Ostriches, Minotaurs, Ghosts and Fossils in the Brave New Metadata World: Categorization & Linked Data
Online Northwest, March 31, 2017
Kelley McGrath

Linked data is the hot topic in the library metadata world. I'm going to talk a little bit about what happens when linked data meets the real world.

Semantic Web

Combination of
- web technology (linking and identifying)
- artificial intelligence

Semantic in semantic web
not = meaning
= computable axiom (If A=B and B=C then A=C)

I couldn't find a definitive take on the relationship between the semantic web and linked data. Opinions on this topic differ somewhat. A widely-held view is that the Semantic Web is made up of Linked Data; i.e. the Semantic Web is the whole, while Linked Data is the parts. Tim Berners-Lee, the person credited with coining the terms Semantic Web and Linked Data has described Linked Data as "the Semantic Web done right.". Often the two terms are used interchangeably. Linked data seems to have grown out the semantic web, which grew out of web technology and a certain type of artificial intelligence. One of the basic ideas linked data took from this strand of artificial intelligence is that if you feed the system a lot of facts in the right way, the system can use logic to uncover new facts that it wasn't explicitly told.

Inference to uncover new facts

Flipper is a dolphin
Dolphins are mammals

→ Flipper is a mammal

This potential to reveal new and useful information thru inference is one of the great promises of linked data. Unfortunately, I wasn't able to find any nontrivial examples, but presumably if you had a large enough pool of data, you could find some interesting things.

Inference to uncover new facts

Clint Eastwood directed *Million Dollar Baby*
Million Dollar Baby won the Academy Award for Best Picture

→ Clint Eastwood directed a film that won the Academy Award for Best Picture

Here is a made-up example (with a few implicit steps), still not very exciting, that you could imagine occurring in library data.

Inference to uncover new facts

Library data is often free-text with minimal mark-up

245 **Million Dollar Baby**

586 Academy Award, 2005: Best achievement in directing, **Best motion picture of the year**, Best performance by an actor...

One of the biggest barriers to creating a system capable of making these kinds of inferences is that library data is often free text. Fields may not even be clearly labeled as to what type of content they contain, such as generic general notes that are tagged as 500 in MARC.

Inference to uncover new facts

Structured data with identifiers

Film identifier	Film title	Year
no2005028500	Million dollar baby (Motion picture : 2004)	2004

Award identifier	Award name
Q102427	Academy Award for Best Picture

Award identifier	Film identifier	Year
Q102427	no2005028500	2005

Structured data just means data that is labeled and formed consistently according to rules. The data on this slide is in a form that you might see in a relational database. If library data looked like this, library catalogs could answer questions like "What film won Best Picture in 2005?" and "What films does the library have that won Best Picture?" In theory, you could also answer the question, "How many Academy Awards did Million Dollar Baby win?"

Inference to uncover new facts

How much do you want to record?

Best Picture
Best Director
Best Actor in a Leading Role
Best Actor in a Supporting Role
Best Actress in a Leading Role
Best Actress in a Supporting Role
Best Animated Feature
Best Animated Short Film
Best Cinematography
Best Costume Design
Best Documentary Feature
Best Documentary Short Subject
Best Film Editing
Best Foreign Language Film
Best Live Action Short Film
Best Makeup and Hairstyling
Best Original Score
Best Original Song
Best Production Design
Best Sound Editing
Best Sound Mixing
Best Visual Effects
Best Writing (Adapted Screenplay)
Best Writing (Original Screenplay)

However, the question "How many Academy Awards did Million Dollar Baby win?" runs up against practical constraints. Library catalogers have often only recorded awards that are considered major. In this age of economic austerity, do we have the time to spend to record every award? Is this a cost-effective use of catalogers' time?

