

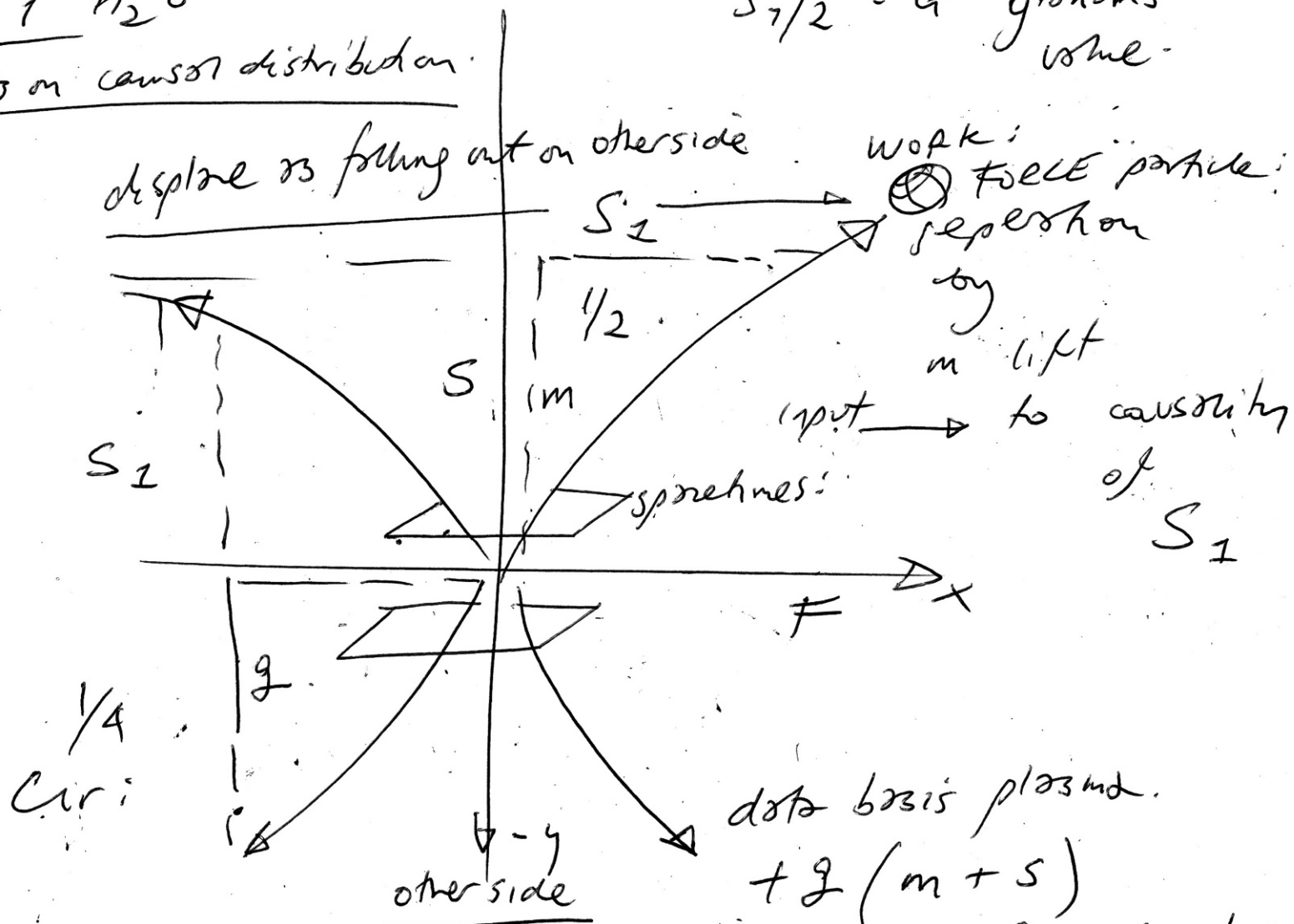
Big Data and You

Preparing Current & Future Information Specialists

Sands Fish
Data Scientist / MIT Libraries
@sandsfish
sands@mit.edu

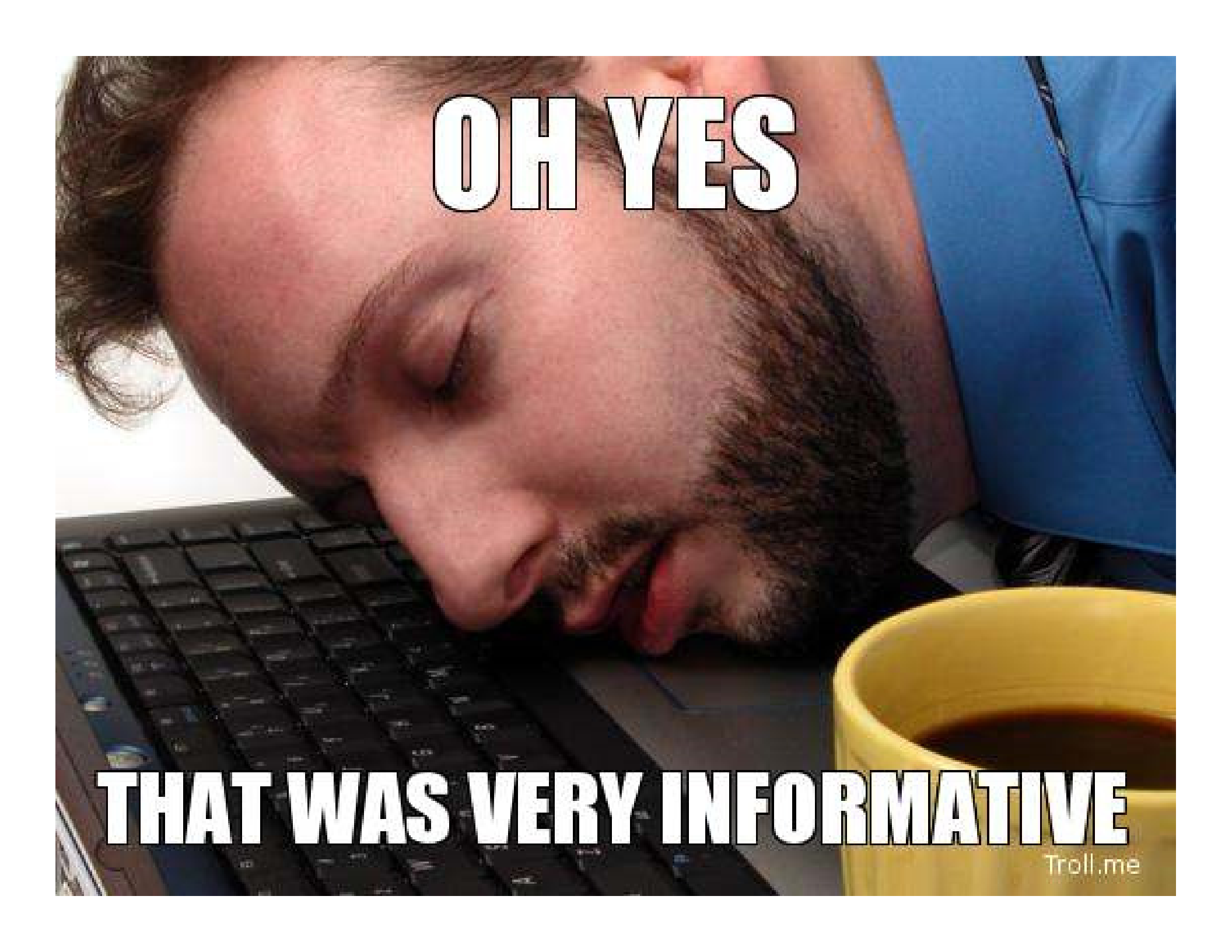
type 1 H_2O
effects on cancer distribution.

$S_{7/2} = G$ grams
volume



other side is where failure -
minimal restriction took into
the universes force warp drive

as formats for S input into
 $\frac{1}{4}$ Cir = $\frac{1}{2}$ displaced
 intimer
 = continue by S, $\frac{1}{2}$ F; values: 50% chance.
 out.

