CHENNAI MATHEMATICAL INSTITUTE

Presented at
Qualcomm
on 24th June 2021

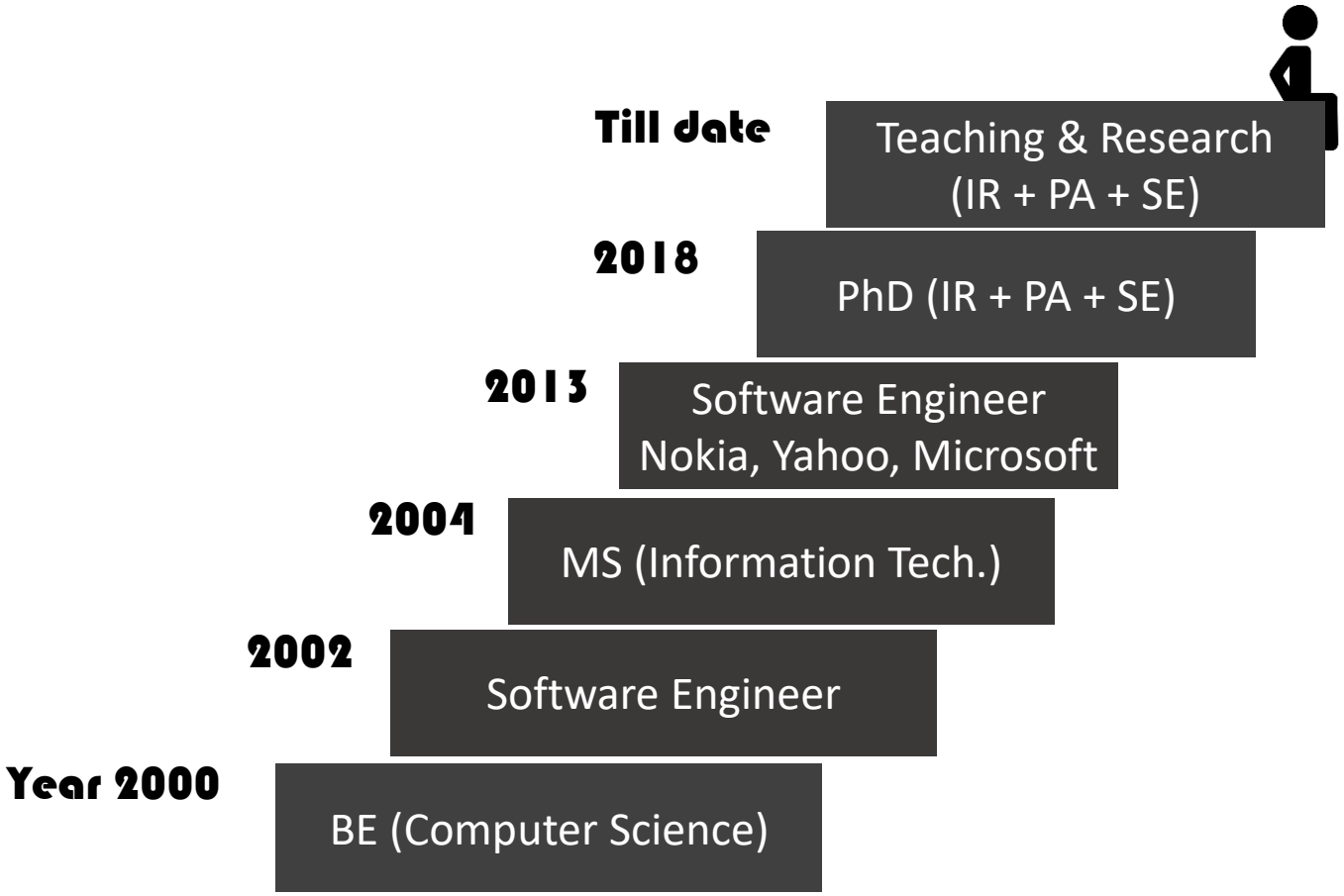
Knowledge Graphs - Going Beyond Data!

Venkatesh Vinayakarao
venkateshv@cmi.ac.in
http://vvtesh.co.in

Chennai Mathematical Institute

To know that we know what we know, and to know that we do not know what we do not know, that is true knowledge—Nicolaus Copernicus

About Me


Till date
Teaching & Research
(IR + PA + SE)
2018
PhD (IR + PA + SE)
2013
Software Engineer
Nokia, Yahoo, Microsoft
2004
MS (Information Tech.)
2002
Software Engineer
Year 2000
BE (Computer Science)

Agenda

Knowledge Graphs – Going Beyond Data!

Will Discuss	Will not Discuss
✓ Concepts	⦸ Details
✓ Illustrations	⦸ Definitions
✓ Intuitions	⦸ Formalism
✓ Purpose	⦸ Derivations
✓ Properties	⦸ Proofs

Three Stories

(1) Knowledge Graphs (2) Tools and Techniques

(3) Algorithms (if time permits)

