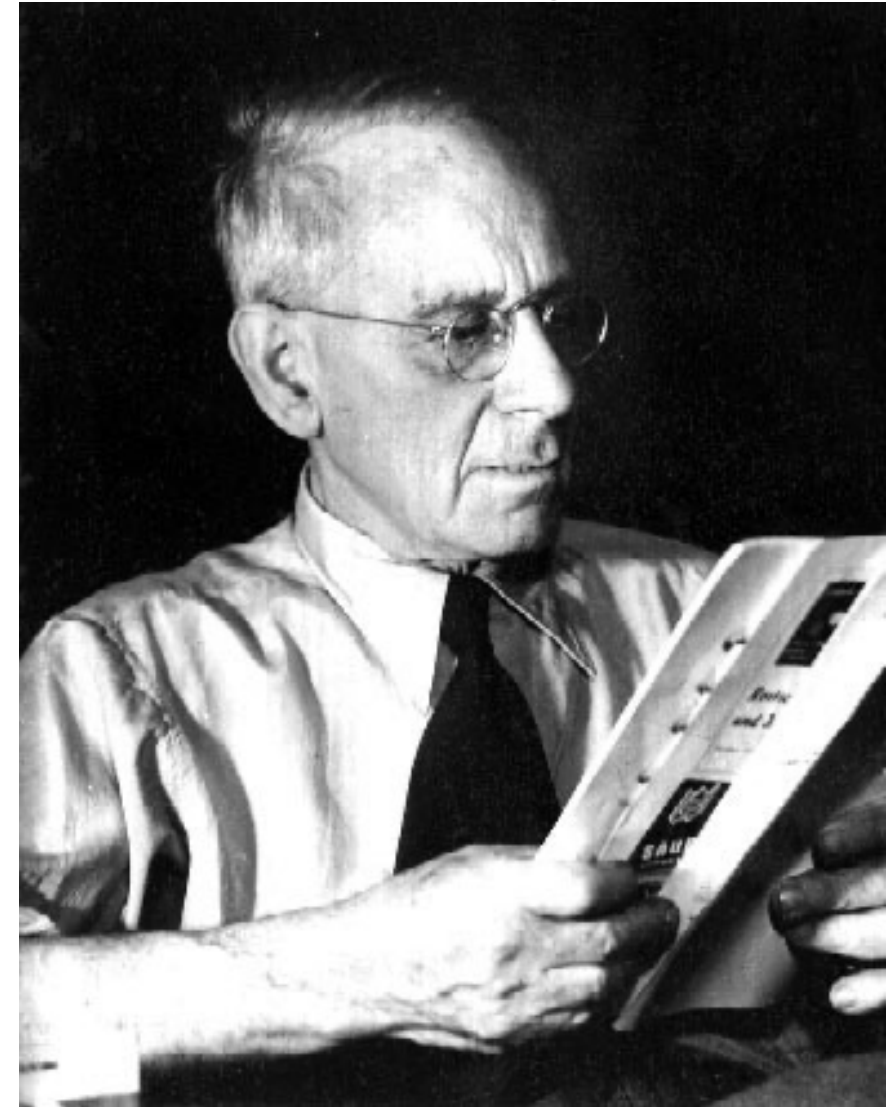
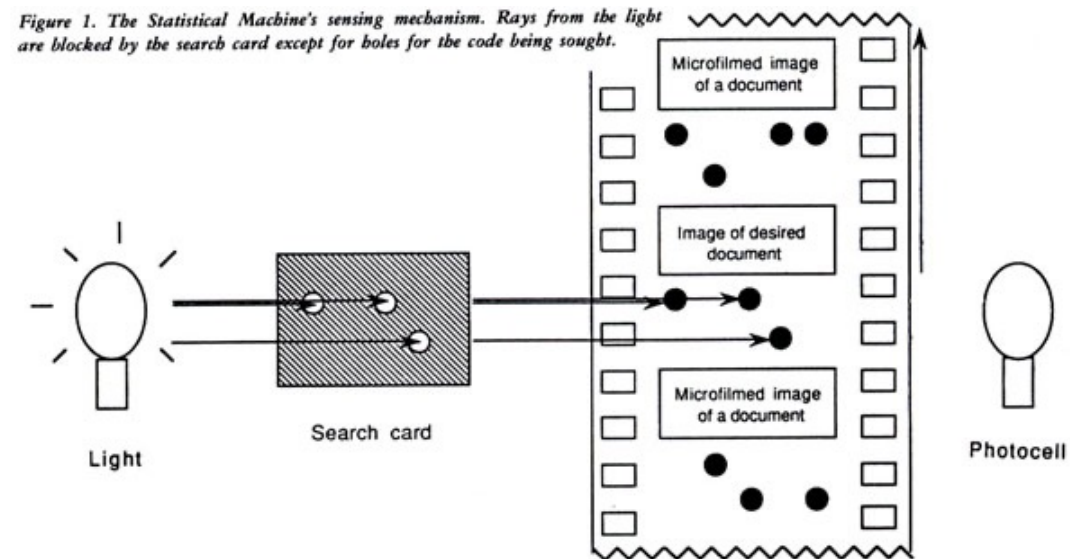
Lots of research on finding things in catalogs

