

Applications of Indexing

Lecture 16



Overview

- We've talked about how indexes can improve performance inside a DBMS
- But these same data structures/techniques can be useful on other systems as well!
- Let's look at a couple examples :)



Outline

1. Problem #1: What is **Full-Text Search**?
 - Can it be done naively? (*spoiler: not for Google!*)
2. To the Rescue: The Inverted Index
 - Design a relational index
 - Advanced Issues
 - **Research Application: Cognitive Modeling**
3. The R-tree
 - Overview
 - **Research Application: Optimization**



What is Full-Text Search?

Given

- A set of “documents”
 - Each containing “words”

Problem

- Find the “best” document(s) that satisfy a query
 - Simplest: set of words

Requirements

- Fast & scalable
- Relevant results
- Expressive queries
- Up-to-date



Example: Web Search

The image shows a Google search interface with the query 'northeastern'. A blue callout box points to the search bar with the text 'Document = web page/map listing'. Another blue callout box points to the search results with the text 'Document = web page/map listing'. A black-bordered box highlights the search results summary: 'About 53,200,000 results (0.61 seconds)'. The search results include a link to 'Northeastern University: a leader in global experiential ... - Boston' with the URL 'www.northeastern.edu/'. Below this, there are several links: 'Undergraduate Admissions', 'News @ Northeastern', 'Northeastern 2025', 'Northeastern College of ...', 'myNortheastern', and 'Graduate Degrees'. To the right, there is a map showing the location of Northeastern University in Boston, with a red pin and a label 'Northeastern'. Below the map, there are buttons for 'See photos' and 'See outside'. The main heading 'Northeastern University' is displayed, followed by buttons for 'Website', 'Directions', and 'Save'. Below these buttons, it says 'Private university in Boston, Massachusetts'. At the bottom, there is a paragraph: 'Northeastern University is a private research university in Boston, Massachusetts, established in 1898. It is categorized as an R1 institution by the Carnegie Classification of Institutions of Higher Education.'

Google northeastern

About 53,200,000 results (0.61 seconds)

Northeastern University: a leader in global experiential ... - Boston
www.northeastern.edu/

Northeastern is a global, experiential, research university built on a tradition of engagement with the world, creating a distinctive approach to education.

Undergraduate Admissions
Schedule a Visit - Contact - Areas of Study - Freshman Admissions

News @ Northeastern
The news@Northeastern team is looking for a creative, driven ...

Northeastern 2025
NORTHEASTERN2025. Learning and discovery in the. Age of ...

Northeastern College of ...
Welcome to the College of Professional Studies. We offer ...

myNortheastern
myNortheastern · Find Faculty & Staff · Find A-Z · Emergency ...

Graduate Degrees
Admissions - Programs - Financial Aid - Contact Information - Boston

More results from northeastern.edu »

See photos See outside

Northeastern University

Website Directions Save

Private university in Boston, Massachusetts

Northeastern University is a private research university in Boston, Massachusetts, established in 1898. It is categorized as an R1 institution by the Carnegie Classification of Institutions of Higher Education.



Other Examples

Facebook search results for "northeastern".

Filter Results

POSTS FROM

- Anyone
- You
- Your Friends and Groups
- Choose a Source...

POSTED IN GROUP

- Any group
- Your Groups
- Choose a Group...

TAGGED LOCATION

- Anywhere

Pages

Northeastern University  Boston, Massachusetts
News & Media Website
106,482 like this

Places

Northeastern University Global Experience Office  5.0 ★★★★★ (9) College & University
0.4 miles · 403 Richards Hall, Boston · [Open Now](#)

search

northeastern

advanced search: by author, subreddit...

subreddits

Northeastern University

[subscribe](#) [r/northeastern](#) 190 subscribers, a community for 7 years
[search within r/northeastern](#)

Northeastern University (NU/NEU)

[subscribe](#) [r/NEU](#) 5,348 subscribers, a community for 8 years
News and discussion of interest to students, faculty, employees, and neighbors of **Northeastern** University in Boston, MA.
[search within r/NEU](#)

Amazon search results for "northeastern".

Showing most relevant results. See all results for **northeastern**.

NCAA Tradition Short sleeve Tri-Blend T-Shirt
by Image One
\$6⁸² - \$29⁹⁹ 
Some sizes are Prime eligible

Northeastern University Pennant Full Size Felt
by WinCraft
\$8⁹⁵
FREE Shipping on eligible orders

NCAA Northeastern Huskies 50/50 Blended 8-Ounc
by SDI
\$31⁴⁰ 
Some sizes/colors are Prime eligible



Fast & Scalable: # of Documents

	> 130 Trillion Pages
	~ 2 Billion Users
	~ 400 Million Products



Fast & Scalable: Queries/sec

	> 63K
	~ 6K



Big-O Review

- What does it mean for an algorithm to scale in linear time (i.e. $O(n)$)?
- Describe a linear-time algorithm for full-text search of keywords in webpages
 - Find all documents that contain “northeastern”



Is Linear-Time Google Possible?

- Assume simple query:

northeastern

~10.4M blu ray
~ 7.75 miles
~ 15 x Burj Khalifa

- Single repo of all pages
 - $130T * 250 \text{ w/page} * 8 \text{ bytes/w} \sim \mathbf{260 \text{ PB}}$
- Require 1s response time
 - $8M * 3\text{GHz } 64\text{-bit CPU (assume 1 cycle/w)}$
 - $(8M * 63K) \text{ CPUs} * 35\text{W/CPU} * \$0.14/\text{kWH}$
 $\sim \mathbf{\$0.7M/sec}$