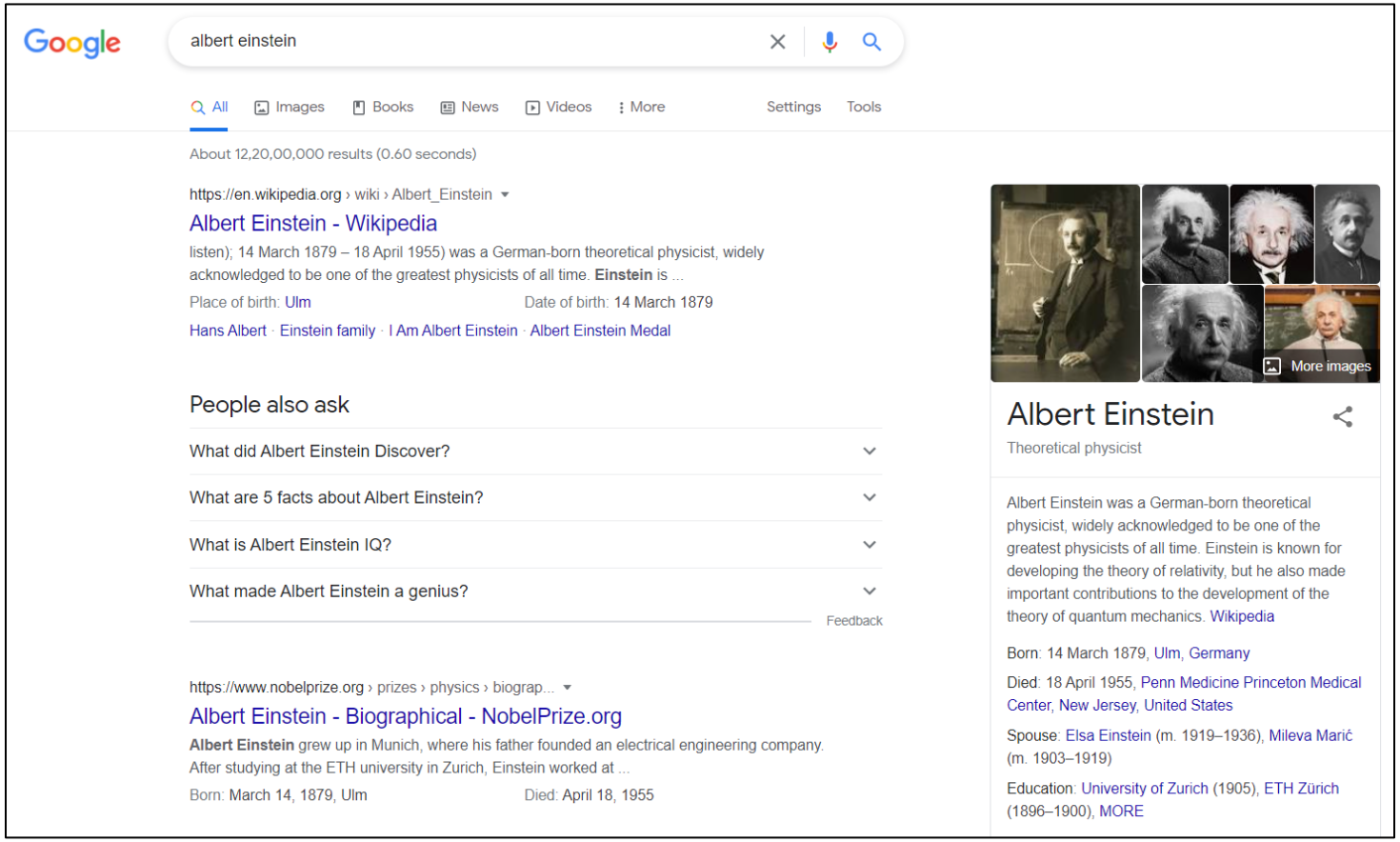
Story 1

Knowledge Graphs

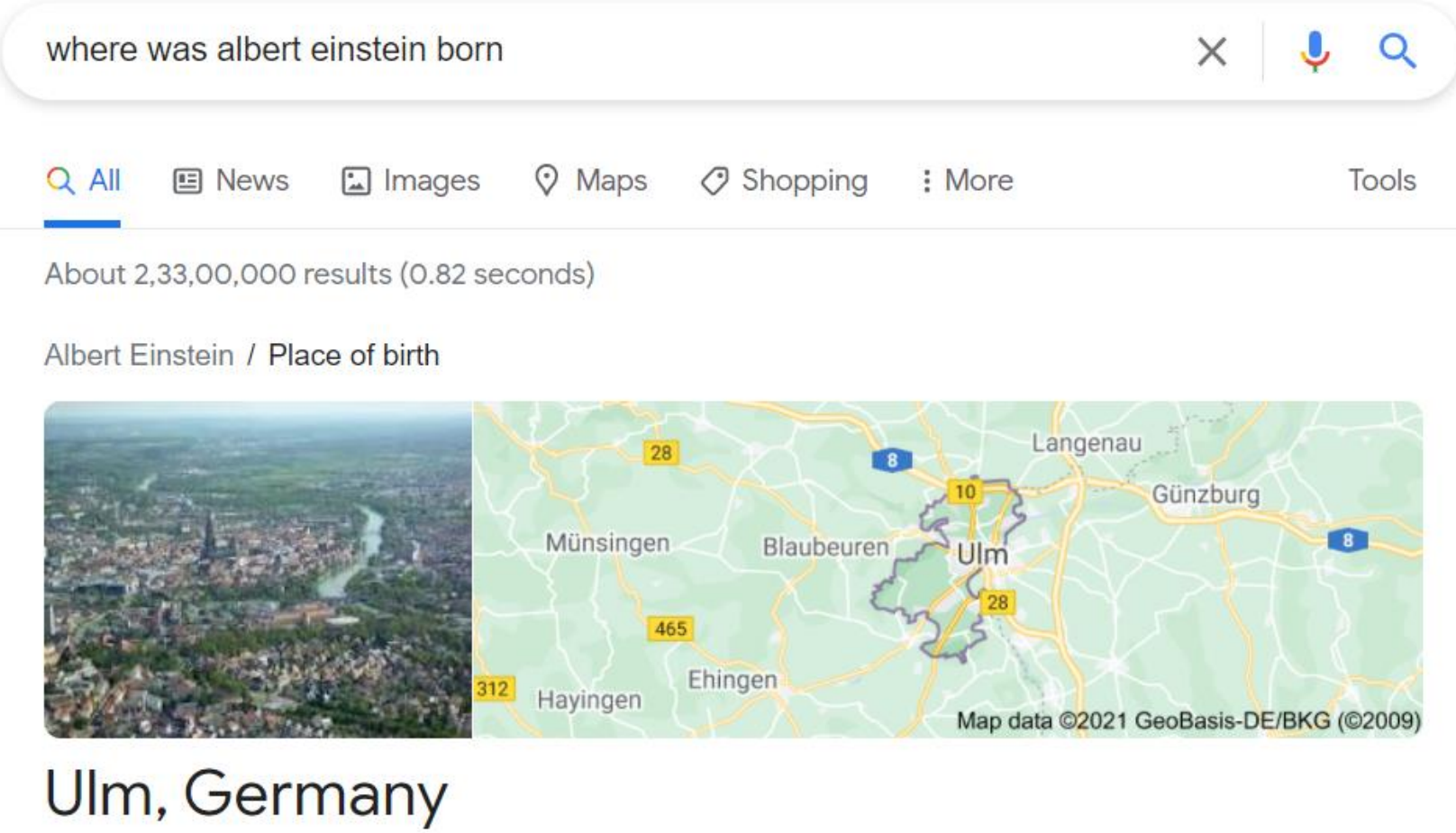
Image Source: https://neo4j.com/blog/neo4j-bloom-1-1-release/

Beyond Searching for Documents

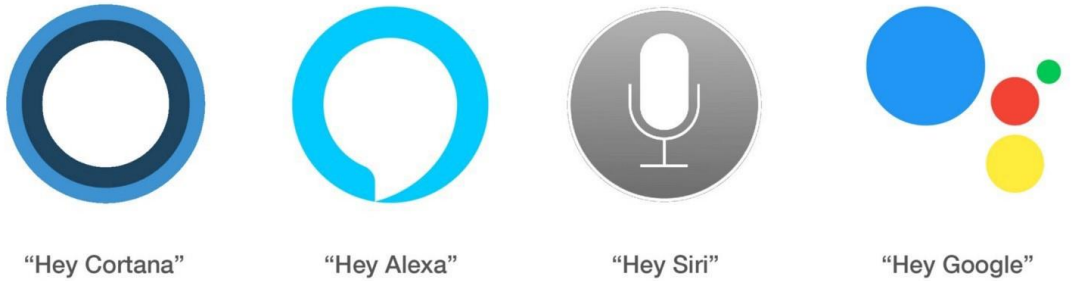
We search for entities. Not always documents!



We Need Answers!



In general, humans make decisions based on the world knowledge they have gathered over time. Why not machines?



Exploring data with a graphical representation

Demo 1

How do we capture knowledge?

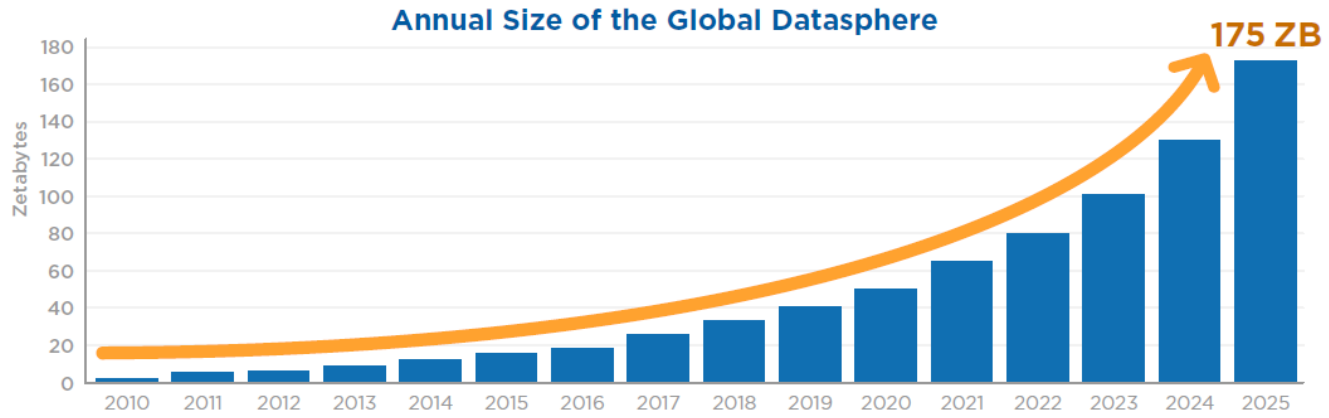
How do we represent this knowledge?

How do we access this knowledge?

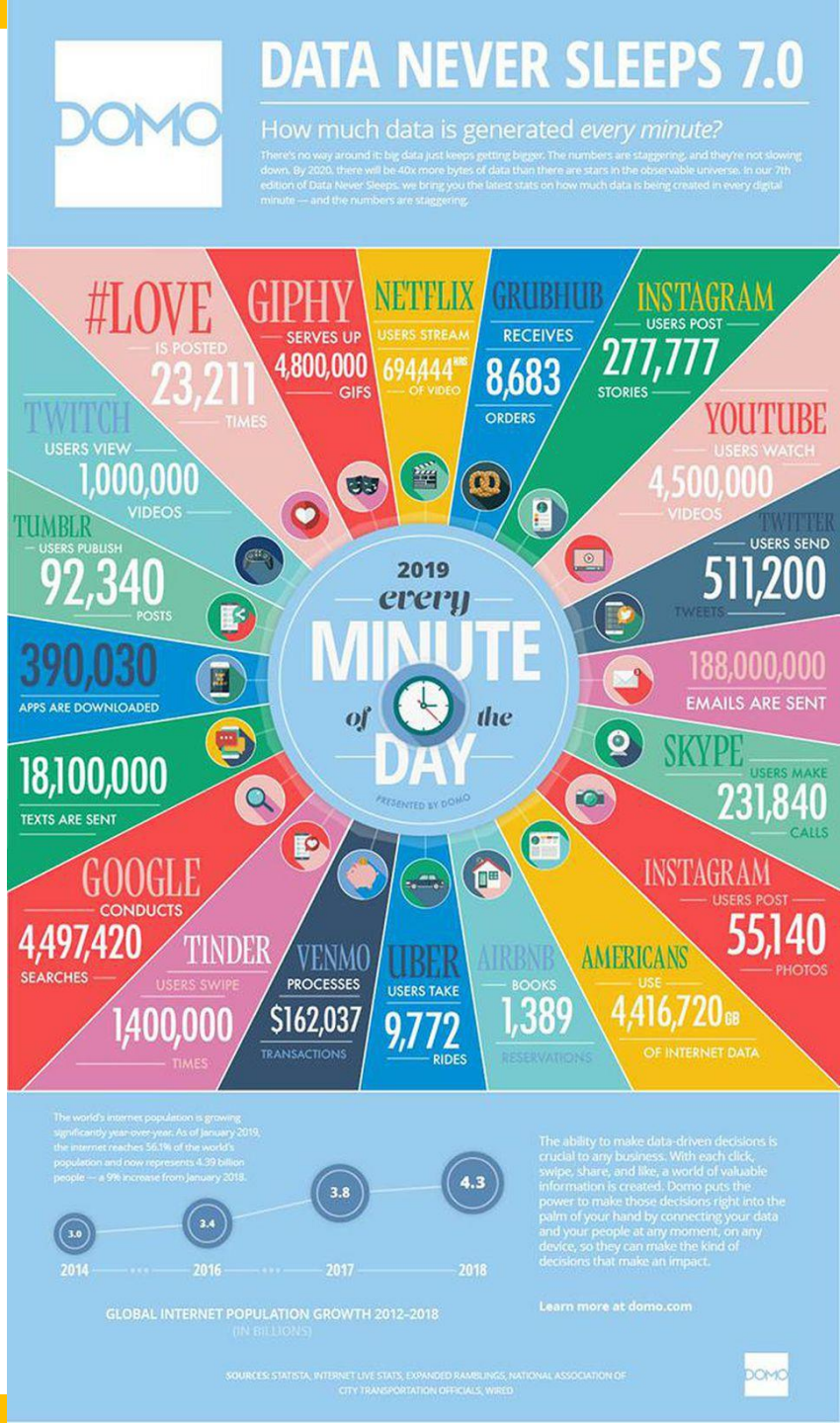
Data is Ubiquitous

But, how do we organize this data?

