Intelligent Information Retrieval

Course Overview

Information overload problem

- *A state of having more information available that one can readily assimilate, that is, people have difficulty absorbing the information into their base of knowledge. This hinders decision-making and judgment by causing stress and cognitive impediments such as confusion, uncertainty and distraction.*
- *A newer definition focuses on time and resources aspects. When a decision-maker is given many sets of information, the quality of its decision is decreased because of the individual's limitation of scarce resources to process all the information and optimally make the best decision.*

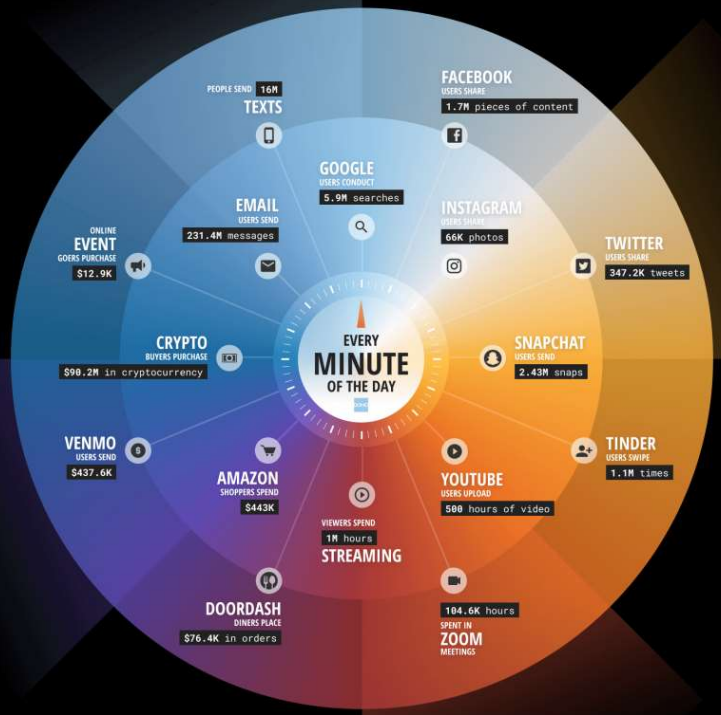
How much information? (2022)

Every minute:

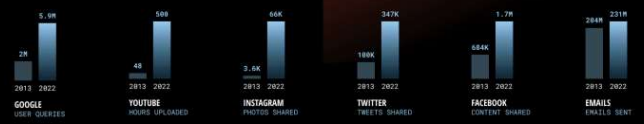
- GOOGLE users conduct 5.9M searches
- TWITTER users share 347.2k tweets
- INSTAGRAM users share 66k photos
- FACEBOOK users share 1.7M pieces of content
- AMAZON shoppers spend $443k
- SNAPCHAT users send 2.43M snaps
- EMAIL users send 231.4M messages
- People send 16M texts

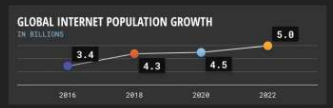
DATA NEVER SLEEPS 10.0

Over the last ten years, digital engagement through social media, streaming content, online purchasing, peer-to-peer payments and other activities has increased hundreds and even thousands of percentage points. While the world has faced a pandemic, economic ups and downs, and global unrest, there has been one constant in society: our increasing use of new digital tools to support our personal and business needs, from connecting and communicating to conducting transactions and business. In this 10th annual "Data Never Sleeps" infographic, we share a glimpse at just how much data the internet produces each minute from some of this activity, marveling at the volume and variety of information that has been generated.