~67 CPU/person

~ \$21.6T/year
~ 1.25 x US GDP



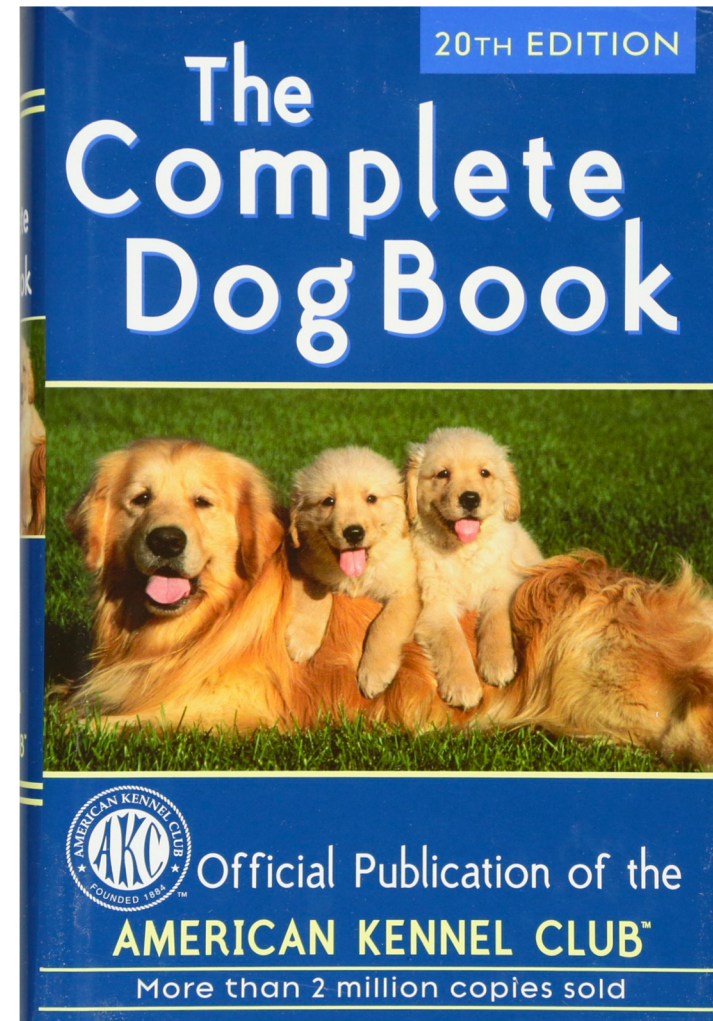
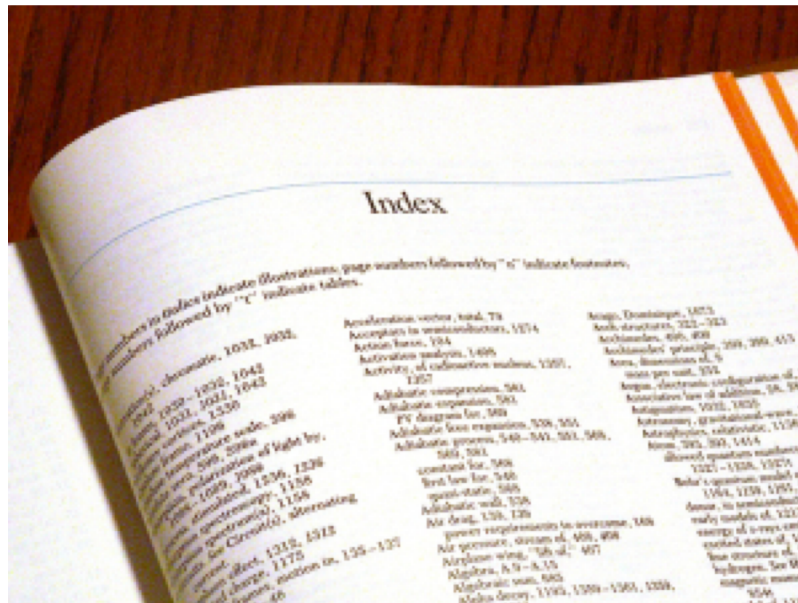
Indexing

- Improve search speed at the cost of extra...
 - **Memory** for data structure(s)
 - **Time** to update the data structure(s)
- Backbone of search engines, databases, graphics/game engines, simulation software, ...



Consider a Physical Textbook

- Find all pages of the book that contain the word “husky”



Analyze a Physical (Inverted) Index

1. Find the term

- Alphabetical search
 - More generally: binary
- Time: $O(\log_2 \text{terms})$

2. Enumerate the pages

- Linear scan
 - Page 1, 2, ..., k
- Time: $O(\text{pages})$

Space: $O(\text{terms} + \text{pages})$

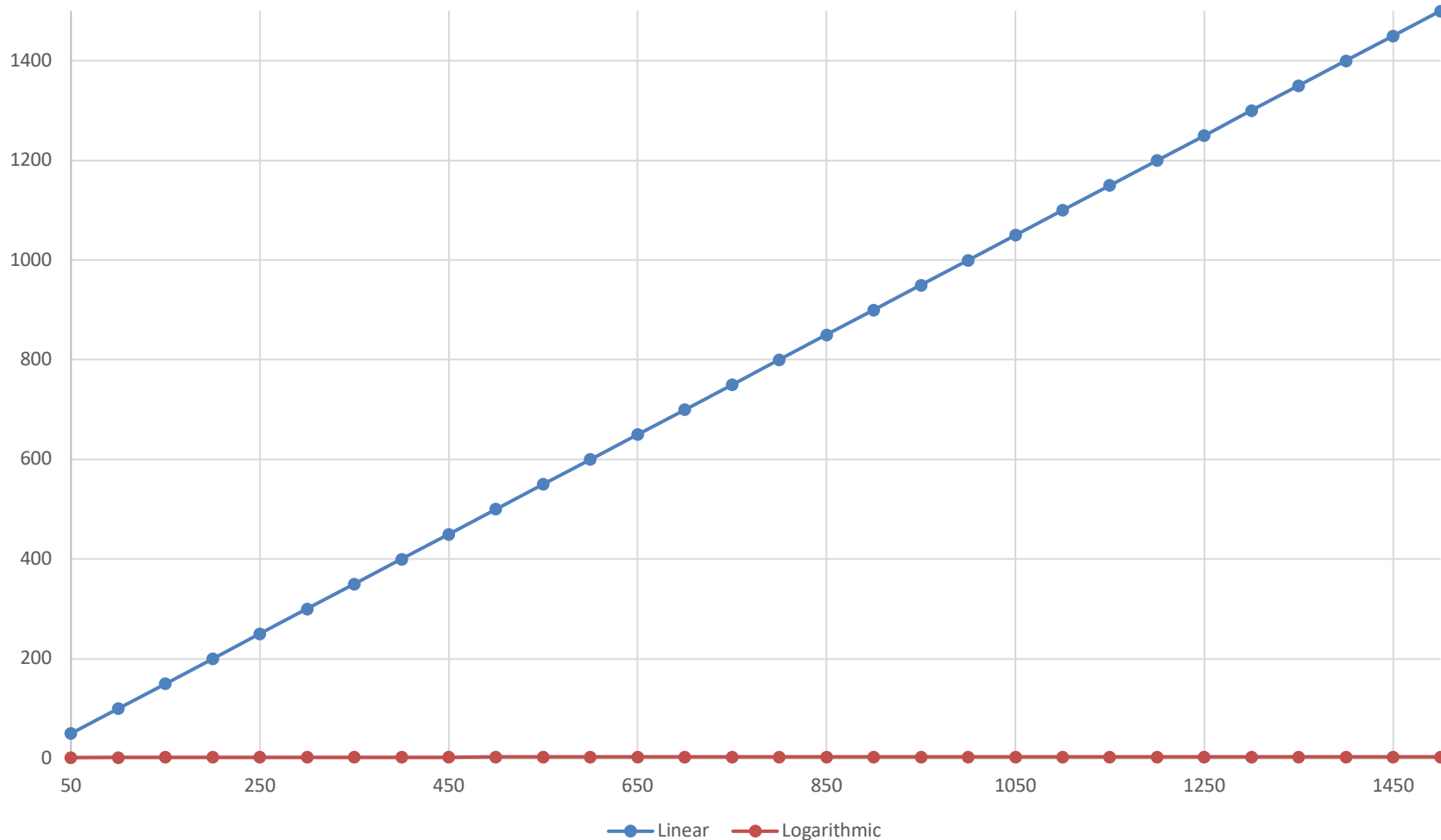


Linear vs Logarithmic (1)