Inference to uncover new facts

Linked data

http://id.loc.gov/authorities/names/no2005028500	http://rdaregistry.info/Elements/w/titleOfWork.en	Million dollar baby (Motion picture : 2004)
http://id.loc.gov/authorities/names/no2005028500	http://rdaregistry.info/Elements/w/dateOfWork.en	2004
http://id.loc.gov/authorities/names/no2005028500	https://www.w3.org/2002/07/owl#sameas	https://www.wikidata.org/wiki/Q10242
https://www.wikidata.org/wiki/Q10242	https://www.w3.org/2000/01/rdf-schema#label	Academy Award for Best Picture
https://www.wikidata.org/wiki/Q184255	https://www.w3.org/2000/01/rdf-schema#label	Million Dollar Baby
https://www.wikidata.org/wiki/Q184255	https://www.wikidata.org/wiki/Property:P166	https://www.wikidata.org/wiki/Q10242

One of the big advantages of linked data is that libraries don't necessarily have to record everything that our users might be interested in. We can connect our data with data provided by other organizations, especially sources that we trust. In this example, the top three rows are hypothetical linked data statements (triples) about http://id.loc.gov/authorities/names/no2005028500, which is the Library of Congress' URI for the film Million Dollar Baby, that a library might make. They just say that the work title is "Million dollar baby (Motion picture : 2004)," the date of the work is 2004 and that the entity described by this URL (Million

Dollar Baby) is the same as the entity described by the Wikidata URI https://www.wikidata.org/wiki/Q10242. Wikidata has linked data statements that list all of the Academy Awards and all of the Academy Award winners. Since Wikidata is likely to be a reliable source on this topic, just by connecting library URIs to Wikidata URIs for films, we can use all the Wikidata information about Academy Awards without doing any of the work to record or maintain this data. Wikidata probably has other data about movies that would also be beneficial to incorporate into the library discovery process.

Classical Theory of Categories

Does it have feathers?
Can it fly?

- clearly-defined
- mutually exclusive
- collectively exhaustive

Defined by necessary and jointly sufficient conditions

Linking data requires structured data that is formed consistently, according to rules. Linked data classes are close to Aristotle's classical view of categories. Classical categories are defined by properties and have hard and fast boundaries. To get raw information into a form that is useful for linked data, it has to be put into clearly-defined boxes. As anyone who has spent much time trying to create standardized metadata knows, this can be tricky.

The Unknowable

London After Midnight, one of the most sought after lost films, had its last surviving reel copy destroyed in a devastating studio vault fire and only exists in the form of film stills, as shown in an original 1927 theatrical release poster.

https://en.wikipedia.org/wiki/Lost_film

The first hurdle is that you have to actually have data to work with. There will be gaps in your linked data graph where data is unknowable or currently unknown.

The Project

Directed by Clint Eastwood

"Directed by"
"Clint Eastwood"

Directed by → direction

Top 11 of 76 categories

production	14.4%
acting	13.6%
direction	10.8%
cinematography	10.1%
writing	9.2%
editing	8.4%
music	7.0%
presents	3.7%
composers	2.4%
other/unsure	**2.0%**
a film by /a ... film	2.0%
	83.5%

Thank you to Alden Lee, Jaroslaw Szurek and many others for help with categorization

Even if you have data, you can still have many problems. Most of the examples in this presentation come from a project to analyze statements about responsibility in a large number of MARC records for film and video. We took all the text from MARC 245$c (statement of responsibility), 508 (creation/production credits note) and 511 (participant or performer note). We split this text into separate statements based on semi-colons, which are the standard ISBD punctuation used to demarcate grammatically complete phrases. We recruited volunteers who used a Web form to identify and separate the functions and names in the statements. For example, the statement "directed by Clint Eastwood" was divided into "directed by" (function) and "Clint Eastwood" (name). We then mapped all the functions to categories. We started with categories based on RDA relationship designators and supplemented those by what seemed to be sensible groupings. Not surprisingly, we ended up with an 80-20 type of distribution, where most of the function statements fall into a few categories. Notice that even though there's a long tail that goes way off this page, we still ended up with 2% of the roles not falling into easily-identifiable clusters.