DATA NEVER SLEEPS 1.0 VS. 10.0





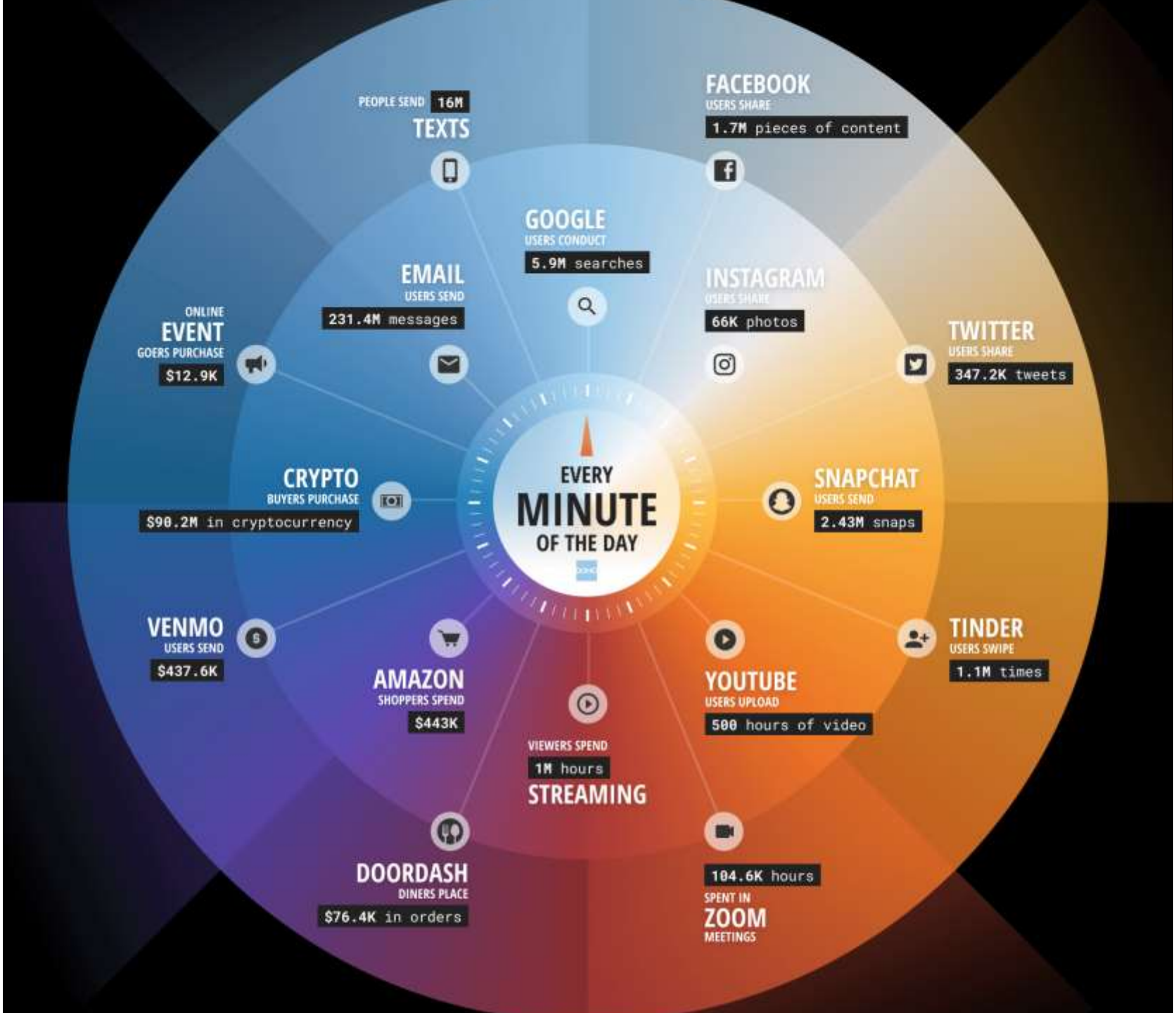
As of April 2022, the internet reaches 63% of the world's population, representing roughly 5 billion people. Of this total, 4.65 billion - over 93 percent - were social media users. According to Statista, the total amount of data predicted to be created, captured, copied and consumed globally in 2022 is 97 zettabytes, a number projected to grow to 181 zettabytes by 2025.

To succeed in an increasingly digital world where the volume of data created keeps accelerating, businesses need the right tools to put that data to work right where work gets done. Domo gives you the power to rapidly unlock value from all your data, regardless of where it lives, and drive actions across your organization that will improve business outcomes. Every click, swipe, share, or like tells a story, and Domo helps you do something powerful with it.

