THE ROLE OF METADATA IN THE SECOND MACHINE AGE

DC-2016 / KØBENHAVN / 13 OCTOBER

Bradley P. Allen

Chief Architect, Elsevier

@bradleypallen

DUBLIN CORE IN ITS THIRD DECADE

A basic description mechanism for digital information that:

- can be used in all domains
- can be used for any type of resource
- is simple, yet powerful
- can be extended and can work with specific solutions.

Dekkers, M. (2009). History, objectives and approaches of the Dublin Core Metadata Initiative. Retrieved September 29, 2016, from http://dublincore.org/resources/training/frd_20091217/Tutorial_FRD_dekkers-1.pdf

DCMI IN ITS THIRD DECADE

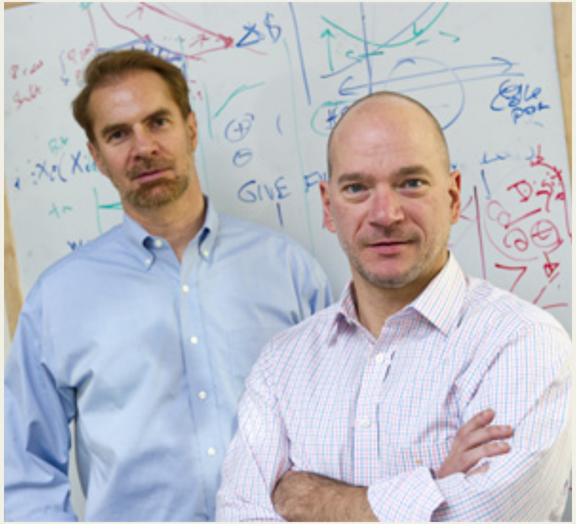
Since its early days in the mid-1990s, DCMI's founding principle has been the discovery and management of resources through metadata across the boundaries of information silos on the Web and within intranets.

Mission and Principles. (n.d.). Retrieved September 29, 2016, from http://dublincore.org/about-us/

THE IMPACT OF DC AND DCMI

- Arguably the most impactful of many efforts to bring to bear the library science approach to information organization on the Web at large
- Simple, easy to understand, easy to apply, well documented
- While the user experience of discovery has come to be dominated by search engines such as Google, metadata standards are pervasive in the infrastructure of content curation and management, and underpins search infrastructure
- All of this has been a huge contribution to the Web
- It established the approach to bringing the Semantic Web and its associated vocabularies and ontologies online as the Linked Open Data Cloud

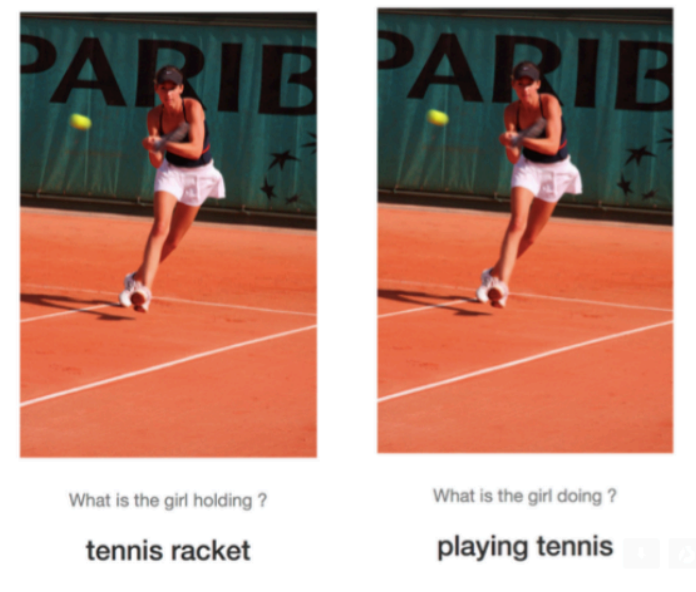
THE NEXT 20 YEARS: THE SECOND MACHINE AGE

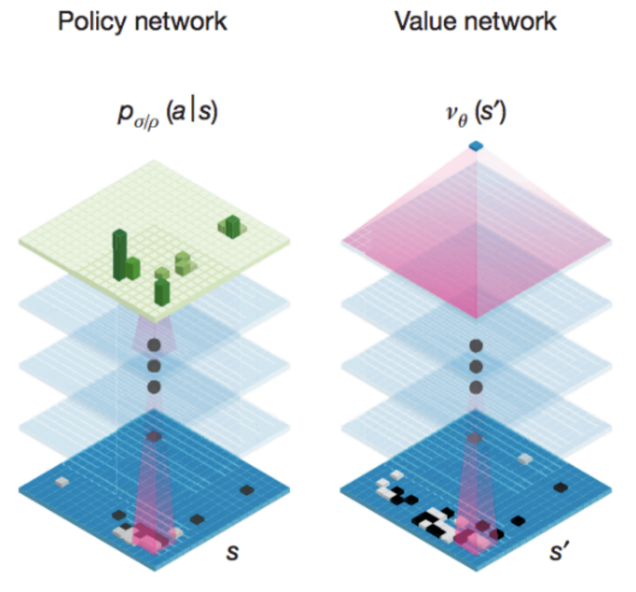


The second machine age will be characterized by countless instances of machine intelligence and billions of interconnected brains working together to better understand and improve our world.

Brynjolfsson, E., & McAfee, A. (2014). The second machine age: Work, progress, and prosperity in a time of brilliant technologies. New York: W.W. Norton & Company.

MACHINE INTELLIGENCE IS ON THE RISE





WHAT CAN MACHINE INTELLIGENCE DO TODAY?

If there's a task that a normal person can do with less than one second of thinking, there's a very good chance we can automate it with deep learning.

Andrew Ng, Chief Scientist, Baidu (lecture at Bay Area Deep Learning School, Stanford, CA, September 24, 2016)

MACHINE INTELLIGENCE IN USE

	Application areas	Specific examples	Recent M&A and hiring
	• Speech understanding • Web search • Image search • Machine translation • Personalization and contextual search • Logistics	• "Hummingbird" release contextual search • Google Now intelligent assistant • Knowledge Graph/Vault • Spam filtering in Gmail • Self-driving cars	• Deepmind • Dark Blue Labs • Vision Factory • Timeful • DNNResearch • Hinton (U. Toronto)
	• Speech understanding • Cloud services • Personalization and contextual search	• Bing contextual search • Cortana intelligent assistant • Azure Machine Learning Services • Satori Knowledge Base • Microsoft Cognitive Services	• Merging Bing, Cortana and MSR into ~5,000-person AI division
	• Speech understanding • Personalization and contextual search	• Siri intelligent assistant • iOS9 Proactive Suggestions	• Turi • Establishing AI division in Seattle
IBM Watson	• Question answering • Cloud services • Healthcare decision support	• Watson Discovery for Life Sciences • Watson Discovery Services • Watson Health Cloud • Watson for Health	• AlchemyAPI • MergeHealth • Truven
	• Face detection & recognition • Personalization and contextual search • Question answering	• Facebook Open Graph • Facebook Graph Search • Face recognition in Facebook, Instagram	• Wit.ai • LeCun (NYU) • Bottou (Microsoft Research)

WHAT DOES THIS MEAN FOR THE FUTURE OF METADATA?

- Our focus as a community has been on helping people find and use information on the Web
- Metadata standards and machines have been a means to that end
- The emergence of machine intelligence and machine reading in the second machine age will make it even easier to automate the production of metadata to help people find, filter and organize information
- Things information workers do in less than a second that we can train machines to do: read a page for concepts and facts, recognize the type of image on a page, …

ELSEVIER'S BUSINESS: PROVIDING ANSWERS FOR RESEARCHERS, DOCTORS AND NURSES

My work is moving towards a new field; what should I know?

