
Foundations of Social and Cultural Data Analysis

Dr. Nanne van Noord & Dr. Melvin Wevers



Recap



Assignment

- Plot types
- Comment quality
- Boolean expressions!



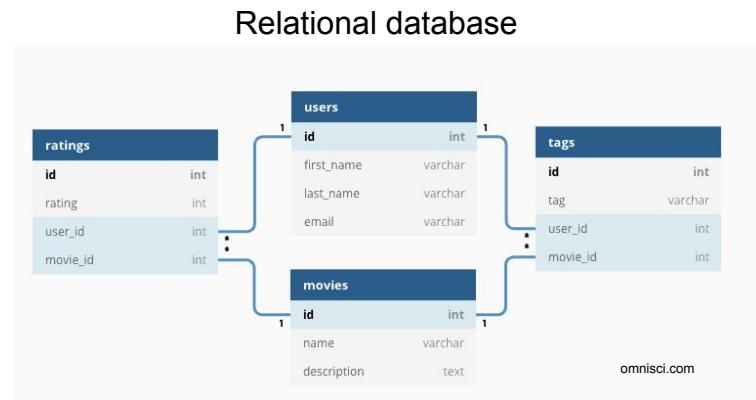
What is data?

- Data can be categorised based on different schemes:
 - **Quantitative** / qualitative
 - **Structured** / semi-structured / unstructured

Structured data

- Has a clearly defined and rigid format
- Easily searchable
- Most common format: table structure

ID	Shape	Colour
1	Round	Red
2	Square	Green



Semi-structured data

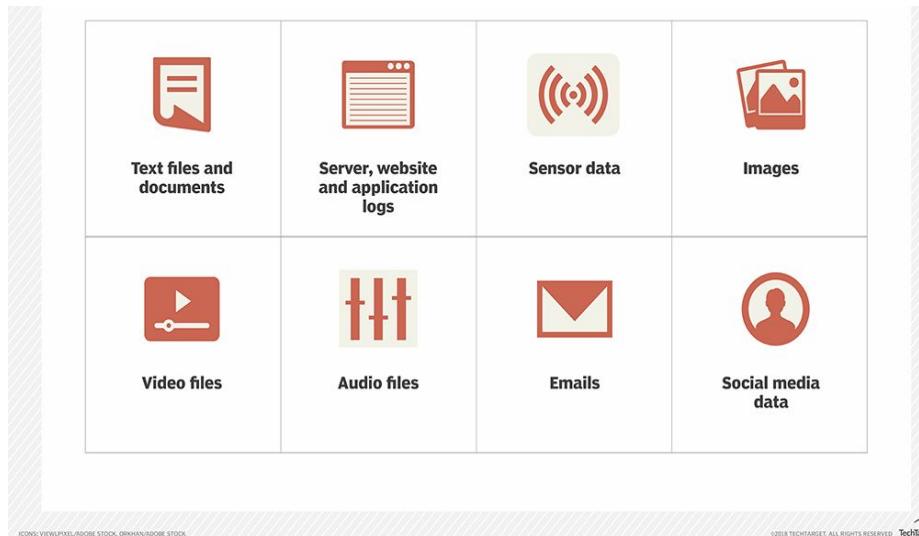
- Less rigid form of structured data
- ‘Self-describing’: entries contain information about their structure
- Prominent examples: XML and JSON

```
<events>
  <event>
    <date>2013-11-05</date>
    <locations>
      <location id="4">United States</location>
      <location id="8">Germany</location>
    </locations>
  </event>
  <event>
    <date>2013-11-11</date>
    <locations>
      <location id="9">South Africa</location>
    </locations>
  </event>
  <event>
    <date>2013-11-12</date>
    <locations>
      <location id="6">Japan</location>
    </locations>
  </event>
</events>
```

```
[
  {
    "date": "2013-11-05",
    "locations": {
      "United States": 4,
      "Germany": 8
    }
  },
  {
    "date": "2013-11-11",
    "locations": {
      "South Africa": 9
    }
  },
  {
    "date": "2013-11-12",
    "locations": {
      "Japan": 6
    }
  }
]
```

Unstructured data

- Lacks easily identifiable structure
- Perhaps better described as: “Everything that isn’t structured or semi-structured”



Data terminology

- (Independent) Variable / Column / Feature / attribute
- Instance / Row / Record / Entry

ID	X1	X2	X3
1	4.5	Square	A
2	9	Round	B

Data terminology: variables

- For any given variable we can determine:
 - The variable type
 - The measurement scale

Variable Types

- **Discrete** variables can only take certain values, and no in-between values
 - Things we can count
 - Number of students in class
 - View count of a video

Variable Types

- **Continuous** variables can be any value (in a range)
 - Uncountable
 - Temperature

Variable types

- Knowing if a variable is discrete or continuous gives us a lot of information about how we can interpret it, and how we should treat it when doing:
 - Visualisations
 - Interactions
 - Calculations
 - Analysis



Measurement Scales

- Another aspect of a variable we can use to describe it is the measurement scale
- The measurement scale tells us:
 - How to interpret the values on the scale in relation to the other values
 - How to compare the values

Measurement Scales

Categorical / Nominal

- Values that have no intrinsic ordering and we can only compare based on equality
 - Names
 - Species of animals or plants



- We can say that:
 - The species of animal 1 and 2 are the same
 - The species of animal 2 and 4 are different
- Cannot say:
 - The species of animal 5 is greater than the species of animal 2

Measurement Scales

Ordinal

- Values have an intrinsic ordering, can compare based on equality and ordering (Military ranks, Ratings in stars)

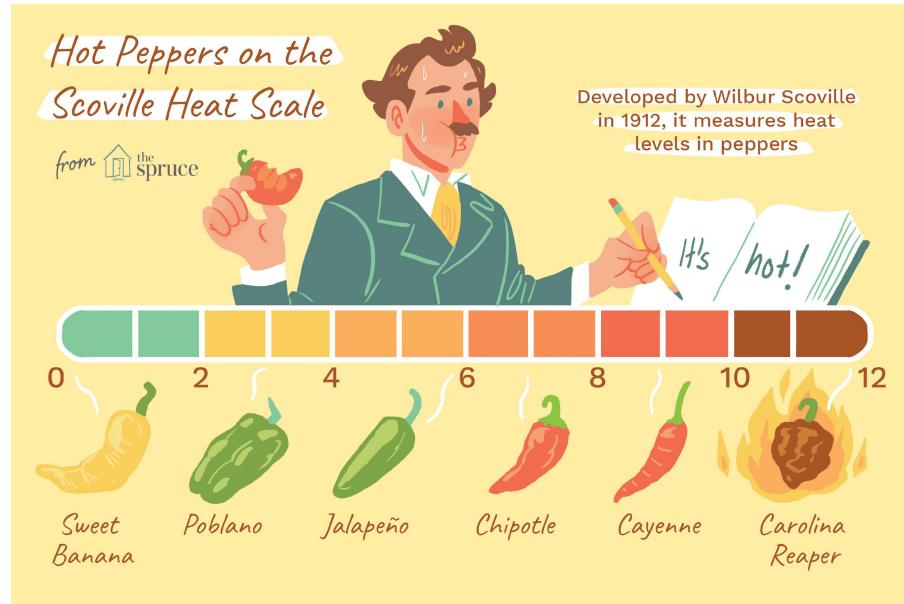


A movie with 4 stars is rated higher than a movie with 2 stars.