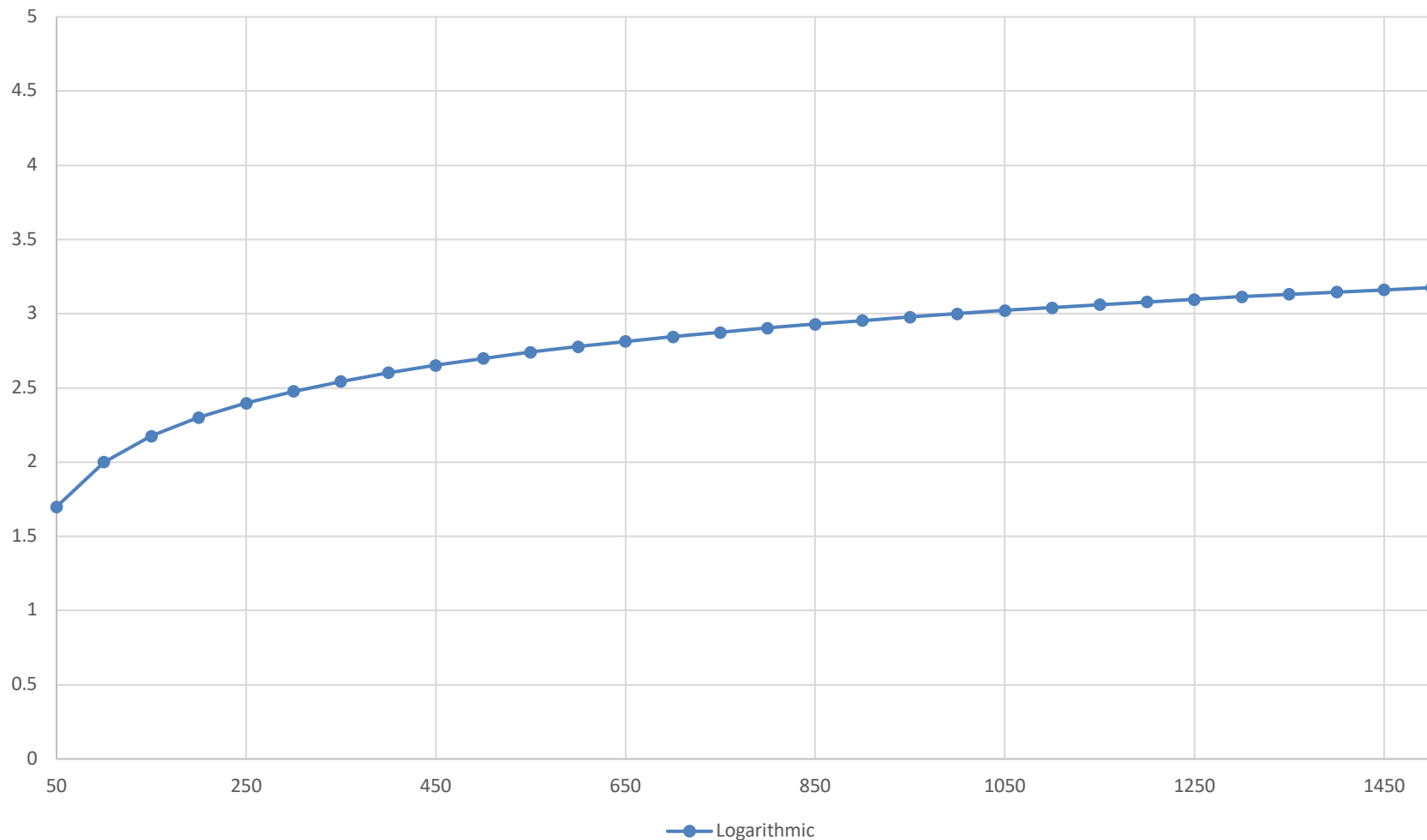
	N=10	N=100	N=1000
Logarithmic	1s	2s	3s
Linear	1s	10s	100s



Linear vs Logarithmic (2)



Linear vs Logarithmic (3)

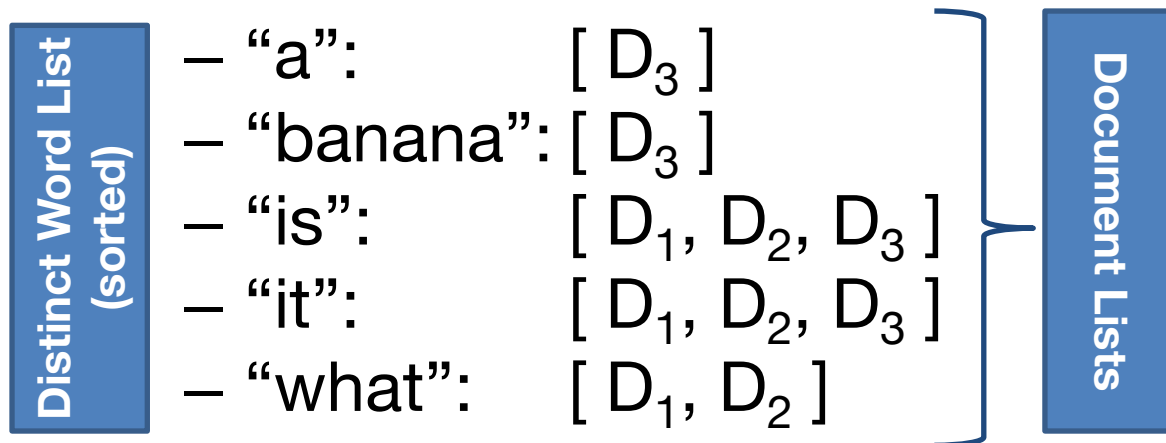


Inverted Index by Example

Given documents $\{ D_1, D_2, D_3 \}$:

- D_1 = “it is what it is”
- D_2 = “what is it”
- D_3 = “it is a banana”

Inverted Index:



Design a **Relational** Inverted Index

Develop a set of table(s) and index(es) that support efficient construction and querying of an inverted index

Documents

<u>DocId</u>	DocPath
1	/path/to/doc

Contents

<u>Word</u>	<u>DocId</u>
Foo	1

```
SELECT DocId FROM Contents WHERE Word=? ORDER BY DocId
```