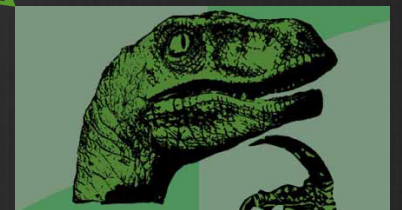
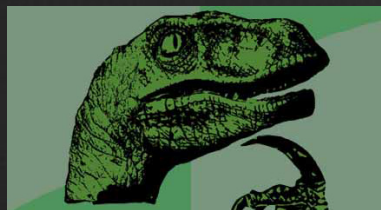
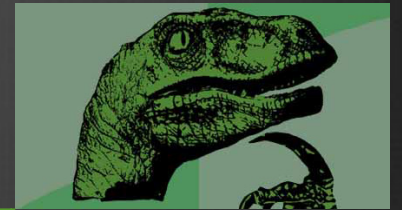
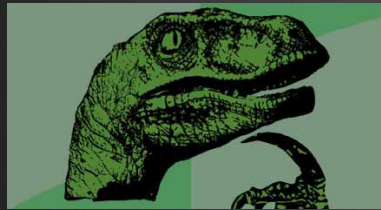
A close-up photograph of a man with a beard and closed eyes, resting his head on a black computer keyboard. He is wearing a blue button-down shirt. To his right is a yellow mug filled with dark coffee. The image is a popular internet meme.