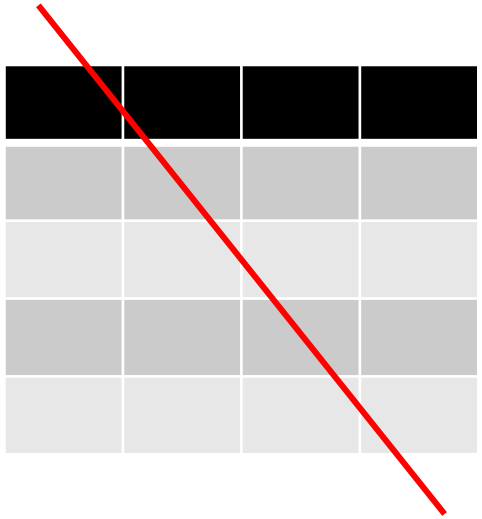
Figure 1 – Annual Size of the Global Datasphere

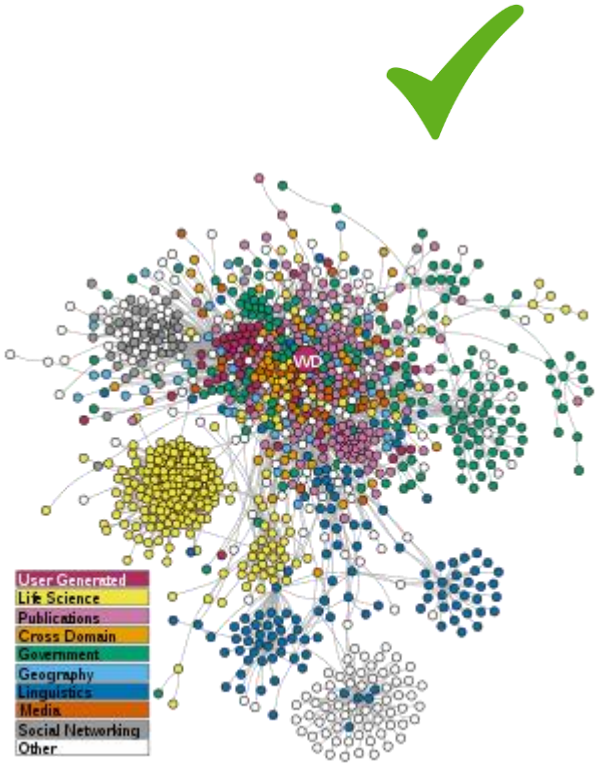


Source: IDC DataAge 2025 whitepaper, and DOMO.



A Graph Data Model





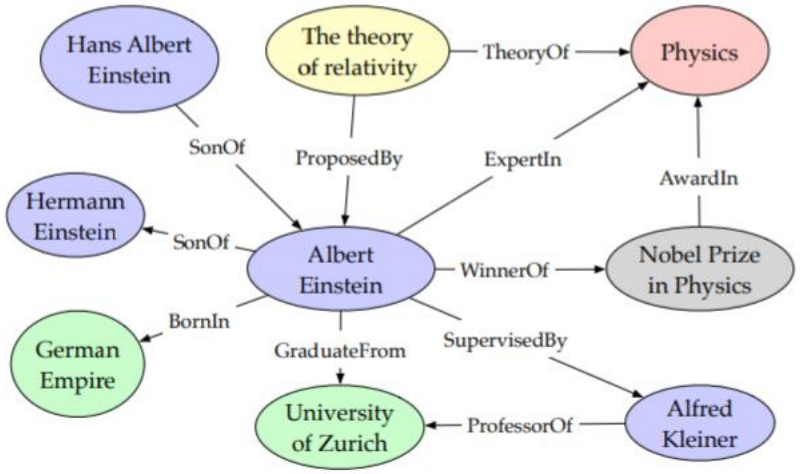
Real-world entities and relationships are better captured through graphs.

Knowledge Graphs

Subject – Predicate – Object Triples

(Albert Einstein, **BornIn**, German Empire)
(Albert Einstein, **SonOf**, Hermann Einstein)
(Albert Einstein, **GraduateFrom**, University of Zurich)
(Albert Einstein, **WinnerOf**, Nobel Prize in Physics)
(Albert Einstein, **ExpertIn**, Physics)
(Nobel Prize in Physics, **AwardIn**, Physics)
(The theory of relativity, **TheoryOf**, Physics)
(Albert Einstein, **SupervisedBy**, Alfred Kleiner)
(Alfred Kleiner, **ProfessorOf**, University of Zurich)
(The theory of relativity, **ProposedBy**, Albert Einstein)
(Hans Albert Einstein, **SonOf**, Albert Einstein)

Knowledge Graph



These facts are stored in a "knowledge base".

Read more at https://arxiv.org/pdf/2002.00388.pdf

Wikidata: A Free Open Knowledgebase

- Data
  - 94M items! Anyone can edit.
- Community Control
  - Contributors edit *the population number of Rome* but also decide whether there is such a number in the first place.
- Conflicting Data
  - Many facts are disputed. Several details are uncertain.
  - Allows conflicting data to co-exist.
  - There is no *true population of Rome*. There is a *population of Rome as published by the city of Rome in 2011*.
- Multilingual
- Easy Access
- Continuous Evolution

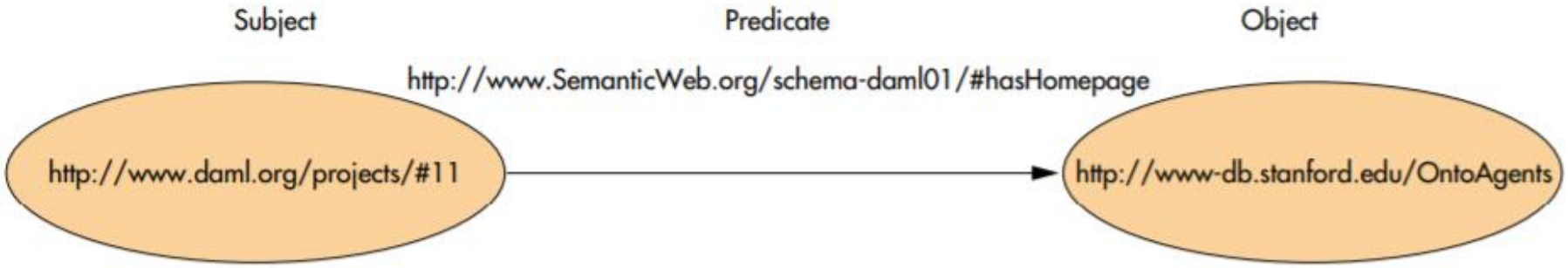
We have many knowledgebases.
Interoperability is a concern ☹

Solution: Resource Description Framework (RDF)

RDF is a directed, labeled graph data format for representing information in the Web.

Resource Description Framework (RDF)

- Made of Subject-Predicate-Object Triples
- Uniform Resource Identifier (URI) disambiguates entities.