Advanced Issues

- More expressive query semantics
 - Multi-word
 - Locality: [“what is it”] vs. [“what it is”] vs. [“what”, “is”, “it”]
- Ranked results
 - Document-ranking algorithm (e.g. PageRank)
 - Efficient ranked retrieval
- Dynamics
 - Document addition/removal/modification
 - Rank
 - Document changes
 - Integration of real-time variables (e.g. location)

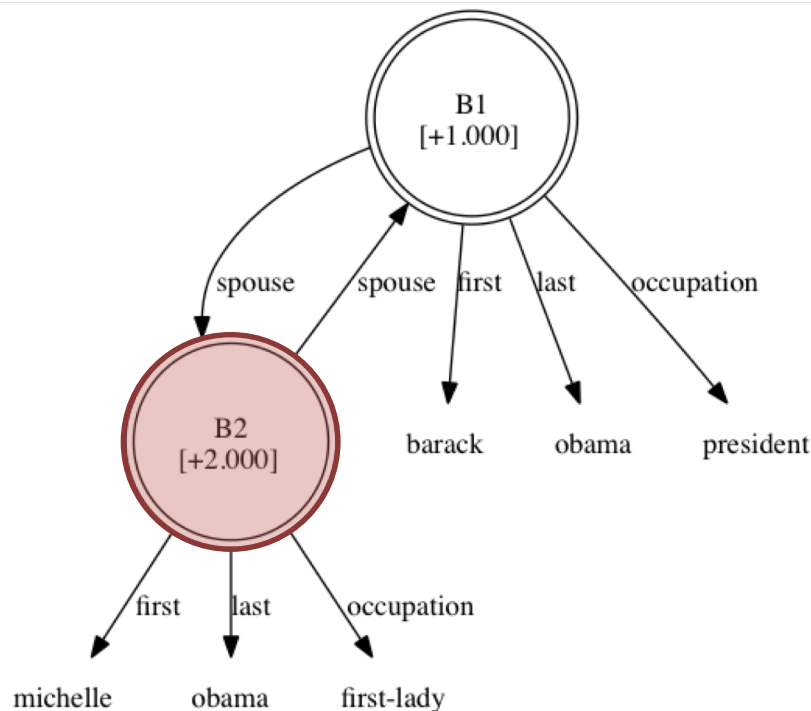


Research: Cognitive Modeling

- **Semantic Memory** is a human's long-term store of facts about the world, independent of the context in which they were originally learned
- The **ACT-R** (<http://act-r.psy.cmu.edu>) model of semantic memory has been successful at explaining a variety of psychological phenomena (e.g. retrieval bias, forgetting)
- The model does not scale to large memory sizes, which hampers complex experiments



Memory Representation



- Document = Node
- Word = edge
- *Office hours: ask me about the number inside each node :)*

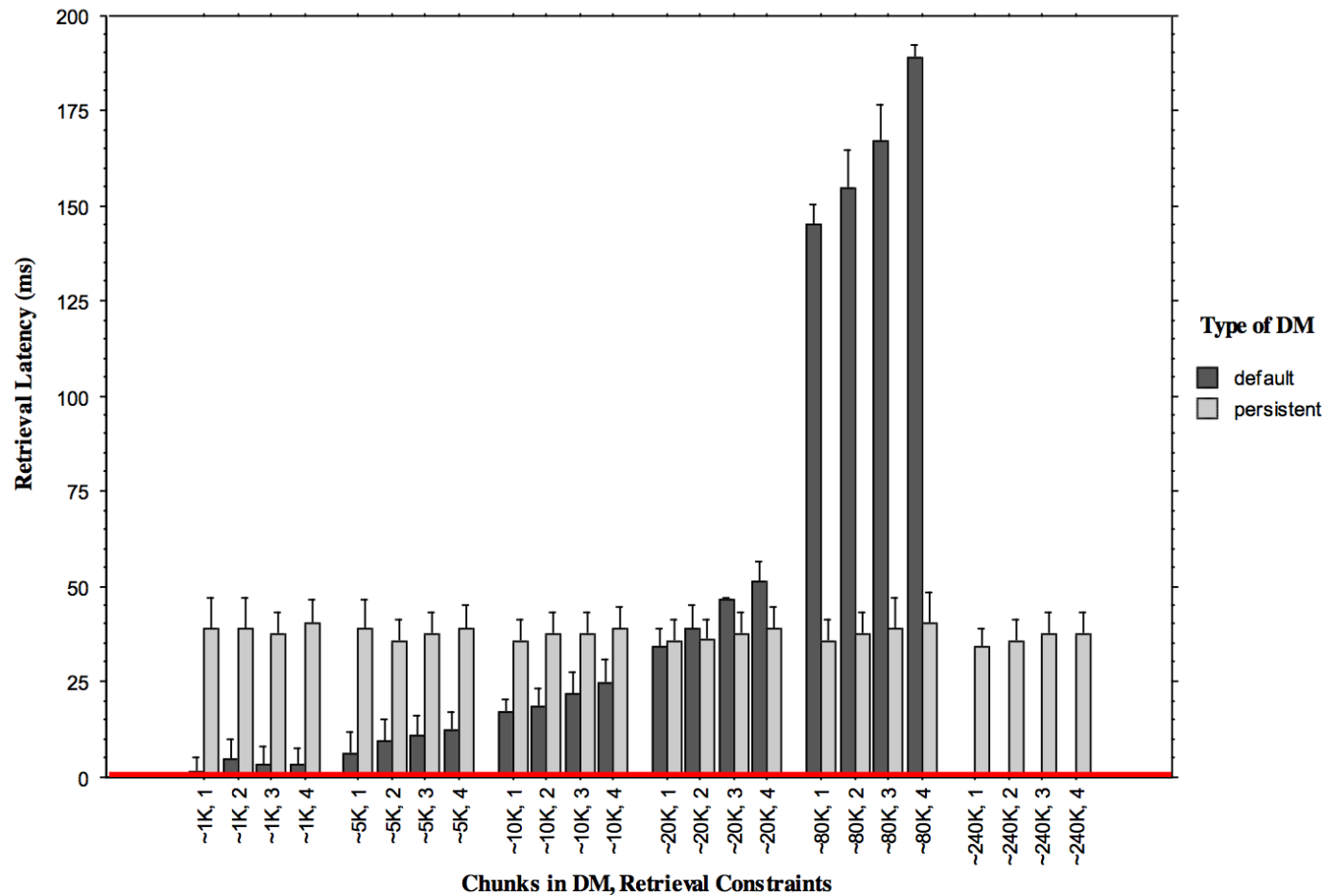
Example cue:

last (obama) , spouse (X)