But we can't definitely express this difference, how much '2 stars better' is, is subjective.

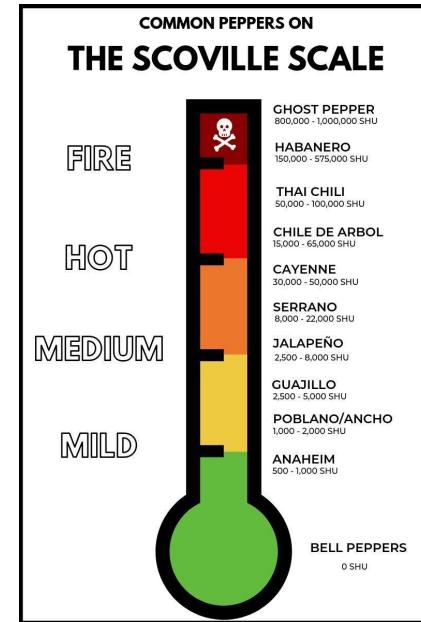
Nor does the rating mean that a movie which is twice as good has twice as many stars.

Scoville heat scale



Ordinal

High-performance liquid chromatography (HPLC)



isabeleats.com/scoville-scale/

Interval

Measurement Scales

Interval

- Meaningful difference (which can be expressed in a unit)
 - For example: the difference between 10 Celsius and 15C is the same as the difference between 20C and 25C
 - **But!** 20C is not twice as hot as 10C



Measurement Scales

Ratio

- Meaningful proportions (0 is the lowest value)
 - Only scale where 0 has an intrinsic meaning, and where 0 is 'absolute'
 - 20 years is twice as long as 10 years
 - 20 meter is twice as far as 10 meters

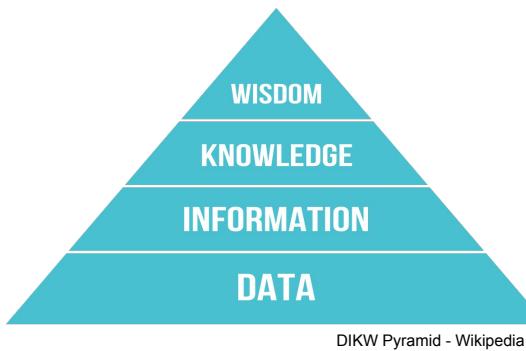


Binary variables

- Special case of discrete categorical variable
- Can only take on two values
- Very common in datasets
- Doesn't have to be 0/1, but can always be represented as 0/1
 - Yes/no
 - On/off
 - Sex: male/female

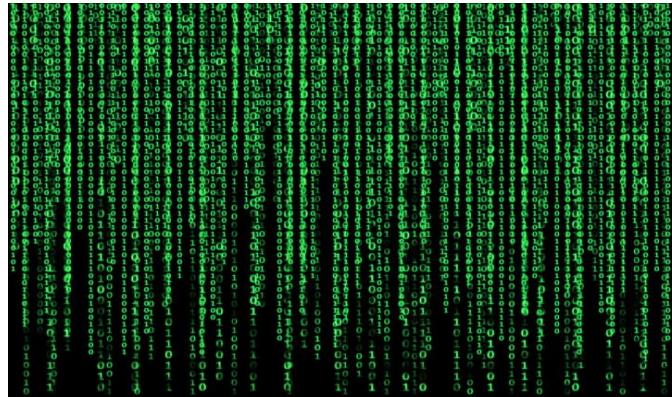
Data vs. Information

- Information has a context; we can interpret and communicate it, and it can be subjective

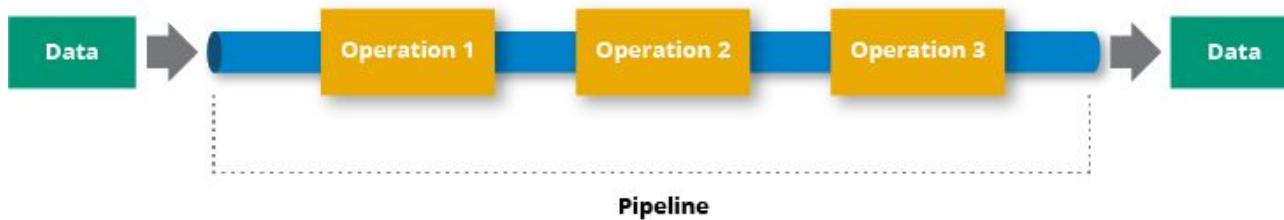


What is data?

- Raw values/facts without context, but can have structure
- Clear terminology
- Something that needs interpretation before you can ‘see the woman in the red dress’



Retrieving Data

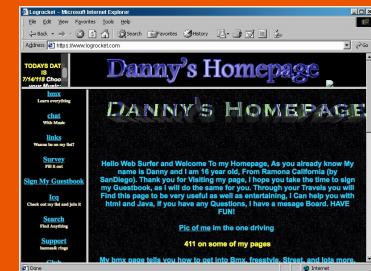
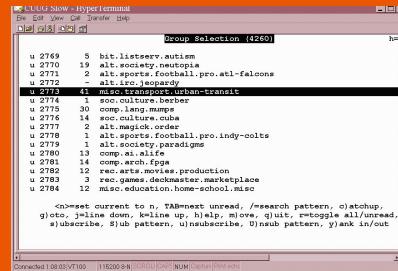