The Vague

responsible, Lê Mỹ Phương

Once you have information to work with, next you have to put it into a consistent, rule-based form. Mapping functions or roles to categories or RDA relationship designators is similar to subject analysis; first you have to figure out what you have and then you have to match it to controlled terms or categories. Sometimes, this breaks down in the first part of the process, as in the category that I call the vague. This is my favorite example. It's the quintessential statement of responsibility, but who knows what this person did.

The Vague

responsible, Lê Mỹ Phương

action, Shyam Kaushal
associate, Ved M Rao
Series proposed by Benoit Peeters
supervisor, Dr. Nurdin Perdana
Team works, Kartawijaya ... [et al.]

There are many more of these.

The Ambiguous

By Shakespeare [author of play]
By W.A. Mozart [composer of opera]
By John Cleese and Connie Booth

Bass [guitar or vocalist?]
Music [performer or composer?]
Songs [writer or singer?]

Another challenge is what I think of as the ambiguous. The types of examples on this slide are usually resolvable with more context or external research, but you can't conclude anything from the bare statement.

The Ambiguous

Musical adaptation and direction, Penella [musical direction or direction?]

associate director and editor, Alexander Hammid [associate editor or editor?]

author and singer of songs, Vladimir Vysotski [author or author of songs?]

graphics and video editor, Michael Seibert [graphics or graphics editor?]

Some statements are grammatically ambiguous. For example, does the initial adjective apply to the whole phrase "musical adaptation and direction" or just to adaptation? Often it is clear what the intended meaning is, although some contextual knowledge may be required. In this case, the first statement is almost certainly referring to "musical direction" since it is unlikely that a major role like direction would appear as the second function in a statement like this. However, in many cases the meaning is not so obvious.

The Incoherent

Winstar Cinema release, Shochiku Co., Ltd. presents a 3H Film Productions of a film by Hou Hsiao-Hsien

an Old Photo film presents

A Sony Pictures Classics release of Vans "Off the Wall" Productions presents an ADP Productions

Villealfa Filmproductions esittää ; Aki Kaurismäen elokuva

The incoherent can also be an obstacle. In our project to analyze existing statements of responsibility, we encountered statements that cannot be parsed grammatically. Some of these have obvious interpretations and are clearly caused by transcription errors, but given all the other odd things publishers do, it would be surprising if publishers don't also create these kinds of statements.

The Unique

a film by Wes Craven
This is not a film by Khvan
a média-stylo by Renée Gosson
the immortal classic by S. Anski

a Park Chan Wook film
a Spike Lee joint
a Danny Shechter dissection

相田和弘**観察映画**

Sometimes credits use unique terms. Some of these special snowflakes are clever spins on standard phrases, such as these variations on a film by so-and-so or a so-and-so film, so they're not hard to map. You do have to be careful not to over-extrapolate. The Anski example looks similar to the others, but is referring to the work that the film was based on.

The Unique

a film by Wes Craven
This is not a film by Khvan
a média-stylo by Renée Gosson
the immortal classic by S. Anski

a Park Chan Wook film
a Spike Lee joint
a Danny Shechter dissection

相田和弘**観察映画**

кино-разведчик Дзига Вертов*

*Thanks to Tom Dousa for researching this term.

Some terms really are in a category of one. The term "kino-razvedchik" was very specifically chosen by the experimental filmmaker D'ziga Vertov. It literally means "film-scout" or "film-spy" and has associations or overtones that would be lost if it were mapped to a more general term.

The Unnamed

Showrunner

"responsible for all creative aspects of the show"

"makes all important decisions about the series' scripts, tone, attitude, look and direction… oversees casting, production design and budget… chooses directors and guest stars…"

Sometimes there is information that isn't explicitly named. For example, the person who provides the overall vision for a television drama series is called the showrunner.

The Unnamed

Showrunner

Usually credited as

- Executive producer
- Created by

However, no one is ever listed in the credits as a showrunner. This person is usually credited as an executive producer or creator. You need external knowledge to identify people performing this role.