- Journal articles, reference works, profiles of researchers, funders & institutions
- Recommendations of people to connect with, reading lists, topic pages

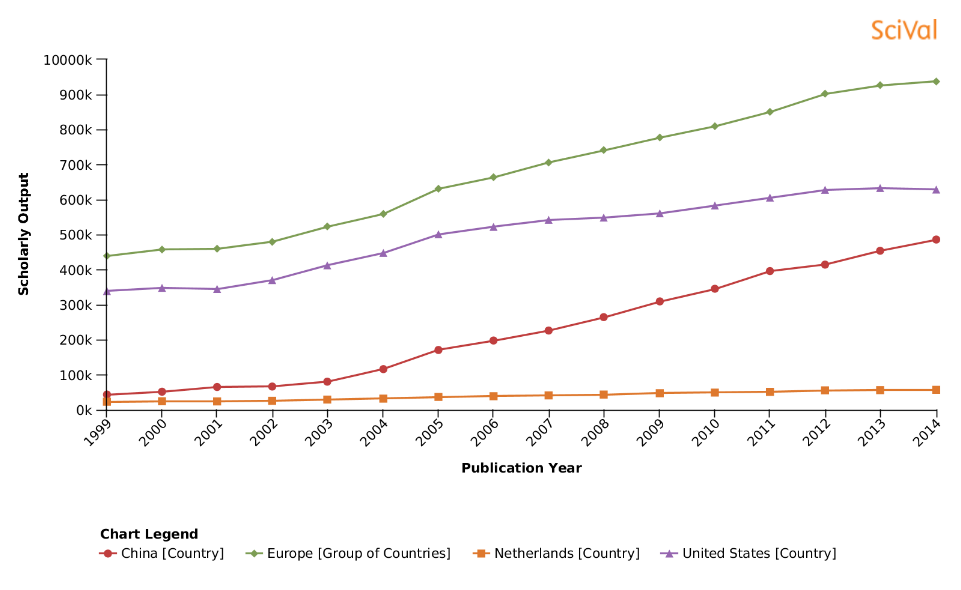
How should I treat my patient given her condition & history?

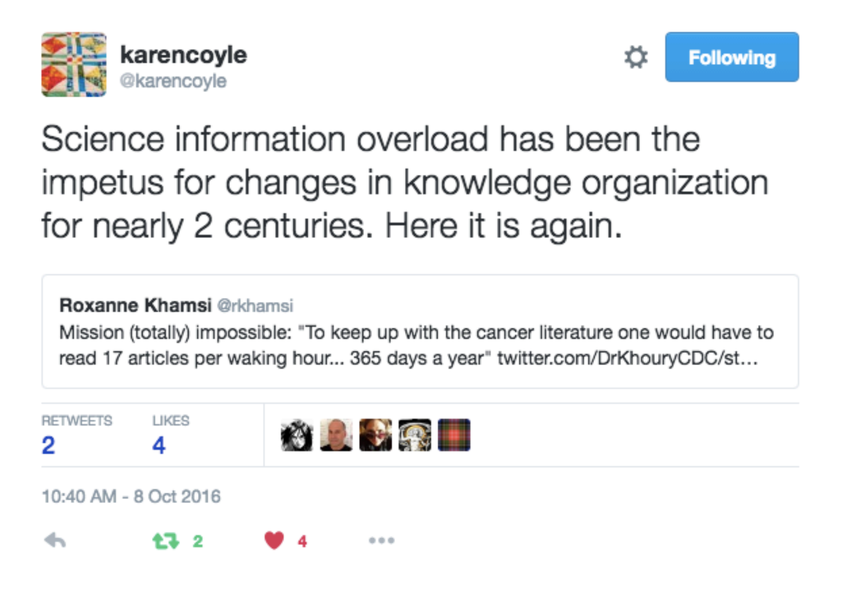
- Journal articles, reference works, medical guidelines, electronic health records
- Treatment plan with alternatives personalized for the patient

How can I master the subject matter of the course I am taking?

- Course syllabus, reference works, course objectives, student history
- Quiz plan based on the student's history and course objectives

THE GROWTH OF SCIENCE COMPLICATES OUR EFFORTS





ANSWERS ARE ABOUT THINGS, NOT JUST WORKS

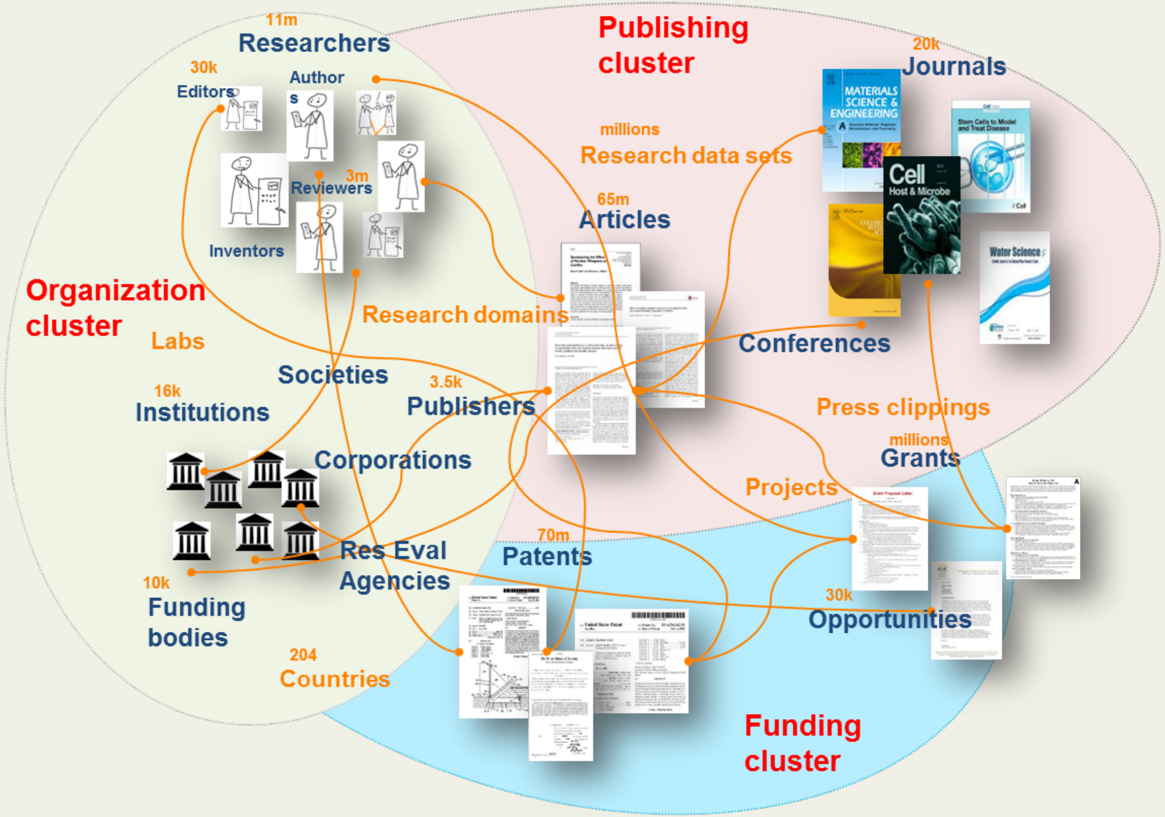
Why shouldn’t a search on an author return information about the author, including the author’s works? Where was the author born, when did she live, what is she known for? … All of this is possible, but only if we can make some fundamental changes in our approach to bibliographic description. ... The challenge for us lies in transforming what we can of our data into interrelated “things” without overindulging that metaphor.

Coyle, K. (2016). FRBR, before and after: a look at our bibliographical models. Chicago: ALA Editions.

KNOWLEDGE GRAPHS AND MACHINE READING TURN CONTENT INTO ANSWERS

- Knowledge graphs are "graph structured *knowledge bases* (KBs) which store factual information in form of relationships between entities” (Nickel, M., Murphy, K., Tresp, V. and Gabrilovich, E. (2015). A review of relational machine learning for knowledge graphs. arXiv:1503.00759v3)
- Knowledge graphs are metadata evolved beyond the focus on the work, linking people, concepts, things and events
- Knowledge graphs organize data extracted from content through machine reading so that queries can provide answers

ELSEVIER: KNOWLEDGE GRAPHS FOR RESEARCH



ELSEVIER: KNOWLEDGE GRAPHS FOR LIFE SCIENCES

Biological Pathways extracted via semantic text mining	Bioactivities through text analysis	Chemical Structures And Properties