Technological Innovation

2000 BC



Now



1890

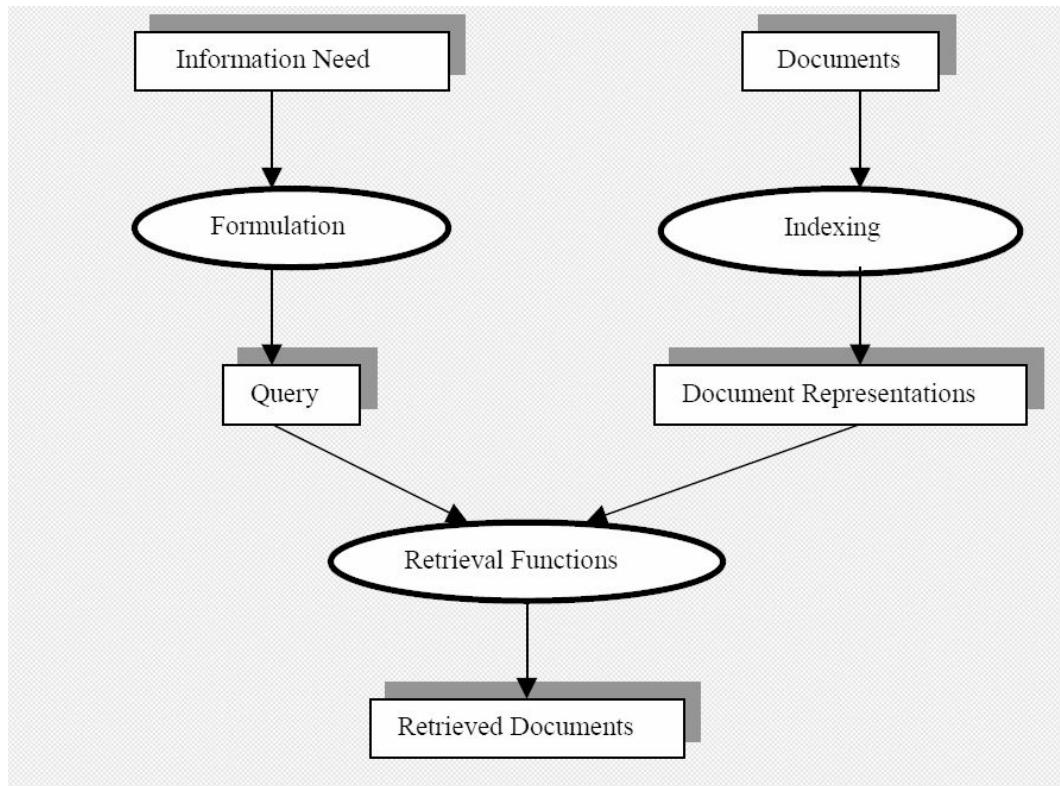
1979

1983

~2000

2023

Sharing/Search Relation

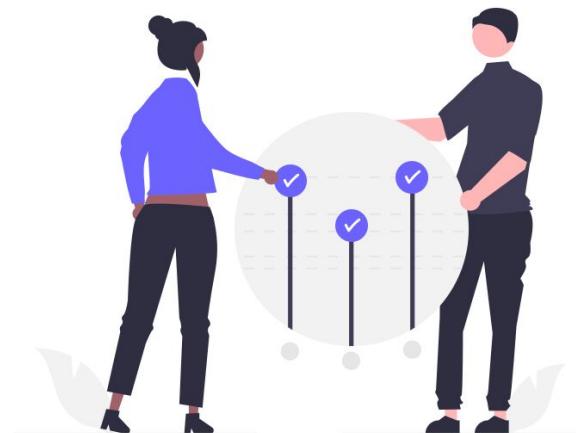


Touches on Various Disciplines

- Data Science / AI / Machine Learning
 - Discover insights in/from data
- Data Engineering
 - Develop systems for data at scale
- Knowledge Engineering
 - Systems and expertise surrounding knowledge as “data”
- Information Retrieval
 - Systems and methods for finding/searching information
- Library Science
 - Focuses on how to classify and use data or objects and emphasizes the importance of preserving knowledge and promoting literacy.
- etc.

Components of Data Retrieval Systems

- Storage
- Representation/Encoding
- Indexing
- Querying
- Ranking
- Presentation



How to store the data depends on **how** it will be retrieved.

Storage



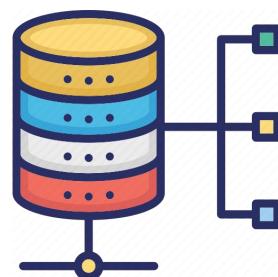
Library:

- Manual access
- Browsing
- Catalogued



File storage:

- Digital access
- Browsing
- Hierarchical



Database:

- Digital access
- Query-based
- Schema



Data Warehouse:

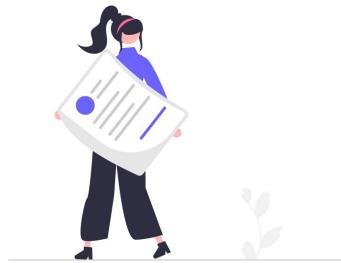
- Digital access
- Analytics
- Schema

Data Structure Complexity

Data Structure	Time Complexity								Space Complexity	
	Average				Worst					
	Access	Search	Insertion	Deletion	Access	Search	Insertion	Deletion		
Array	$\Theta(1)$	$\Theta(n)$	$\Theta(n)$	$\Theta(n)$	$O(1)$	$O(n)$	$O(n)$	$O(n)$	$O(n)$	
Stack	$\Theta(n)$	$\Theta(n)$	$\Theta(1)$	$\Theta(1)$	$O(n)$	$O(n)$	$O(1)$	$O(1)$	$O(n)$	
Queue	$\Theta(n)$	$\Theta(n)$	$\Theta(1)$	$\Theta(1)$	$O(n)$	$O(n)$	$O(1)$	$O(1)$	$O(n)$	
Singly-Linked List	$\Theta(n)$	$\Theta(n)$	$\Theta(1)$	$\Theta(1)$	$O(n)$	$O(n)$	$O(1)$	$O(1)$	$O(n)$	
Doubly-Linked List	$\Theta(n)$	$\Theta(n)$	$\Theta(1)$	$\Theta(1)$	$O(n)$	$O(n)$	$O(1)$	$O(1)$	$O(n)$	
Skip List	$\Theta(\log(n))$	$\Theta(\log(n))$	$\Theta(\log(n))$	$\Theta(\log(n))$	$O(n)$	$O(n)$	$O(n)$	$O(n)$	$O(n \log(n))$	
Hash Table	N/A	$\Theta(1)$	$\Theta(1)$	$\Theta(1)$	N/A	$O(n)$	$O(n)$	$O(n)$	$O(n)$	
Binary Search Tree	$\Theta(\log(n))$	$\Theta(\log(n))$	$\Theta(\log(n))$	$\Theta(\log(n))$	$O(n)$	$O(n)$	$O(n)$	$O(n)$	$O(n)$	
Cartesian Tree	N/A	$\Theta(\log(n))$	$\Theta(\log(n))$	$\Theta(\log(n))$	N/A	$O(n)$	$O(n)$	$O(n)$	$O(n)$	
B-Tree	$\Theta(\log(n))$	$\Theta(\log(n))$	$\Theta(\log(n))$	$\Theta(\log(n))$	$O(\log(n))$	$O(\log(n))$	$O(\log(n))$	$O(\log(n))$	$O(n)$	
Red-Black Tree	$\Theta(\log(n))$	$\Theta(\log(n))$	$\Theta(\log(n))$	$\Theta(\log(n))$	$O(\log(n))$	$O(\log(n))$	$O(\log(n))$	$O(\log(n))$	$O(n)$	
Splay Tree	N/A	$\Theta(\log(n))$	$\Theta(\log(n))$	$\Theta(\log(n))$	N/A	$O(\log(n))$	$O(\log(n))$	$O(\log(n))$	$O(n)$	
AVL Tree	$\Theta(\log(n))$	$\Theta(\log(n))$	$\Theta(\log(n))$	$\Theta(\log(n))$	$O(\log(n))$	$O(\log(n))$	$O(\log(n))$	$O(\log(n))$	$O(n)$	
KD Tree	$\Theta(\log(n))$	$\Theta(\log(n))$	$\Theta(\log(n))$	$\Theta(\log(n))$	$O(n)$	$O(n)$	$O(n)$	$O(n)$	$O(n)$	