- Early 20th C: Hollerith punched cards
- But not yet mechanical devices that searched a catalogue for a particular entry

First electronic document retrieval system

Emanuel Goldberg (1881-1970)

- Founder of Zeiss Ikon
- 1928 microfilm reader



"on his desk... he would dial a number, press a button and after three seconds the document would be projected".

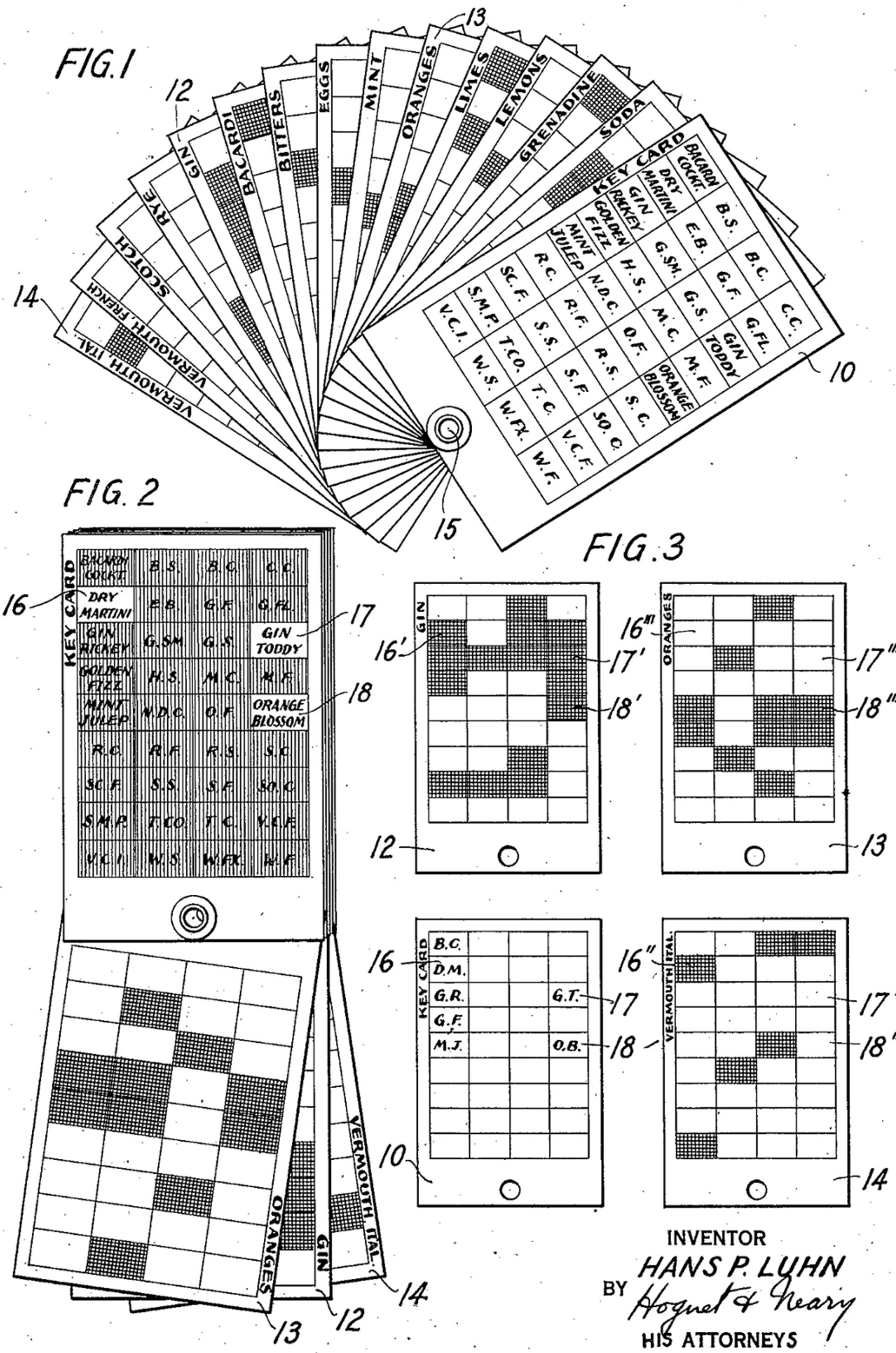
1950s: Computers begin to be used for IR