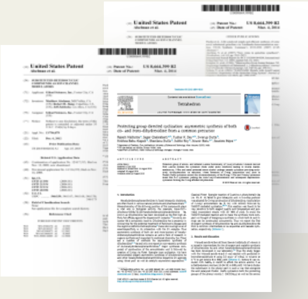
A upregulates B

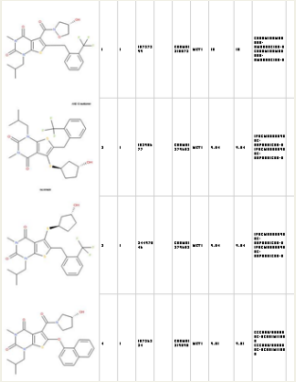
B upregulates C

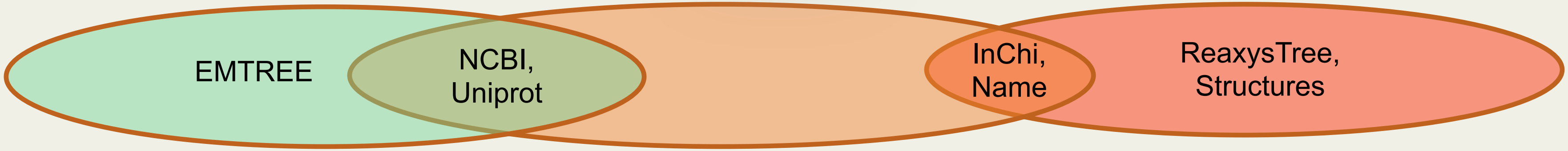
C increases disease D

A ⇨ B ⇨ C ⇨ D

IC_{50} 6.3nM, kinase binding assay
10mM concentration

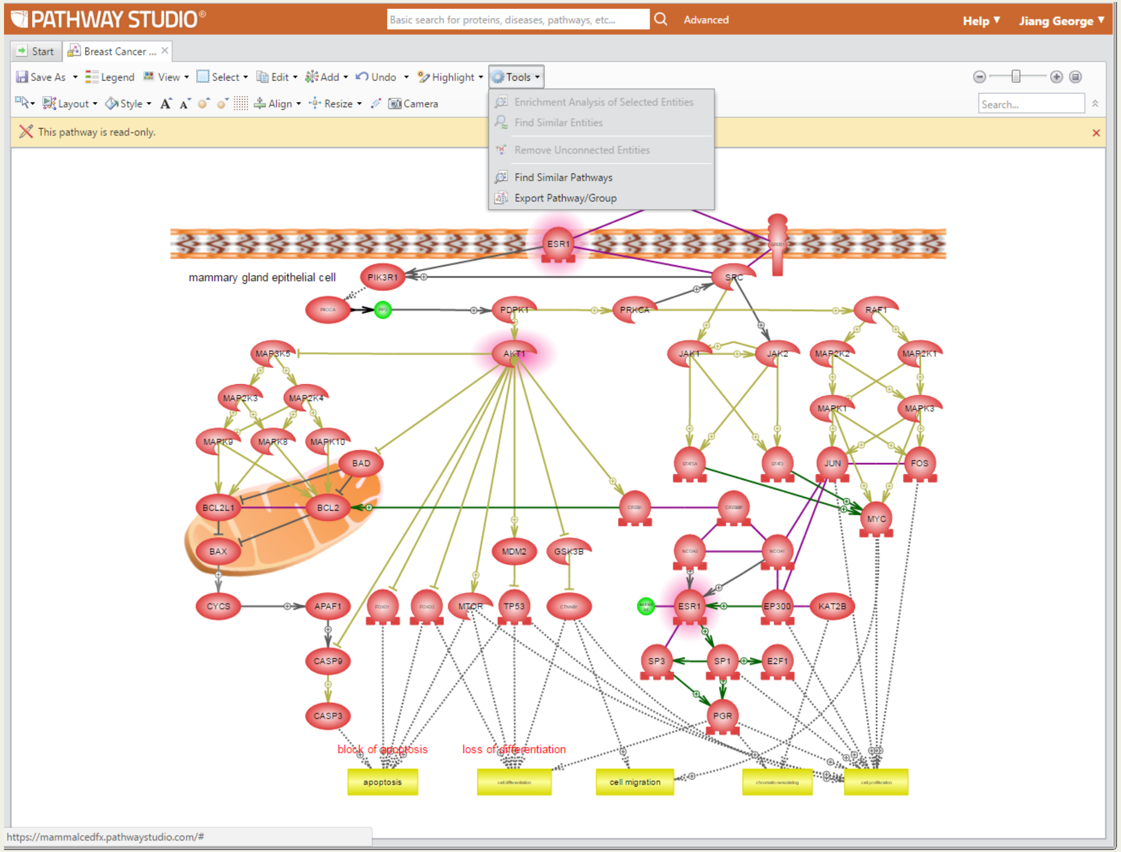




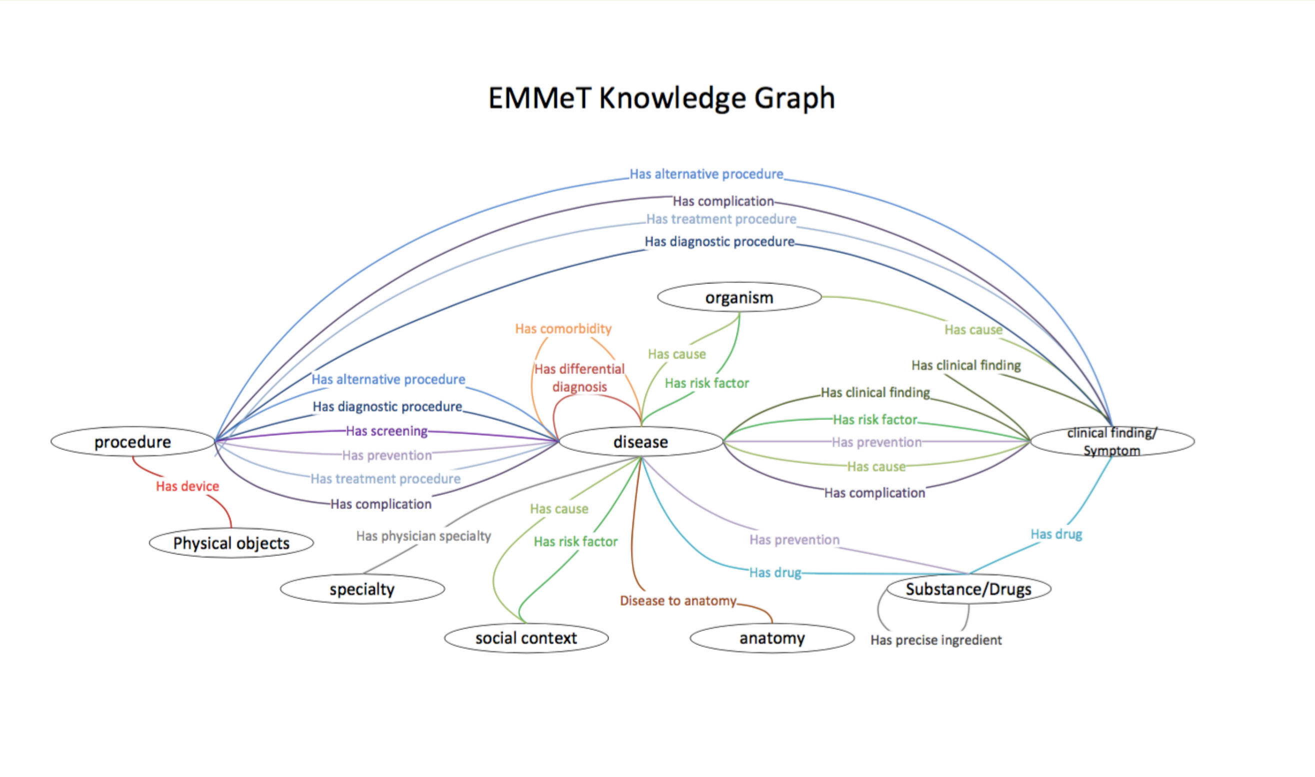


Normalizing vocabularies required: proteins, diseases, drugs, chemicals

ELSEVIER: KNOWLEDGE GRAPHS FOR LIFE SCIENCES



ELSEVIER: KNOWLEDGE GRAPHS FOR HEALTHCARE



ELSEVIER: USING JSON-LD FOR ANNOTATIONS

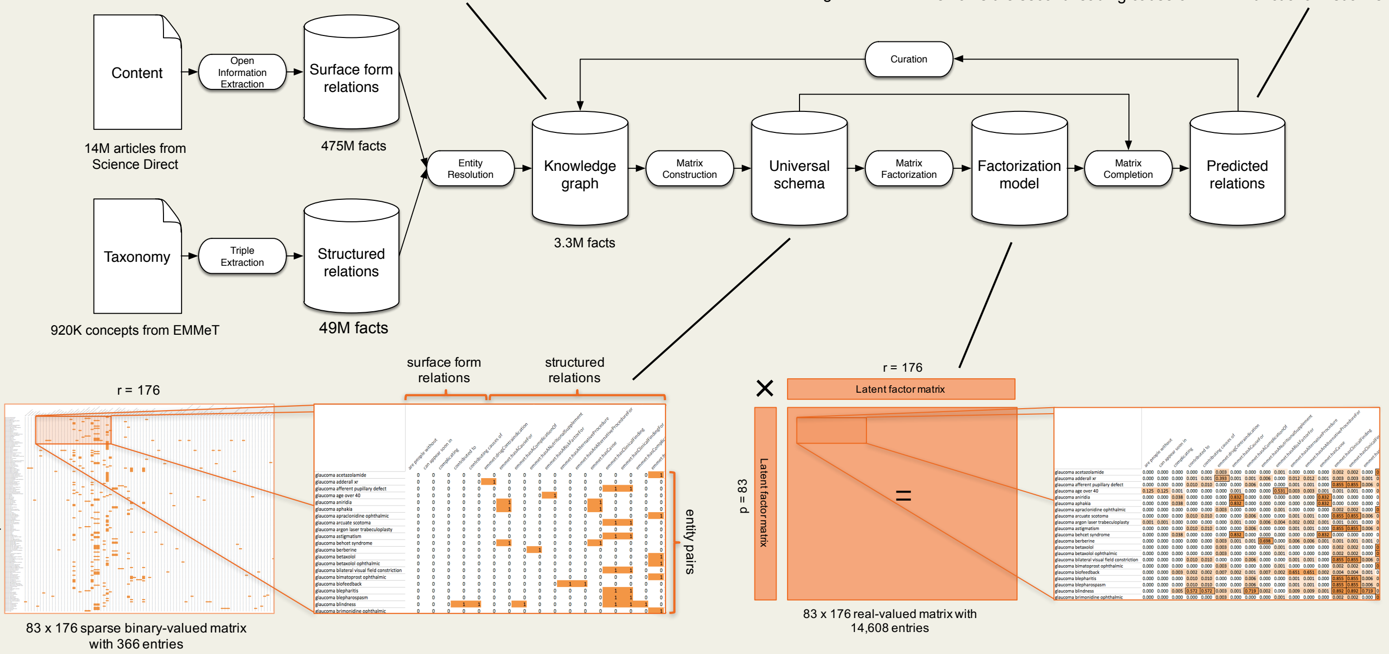

```
{
    "@context": ["
http://vtw.elsevier.com/metadata/context.jsonld",
                 {"eaf": "
http://vtw.elsevier.com/data/ns/Formats/Annotation",
                  "oca":"http://ontochem.com/ns/ann"}
                ],
    "@type": "eaf:AnnotationDocument",
    "eaf:annotatedSourceInfo": {
        "dct:identifier": "E-602148748",
        "dct:title": "A rare mutation in MYH7 gene occurs with
overlapping phenotype",
        "dct:type": "
http://data.elsevier.com/vocabulary/ElsevierContentTypes/1.3"
    },
    "eaf:annotationProvenance": {
        "dct:format" : "application/xhtml+xml",
        "cef:enrichmentTool":"OCMiner",
        "cef:enrichmentToolVersion": "RxT edition
2.2.Milestone6",
        "cef:dictionary": "ReaxysTree",
        "cef:dictionaryVersion": "13.0",
        "prov:generatedAtTime": "2016-02-16T16:11:07Z"
    },
```

```
"eaf:annotations": [
       {
           "eaf:localID": "1",
           "eaf:sourceText": "beta-myosin",
           "eaf:conceptID": "",
           "eaf:annotationType": "concept",
           "eaf:annotationScope": "text range",
           "eaf:range": "2726-2737",
           "eaf:confidence": "77",