OH YES

THAT WAS VERY INFORMATIVE

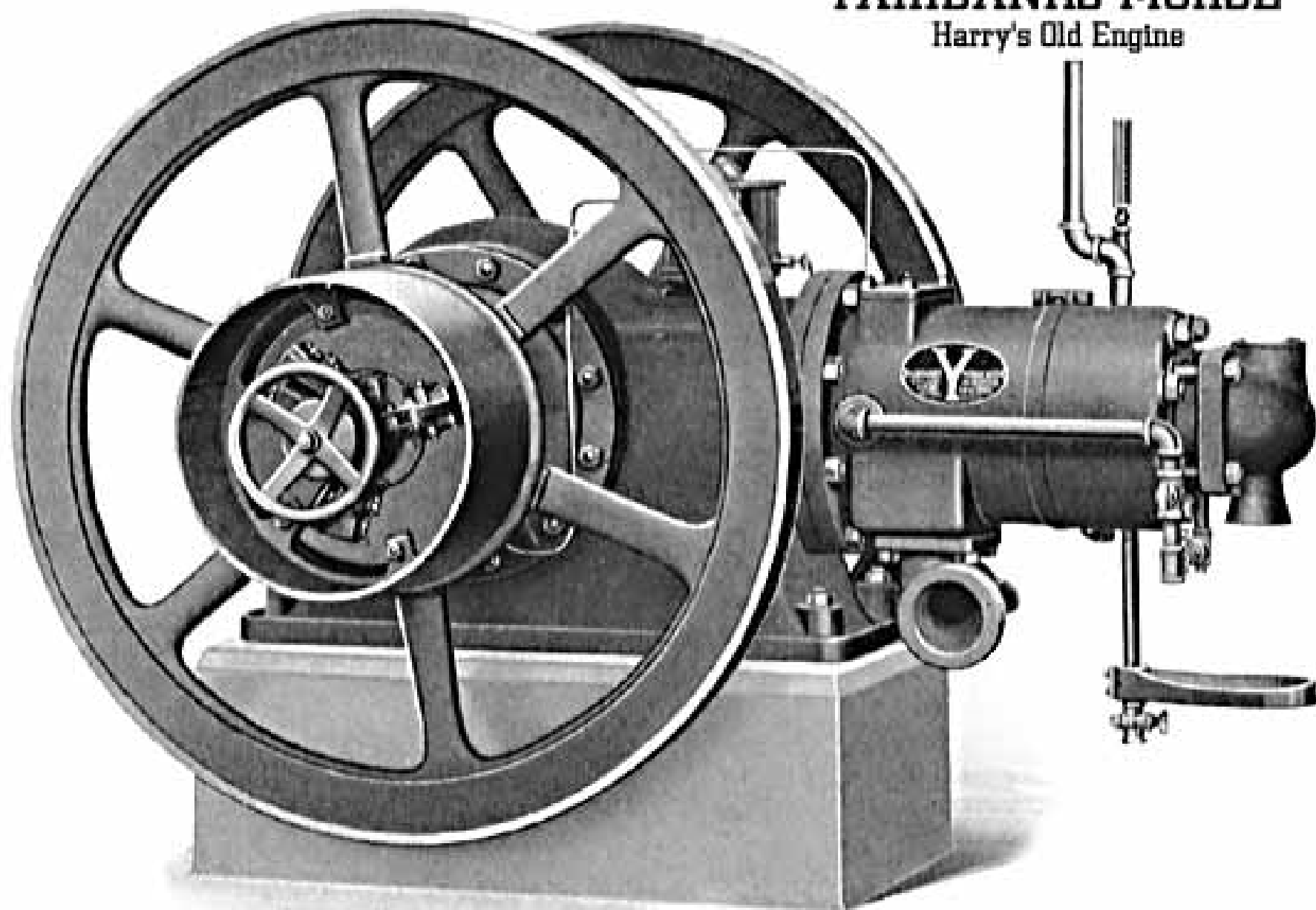
Troll.me

Knowing in the Age of Networked Knowledge

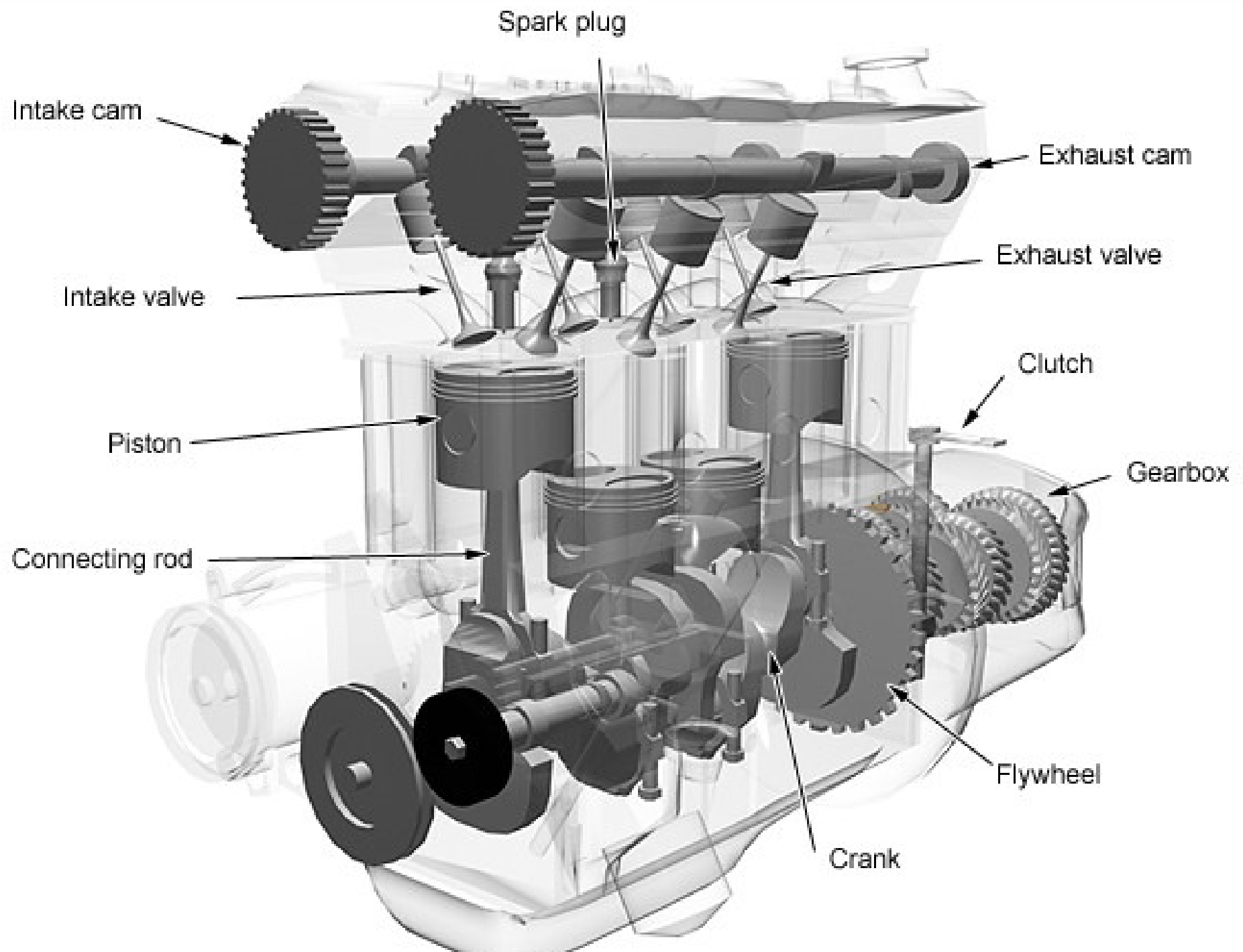


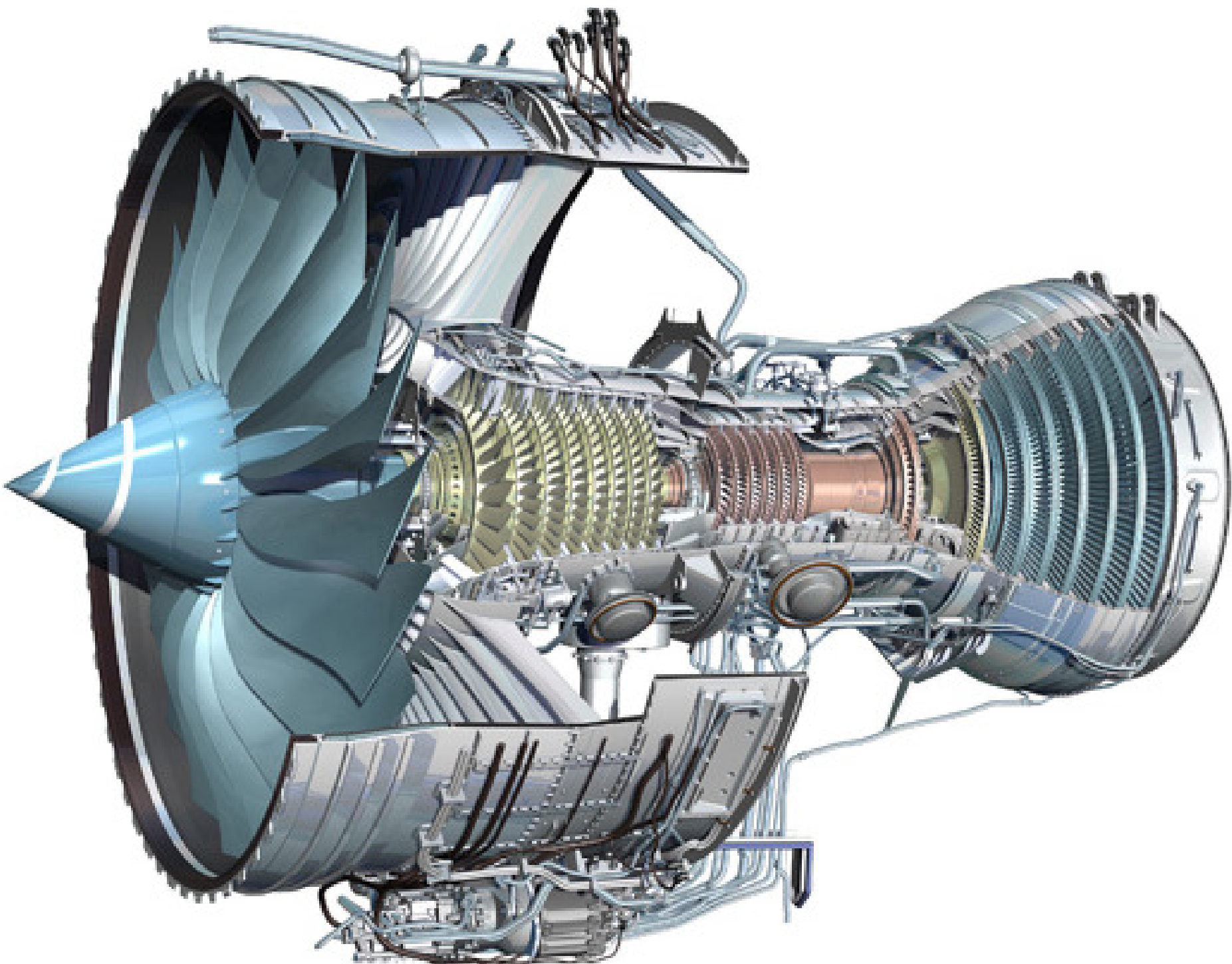
FAIRBANKS MORSE

Harry's Old Engine

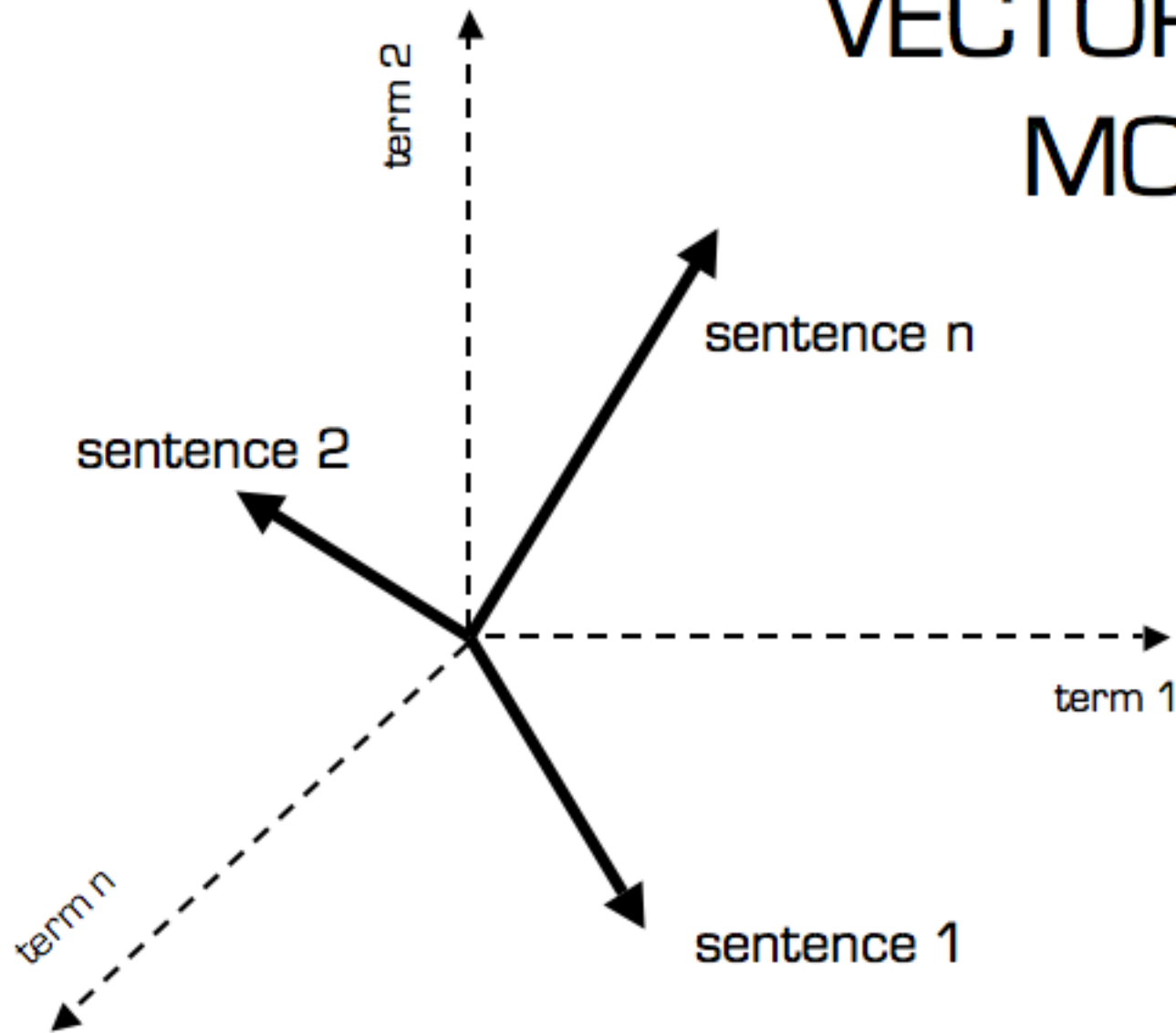


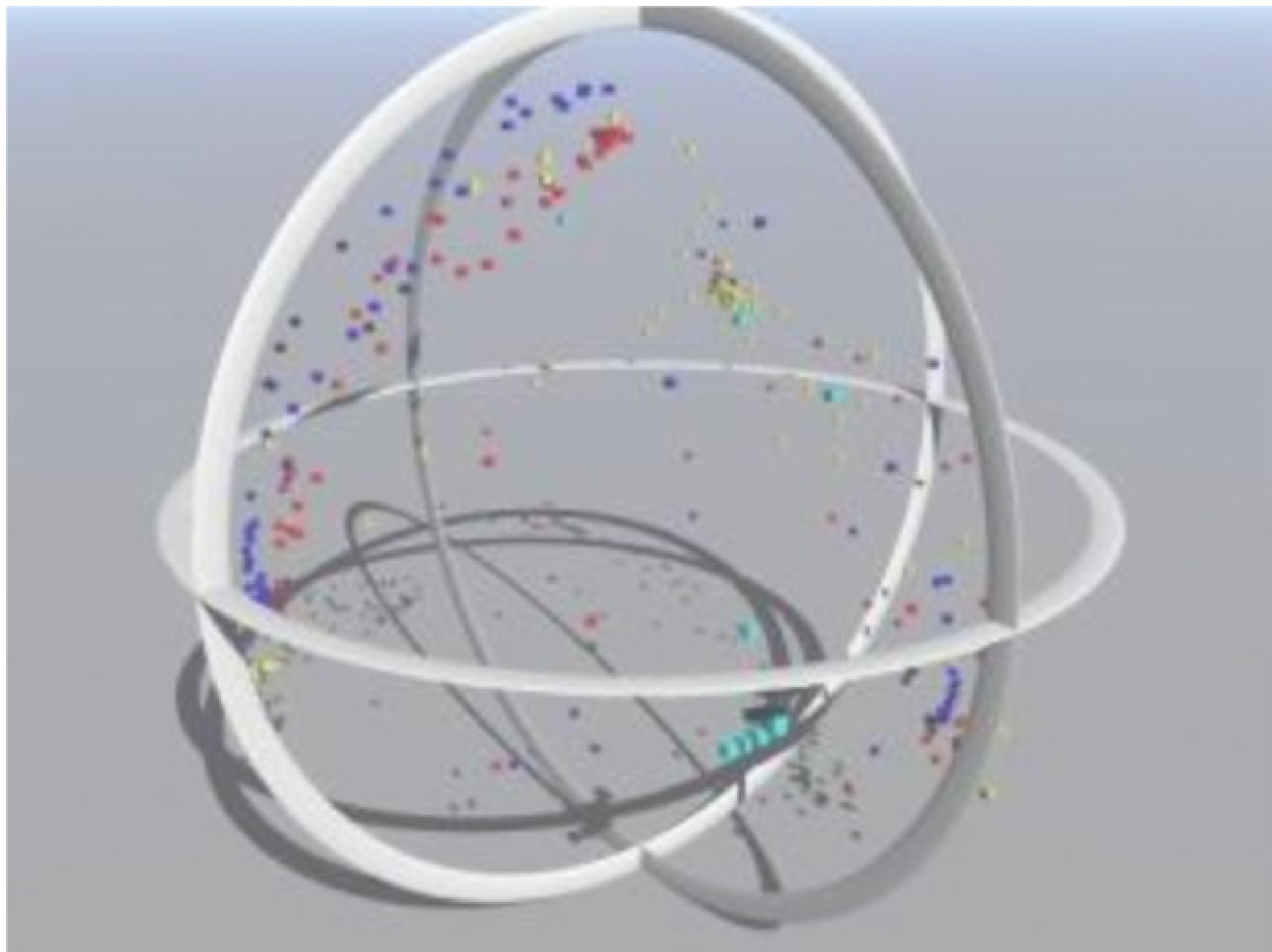
15 H.P. "Y" Oil Engine, Style "H" with "Z" Clutch Pulley

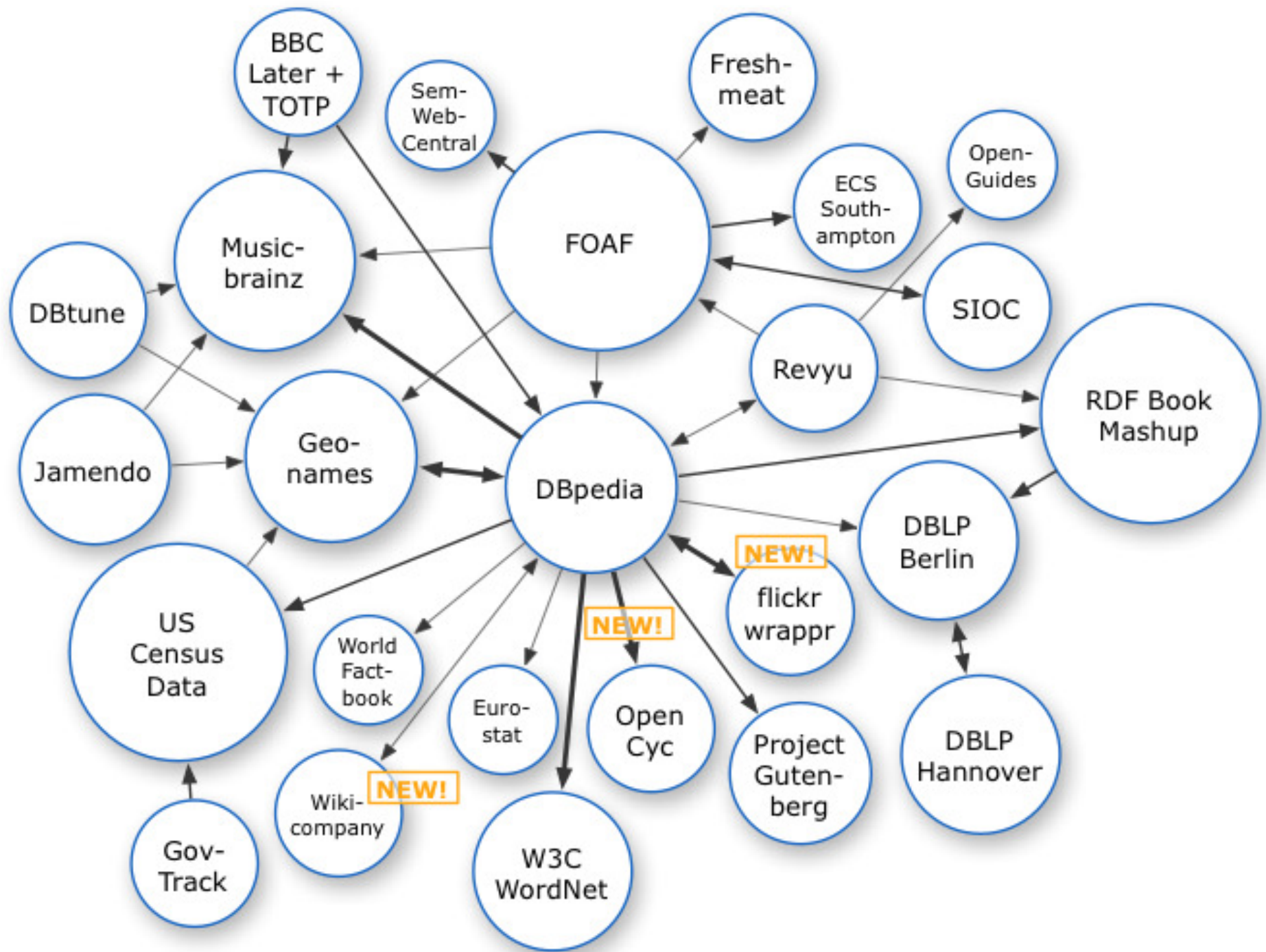




VECTOR SPACE MODEL







Nothing is static.

Everything is connected.

Knowledge representation
is now complex

Scholarly Primitives

- Discovering
- Annotating
- Comparing
- Referring
- Sampling
- Illustrating
- Representing

John Unsworth, 2000.

<http://people.brandeis.edu/~unsworth/Kings.5-00/primitives.html>

Complex Knowledge Objects

- Have multiple representations & ways of being consumed
- Can be a link in a chain, node in a graph, or ecosystem of knowledge.
- Allow different perspectives or ways to ask questions of.