Source: http://www.contrib.andrew.cmu.edu/~mm6/95-733/PowerPoint/rdftutorial.pdf

RDF

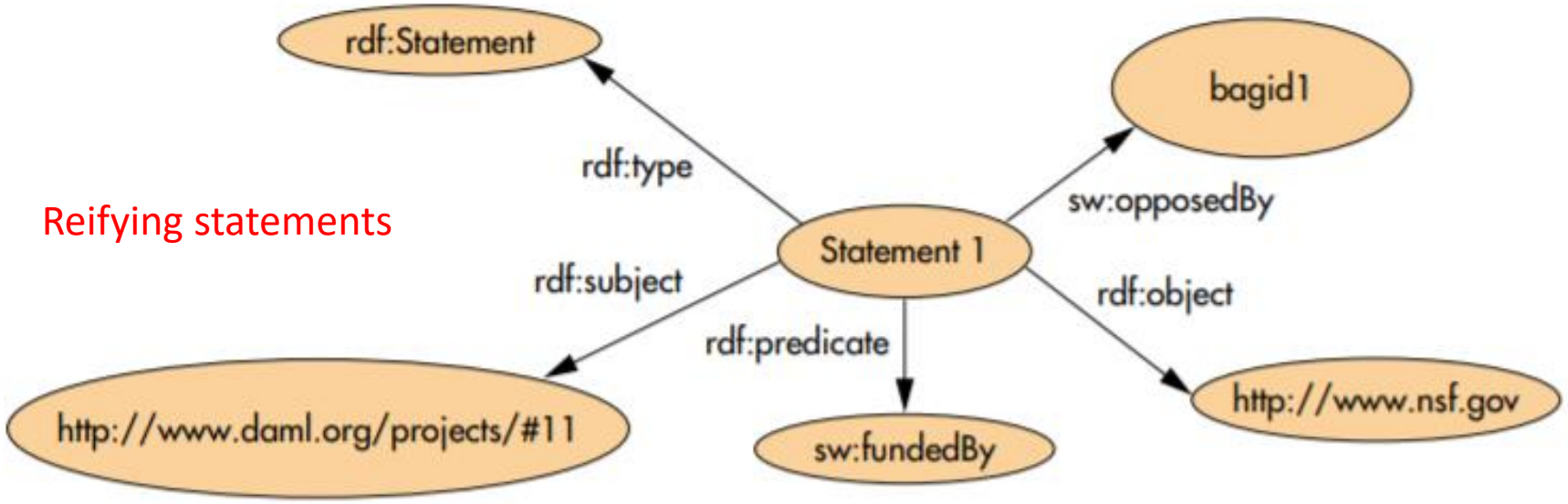
- Partial RDF Snippet

```
<Project rdf :about="http://www.daml.org/projects/#11">
    <hasHomepage>
        <rdfs:Resource rdf:ID="http://www-db.stanford.edu/OntoAgents">
            <dc:Creator>Stefan Decker</dc:Creator>
        <rdfs:Resource>
    </hasHomepage>
</Project>
```

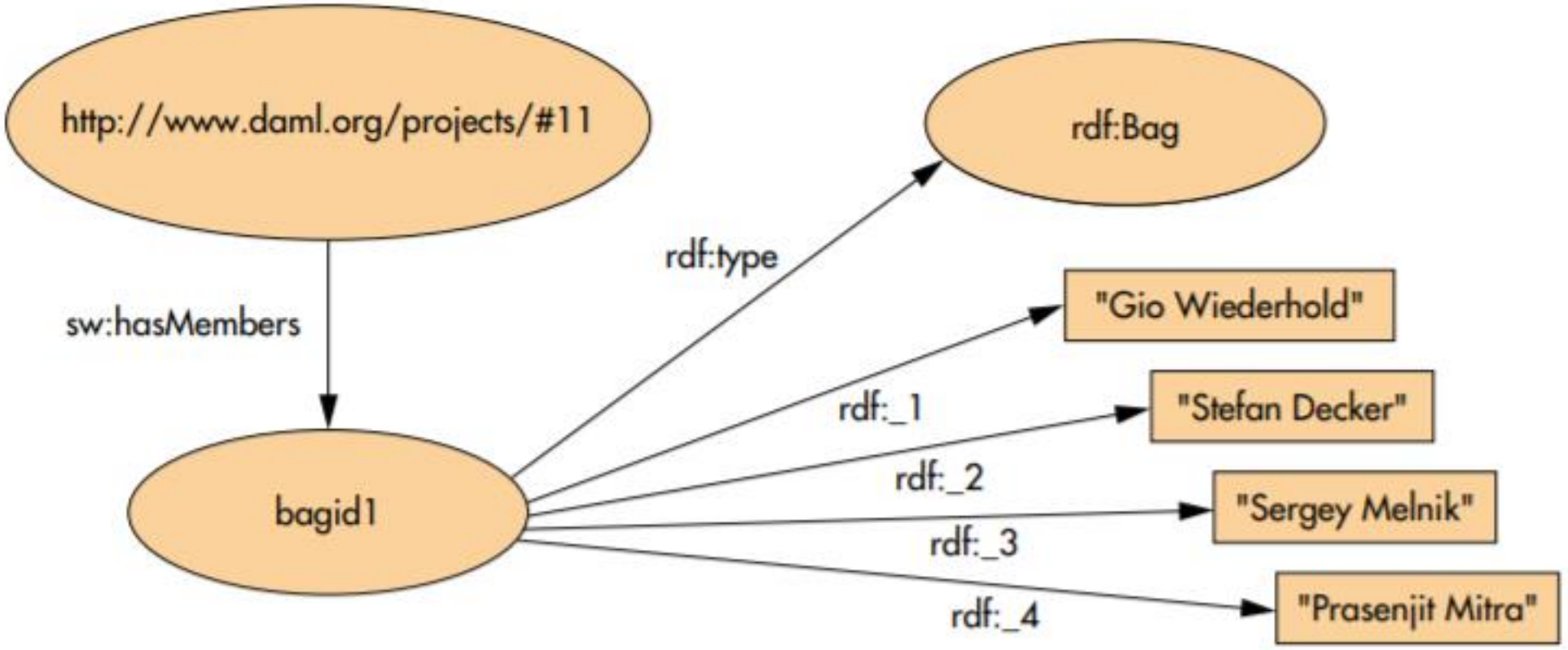
- These RDF Triples can be stored in a Triplestore
  - Such as RDFox and Jena TDB.
- These RDFs can get complex!

Reifying Statements in RDF

Reifying statements



RDF


http://www.daml.org/projects/#11
sw:hasMembers
bagid1
rdf:type
rdf:Bag
rdf:_1
"Gio Wiederhold"
rdf:_2
"Stefan Decker"
rdf:_3
"Sergey Melnik"
rdf:_4
"Prasenjit Mitra"

SPARQL

- Assume a data graph:

  <http://.../book1> <http://... /title> "SPARQL Book"

- Then, the SPARQL query:

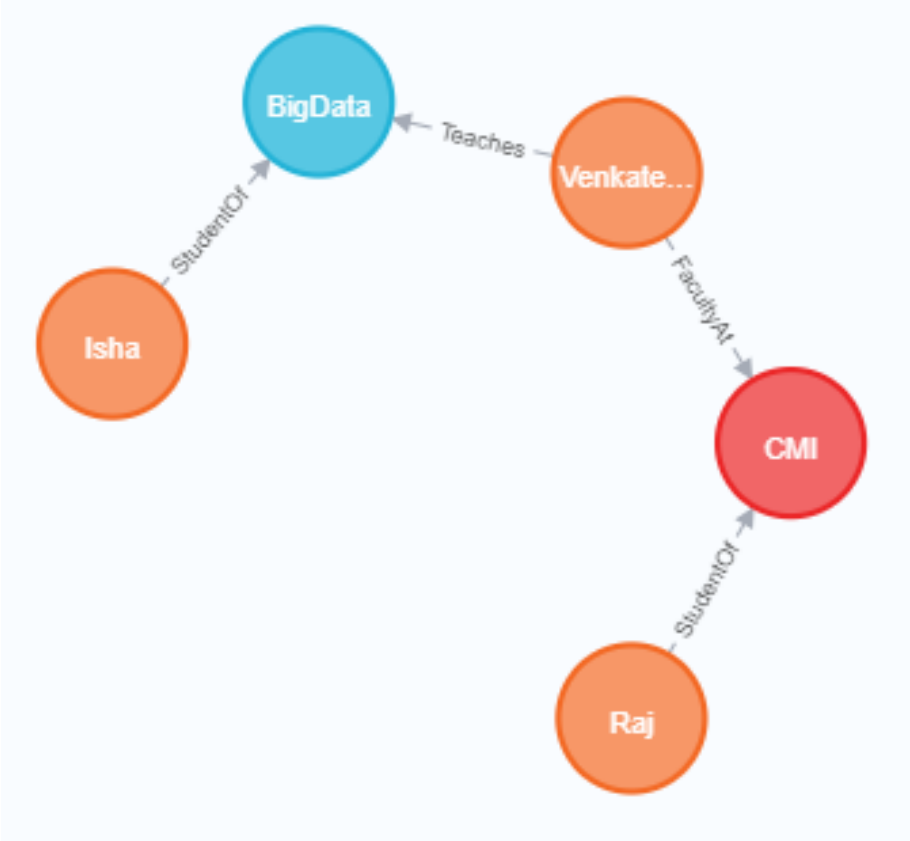
```
SELECT ?title WHERE {
  <http://.../book1> <http://.../title> ?title .
}
```

- Results in "SPARQL Book".
- Several tools support SPARQL
  - such as Apache Jena.

Try few SPARQL queries here - https://dbpedia.org/sparql/.