           "eaf:relevance": "72",
           "oca:d": "chemCompound",
           "oca:CNType": "Compound",
           "oca:CNHasStructure": "NO",
           "oca:CNSubType": "Trivial"
       },{
           "eaf:localID": "2",
           "eaf:sourceText": "liquid",
           "eaf:conceptID": "817100004230",
           "eaf:annotationType": "concept",
           "eaf:annotationScope": "text range",
           "eaf:range": "7448-7454",
           "eaf:confidence": "65",
           "eaf:relevance": "22",
           "oca:d": "reaxystree",
```

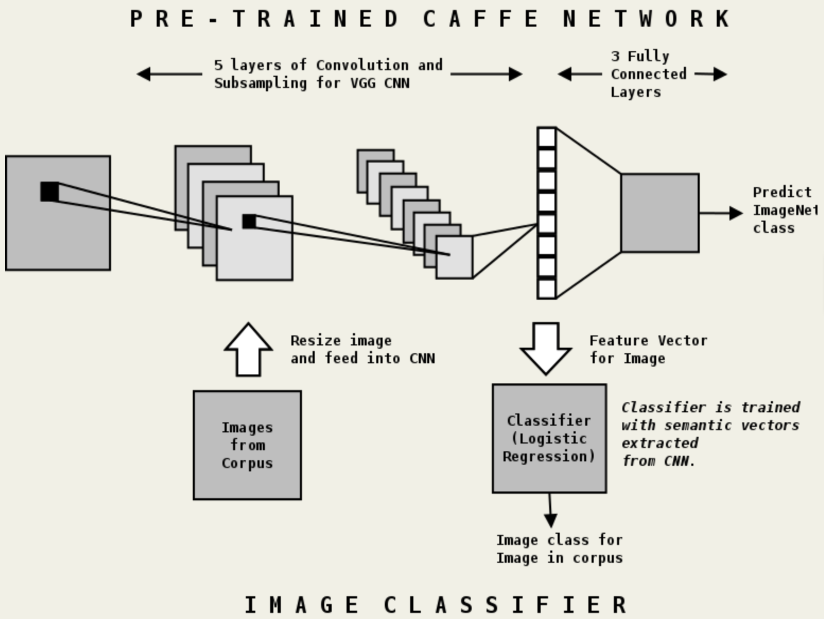
ELSEVIER: MACHINE LEARNING FOR LINK DISCOVERY

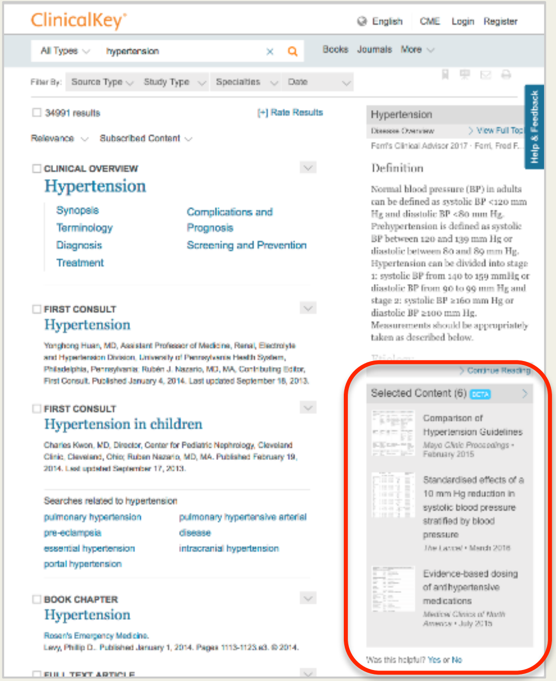
diseases	2791370	glaucoma	have been documented to cause	contact dermatitis	3815093	diseases
diseases	2791370	glaucoma	is assessed through	evaluation	5415395	qualifier
diseases	2791370	glaucoma	progresses more rapidly than	primary open-angle glaucoma	8247149	diseases
diseases	2791370	glaucoma	recommend	treatment	5216597	procedures
diseases	2791370	glaucoma	supports the assumption that	oxidative stress	8184588	diseases
diseases	2791370	glaucoma	is the death of	retinal ganglion cells	8002088	anatomy

glaucoma	developed many years after	chronic inflammation of uveal tract
glaucoma	develop following	chronic inflammation of uveal tract
glaucoma	can appear soon in	family history of glaucoma
glaucoma	can appear soon in	age over 40
glaucoma	the risk of	functional visual field loss
glaucoma	contributing causes of	functional visual field loss
glaucoma	contributed to	functional visual field loss
glaucoma	is considered the second leading cause of	functional visual field loss
glaucoma	remains the second leading cause of	functional visual field loss