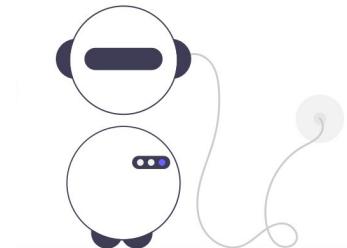
Storage

How to store the data depends on **who** will retrieve it.



Human:

- Human-readable
- Browsing
- Information-focused



Machine:

- Machine-readable
- Search
- Data-focused

Human vs Machine-readable

ISBN 978-3-16-148410-0



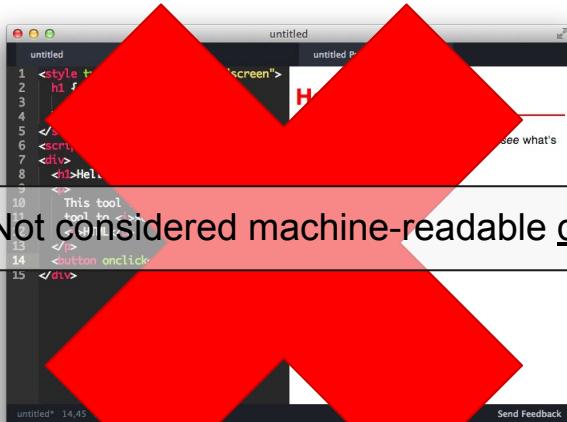
9 783161 484100

The main information in the Machine Readable Zone



1. Surname/name
 2. Nationality
 3. Date of birth
 4. Sex
 5. Document date of expiry

Format vs. Content



Not considered machine-readable data

	A	B	C	D
1	ID	Gender	City	Monthly_Income
2	ID000002C	Female	Delhi	20000
3	ID000004E	Male	Mumbai	35000
4	ID000007H	Male	Panchkula	25000
5	ID000008I	Male	Saharsa	35000
6	ID000009J	Male	Bengaluru	100000
7	ID000010K	Male	Bengaluru	45000
8	ID000011L	Female	Sindhudhu	70000
9	ID000012N	Male	Bengaluru	20000
10	ID000013M	Male	Kochi	75000
11	ID000014C	Female	Mumbai	30000
12	ID000016C	Male	Mumbai	25000
13	ID000018S	Female	Surat	25000
14	ID000019T	Female	Pune	24000
15	ID000021V	Male	Bhubanes	27000
16	ID000022V	Female	Howrah	28000

CSV

JSON

```
"Employee": [  
    {  
        "id": "1",  
        "Name": "Ankit",  
        "Sal": "1000",  
    },  
    {  
        "id": "2",  
        "Name": "Faizv"
```

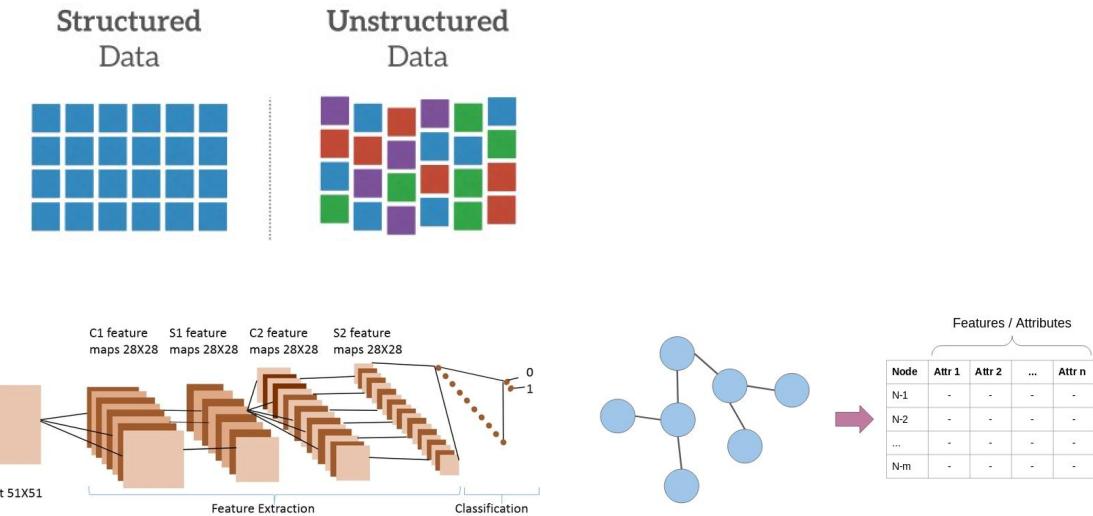
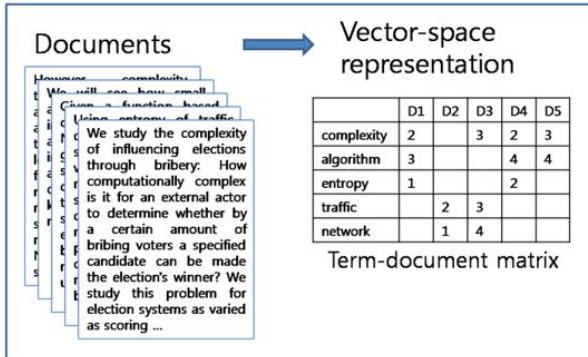
XML

```
<?xml version="1.0"?>  
  
<contact-info>  
  
<name>Ankit</name>  
  
<company>Analytics Vidhya</company>  
  
<phone>+9187654321</phone>  
  
</contact-info>
```