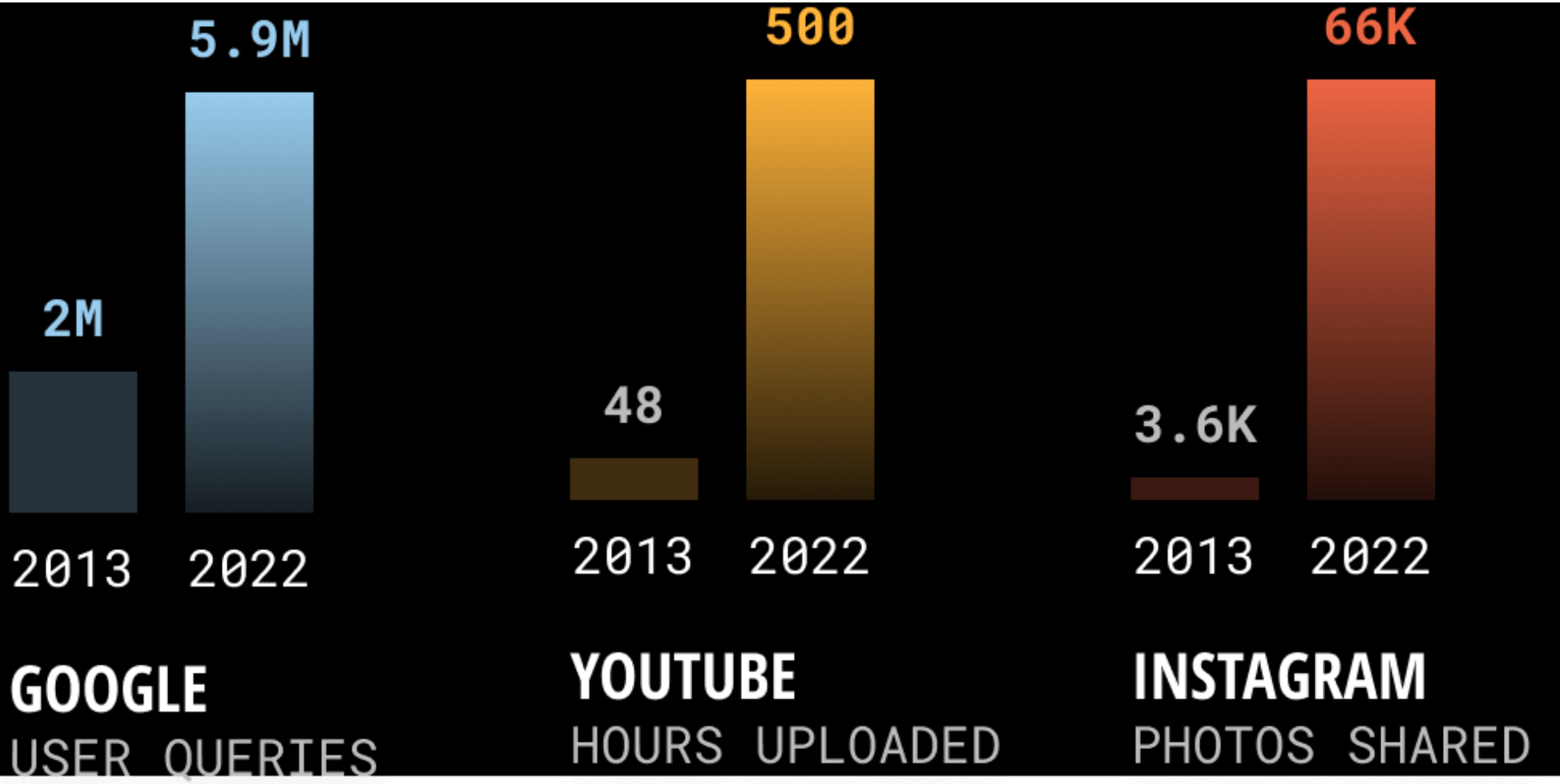
LEARN MORE AT DOMO.COM

SOURCES

Global Media Insight, Oberlo, Hootsuite, Earthweb, Matthew Woodward.co.uk, Web Tribunal, Deadline.com, Local IQ, Business of Apps, Query Sprout, Young and the Invested, Dating Zest, IBIS World, DoorDash, TechCrunch, Statista, Data Never Sleeps 1.0


EVERY
MINUTE
OF THE DAY
PEOPLE SEND 16M
TEXTS
GOOGLE
USERS CONDUCT
5.9M searches
FACEBOOK
USERS SHARE
1.7M pieces of content
EMAIL
USERS SEND
231.4M messages
INSTAGRAM
USERS SHARE
66K photos
ONLINE
EVENT
GOERS PURCHASE
$12.9K
TWITTER
USERS SHARE
347.2K tweets
CRYPTO
BUYERS PURCHASE
$90.2M in cryptocurrency
SNAPCHAT
USERS SEND
2.43M snaps
VENMO
USERS SEND
$437.6K
AMAZON
SHOPPERS SPEND
$443K
TINDER
USERS SWIPE
1.1M times
YOUTUBE
USERS UPLOAD
500 hours of video
VIEWERS SPEND
1M hours
STREAMING
DOORDASH
DINERS PLACE
$76.4K in orders
104.6K hours
SPENT IN
ZOOM
MEETINGS