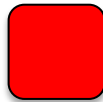
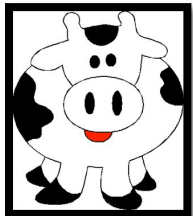
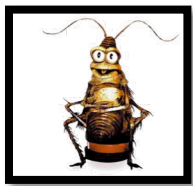
Scale Fail [AFRL '09]

Retrieval Latency: Chunks in DM x Retrieval Constraints x Type of DM
(Error Bars: 95% Confidence Interval)



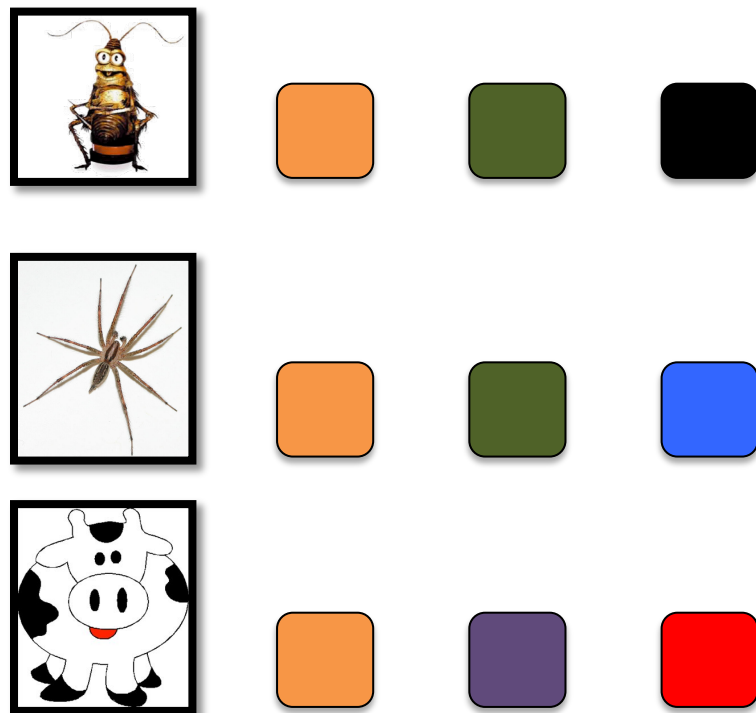
Example

Semantic Objects: Features

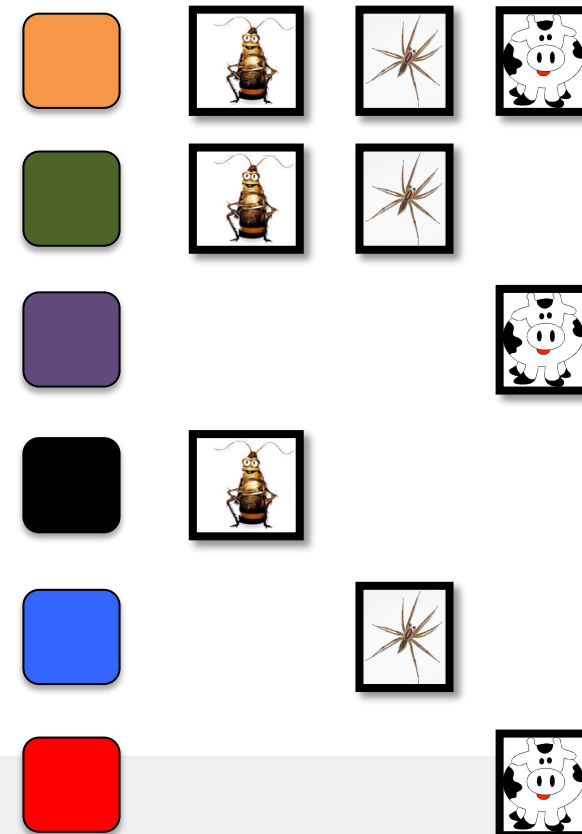