A Graph DB: Neo4j

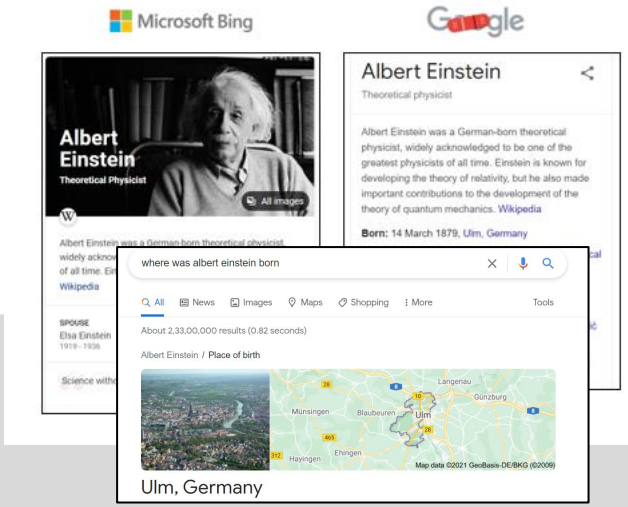
Demo 2



Any Graph DB can be used to store the triples.

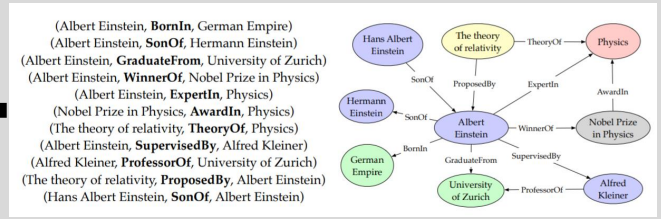
Summary

Applications



Triplestore
(E.g.: Jena TDB, Neo4j with SPARQL/Cypher interface**)**

RDF



Triples

World Knowledge

Story 2

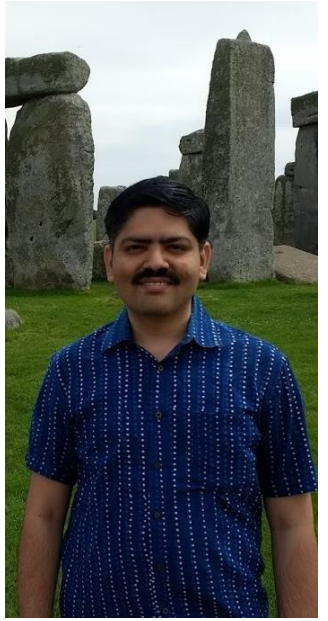
Tools and Technologies

Image Source: https://neo4j.com/blog/neo4j-bloom-1-1-release/

Match the Text with the Image

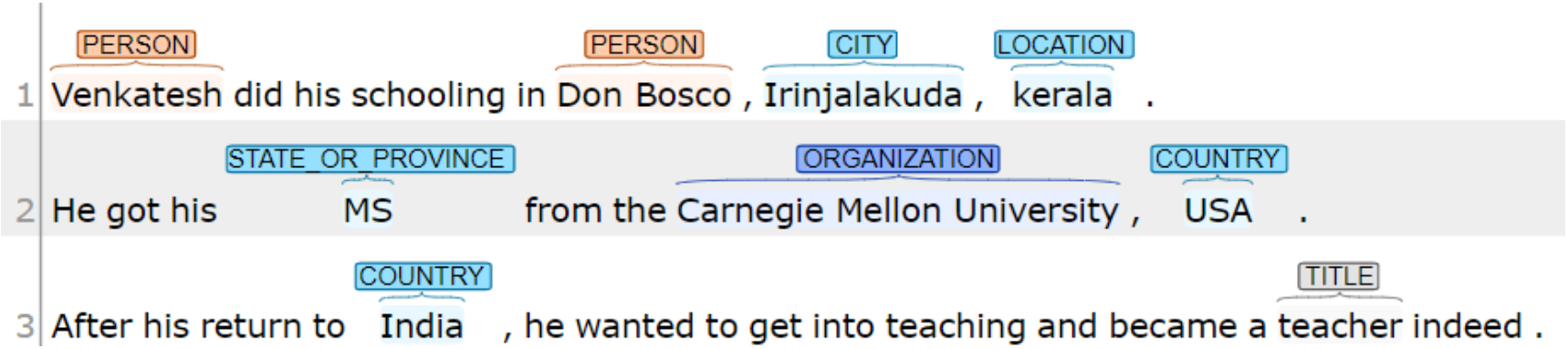
Venkatesh did his schooling in Don Bosco, Egmore, Chennai. He got his MBA from the Middlebury Institute of International Studies at Monterey, USA. After his return to India, he wanted to get into film production but instead, became an actor in Telugu films.

Venkatesh did his schooling in Don Bosco, Irinjalakuda, kerala. He got his MS from the Carnegie Mellon University, USA. After his return to India, he wanted to get into teaching and became a teacher indeed.



We Have Two Problems

- Given the input text
  - Recognize the entities – NER.
  - Disambiguate them – Entity Linking.



Try it at https://corenlp.run/

CoreNLP

- Extracted the following triples:

Subject	Predicate	Object
Venkatesh	per:statesorprovinces_of_residence	MS
Venkatesh	per:schools_attended	Carnegie Mellon University

Codeq NER and Linking

Venkatesh did his schooling in Don Bosco , Irinjalakuda , kerala . He got his MS from the Carnegie Mellon University , USA . After his return to India , he wanted to get into teaching and became a teacher indeed .

Venkatesh Daggubati score: 0.0313
Indian actor
https://en.wikipedia.org/wiki/Venkatesh_Daggubati

John Bosco score: 0.7517
Italian Roman Catholic priest, educator and writer
https://en.wikipedia.org/wiki/John_Bosco

Irinjalakuda score: 0.9899
human settlement
https://en.wikipedia.org/wiki/Irinjalakuda

Microsoft score: 0.7425
American multinational technology corporation
https://en.wikipedia.org/wiki/Microsoft

https://api.codeq.com/demo-nel

More Tools

- CYC – A Machine Reasoning Platform
  - OpenCYC (KB + Reasoning Engine)
    - 239K terms, 2M triples
  - ResearchCYC
    - 500K concepts, 26K relations
- Stanford Open Information Extraction (OpenIE)
  - Refers to the extraction of relation tuples, typically binary relations, from plain text
  - *Barack Obama was born in Hawaii* would create a triple *(Barack Obama; was born in; Hawaii)*

https://cyc.com/
https://www.ime.usp.br/~fr/opencyc/ OpenCyC does not exist anymore.
https://en.wikipedia.org/wiki/Cyc#ResearchCyc
https://nlp.stanford.edu/software/openie.html

Tools Need Resources and We Have Many

- Dbpedia (https://www.dbpedia.org/)
  - 228 Million entities
- Wiktionary (https://www.wiktionary.org/)
  - 6.6M dictionary entries
- Yago (https://yago-knowledge.org/)
  - A large knowledge base about people, cities, countries, movies, and organizations.
- Freebase
  - Used by Google's Knowledge Graph.
  - Does not exist now.

ConceptNet

- A freely-available semantic network, designed to help computers understand the meanings of words.
- Has ~15 million facts in English
- Uses Crowdsourced knowledge
  - Open Mind Common Sense, Wiktionary, DBPedia, Yahoo Japan / Kyoto University project

[Havasi et al., RANLP '07; Speer and Havasi, LREC '12]

Summary

Demo 3

- Several tools and technologies exist to help you build the knowledge graph.
  - CoreNLP
  - Codeq
  - ConceptNet
  - CyC
  - OpenIE
- Caveat: They may not be perfect.

Story 3

Two Algorithms

(related to Entity Retrieval and Knowledge Graphs)