Data Never Sleeps 1.0
vs.
Data Never Sleeps 10.0



The special role of textual information

- The most natural way of encoding knowledge
  - Think about scientific literature
- The most common type of information
  - Think about the amount of textual information we produce and consume every day
- A universal representation language
  - It can be used to describe other media of information

We mostly focus on text information in this course

Motivation: Harnessing big text data

- Text data is **ubiquitous** and **growing rapidly**
  - Internet
  - Blogs
  - News
  - Email
  - Literature
  - Twitter
  - …

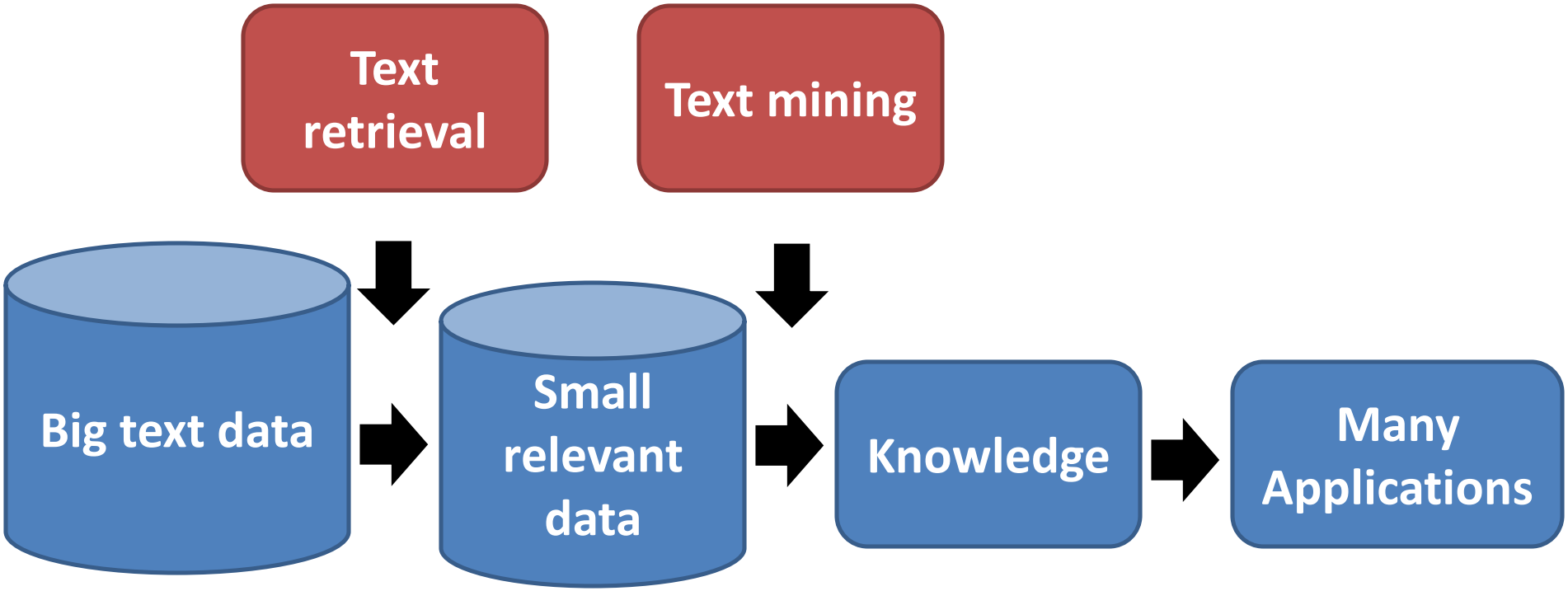
→ **Knowledge** → **Many applications!**

Many challenges

- How can we
  - find useful information?
  - organize information automatically?
  - extract patterns?
  - …

How to manage text information effectively and efficiently?