Representation/Encoding

Machine/Human readable is a representation choice

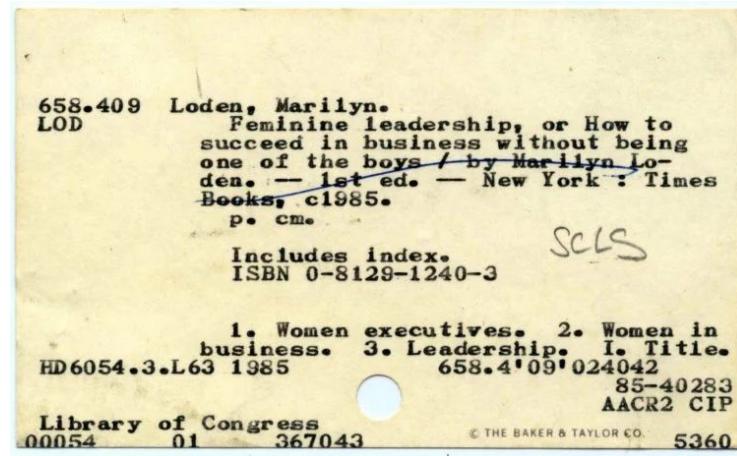
- Possible to have parallel representations



Representation/Encoding

It is also a choice of what to represent (abstraction/modelling)

- Informed by how we might want to retrieve



Indexing

To aid search we may want to construct a representation of the data that is specifically tailored to enable search.

Index

Entry titles are printed in small capitals. Bold page-numbers indicate a sustained discussion of a topic, whether or not it features as an entry.

ÄSTHETIKS 18, **31–5**, 123, 251–3; and
ETHIKS 28, 31–2, 107
ability *see* disposition
abstractionism 314
acquaintance 43, 102, 160, 208, 212,
254, 269–70, 277, 299, 310, 348–9,
353
agreement 128–9, 135–6, 328, 368
ambiguity/synonymy 40, 122, 240
analytic/synthetic 18, 20, 131, 199–200,
202, 353, 356–7
analyticity 26, 33, 35, **113–14**,
120–4, 152
ancestral relation 266
'and so on' 149, 265, 328
Anscombe, G.E.M. 29, 74, 75
antecedent 95, 301, 384
ANTHROPOLOGY **35–6**, 126, 128, 236; *se*
also HUMAN BEING
a priori *see* analytic/synthetic; philosophy;
synthetic a priori
Aquinas, T. 323
approximation 303
Aristotle 29, 43, 124, 199, 212, 220, 226,
241, 292, 294, 300, 318, 340, 354, 362
arithmetic 20, 24, 234
argument 34–5, 49
aspect-dawning 38–9
ASPECT-PERCEPTION 27, 34, **36–40**, 57,
120, 170; continuous 40
assertion **60–3**, 301–2
assumption 61–3
Augustine 23, 24, 277, 285, 295
AUGUSTINIAN PICTURE OF LANGUAGE 25,
41–5, 144, 175, 195, 211, 238, 255–6,
274, 277, 310, 376

Doc 1:
I did enact Julius
Caesar: I was killed
't the Capitol,Brutus
killed me.

⇒ Tokenisation

Doc 2:
© let it be with
Caesar. The noble
Brutus hath told
you Caesar was
ambitious.

⇒ Tokenisation

Term	docID
I	1
did	1
enact	1
julius	1
caesar	1
I	1
was	1
killed	1
i'	1
the	1
capitol	1
brutus	1
killed	1
me	1
so	2
let	2
it	2
be	2
with	2
caesar	2
the	2
noble	2
brutus	2
hath	2
told	2
you	2
caesar	2
was	2
ambitious	2

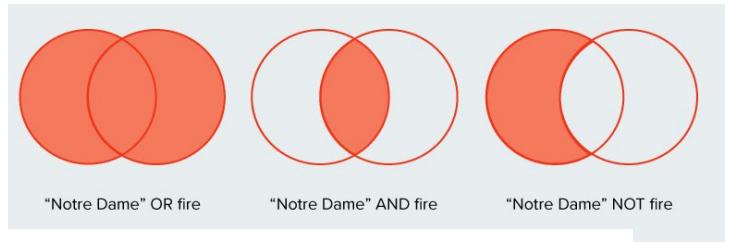
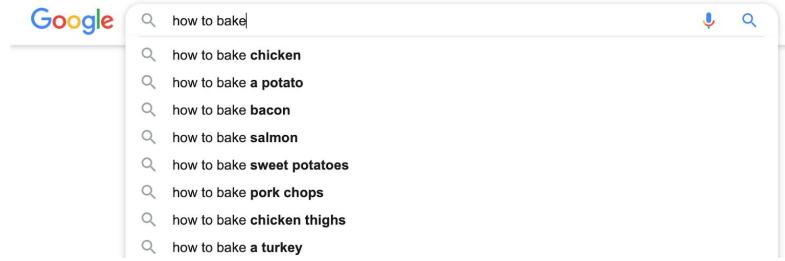
Term (sorted)	docID
ambitious	2
be	2
brutus	1
brutus	2
capitol	2
caesar	1
caesar	2
caesar	2
did	1
enact	1
hath	1
I	1
i'	1
it	2
julius	1
killed	1
let	2
me	1
noble	2
so	2
the	1
the	2
told	2
you	2
was	1
was	1
with	2

Querying

How to enable querying depends on needs and expertise of users.