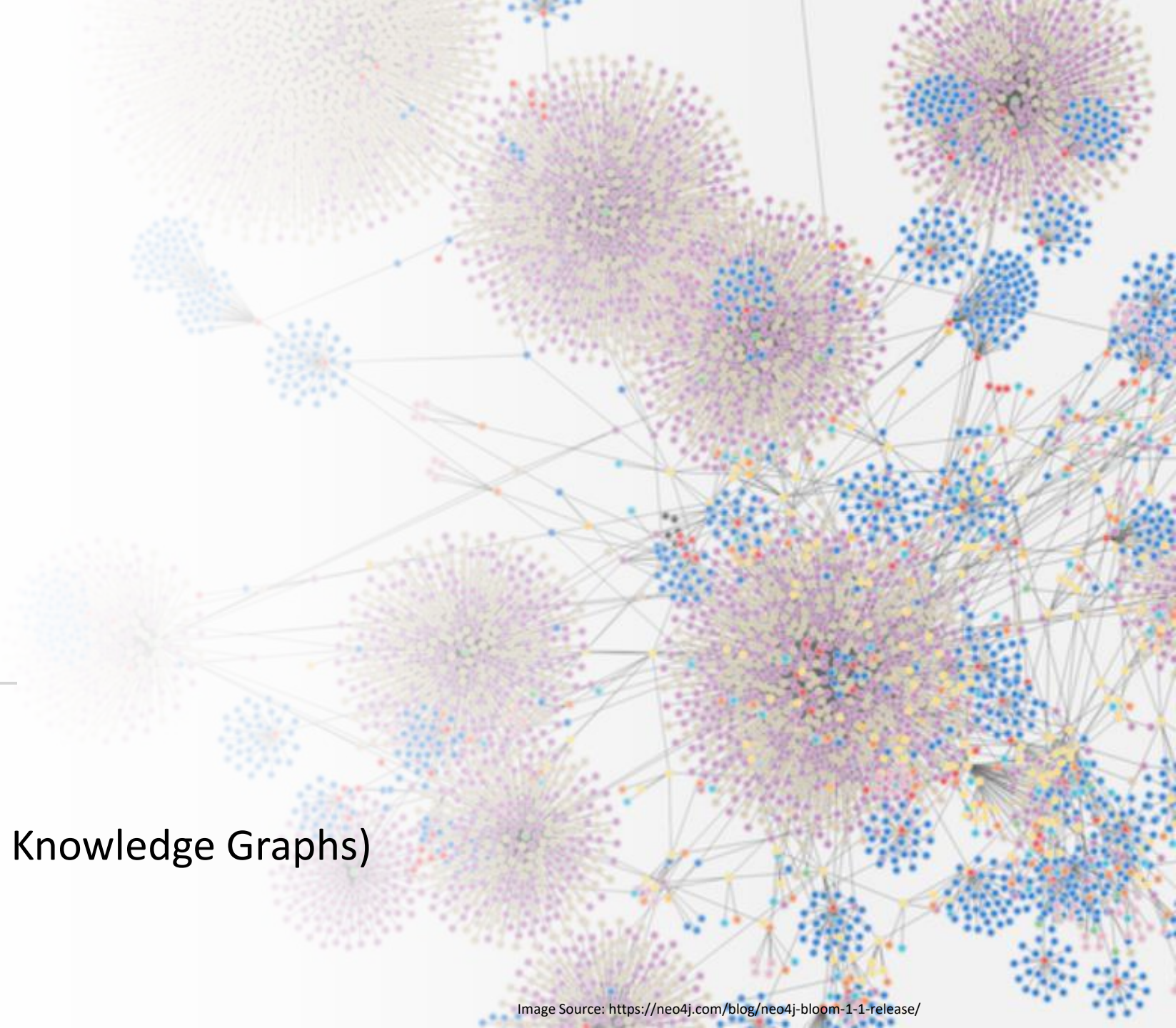
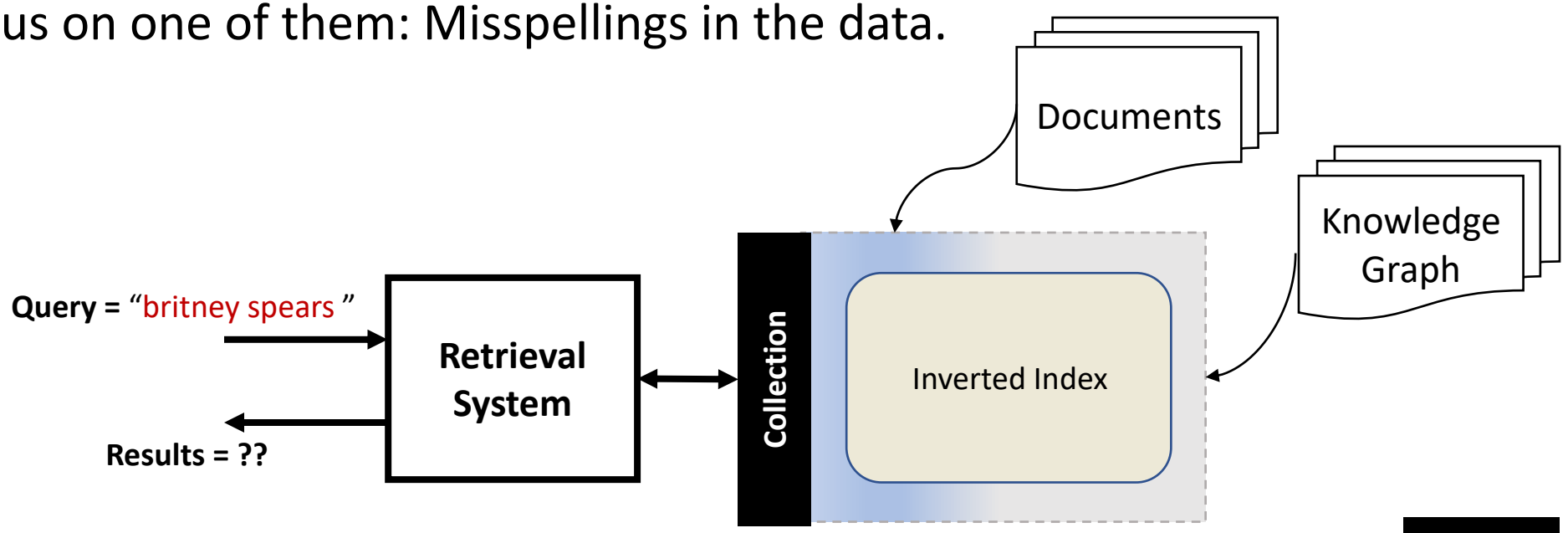


Image Source: https://neo4j.com/blog/neo4j-bloom-1-1-release/

Entity Retrieval

- Retrieving entities – not documents.
- Knowledge graphs are very useful for this purpose.
- However, both KG construction and ER have common problems
  - We focus on one of them: Misspellings in the data.



Notorious Britney

The data below shows some of the misspellings detected by our spelling correction system for the query [britney spears], and the count of how many different users spelled her name that way. -- Google.

488941 britney spears	29 britent spears	9 brinttany spears	5 brney spears	3 britiy spears	2 brirreny spears
40134 brittany spears	29 brittnany spears	9 britanay spears	5 broitney spears	3 britmeny spears	2 brirtany spears
36315 brittney spears	29 britttany spears	9 britinany spears	5 brotny spears	3 britneeey spears	2 brirttany spears
24342 britany spears	29 btiney spears	9 britn spears	5 bruteny spears	3 britnehy spears	2 brirttney spears
7331 britny spears	26 birttney spears	9 britnew spears	5 btiyney spears	3 britnely spears	2 britain spears
6633 briteny spears	26 breitney spears	9 britneyn spears	5 btrittney spears	3 britnesy spears	2 britane spears
2696 britteny spears	26 brinity spears	9 britrney spears	5 gritney spears	3 britnetty spears	2 britaneny spears
1807 briney spears	26 britenay spears	9 brtiny spears	5 spritney spears	3 britnex spears	2 britania spears
1635 brittny spears	26 britneyt spears	9 brtittney spears	4 bittny spears	3 britneyxxx spears	2 britann spears
1479 brintey spears	26 brittan spears	9 brtny spears	4 bnritney spears	3 britnity spears	2 britanna spears
1479 britanny spears	26 brittne spears	9 brytny spears	4 brandy spears	3 britntey spears	2 britannie spears
1338 britiny spears	26 btittany spears	9 rbitney spears	4 brbritney spears	3 britnyey spears	2 britannt spears
1211 britnet spears	24 beitney spears	8 birtiny spears	4 breatiny spears	3 britterny spears	2 britannu spears
1096 britiney spears	24 birteny spears	8 bithney spears	4 breetney spears	3 brittneey spears	2 britanyl spears
991 britaney spears	24 brightney spears	8 brattany spears	4 bretiney spears	3 brittnney spears	2 britanyt spears
991 britnay spears	24 brintiny spears	8 breitny spears	4 brfitney spears	3 brittnyey spears	2 briteeny spears
811 brithney spears	24 britanty spears	8 breteny spears	4 briattany spears	3 brityen spears	2 britenany spears
811 brtiney spears	24 britenny spears	8 brightny spears	4 brieteny spears	3 briytney spears	2 britenet spears
664 birtney spears	24 britini spears	8 brintay spears	4 briety spears	3 brltney spears	2 briteniy spears
664 brintney spears	24 britnwy spears	8 brinttey spears	4 briitny spears	3 broteny spears	2 britenys spears
664 briteney spears	24 brittni spears	8 briotney spears	4 briittany spears	3 brtaney spears	2 britianey spears
601 bitney spears	24 brittnie spears	8 britanys spears	4 brinie spears	3 brtiiany spears	2 britin spears
601 brinty spears	21 biritney spears	8 britley spears	4 brinteney spears	3 brtinay spears	2 britinary spears
544 brittaney spears	21 birtany spears	8 britneyb spears	4 brintne spears	3 brtinney spears	2 britmy spears
544 brittnay spears	21 biteny spears	8 britnrey spears	4 britaby spears	3 brtitany spears	2 britnaney spears

Source: https://archive.google.com/jobs/britney.html

Correcting Misspellings

- There are many approaches. We focus on two major approaches:
  - finding "**nearest**" term using ***Edit Distance.***
  - Finding "**similar sounding terms**" using ***Phonetic Hash***.

Algorithm 1: Finding Nearest Term with Edit Distance

- Given two strings S_1 and S_2, edit distance is the minimum number of operations to convert one to the other.
- Operations are typically character-level.
  - Insert, Delete, Replace
  - E.g., the edit distance from ***dof*** to ***dog*** is 1
  - From ***cat*** to ***act*** is 2
  - From ***cat*** to ***dog*** is 3.

*Here, we do not consider transposition.

Live quiz

Join at

slido.com

#quriosity

What is the edit distance between Sunday and Saturday?

- 4
- 2
- 3
- 1

Quiz

What is the edit distance between Sunday and Saturday?

*You are allowed to perform only Insert, Delete, and Replace operations.