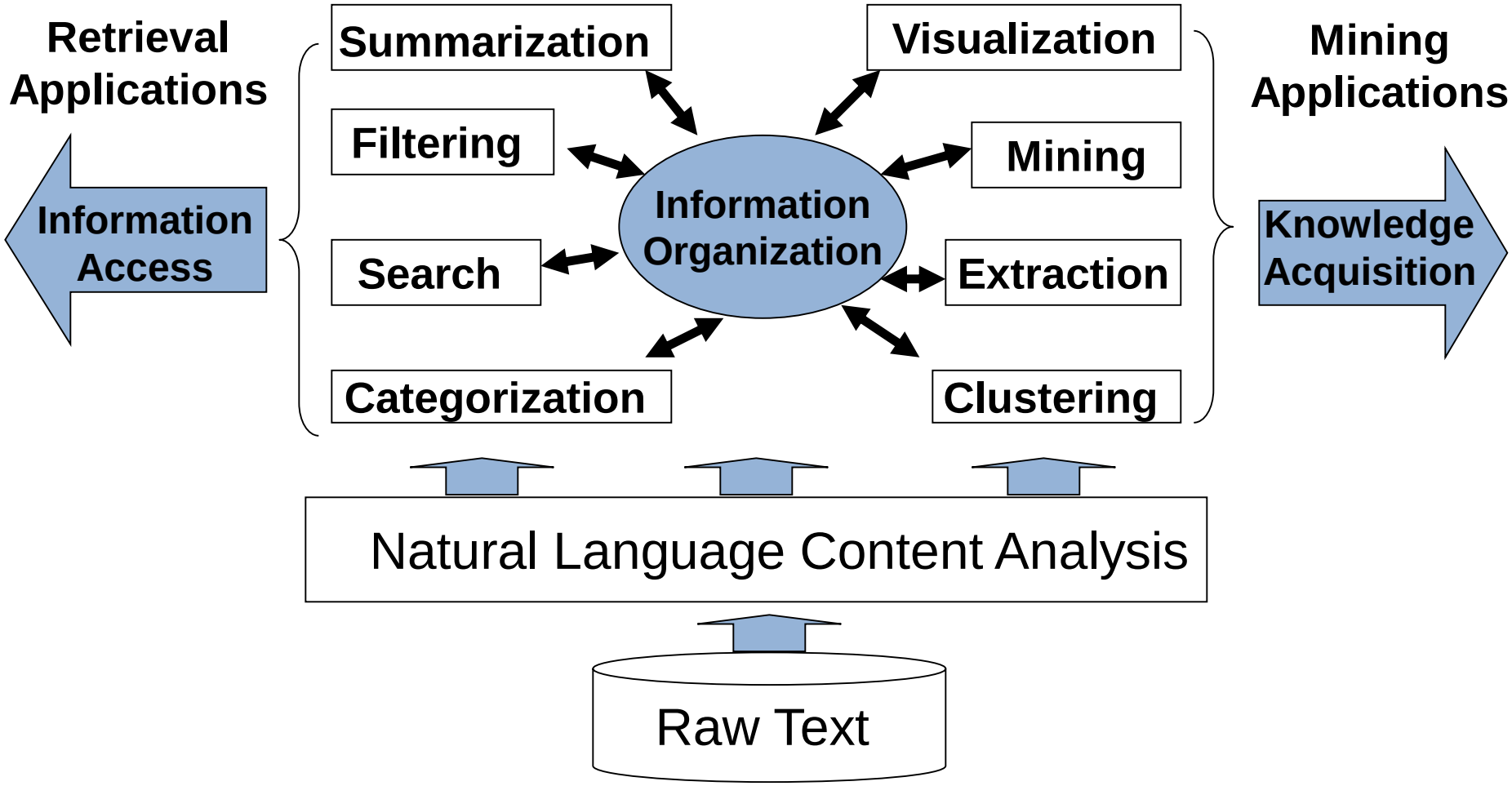
Main techniques for harnessing big text data: text retrieval + text mining



Examples of text management applications

- Search
  - Web search engines (Google, Bing, …)
  - Library systems
- Filtering/Recommendation
  - News filter
  - Spam email filter
  - Literature/movie recommender
- Categorization
  - Automatically sorting emails
- Mining/Extraction
  - Discovering major complaints from email in customer service
  - Bioinformatics
- Many others…