Inverted Index

Semantic Objects: Features

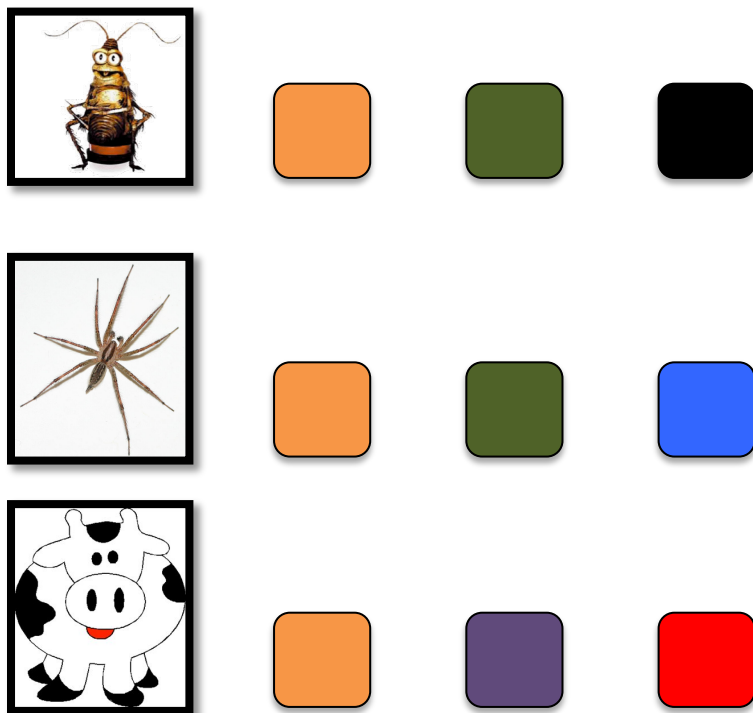


Inverted Index

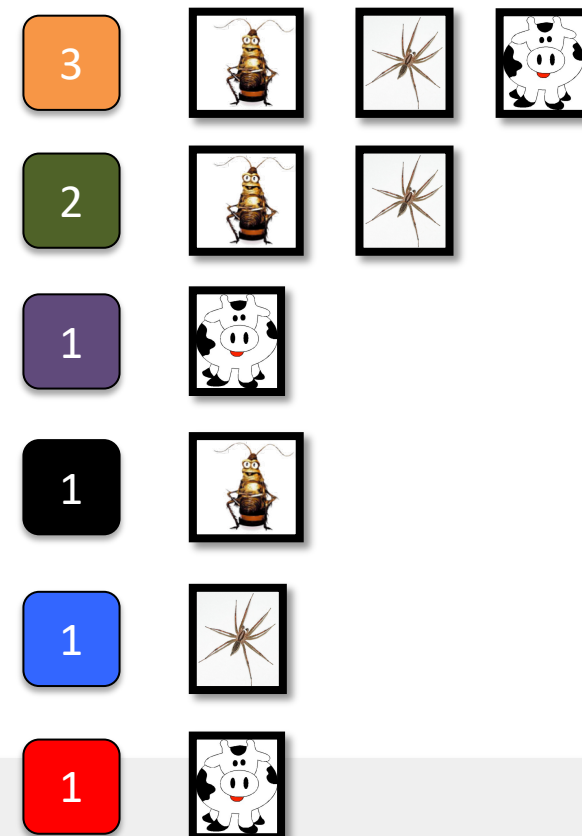


Index Statistics

Semantic Objects: Features

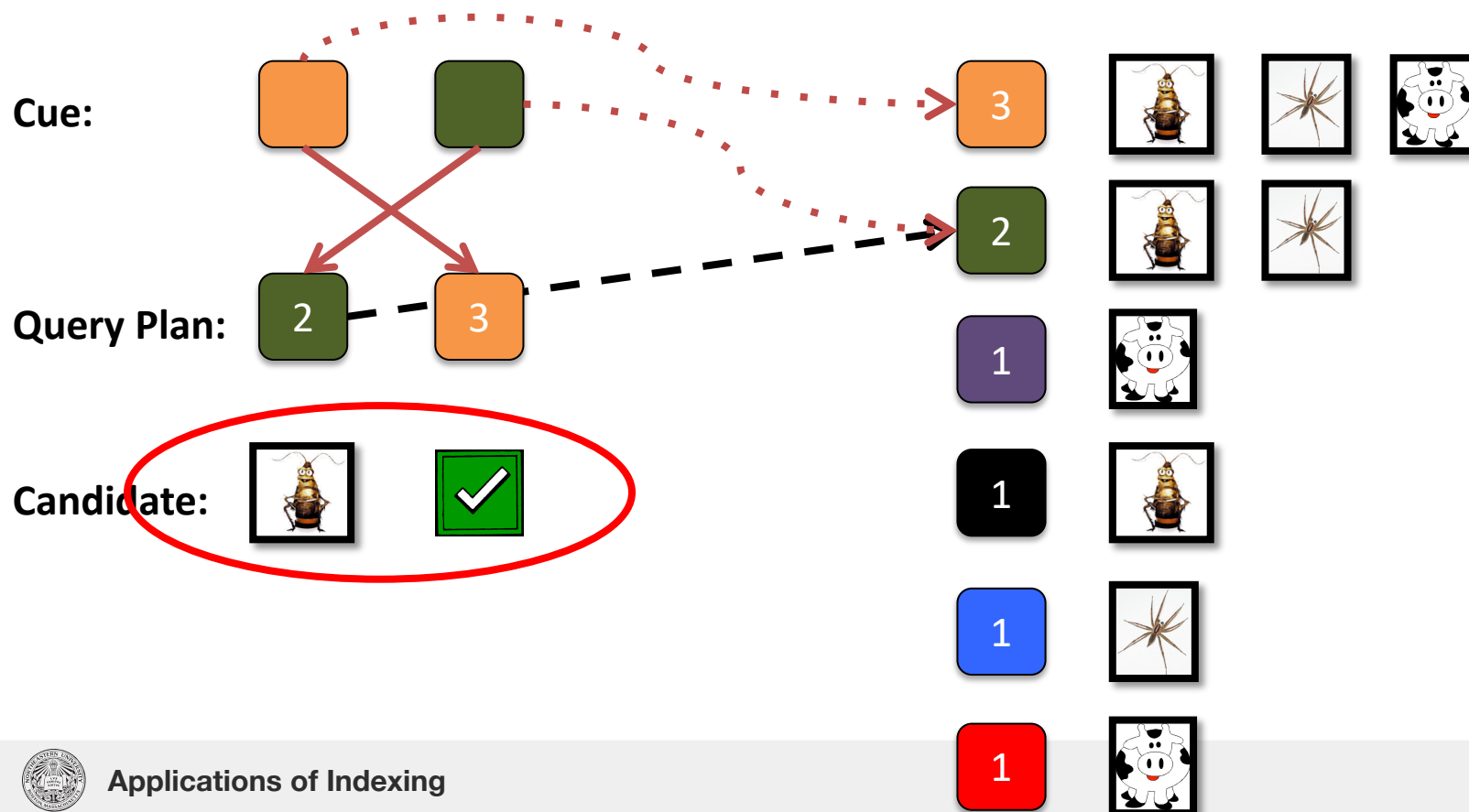


Inverted Index



Top-1 Retrieval

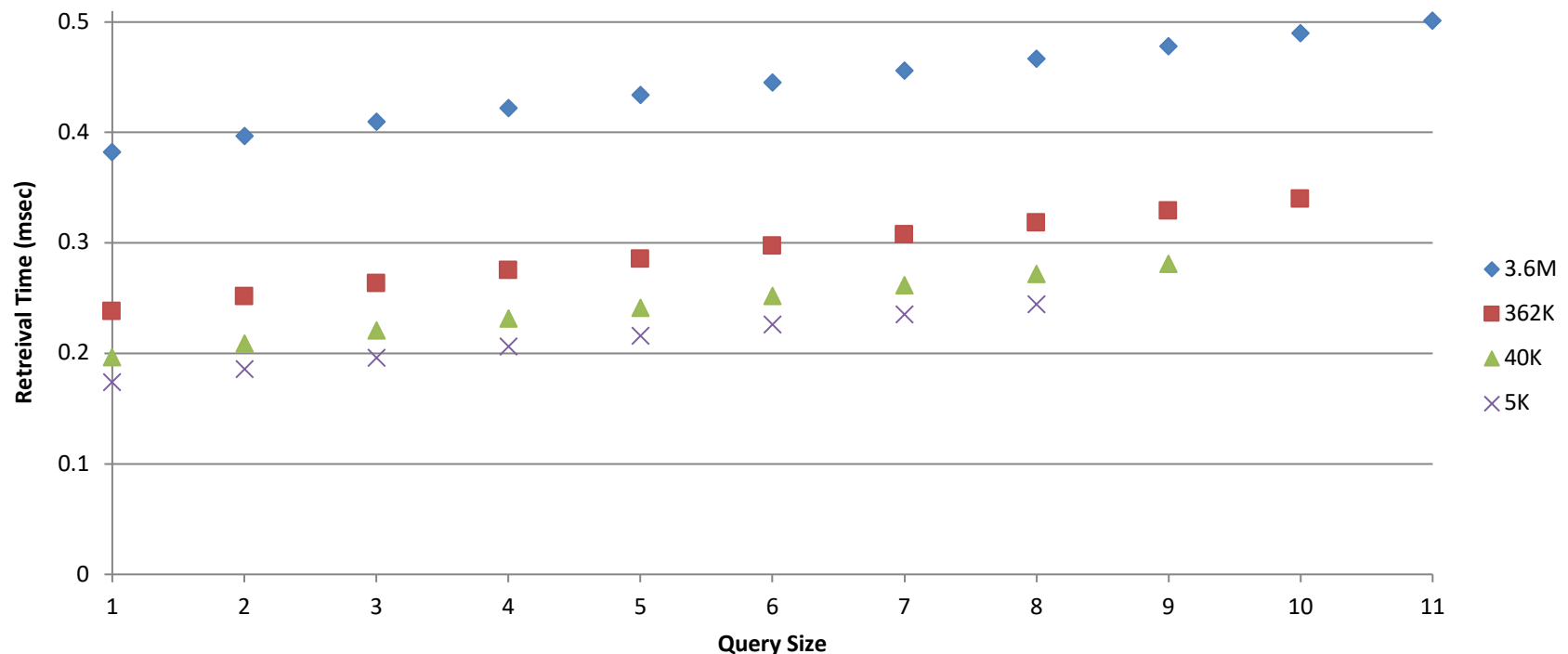
Inverted Index



Some Results [ICCM '10]