(none of these are true of physical books)

Complex Knowledge Objects

Data Examples:

- JSON, XML, etc. esp. from a URL that allows it to be updated
- Visualizations, sonifications, etc. (mind-maps, interactives)
- Geospatial data, layered, constrained by area
- APIs
- Linked Data, integrated with many other resources

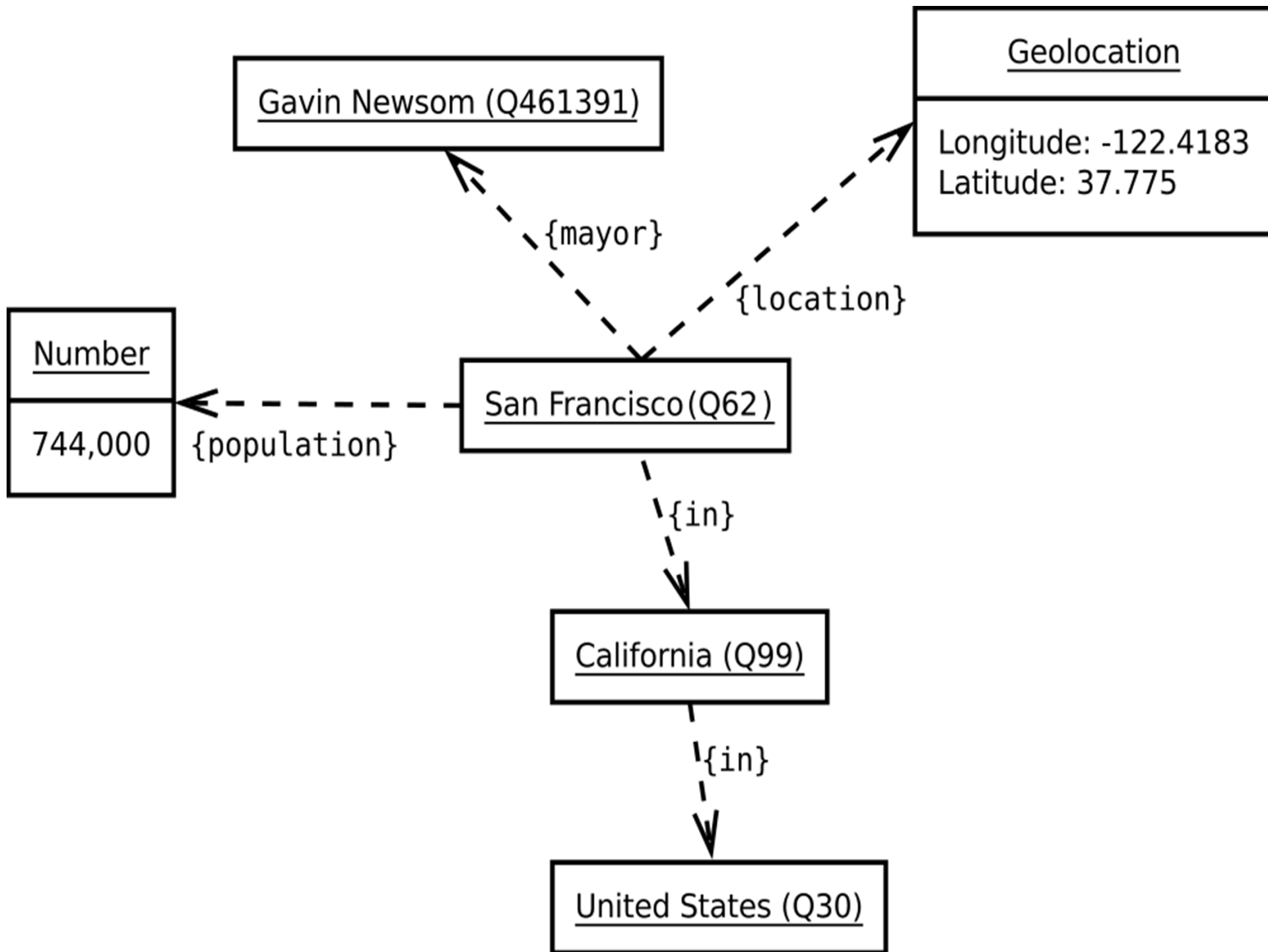
Complex Knowledge Objects

- Tool / Platform Examples:
 - Integrated Data Platforms
 - Courseware
 - MOOCs
 - Interactive Visualizations
 - Commons-based peer production (wikis, reviews, software, etc.)
 - Tweets
 - Data analysis tools
 - Data Enclaves (limited access processing endpoints)

FAA AIRCRAFT REGISTRY + Add filter... 660 OF 360,822 ROWS EXPORT 						
	Name	Address	City	Aircraft Model	Year Manufactured	Aircro
	Federal Express	3131 Democra..	Memphis	190030	1995	5
	Federal Express	2461 Democra..	Memphis	1390006	2000	5
	Federal Express	3101 Tchulaho..	Memphis	5170805	1999	5
	Federal Express	2461 Democra..	Memphis	1390006	2001	5
	Federal Express	3131 Democra..	Memphis	5170805	2008	5
	Federal Express	2461 Democra..	Memphis	1390044	2012	5
	Federal Express	3101 Tchulaho	Memphis	1390044	2012	5

EXPOSE WHAT'S HIDDEN

Enigma empowers the discovery of hidden facts and connections across the universe of big public data.



▼ {

```
"completed_in": 0.09,
"max_id": 335187769854926850,
"max_id_str": "335187769854926848",
"next_page": "?page=2&max_id=335187769854926848&q=swanros",
"page": 1,
"query": "swanros",
"refresh_url": "?since_id=335187769854926848&q=swanros",
"results": [
```