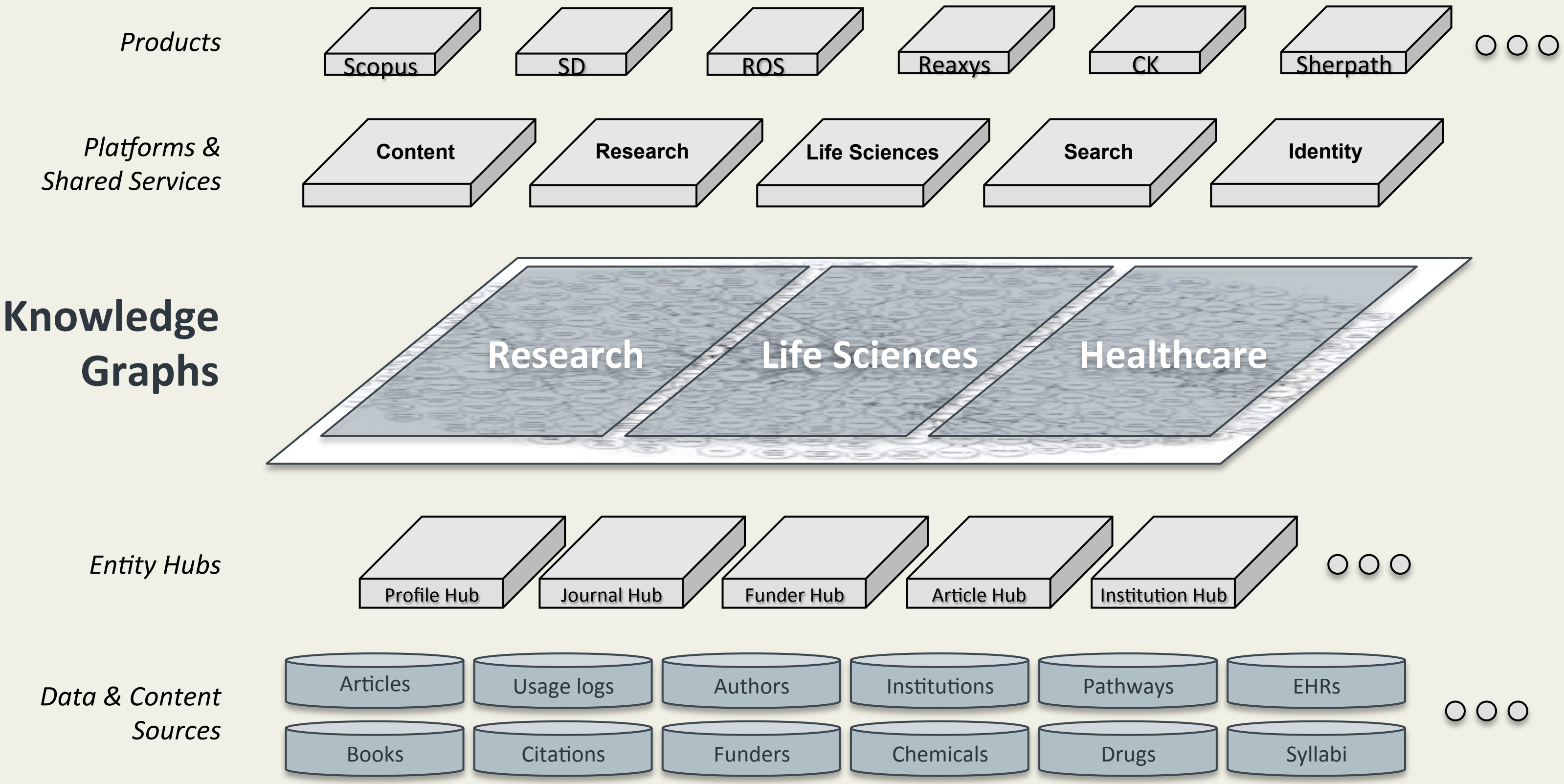


ELSEVIER: DEEP LEARNING FOR IMAGE ANNOTATION





ELSEVIER'S KNOWLEDGE PLATFORM



MOVING FORWARD: QUESTION ANSWERING AS AN AI-COMPLETE PROBLEM

Question answering (QA) is a complex natural language processing task which requires an understanding of the meaning of a text and the ability to reason over relevant facts. Most, if not all, tasks in natural language processing can be cast as a question answering problem …

Kumar, A., Ondruska, P., Iyyer, M., Bradbury, J., Gulrajani, I., Zhong, V., Paulus, R. & Socher, R. (2016). Ask Me Anything: Dynamic Memory Networks for Natural Language Processing. Proceedings of the 33rd International Conference on Machine Learning (ICML 2016).

THE BATTLE FOR THE KNOWLEDGE GRAPH

I really believe that the key battleground in any industry is that of its knowledge graph. Google has it for media/advertising, Netflix has it for filmed entertainment, Uber has it for inner city transportation, Facebook has it across social media as well as messaging and the multiples speak for themselves.

Tony Askew, Founder/Partner at REV (personal communication, September 29, 2016)

METADATA STANDARDS ARE AND WILL CONTINUE TO BE FOUNDATIONAL FOR KNOWLEDGE GRAPHS

- Semantic Web and Linked Data standards are the consensus approach for representing and sharing knowledge graphs over the Web
- There is a single thread from the establishment of the Dublin Core through Open Linked Data to the emergence of knowledge graphs
    - Linked data principles
    - Standard vocabularies, taxonomies and ontologies

BUT WHAT DO METADATA STANDARDS DO FOR MACHINE READING?

- Machines' proficiency in constructing knowledge graphs from text, audio, images and video will depend on our ability to train them effectively to read information from the Web
- How machines read the Web today
    - Crawling and indexing Web resources, possibly semantically tagged (e.g. using schema.org)
    - Find-and-follow crawling of open linked data resources for ontology and data sharing and reuse
    - Programmatic access to APIs mediated through HTTP/S and other Internet protocols

THE SEMANTIC WEB WAS INTENDED FOR MACHINE READING

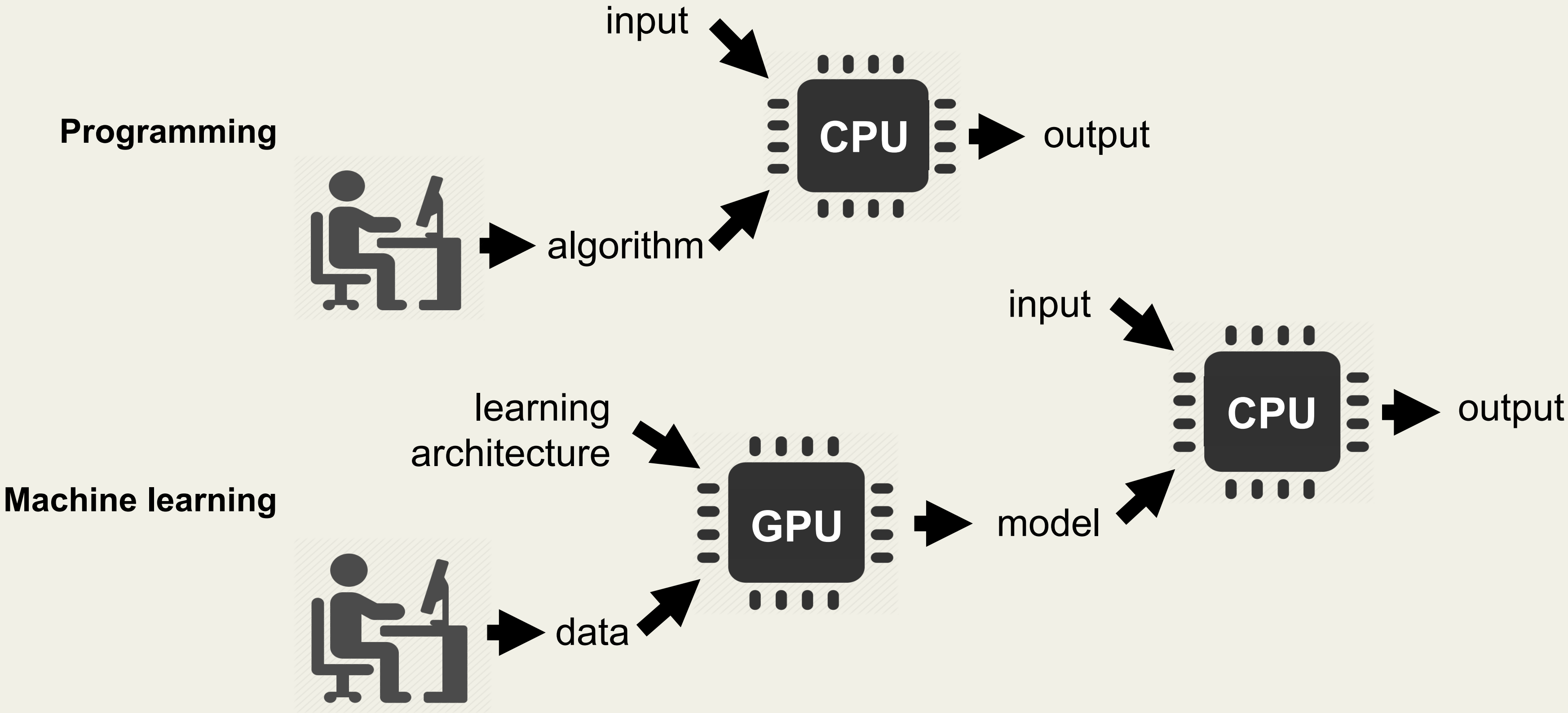
… that's the real idea behind the Semantic Web: letting software use the vast collective genius embedded in its published pages.