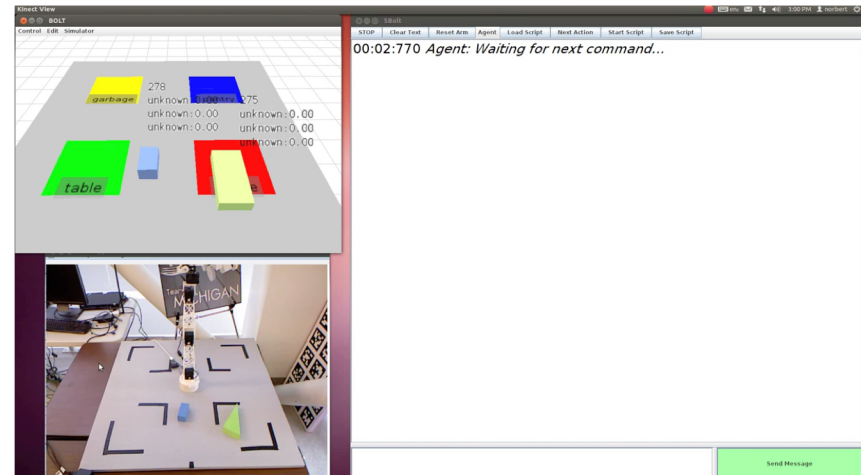
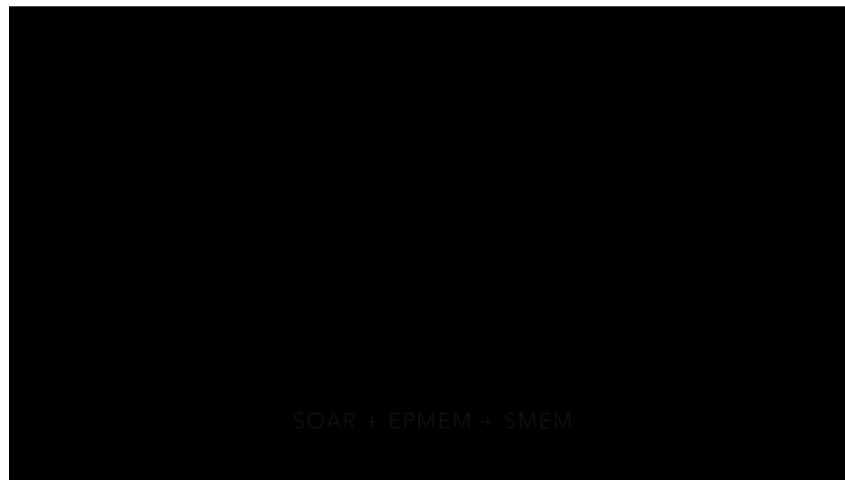
Inverted index (via SQLite) + new approach was **fast** and **scaled**!

>30x faster than off-the-shelf database (on >3x data)!



Applications: AI + Inverted Index

Learned Navigation

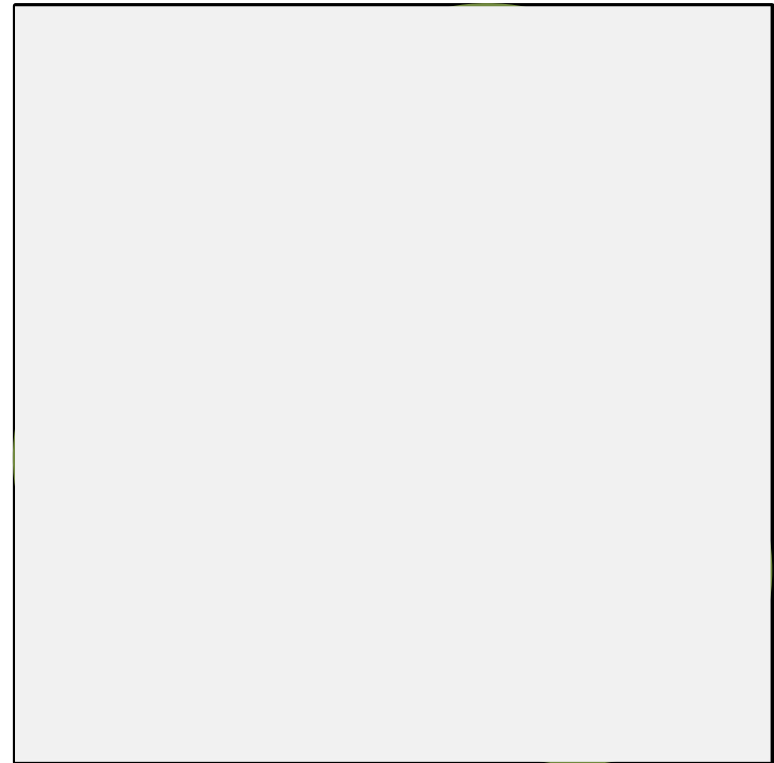
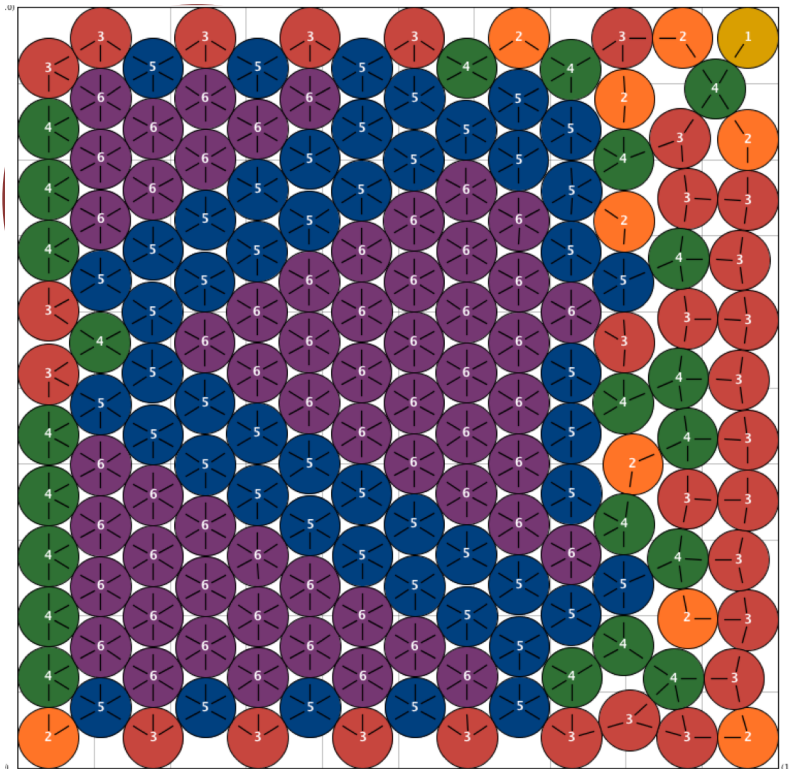