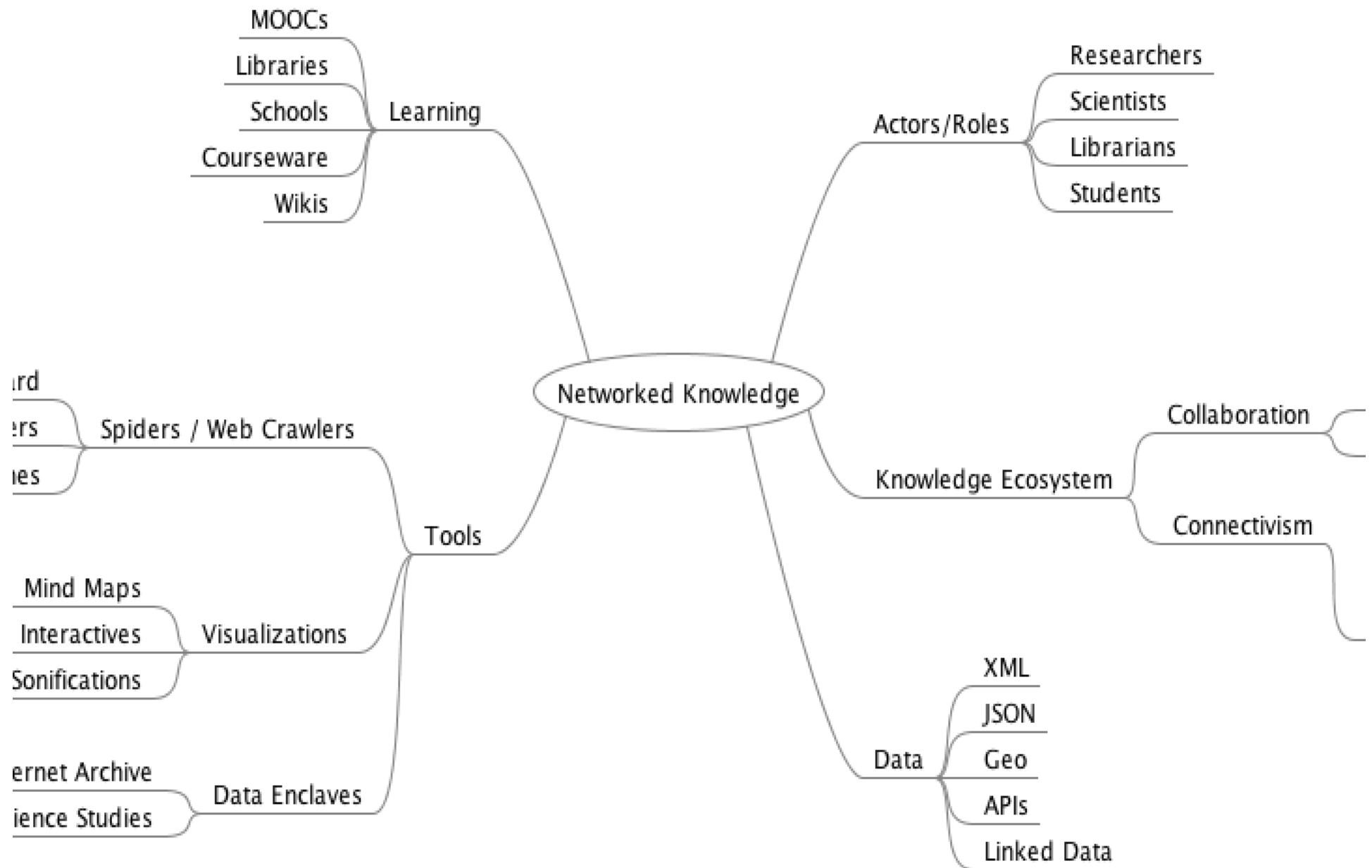
▼ {

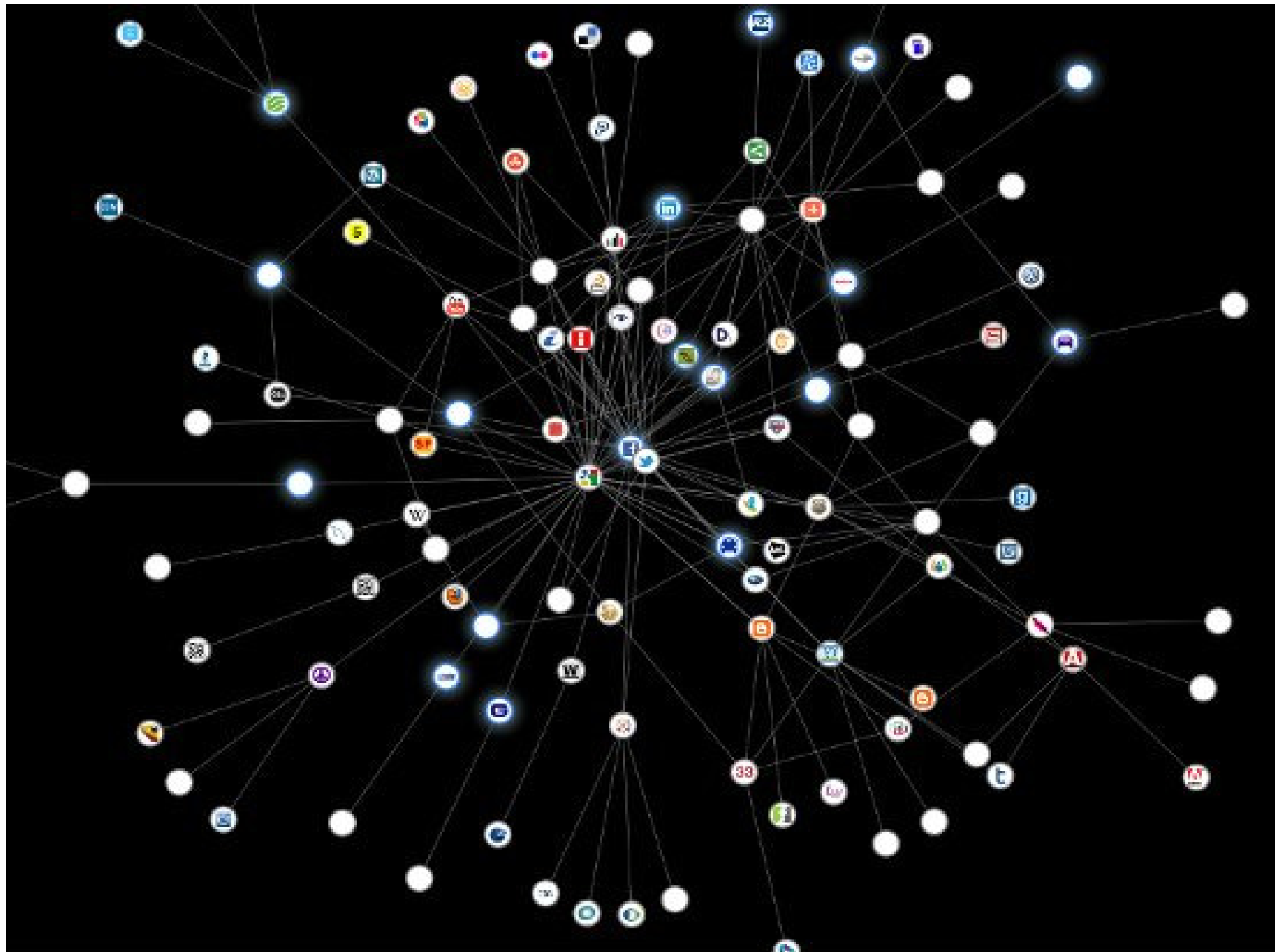
```
  "created_at": "Fri, 17 May 2013 00:19:26 +0000",
  "from_user": "Swanros",
  "from_user_id": 229981602,
  "from_user_id_str": "229981602",
  "from_user_name": "Oscar Swanros 🍏",
  "geo": null,
  "id": 335187769854926850,
  "id_str": "335187769854926848",
  "iso_language_code": "en",
  "metadata": {
    "result_type": "recent"
  },
  "profile_image_url": "http://a0.twimg.com/profile_images/3652030195/a7c522814581c3110aa908726d7028b2_normal.jpeg",
  "profile_image_url_https": "https://si0.twimg.com/profile_images/3652030195/a7c522814581c3110aa908726d7028b2_normal.jpeg",
  "source": "&lt;a href=&quot;http://sites.google.com/site/yorufukurou&quot;&gt;YoruFukurou&lt;/a&gt;",
  "text": "@treehouse Your courses are awesome. Almost am a pro Android dev within two days! Thanks a lot! :=]",
  "to_user": "treehouse",
  "to_user_id": 14843763,
  "to_user_id_str": "14843763",
  "to_user_name": "Treehouse",
  "in_reply_to_status_id": 335186686361341950,
  "in_reply_to_status_id_str": "335186686361341952"
```

▼ {

Raw

Parsed





Methods of Exploration

In this diverse ecosystem, there is no one way of exploring a topic.

- Manual Browsing
- Automated Spidering (e.g. Berkman / Media Cloud)
- Collection / Trawling (e.g. Browser Plugins)
- Conventional Big Data (e.g. Hadoop, Map/Reduce)
- Using Linked Data to branch out through related concepts
- Algorithmic Data Processing (e.g. Topic Modeling)

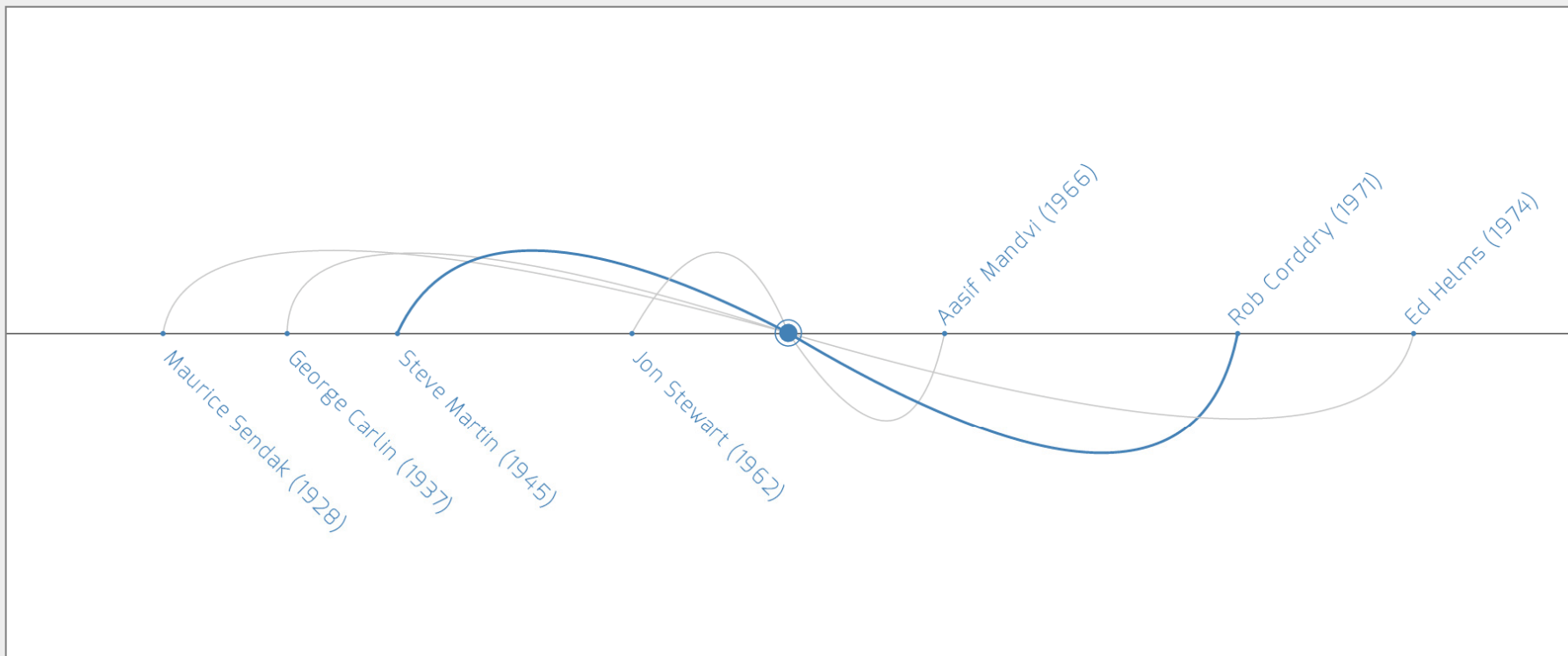
Node: Singing
Node: Ben Lovett
Node: 1978 births
Node: Felipe Claybrooks
Node: Georgia Institute of Technology
Node: William H. Glenn
Node: Cave Spring, Georgia
Node: Floyd County, Georgia
Node: King Archaeological Site
Node: Floyd County, Georgia
Node: U.S. Route 27 in Georgia
Node: Stewart County, Georgia
Node: National Register of Historic Places listings in Stewart County, Georgia
Node: Richland, Georgia
Node: Populated places in Stewart County, Georgia
Node: Lumpkin, Georgia
Node: List of sovereign states
Node: Duncan, West Virginia
Node: Jackson County, West Virginia
Node: Meigs County, Ohio
Node: Tobias A. Plants
Node: Members of the Ohio House of Representatives
Node: Jerry W. Krupinski
Node: Eileen Krupinski
Node: Jerry W. Krupinski
Node: Arthur Bowers
Node: Douglas Applegate
Node: Democratic Party (United States)