Answer

- S~~atur~~day = S~~un~~day = S*day
- Problem is same as
  - What is the edit distance between atur and un?
  - Answer
    - Delete a,t. Replace r with n.
    - 3.

Levenshtein Example

		s	a	t	u	r	d	a	y
	0	1	2	3	4	5	6	7	8
s	1	0	1	2	3	4	5	6	7
u	2	1	1	2	2	3	4	5	6
n	3	2	2	2	3	3	4	5	6
d	4	3	3	3	3	4	3	4	5
a	5	4	3	4	4	4	4	3	4
y	6	5	4	4	5	5	5	4	3

Sunday

↓

Keep s. Insert a, t.

Keep u.

Replace r.

Keep day.

↓

Saturday

$$
\left.
\begin{aligned}
D(i,j) &= \min[D(i-1,j)+w_d, \\
&\quad\ D(i,j-1)+w_i, \\
&\quad\ D(i-1,j-1)+w_r] \\
D(i,0) &= D(i-1,0)+w_d \\
D(0,j) &= D(0,j-1)+w_i
\end{aligned}
\right\} \forall i,j>0
$$

$$
D(0,0) = 0
$$

Note: $w_r = 0$ *if* $a_i = a_j$ i.e., if the characters being compared are the same.

V. I. Levenshtein, Binary codes capable of correcting deletions insertions and reversals. Soviet Physics. 10, 707-710, 1966.

Levenshtein Algorithm

```
EDITDISTANCE(s1, s2)
 1  int m[i, j] = 0
 2  for i ← 1 to |s1|
 3  do m[i, 0] = i
 4  for j ← 1 to |s2|
 5  do m[0, j] = j
 6  for i ← 1 to |s1|
 7  do for j ← 1 to |s2|
 8       do m[i, j] = min{m[i − 1, j − 1] + if (s1[i] = s2[j]) then 0 else 1fi,
 9                        m[i − 1, j] + 1,
10                        m[i, j − 1] + 1}
11  return m[|s1|, |s2|]
```

Algorithm 2: Soundex

- Homophones (sound alike but have different spellings and different meaning)
  - pair, pear
  - break, brake
  - cell, sell
  - cent, scent
  - knight, night

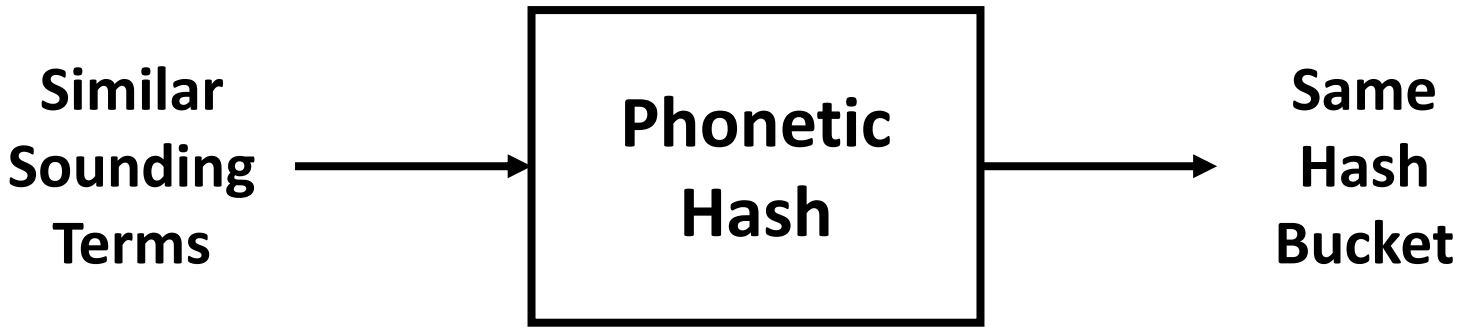

Donald Knuth

Algorithm 2: Soundex

- Homophones (sound alike but have different spellings and different meaning)
  - pair, pear
  - break, brake
  - cell, sell
  - cent, scent
  - knight, night

Soundex algorithm became more popular after it was discussed in “The Art of Programming”!
Find a bug and take home 100_{16} (or 0x00000100) cents! i.e., 256 cents.

Main Idea

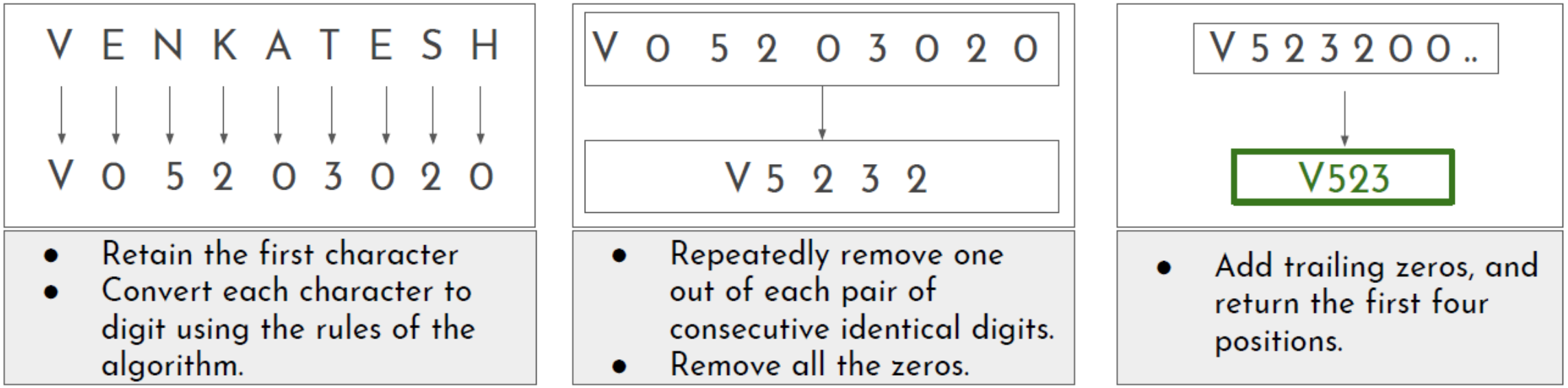

Similar Sounding Terms
Phonetic Hash
Same Hash Bucket
Soundex Algorithms generate phonetic hash

Standard Soundex Algorithm

1. Retain the first character.
2. Convert each character to digit using the rules in the table.
3. Repeatedly remove one out of each pair of consecutive identical digits.
4. Remove all the zeros.
5. Add trailing zeros, and return the first four positions.

Alphabets to be replaced	Digit
A, E, I, O, U, H, W, Y	0
B, F, P, V	1
C, G, J, K, Q, S, X, Z	2
D, T	3
L	4
M, N	5
R	6

An Example



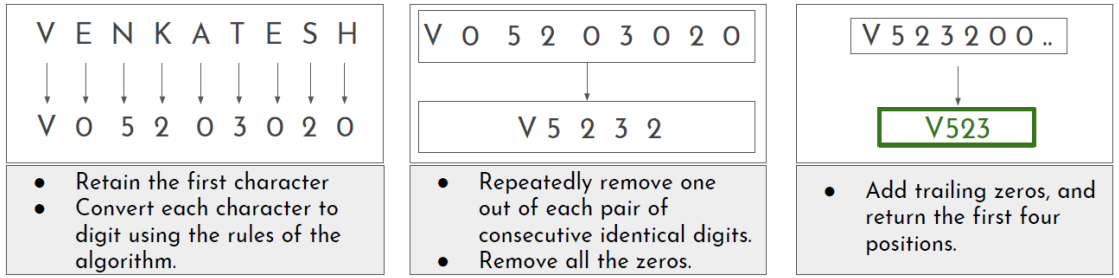
The flow

Summary

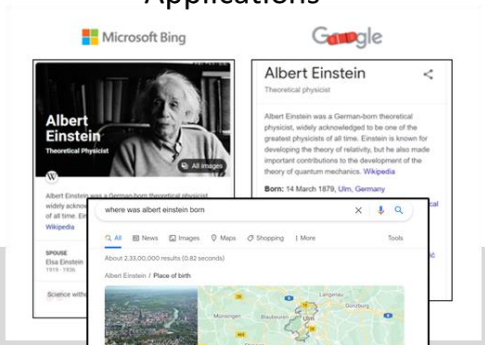
- Finding nearest term – Levenshtein's Algorithm

$$
\left.
\begin{aligned}
D(i,j) &= \min[D(i-1,j)+w_d, \\
&\quad\ D(i,j-1)+w_i, \\
&\quad\ D(i-1,j-1)+w_r] \\
D(i,0) &= D(i-1,0)+w_d \\
D(0,j) &= D(0,j-1)+w_i
\end{aligned}
\right\} \forall i,j>0
$$