Task Learning



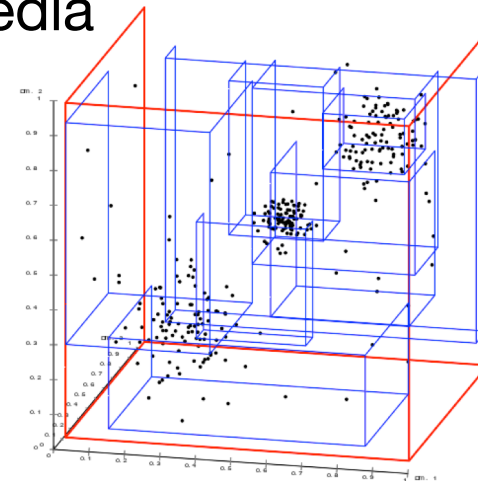
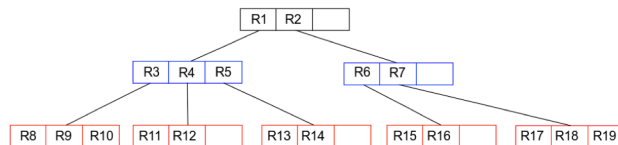
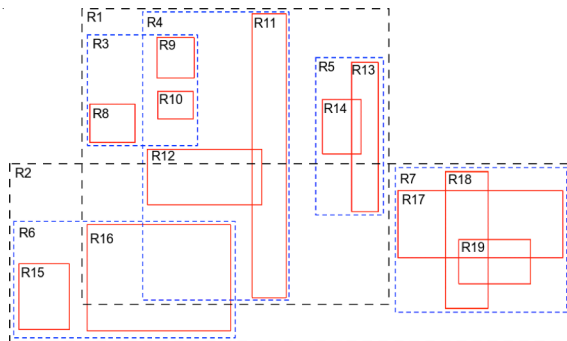
Research: Optimization

Packing. Fit n circles of radius r in a square of side-length s without overlap (non-convex, NP-hard, ∞ solutions). Used in making codes, physical packing, computer-assisted origami.



Another Index: R-tree

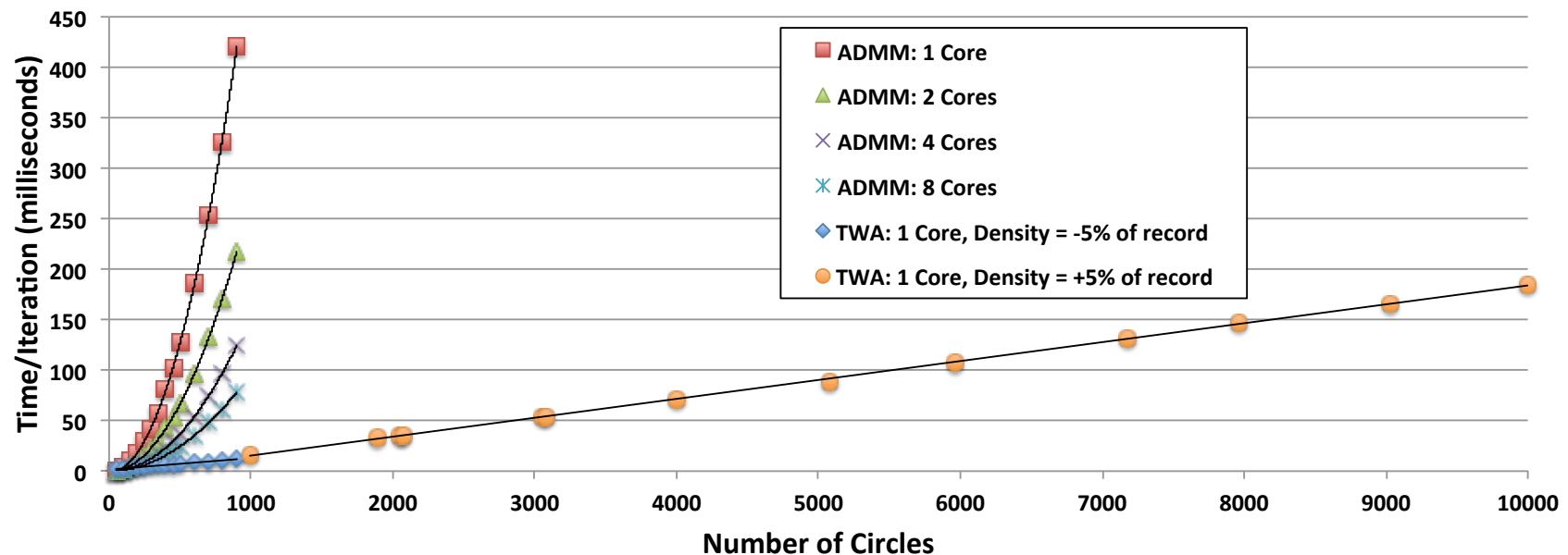
“Group nearby objects and represent them with their **minimum bounding rectangle** in the next higher level of the tree... Since all objects lie within this bounding rectangle, a query that does not intersect the bounding rectangle also cannot intersect any of the contained objects. At the leaf level, each rectangle describes a single object; at higher levels the aggregation of an increasing number of objects.” -- Wikipedia



Core Idea: reduce the constraints via relative distance



Large-Scale Evaluation [BICA '14]



Takeaways

- A common approach to large-scale search is indexing: using data structure(s) to improve access speed
- An inverted index is commonly used for full-text search (even in situations that might not look like it)
 - Inverted indexes are fast, scalable, and straight-forward to implement
- An R-tree is commonly used for spatial queries over objects in 2/3D space (e.g. what is within X miles of Y? are A and B colliding?)
- Know your indexes/data structures - careful problem analysis and algorithm development can often beat generic approaches
 - Even if you don't use a DBMS, DBMS methods can be very useful in a variety of applications!