Swartz, A. (2013). Aaron Swartz's A programmable Web: An unfinished work. San Rafael, Calif.: Morgan & Claypool Publishers.

BUT THE SEMANTIC WEB IS BUILT FOR PEOPLE, NOT MACHINES

- The Semantic Web is largely a logicist take on the way knowledge is to be represented
- The latest advances in machine intelligence are based on a connectionist approach to knowledge representation
- There is a gap between how knowledge is represented in the Semantic Web and what deep learning is exploiting to such good effect
- The Semantic Web is silent about how machines can become better readers, and hence better partners in the second machine age
- How will we evolve metadata standards to better accommodate machines?

MACHINE READING IS ENABLED BY MACHINE LEARNING



MACHINES SEE THINGS DIFFERENTLY THAN PEOPLE

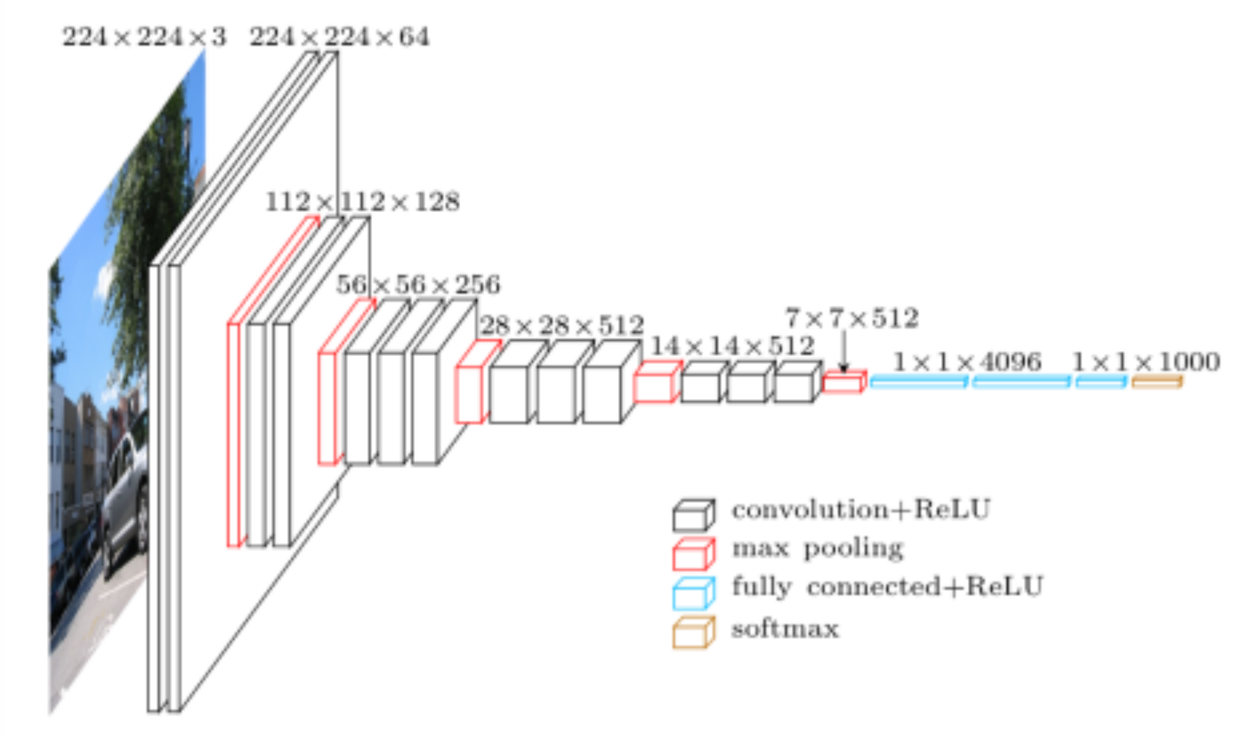
(a) hex dump of picture of a lion

(b) same lion in human-readable format

Figure 1: The hex dump represented on the left has more information contents than the image on the right. Only one of them can be processed by the human brain in time to save their lives. Computational convenience matters. Not just entropy.

From: Alain, G. and Bengio, Y. (2016). Understanding intermediate layers using linear classifier probes. arXiv:1610.01644v1.

MACHINES LEARN THINGS DIFFERENTLY THAN PEOPLE



MACHINE LEARNING DATASETS AND MODELS ARE BECOMING PART OF THE WEB

- Machines need lots and lots of data to learn how to read
- Datasets with ad-hoc formats are being made openly available
    - Open Images "~9 million URLs to images that have been annotated with labels spanning over 6000 categories" (The Open Images Dataset. (n.d.). Retrieved September 29, 2016, from https://github.com/openimages/dataset.)
    - YouTube-8M : "8 million YouTube video URLs (representing over 500,000 hours of video), along with video-level labels from a diverse set of 4800 Knowledge Graph entities" (Vijayanarasimhan S. and Natsev, P. (2016). Announcing YouTube-8M: A Large and Diverse Labeled Video Dataset for Video Understanding Research. Retrieved September 29, 2016, https://research.googleblog.com/2016/09/announcing-youtube-8m-large-and-diverse.html.)
    - Stanford Natural Language Inference: "570k human-written English sentence pairs manually labeled for balanced classification with the labels *entailment*, *contradiction*, and *neutral*, supporting the task of natural language inference" (The Stanford Natural Language Inference (SNLI) Corpus. (n.d.). Retrieved September 29, 2016, from http://nlp.stanford.edu/projects/snli/.)
- Standard architectures for machine (deep) learning are being released as open source
    - Dense neural networks for classification
    - Convolutional neural networks for image, audio and video recognition
    - Recurrent neural networks for sequence processing and generation
- Advances in the field are being published quickly and transferred to industrial application just as quickly