Various types of queries:

- Keyword queries
 - Most common
 - Keywords implicitly connected by AND
- Boolean queries
 - Allow range of logical operators (AND, OR, NOT)
- Phrase queries
 - Search for exact multi-word match



Querying

Various types of queries:

- Keyword queries
 - Most common
 - Keywords implicitly connected by AND
- Boolean queries
 - Allow range of logical operators (AND, OR, NOT)
- Phrase queries
 - Search for exact multi-word match
- Wildcard queries
 - Based on (simple) regular expressions
- Natural language queries
 - System tries to interpret ‘meaning’ of sentence and search based on that (e.g., ChatGPT/Bing)
- Query-by-example
 - Similarity search based on a provide example (e.g., reverse image search)

Who

First name(s)

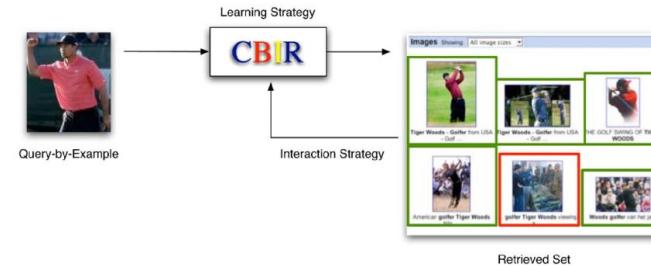
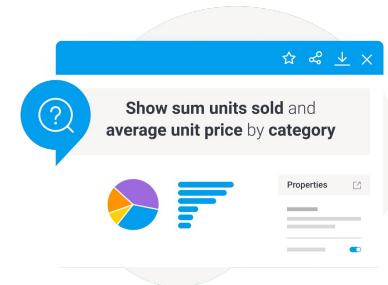
Last name

Fo*kes

Name variants i

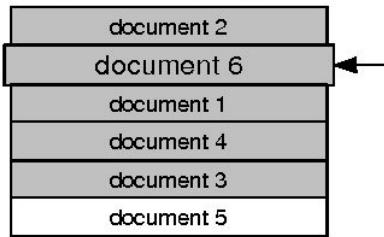
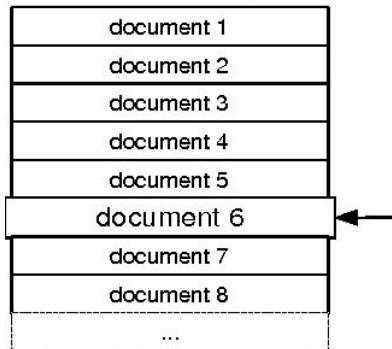
Name variants i

All fields are optional

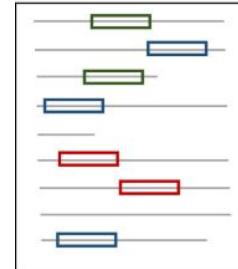


Ranking

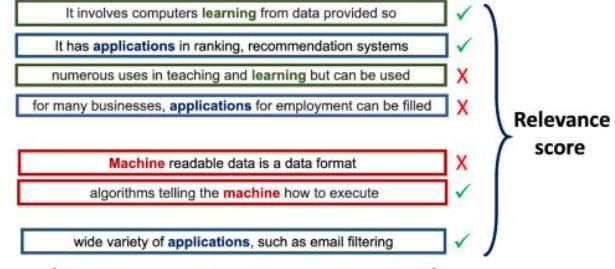
Give a score to each indexed document based on query and return in order