influence discovery

powered by dbpedia.org

http://dbpedia.org/resource/Stephen_Colbert

Analyze Influence

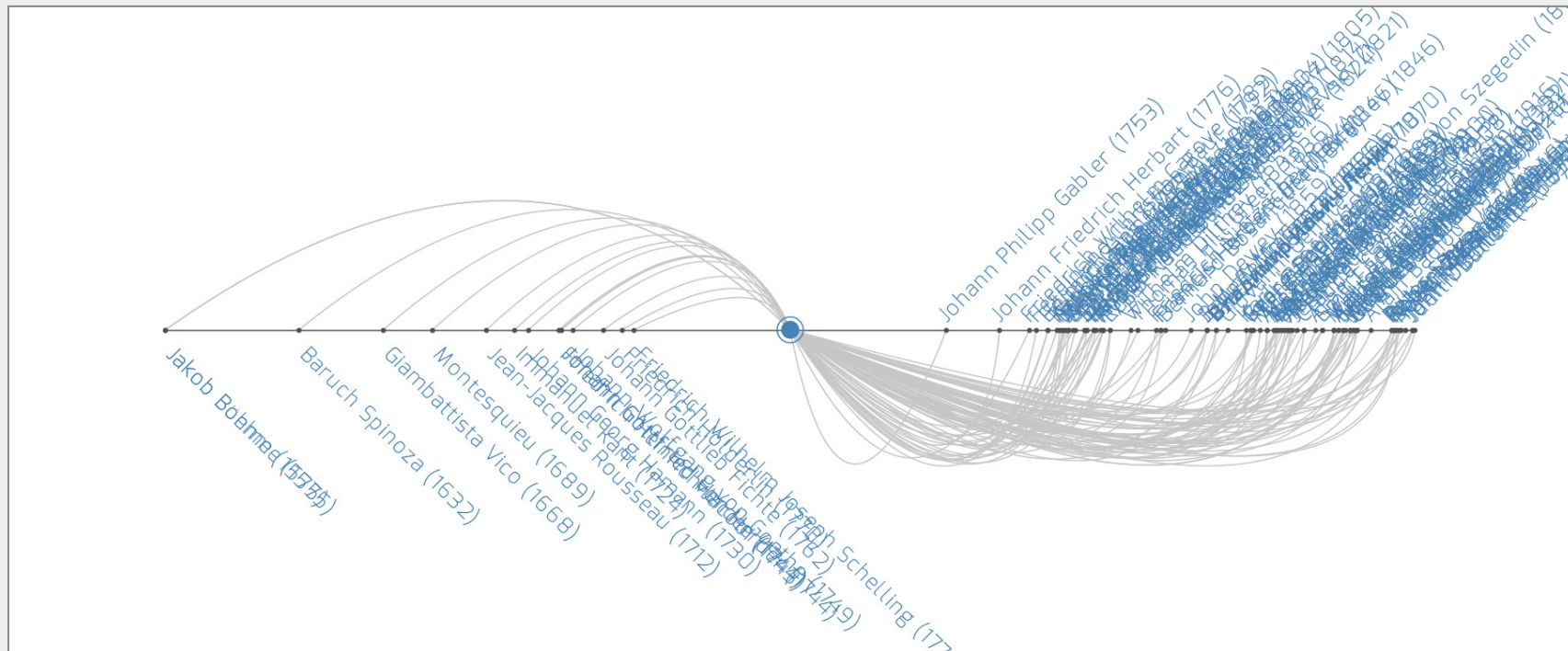


http://dbpedia.org/resource/Stephen_Colbert influenced 3 people:

► http://dbpedia.org/resource/Aasif_Mandvi

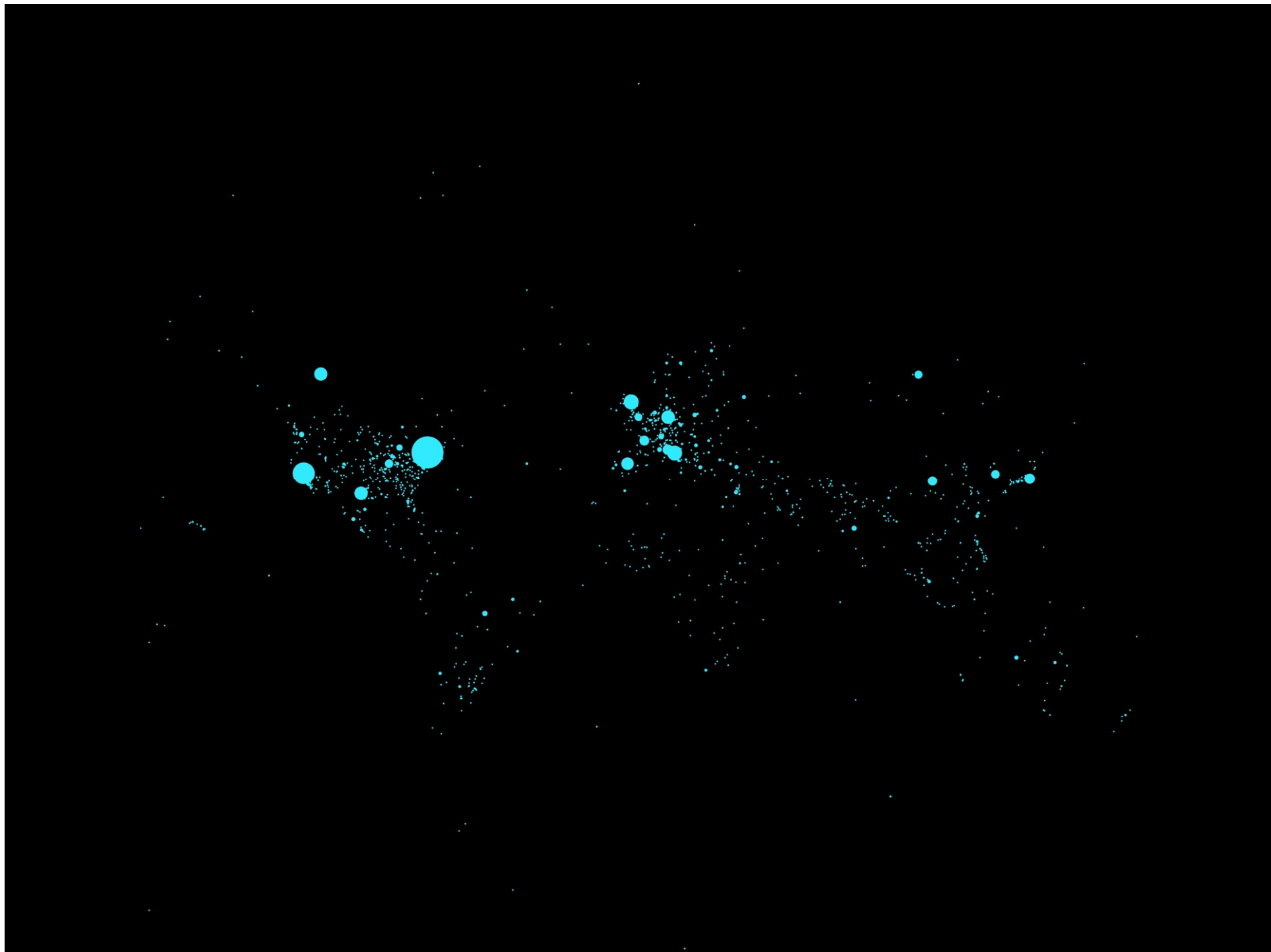
http://dbpedia.org/resource/Georg_Wilhelm_Friedrich_Hegel