VOCABULARIES ARE SETS OF VECTOR EMBEDDINGS

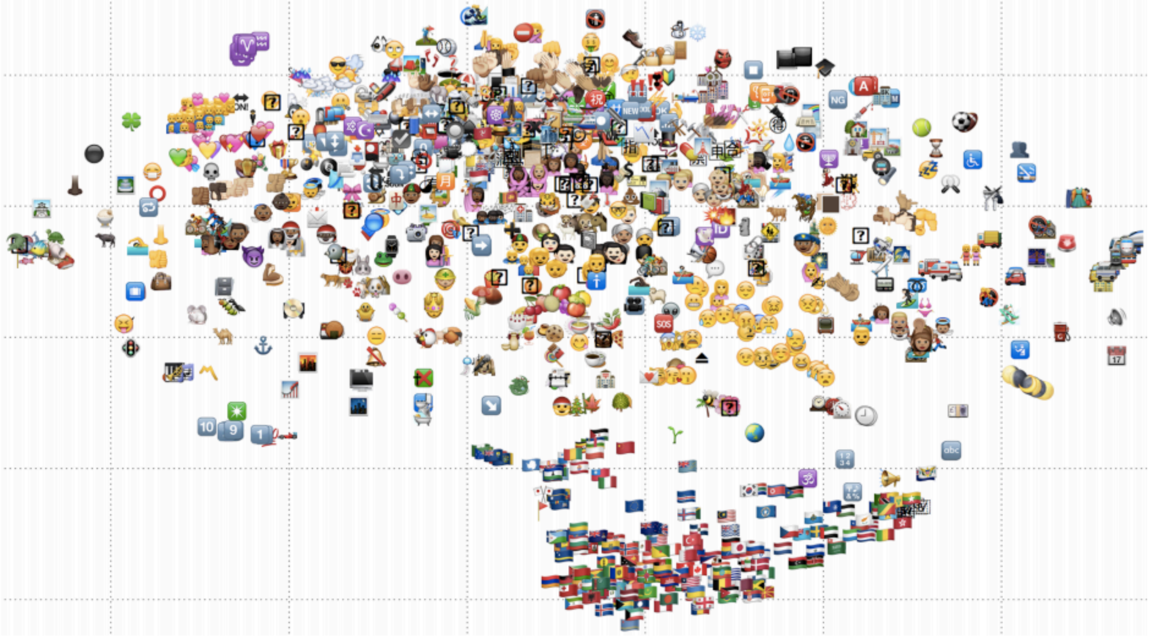


Figure 3: Emoji vector embeddings, projected down into a 2-dimensional space using the t-SNE technique. Note the clusters of similar emoji like flags (bottom), family emoji (top left), zodiac symbols (top left), animals (left), smileys (middle), etc.

From: Eisner, B., Rocktäschel, T., Augenstein, I., Bošnjak, M. and Riedel, S. (2016). Emoji2vec: learning emoji representations from their description. arXiv:1609.08359v1.

TRAINING DATASETS ARE GROWING IN VOLUME AND COVERAGE

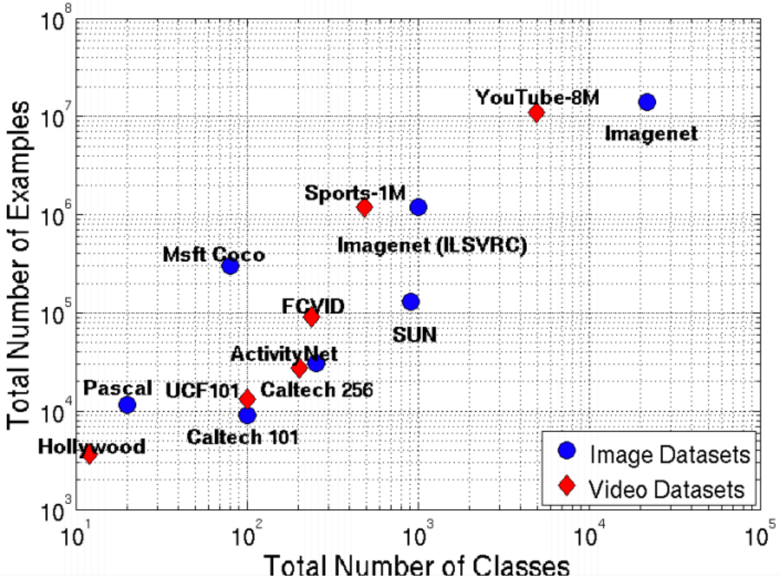
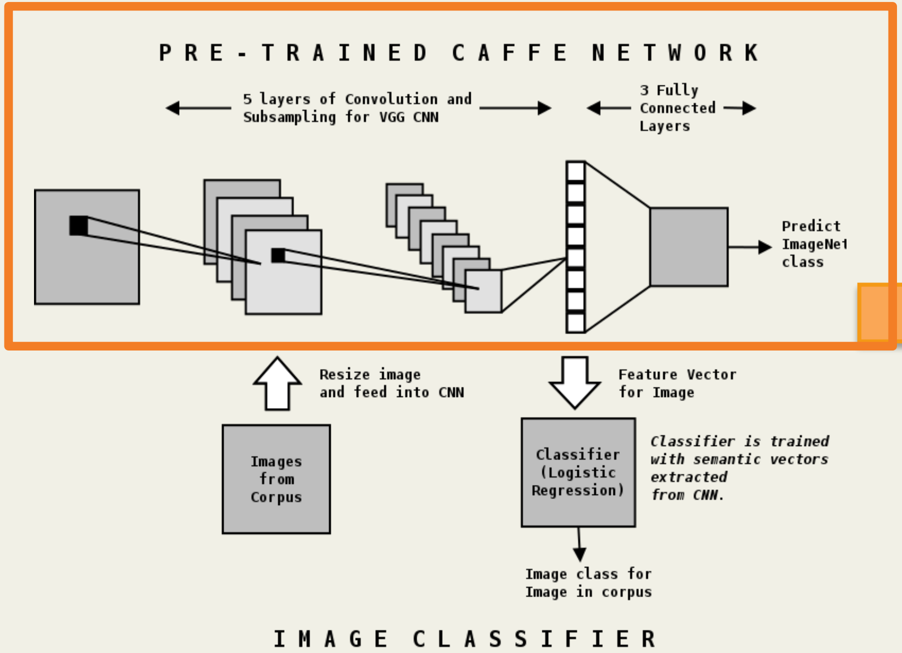
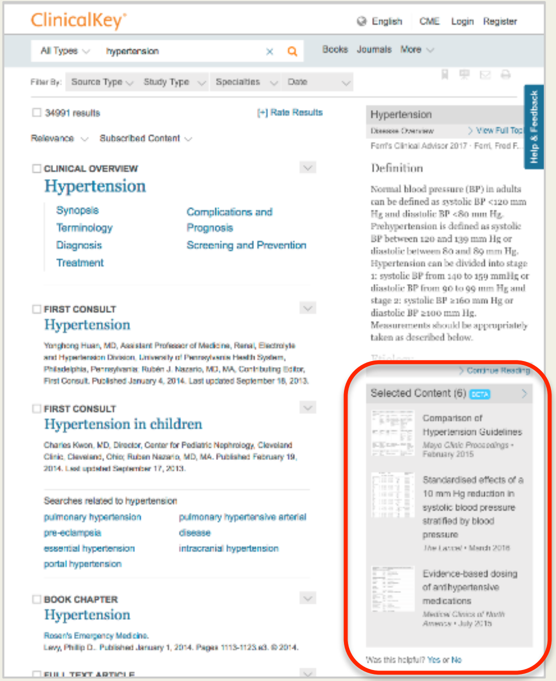


Figure 2: The progression of datasets for image and video understanding tasks. Large datasets have played a key role for advances in both areas.

From: Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., Toderici, G., Varadarajan, B. and Vijayanarasimhan, S. YouTube-8M: a large-scale video classification benchmark. arXiv:1609.08675.