Elements of text information management technologies



Major Research Milestones

- Early days (late 1950s to 1960s): foundation and founding of the field
  - Luhn's work on automatic encoding
  - Cleverdon's Cranfield evaluation methodology and index experiments
  - Salton's early work on SMART system and experiments
- 1970s-1980s: a large number of retrieval models
  - Vector space model
  - Probabilistic models
- 1990s: further development of retrieval models and new tasks
  - Language models
  - TREC evaluation
- 2000s-present: more applications, especially Web search and interactions with other fields
  - Web search
  - Learning to rank
  - Scalability (e.g., MapReduce)
  - Neural IR

Indexing: auto vs. manual

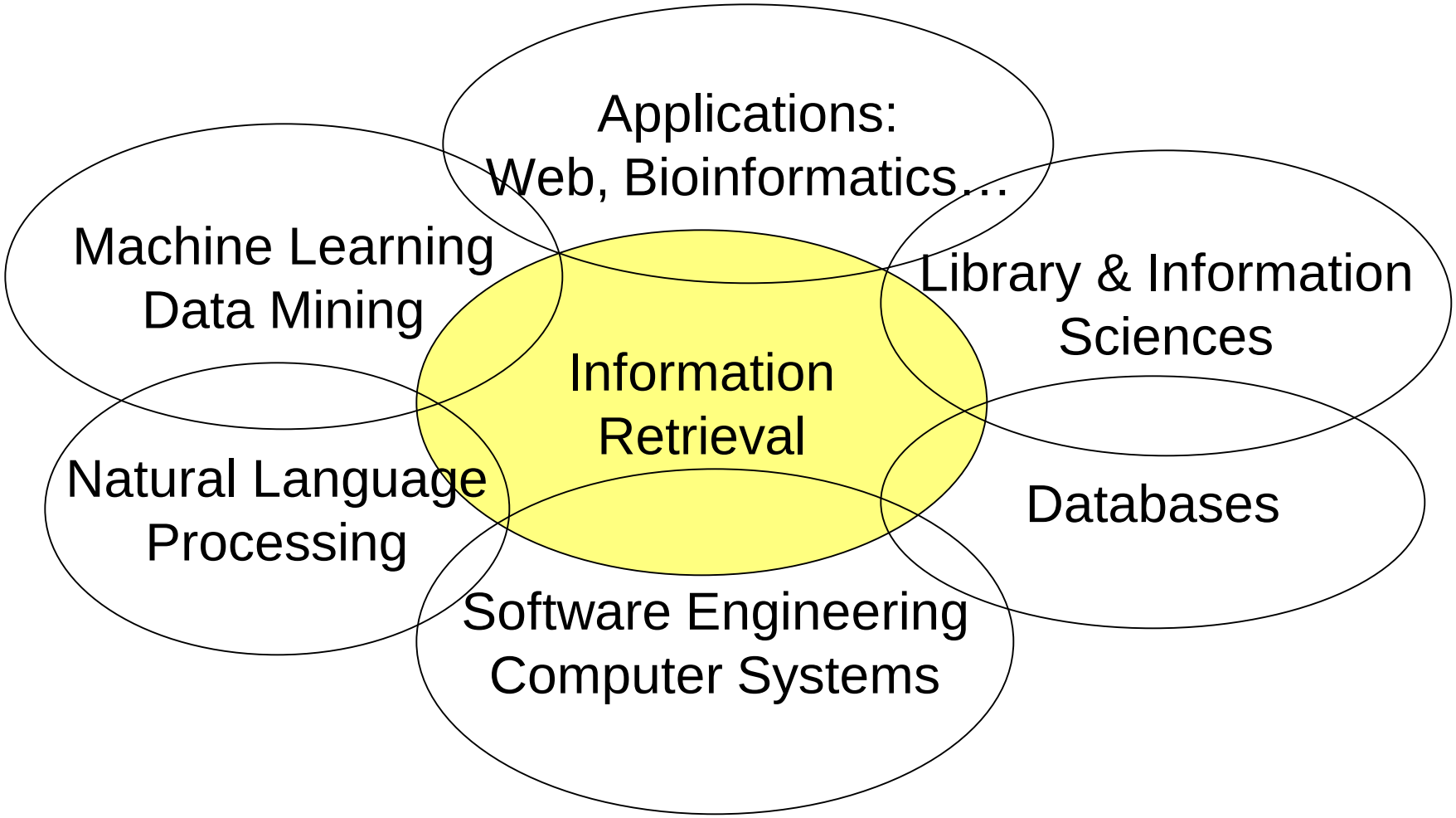
Evaluation System

Indexing + Search Theory

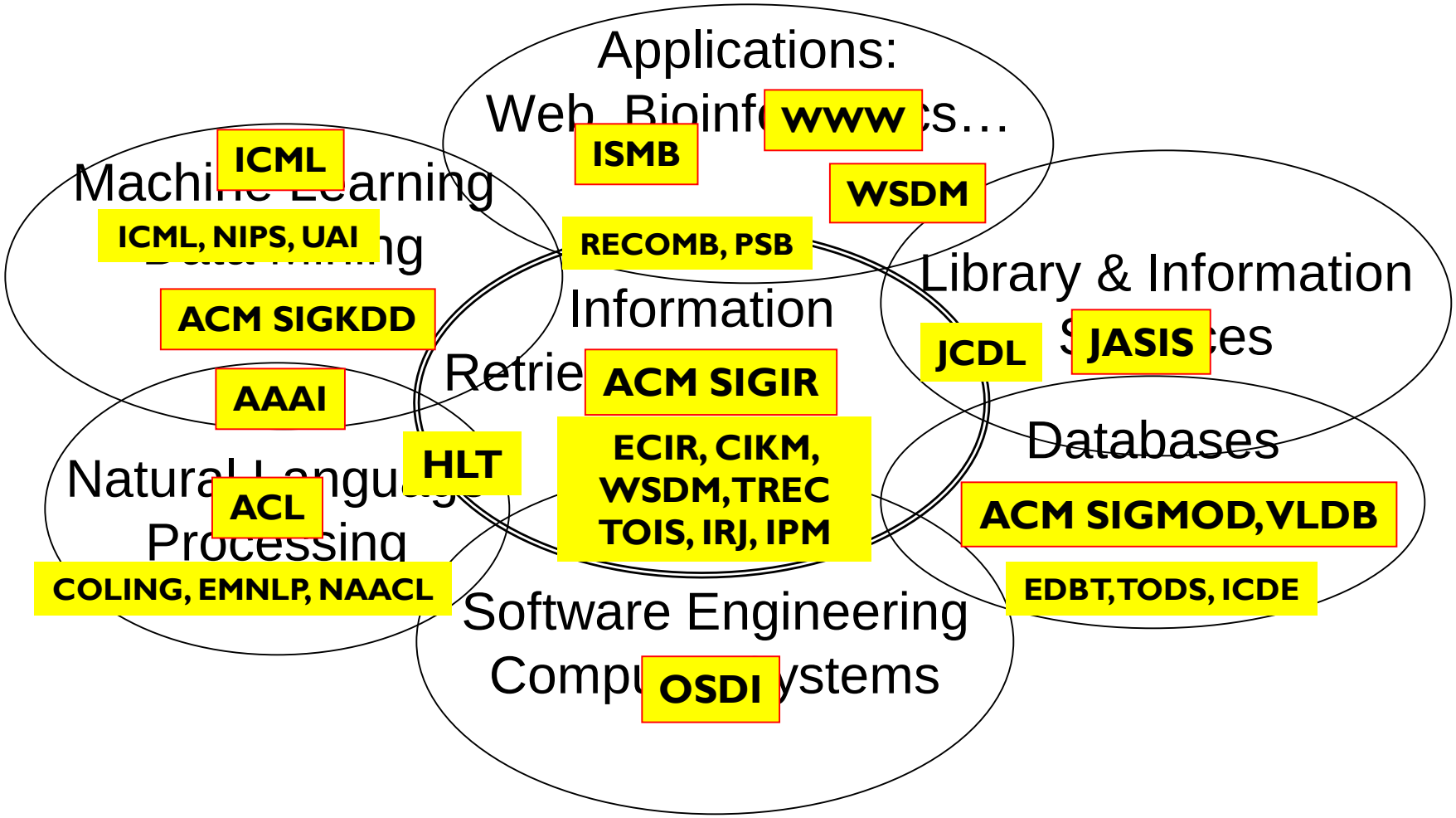
Large-scale evaluation, beyond ad hoc retrieval

Web search
Machine learning
Scalability

Related research areas



Publications/societies



Questions?