- Phonetic Hash – Soundex Algorithm

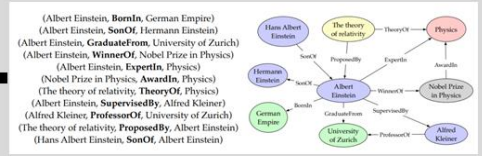


Applications



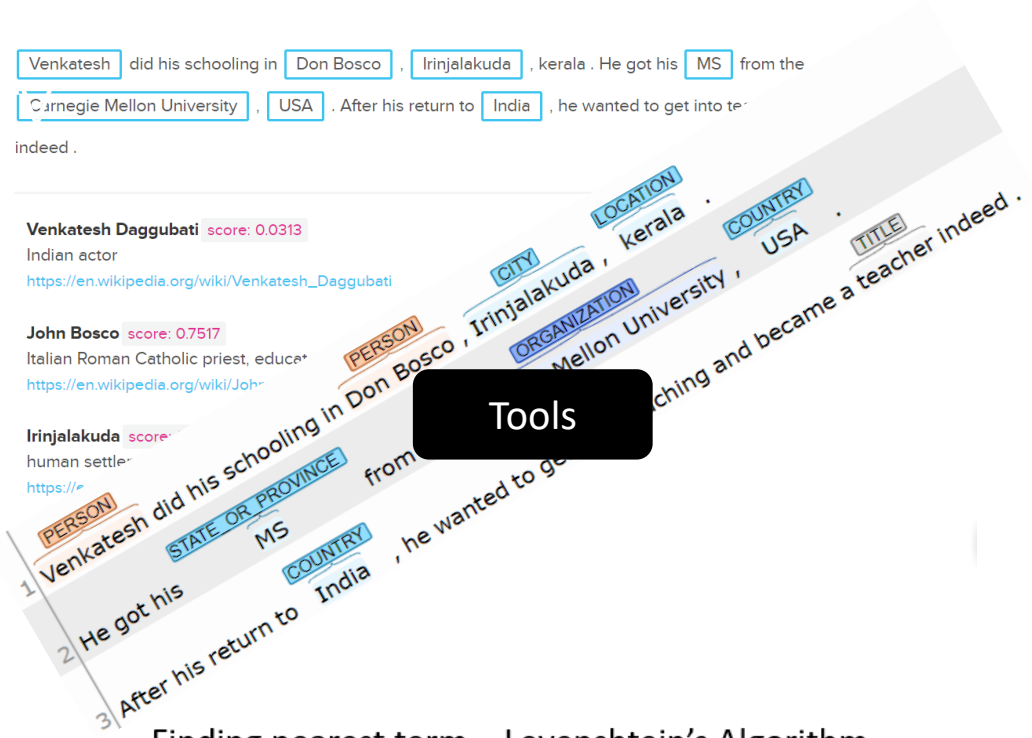
Triplestore
(E.g.: Jena TDB, Neo4j with SPARQL/Cypher interface)

RDF



Triples

World Knowledge

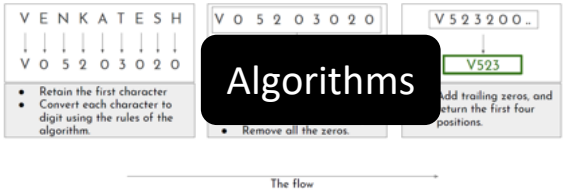


Tools

Finding nearest term – Levenshtein's Algorithm

$$
\begin{aligned}
D(i,j) &= \left.\begin{array}{l}\min[D(i-1,j)+w_d,\\ D(i,j-1)+w_i,\\ D(i-1,j-1)+w_r]\end{array}\right\} \forall i,j>0\\
D(i,0) &= D(i-1,0)+w_d\\
D(0,j) &= D(0,j-1)+w_i
\end{aligned}
$$

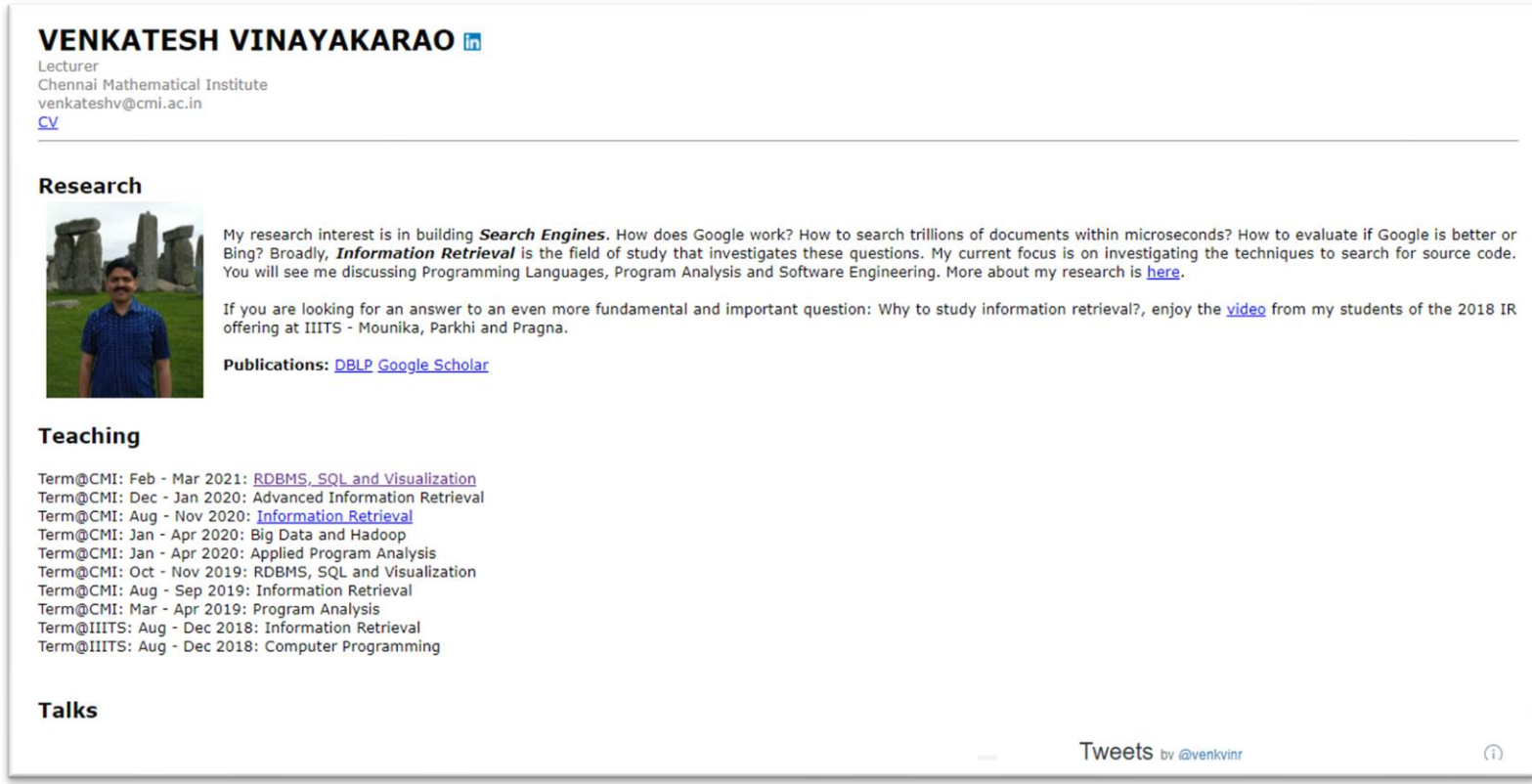
Phonetic Hash – Soundex Algorithm



Algorithms

Future Directions

- Handling uncertain data
  - We do not like “population in rome is …”
  - We like “As per 2012 report, the population in rome is …”
- Knowledge Graph Embeddings
  - Represent entities in a continuous vector space
- Multimodal Knowledge Graphs
- Explainability and Knowledge Graphs
- Relationship Mining
- Interoperability of knowledgebases
- … and many applications in several domains.



VENKATESH VINAYAKARAO

Lecturer
Chennai Mathematical Institute
venkateshv@cmi.ac.in
CV

Research

My research interest is in building ***Search Engines***. How does Google work? How to search trillions of documents within microseconds? How to evaluate if Google is better or Bing? Broadly, ***Information Retrieval*** is the field of study that investigates these questions. My current focus is on investigating the techniques to search for source code. You will see me discussing Programming Languages, Program Analysis and Software Engineering. More about my research is here.

If you are looking for an answer to an even more fundamental and important question: Why to study information retrieval?, enjoy the video from my students of the 2018 IR offering at IIITS - Mounika, Parkhi and Pragna.

Publications: DBLP Google Scholar

Teaching

Term@CMI: Feb - Mar 2021: RDBMS, SQL and Visualization
Term@CMI: Dec - Jan 2020: Advanced Information Retrieval
Term@CMI: Aug - Nov 2020: Information Retrieval
Term@CMI: Jan - Apr 2020: Big Data and Hadoop
Term@CMI: Jan - Apr 2020: Applied Program Analysis
Term@CMI: Oct - Nov 2019: RDBMS, SQL and Visualization
Term@CMI: Aug - Sep 2019: Information Retrieval
Term@CMI: Mar - Apr 2019: Program Analysis
Term@IIITS: Aug - Dec 2018: Information Retrieval
Term@IIITS: Aug - Dec 2018: Computer Programming

Talks

Tweets by @venkvinr

Thank You

venkateshv@cmi.ac.in

Thanks to Snigdha Desiraju for the presentations. She can be reached at snigdha.desiraju@gmail.com.
This slide deck is available at http://vvtesh.co.in/.