The Unnamed

Cast:

Sicilian fishermen
Shaima and Hossein and families
the children of Biviers
a group of African students of the University of Rome
inmates (of the State Prison in Ellis, Texas)

Those are all problems in figuring out what something is. However, even if you know what something is, you still have to decide how to name it. There are people and groups of people identified in credits that haven't been given proper names and possibly can't even be given proper names. How do you record these entities in linked data?

The Untranslatable

realización = production? direction?
Dirección, realización y producción

总导演 = director-in-chief? executive director?

企画 = planning? development? production?

mise en scène = staging? direction?

Sometimes it's hard to map concepts from one language or culture to another one. Even though RDA is an international standard, the relationship designators are largely based on Anglo-American usage. In the first example, most of the people who helped with the project to annotate statements of responsibility from video records translated realización as production or direction. However, what then do we do with the statement "dirección, realización y producción," which already includes references to clear analogs for direction and production. What is realización communicating that isn't already conveyed by direction and production?

The Long Long Tail

- Dream sequences based on designs by Salvador Dali.
- garden designer
- synchronization director
- tiger trainer
- spider web spinner
- floor drop painted by ...
    - body art & locations
    - movement coordinator

As I mentioned, there is a long, long tail. Who knew that a movie might need a spider web spinner? It doesn't seem practical to include all these specific roles in a controlled vocabulary.

Not to Even Start on Instruments

From accordion to whistles
From clarinet to clavinet





From around the world:
bodhrán, conga,
kenkeni, mridangam,
tabla, taiko and so on

Not to even mention the even longer list of musical instruments. A key question is how specific to make your categories. In trying to map narrower categories, there is a trade-off between specificity and consistency or accuracy. The finer distinctions that you try to draw, the more challenging it becomes.

IMDb-Style Solution?

Does the Internet Movie Database (IMDb) provide a potential model?

- Limited list of broad role categories
- Very specific, free-text roles

Perhaps one way to resolve this tension is to use the approach taken by the Internet Movie Database. IMDb uses a limited number of controlled categories for cast and crew roles. It supplements these with free text displays giving a more specific role or qualifying the role in some way.

IMDb-Style Solution?

Romeo and Juliet (1968)

Controlled vocabulary for Broad Terms:

Writing Credits

William Shakespeare	...	(play)
Franco Zeffirelli	...	(screenplay)

Grab Bag Category:

Other crew [=contributor?]

Gabriella Bernardi	...	production secretary
Alberto Testa	...	choreographer

In this example, “writing credits” and ”other crew” are controlled vocabulary labels. These are the categories that could be used for limiting searches or making inferences. Where relevant, names are qualified by more specific information, usually taken from the credits (if the role information is taken from elsewhere, it is labeled “uncredited”). The library world might choose to divide roles in a different way, such as separating credits related to the screenplay from information related to the work on which it’s based or by using a controlled term for choreographer, but the general principle would work. In fact this is similar to what we do now under RDA with free text, often transcribed, statements of responsibility and standardized forms of names associated with relationship designators from a controlled vocabulary. However, MARC data fails to encode the connection between instances of these two types of data so we cannot create displays that are as clear as IMDb’s.

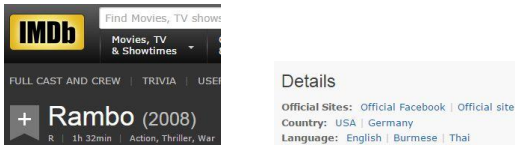
The Variations

Assistant producer
Associate producer
Chief producer
Delegate producer
Deputy producer
Executive producer
DVD producer
Segment producer
Producer in Japan
Series producer
Series senior producer
Producer for BBC
Producer of U.S. release
Executive producer of English version
Line producer

Many of the broad categories that we sorted roles into hide a great deal of variation that may or may not be important to rigorously differentiate.

A case for narrower categories

Most popular Thai language films



- Primary or secondary
- Spoken, sung, signed or written (subtitles, captions or intertitles)

Sometimes defining narrower categories might be helpful or necessary. Linked data is good at modeling hierarchical categories and defining more specific subcategories for a category (the harder problem is when two sets of categories split up the universe in different ways, as in the translation problems shown on a previous slide).