Document



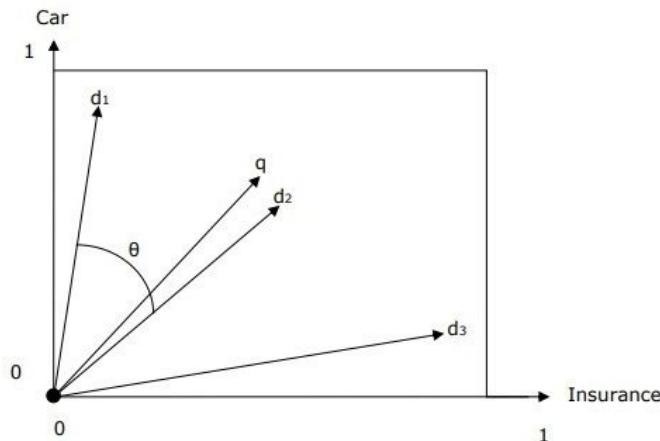
Query: Machine learning applications



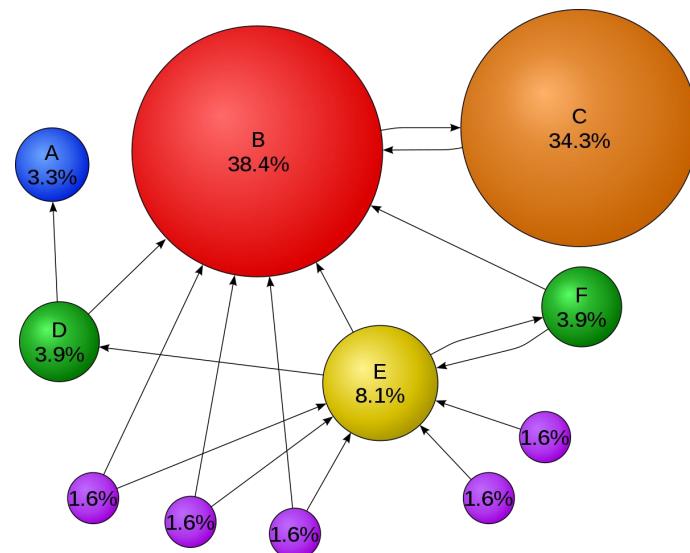
Query-Centric
contexts

Ranking - Relevance Score

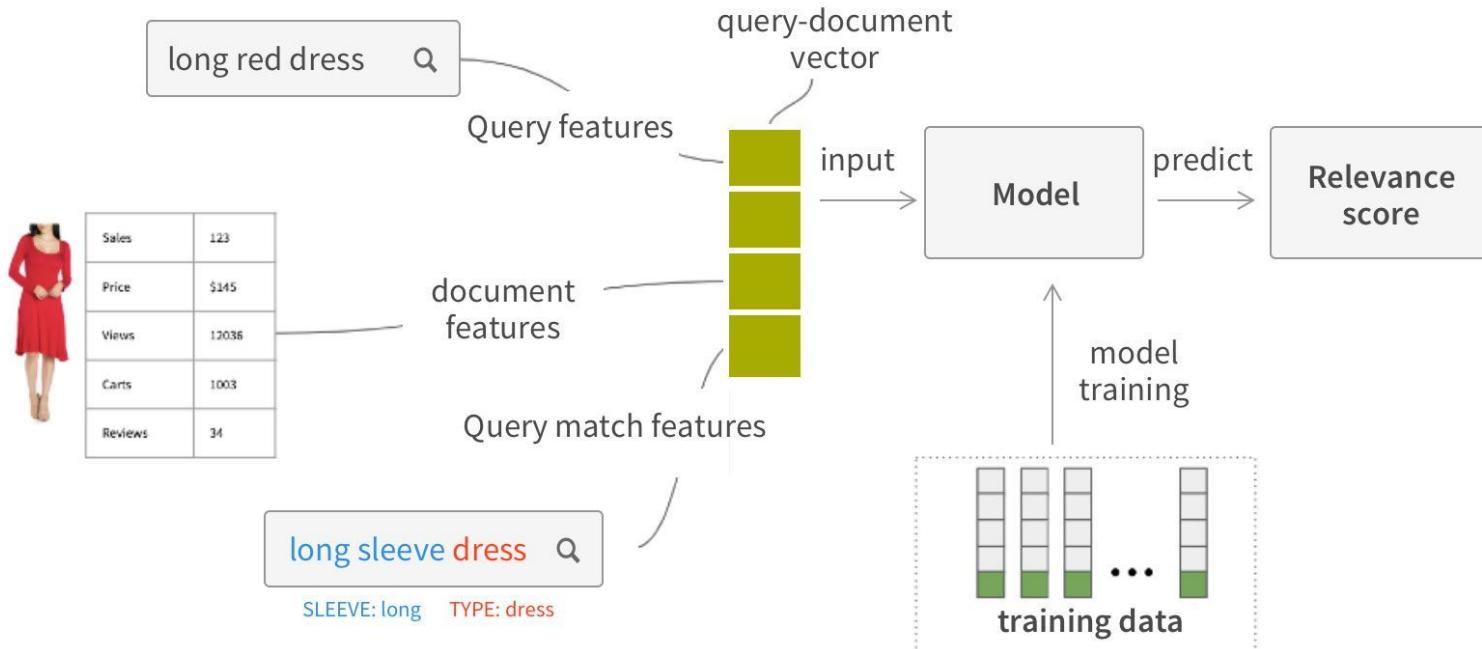
Vector Space Model



PageRank



Learning to Rank



Presentation

Google Search results for "France OR Spain":

- [Spain - Wikipedia](https://en.wikipedia.org/wiki/Spain)
Spain is a country in Southwestern Europe. Spain's conquest by France benefited Latin American anti-colonialists who...
- [France - Wikipedia](https://en.wikipedia.org/wiki/France)
France officially the French Republic, is a transcontinental country spanning Western Europe. Spain and France engaged in an all-out war between 1635 and 1659...
- About 8,120,000,000 results (0.85 seconds)

People also ask:

- Is France in Spain?



Microsoft Bing

SEARCH CHAT

I am planning a trip for our anniversary in September. What are some places we can go that are within a 3 hour flight from London Heathrow?

Congratulations on your anniversary! There are many places you can go that are within a 3 hour flight from London Heathrow. Here are some suggestions based on your preferences and the best destinations in Europe in September:

- If you like beaches and sunshine, you can fly to [Malaga](#) in Spain, which has a pleasant climate, a lively culture, and a beautiful coastline. You can enjoy the sandy beaches, the historic monuments, and the delicious tapas. Malaga is also a great base to explore other nearby attractions, such as Granada, Seville, and Cordoba.
- If you like mountains and lakes, you can fly to [Annecy](#) in France, which is a charming alpine town near the Swiss border. Annecy is known as the "Venice of France" because of its canals and bridges. You can admire the stunning scenery, visit the medieval castle, and try some local cheese and wine. Annecy is also close to Geneva, Chamonix, and Lyon, if you want to see more of the region.
- If you like art and history, you can fly to [Florence](#) in Italy, which is the birthplace of the Renaissance and a UNESCO World Heritage Site. Florence is a treasure trove of artistic and architectural masterpieces, such as the Duomo, the Uffizi Gallery, and the Ponte Vecchio. You can also explore the Tuscan countryside, taste the famous gelato, and shop for leather goods.

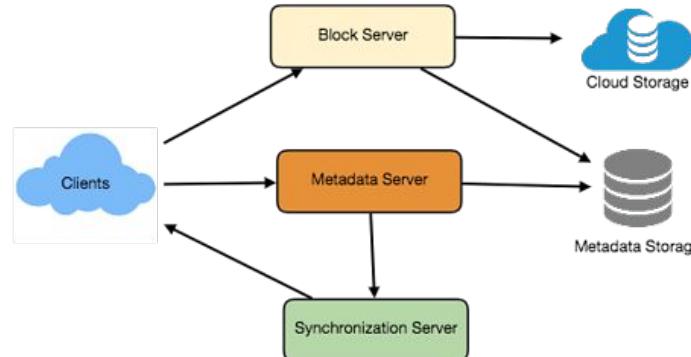
Ask me anything...

Different systems emphasise different components

- Storage
- Representation/Encoding
- Indexing
- Querying
- Ranking
- Presentation

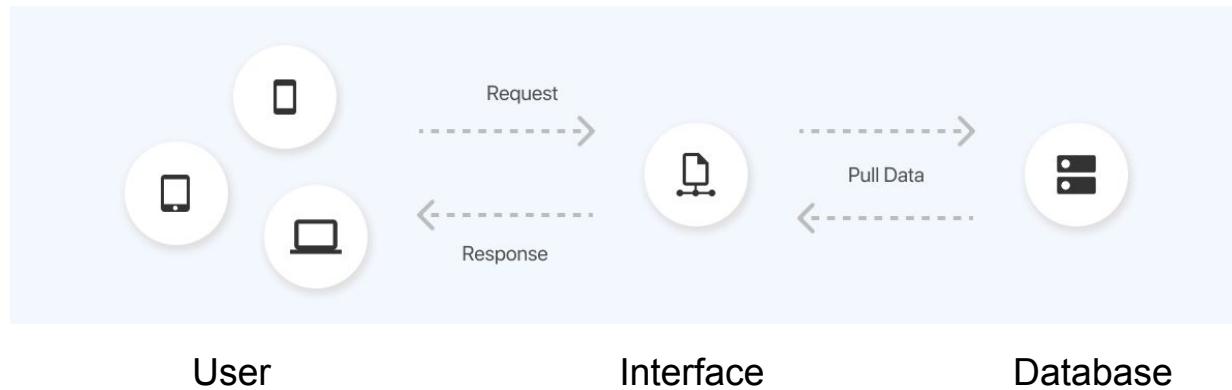
Different systems emphasise different components

- Storage
- Representation/Encoding
- Indexing
- Querying
- Ranking
- Presentation