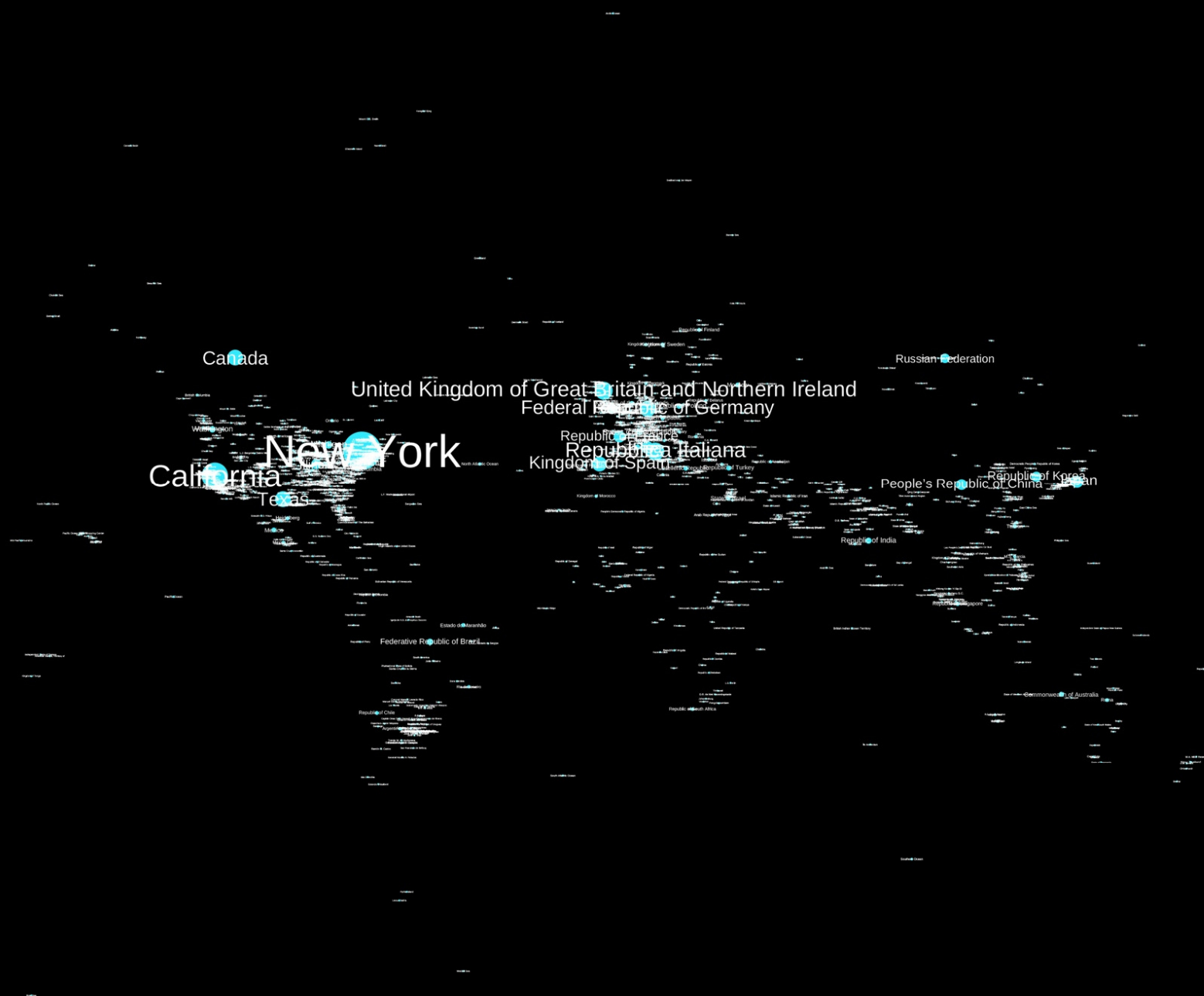
Analyze Influence

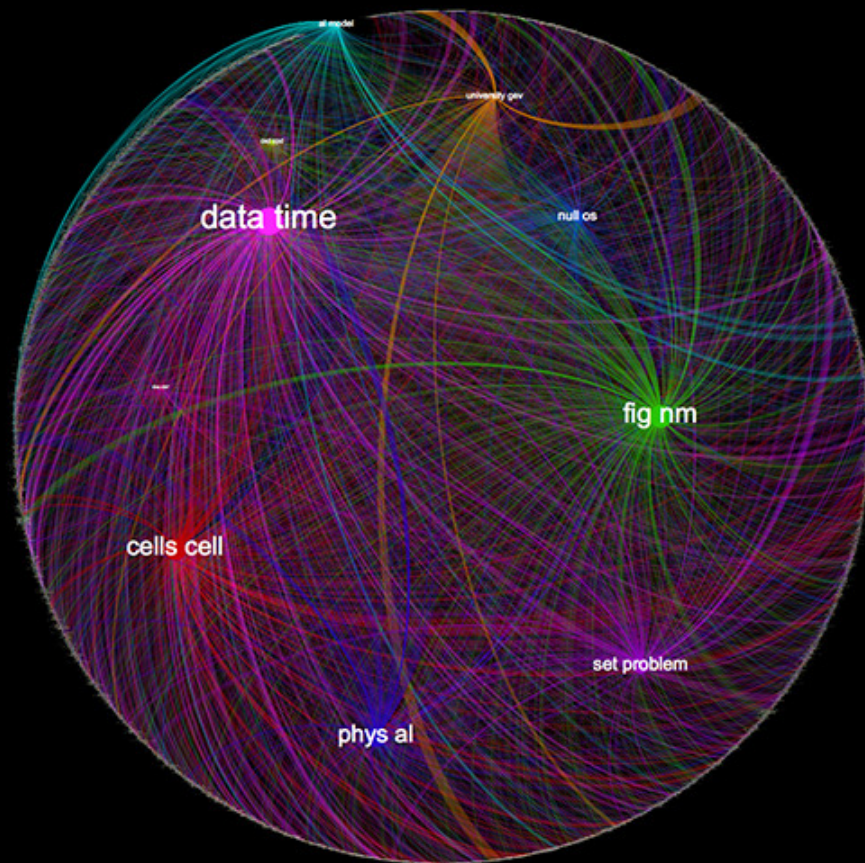


http://dbpedia.org/resource/Georg_Wilhelm_Friedrich_Hegel influenced 83 people:

- ▶ http://dbpedia.org/resource/George_Henry_Lewes
- ▶ http://dbpedia.org/resource/J._M._E._McTaggart
- ▶ http://dbpedia.org/resource/Murray_Bookchin
- ▶ http://dbpedia.org/resource/Robert_B._Pippin
- ▶ http://dbpedia.org/resource/Robert_Brandom
- ▶ http://dbpedia.org/resource/Roberto_Mangabeira_Unger
- ▶ http://dbpedia.org/resource/Gilles_Deleuze
- ▶ http://dbpedia.org/resource/Guy_Debord
- ▶ http://dbpedia.org/resource/Heinrich_Heine
- ▶ http://dbpedia.org/resource/Johann_Friedrich_Herbart
- ▶ http://dbpedia.org/resource/John_Dewey







Problems of Completeness

- When do you know that you have enough information?
- What kind of compromises are made when information is more massive than anyone can consume?

Problems of Integration

When data comes from many different silos, in many different structures and formats, how do you bring all of this knowledge together?

- One solution is RDF, which provides a common data generic data model. Collaborative ontology development can allow communities to work together.
- Open standards.
- Build tools and services that provide easy access to the underlying data.

How To Get A Grip

- Keep abreast of W3C developments and other standards bodies.
- Don't focus too much on single technologies. They will shift quickly.
- Learn at least one data visualization technology.
- Remember to frame questions of data in more than one way.
- Ask your own questions of the data yourself. Understand it from the point of the user.



The most merciful thing in the world, I think, is the inability of the human mind to correlate all its contents.

(H. P. Lovecraft)

izquotes.com