MODELS ARE BECOMING REUSABLE DATA RESOURCES THEMSELVES

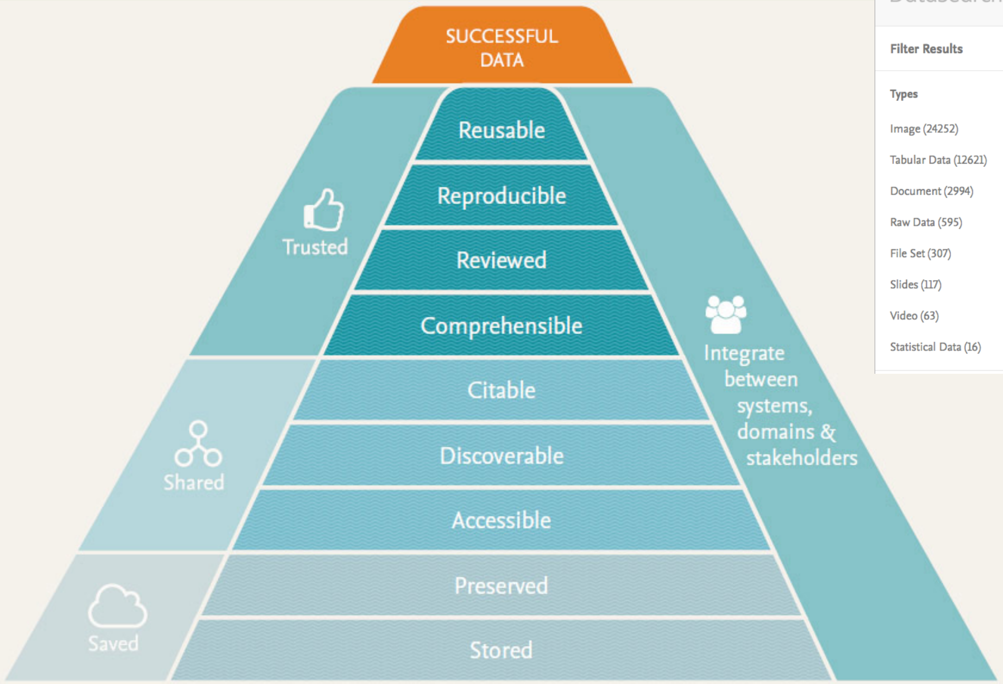


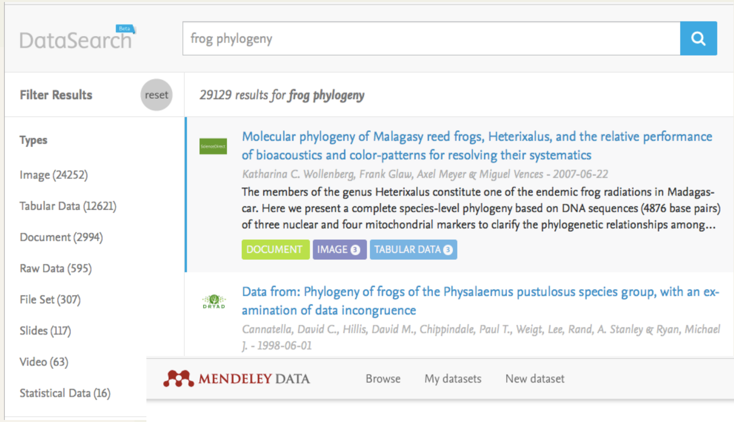


THE CHALLENGE FOR THE METADATA COMMUNITY: LINKED DATA THAT MACHINES CAN LEARN FROM

- We should be able to keep the core linked data principles that allow us to leverage Web architecture
    - URI as identifiers support (re)use of data in place (e.g. as in Google Open Images)
- We need to understand whether linked data formats need to change to support the needs of machine readers
    - The need for n-ary relations
    - The need for efficient indexing
- We need to investigate how vocabulary management can adapt to support the needs of machine reading
    - Vocabularies define what machines can recognize
    - Vocabularies need to accommodate lexical and vector representations together

WORK AT ELSEVIER: RESEARCH DATA MANAGEMENT





THE OPPORTUNITY FOR LIBRARIANS AND PUBLISHERS

As machines become increasingly capable of general-purpose language understanding, the burden of effort in building machine intelligences will shift from software engineering to the acquisition, organization and curation of training content and data.

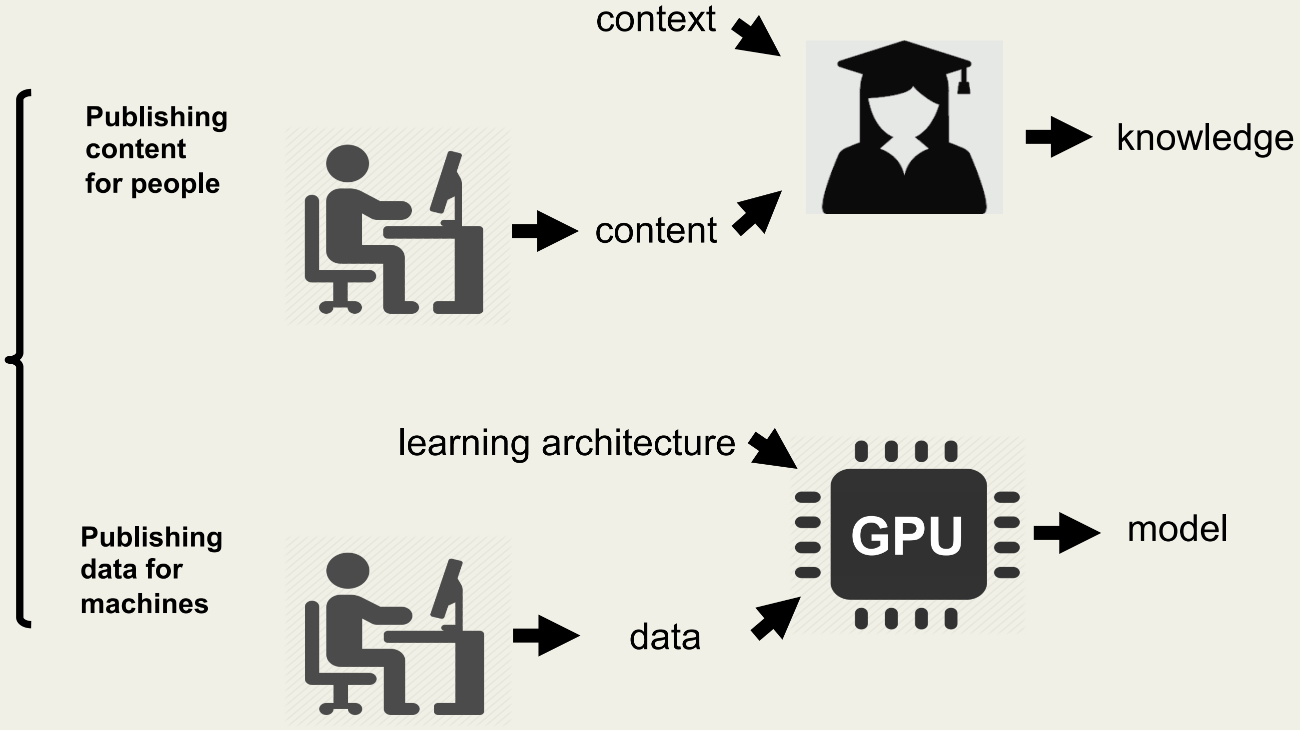
SAVE THE TIME OF THE MACHINE READER

Perhaps this law is not so self-evident as the others. None the less, it has been responsible for many reforms in library administration and has a great potentiality for effecting many more reforms in the future.

Ranganathan, S.R. (1931). The five laws of library science. Madras: The Madras Library Association.

SCHOLARLY PUBLISHING IN THE SECOND MACHINE AGE

ELSEVIER



IN SUMMARY: THE ROLE OF METADATA IN THE SECOND MACHINE AGE

- We need to save the time of the (human) reader by providing answers, not content
- This requires a shift from document-centric to entity-centric management of knowledge
- Knowledge graphs are ”vehicles of communication” between people and machines and the battleground on which dominance in markets will be established
- We can use machine reading to build knowledge graphs
- The Semantic Web falls short in helping machines learn how to read
- Metadata standards to support machine (deep) learning should receive attention from the metadata community
- Our goal: making information resources on the Web discoverable and comprehensible for both people and machines

THANK YOU

Bradley P. Allen

Chief Architect, Elsevier

@bradleypallen

IMAGE ATTRIBUTIONS

- p. 2: https://twitter.com/stuartweibel
- p. 3: https://vimeo.com/66143280
- p. 5: http://www.prospectmagazine.co.uk/other/erik-brynjolfsson-andrew-mcafee
- p. 6: https://arxiv.org/pdf/1411.4555v2.pdf
- p. 6: http://media.wix.com/ugd/142eb4_7581cfcf090e4e31a52599315f77c648.pdf
- p. 6: http://www.nature.com/nature/journal/v529/n7587/full/nature16961.html
- p. 7: https://www.wired.com/2013/05/neuro-artificial-intelligence/
- p. 11: https://twitter.com/karencoyle/status/784810709641158656
- p. 22: http://www.pcworld.com/article/3051631/salesforce-buys-ai-specialist-metamind-to-avoid-being-flanked.html
- p. 23: https://twitter.com/tonyaskew
- p. 26: http://quotesgram.com/aaron-swartz-quotes/
- p. 29: https://arxiv.org/pdf/1610.01644.pdf
- p. 30: https://blog.heuritech.com/2016/02/29/a-brief-report-of-the-heuritech-deep-learning-meetup-5/
- p. 32: https://arxiv.org/pdf/1609.08359v1.pdf
- p. 33: https://arxiv.org/pdf/1609.08675v1.pdf
- p. 38: https://plus.google.com/photos/112452283652920158570/album/6059593062673732337/6059593180461846098
- Save corporate logos on p. 8, all other images © 2016 Elsevier, Inc.