1957-8 Hans Peter Luhn (1896-1964), at IBM
proposes **term frequency** and **ranked retrieval**

He also invented hashing



And something called the Cocktail Oracle



INVENTOR
HANS P. LUHN
BY *Hoguet & Heary*
HIS ATTORNEYS

1963: Cosine and vectors

Gerald Salton (1927-1995)

At Cornell

- The vector model of IR
- The use of the cosine

Similarity matrix in 1963 paper:

$$\left(\mathbf{R}_i^i = \mathbf{R}_i^j = \sum_{k=1}^m \mathbf{C}_k^i \mathbf{C}_k^j / \sqrt{\left(\sum_{k=1}^m (\mathbf{C}_k^i)^2 \sum_{k=1}^m (\mathbf{C}_k^j)^2 \right)} \right)$$



1972: Inverse Document Frequency (IDF)



Karen Spärck Jones
(1935-2007)
Cambridge

A STATISTICAL INTERPRETATION OF
TERM SPECIFICITY AND ITS APPLICATION
IN RETRIEVAL

KAREN SPARCK JONES
University of Cambridge Computer Laboratory

1990s: Web search

1993 the web opens to the public

1993 a few early search indexes

1995 Altavista: natural language web search

1996 Ask Jeeves

1998 Google

2024 Google includes Gemini results



[Editorial Guidelines for Answers](#)
[Privacy Statement](#)
© 1996-2000 Ask Jeeves, Inc.
ASK JEEVES, ASK.COM and the JEEVES DESIGN are service marks of Ask Jeeves, Inc.
All other brands are property of their respective owners. Patent pending.

Today's lab 3

- 1) TF-IDF vector cosines
- 1b) Sparse Retrieval vs Dense Retrieval
- 2) Precision and Recall
- 3) The role of IR/LLM as Mirror vs Aspiration

Part 1: Vector dot products and cosine

Core of IR: dot products of vectors

- Sparse retrieval
- Dense retrieval

A core of LLMs: dot products of vectors

- Next week: embeddings (dense vectors) are how LLMs represent information
- Feedforward network layers
 - Compute dot products between stored prototypes and input data

TF-IDF

As a weighting scheme for measuring importance, in IR and:

- Recommender systems
- Computational social science
- Even for non-text data

Part 2: Precision and Recall

		<i>gold standard labels</i>		
		gold positive	gold negative	
<i>system output labels</i>	system positive	true positive	false positive	precision = $\frac{tp}{tp+fp}$
	system negative	false negative	true negative	
		recall = $\frac{tp}{tp+fn}$		accuracy = $\frac{tp+tn}{tp+fp+tn+fn}$

In lab today

We'll make sure you know how to compute it
And we'll introduce the idea that there can be
a **tradeoff** between precision or recall

Part 3: What about Generative AI?

LLMs create content instead of just retrieving it

Leads to new ethical questions about IR with respect to historical or social facts!!

When should LLMs generating content be:

- **Mirror**: reflect the world as it is
- **Aspiration**: reflect a better world we envision