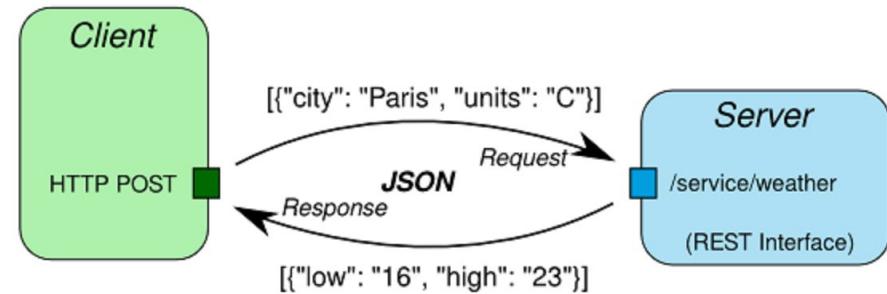
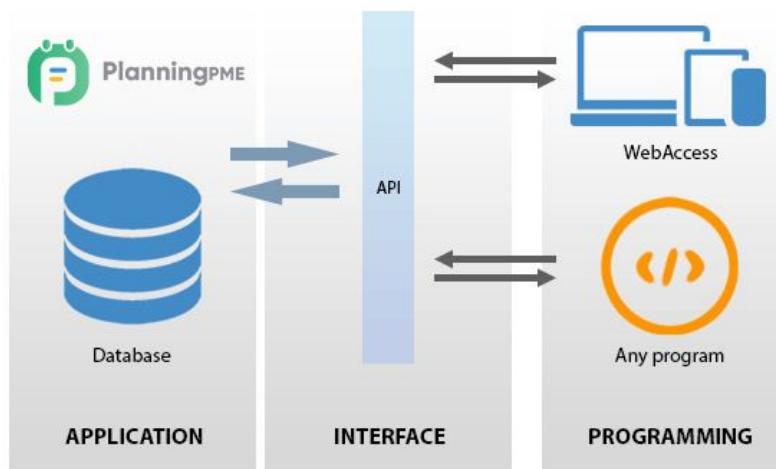
Retrieving Data

Many options for getting data, but generic sharing of data places constraints on how (the interface) a data-provider should/may offer data.



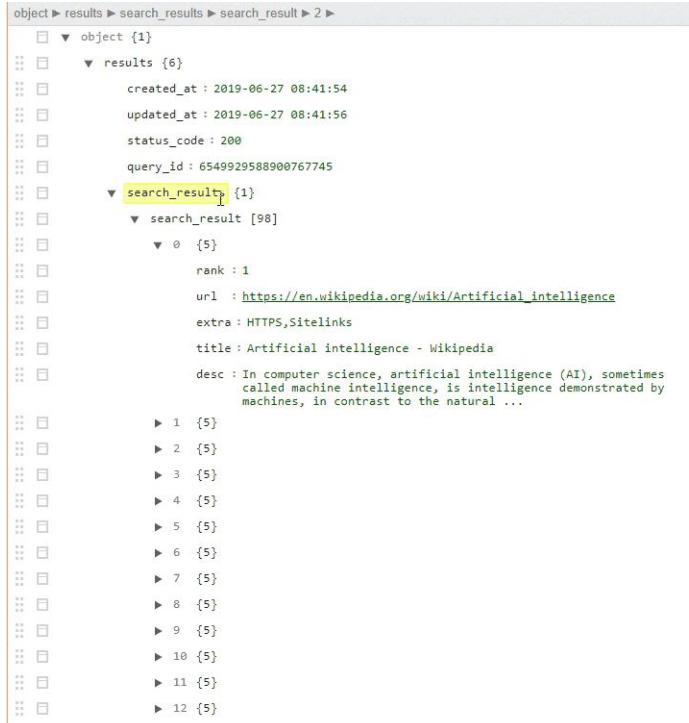
Machine-Readable Data Retrieval

- Source offers Machine-Readable format
 - Via API



Application Programming Interface

```
{  
    "disclaimer": "This lookup service is provided as an information service only and does not constitute legal advice. While it is set up to provide the lat  
    "query_details": {  
        "query_date_time": "2019-06-28T12:45:11",  
        "longitude": "25.04657",  
        "database_last_updated": "2019-06-28T12:38:09",  
        "latitude": "25.8564",  
        "un_url": "https://treaties.un.org/pages/ViewDetails.aspx?src=IND&mtsg_no=XXVII-8-b&chapter=27&clang=_en",  
        "radius": {  
            "unit": "m",  
            "value": "4800"  
        }  
    },  
    "results": [  
        {  
            "nagoya_iso2": "EG",  
            "party_date": "2014-10-12",  
            "marine_regions_country": "Egypt",  
            "wikidata_country": "Egypt",  
            "distance": 0,  
            "absch_profile": "https://absch.cbd.int/countries/EG",  
            "cbd_profile": "https://www.cbd.int/countries/?country=EG"  
        }  
    ]  
}
```



The screenshot shows a JSON viewer interface with a tree-based navigation system. The root node is 'object' which contains a single child 'results'. The 'results' node has six children, each representing a search result. One of these results is expanded to show its details. The expanded result is titled 'search_result_1' and contains a single child 'search_result' with 98 items. The first few items are numbered 0 through 12. Item 0 has a 'rank' of 1, a URL pointing to https://en.wikipedia.org/wiki/Artificial_intelligence, and an 'extra' field containing 'HTTPS,Sitelinks'. The title is 'Artificial intelligence - Wikipedia' and the description starts with 'In computer science, artificial intelligence (AI), sometimes called machine intelligence, is intelligence demonstrated by machines, in contrast to the natural ...'. Other items follow in a similar pattern.

```
object ► results ► search_results ► search_result ► 2 ►  
  object {1}  
    results {6}  
      created_at : 2019-06-27 08:41:54  
      updated_at : 2019-06-27 08:41:56  
      status_code : 200  
      query_id : 654929588900767745  
      search_result_1 {1}  
        search_result [98]  
          0 {5}  
            rank : 1  
            url : https://en.wikipedia.org/wiki/Artificial\_intelligence  
            extra : HTTPS,Sitelinks  
            title : Artificial intelligence - Wikipedia  
            desc : In computer science, artificial intelligence (AI), sometimes called machine intelligence, is intelligence demonstrated by machines, in contrast to the natural ...  
          1 {5}  
          2 {5}  
          3 {5}  
          4 {5}  
          5 {5}  
          6 {5}  
          7 {5}  
          8 {5}  
          9 {5}  
          10 {5}  
          11 {5}  
          12 {5}
```

Data Retrieval

- Long history of organising data/information/knowledge
 - Across many fields
- Technological innovation has had great impact
- Traditionally designed around humans as data consumers
 - See a shift to machines as consumers
- Data standards help ensure FAIR principles
 - Add complexity