For example, IMDb has a page of languages that will take you to lists of movies in that language (http://www.imdb.com/language/). In the past, looking for popular movies in less common languages brought up major Hollywood pictures with some dialog in that language. For example, IMDb might have presented Rambo as a popular Thai-language movie, even though that's unlikely to be what users are looking for. They seem to have solved that problem, but it is still a potential problem for library data. In MARC, it is common practice to only record the major language(s) present on a resource, but there is nothing that prevents catalogers from recording more minor languages and there is no way to distinguish the main language(s) from those that occur only briefly.

A case for narrower categories

Specific types of dates

Ivan the Terrible, Part II (1958)

- Copyright
- Production
- Recording / Filming
- Release / Distribution / Broadcast
- Unknown

Dates can also be misleading in some cases. IMDb uses the date of first public screening to identify a movie (http://www.imdb.com/help/show_leaf?titleformat). The movie Ivan the Terrible, Part II was completed in 1947 and the filmmaker died in 1948. However, Stalin banned the film and it was not shown publicly until 1958. Even though the film is cited by IMDb using the 1958 date, the content of the film dates to 1947. (The recommendations for narrower categories under language and date on these slides are discussed in more detail in the report of OLAC's Moving Image Work-Level Records Task Force at http://olacinc.org/sites/capc_files/MIW_3a.pdf)

It depends...

television producer

producer of *made-for-TV movie*

producer of *TV drama/comedy series*

- writer
- line producer
- no particular reason (vanity credit)

producer of *TV news*

- have editorial and supervisory roles
- select and order stories
- may write or edit stories

There are some other problems with naming things. Sometimes a role can have the same name, but multiple meanings. One of my pet peeves about RDA is the way the it splits up film director and television director, as well as film producer and television producer. It is useful to provide information about the original method of distribution for a moving image, but it isn't a particularly good way to split up roles. For example, the work of a producer of a made-for-TV movie is probably more similar to a feature film producer than to the producer of an episode of a TV drama or comedy series. In fact, the person credited as a producer of an episode of a television program is often a writer or may be performing the narrower role of line producer. Producer is also sometimes used as a vanity credit to reward someone connected to the show. People credited as producers on TV news programs are performing yet another set of tasks. Thus television producer lumps together some disparate roles while simultaneously splitting apart the similar roles of producing a made-for-TV movie and a feature film.

The Fickle

- technical director (some silent films)
- art director
- production designer

Words can be fickle; their meaning may change over time. The term production designer was first used for the film Gone with the Wind. The director wanted to give a shout-out to the production designer. Others imitated this and over time, production designer became the standard term in Hollywood credits. In early film, this person was briefly known as the technical director, a term that has a completely different meaning now. After that, this person was called the art director. As production designer came into come use the meaning of art director morphed into a narrower role under the production designer. This change happened gradually and there was a period of time when art director was used with both meanings. It's also not clear how much Hollywood conventions can be applied to non-Hollywood films.

The Infinite Shades of Gray

Adaptation

"a continuum from works that are clearly adapted from Robinson Crusoe, to works that only belong to the genre known as "Robinsonade" or "desert island fiction," and in which no more than the Robinson motif in a very general sense remains"

Wallheim, Henrik. "From Complex Reality to Formal Description: Bibliographic Relationships and Problems of Operationalization in RDA." *Cataloging & Classification Quarterly* 54.7 (2016).

Real life is often not black and white. Henrik Wallheim wrote an interesting article for Cataloging and Classification Quarterly about problems for operationalizing some relationships in RDA. He talks about work-to-work relationships and uses the example of possible adaptations of Robinson Crusoe, which range from clear-cut cases to works that are only loosely-related. Where and how should you draw the line between an adaptation and a non-adaptation? The optimal place to draw this line depends partially on your use case.

The Infinite Shades of Blue

If you start with a bucket of white paint and add drops of blue paint one-by-one, at some point, the paint become blue. Can you tell which drop it is that pushes the paint over the line from white to blue? Where is the dividing line?

Classical Theory of Categories

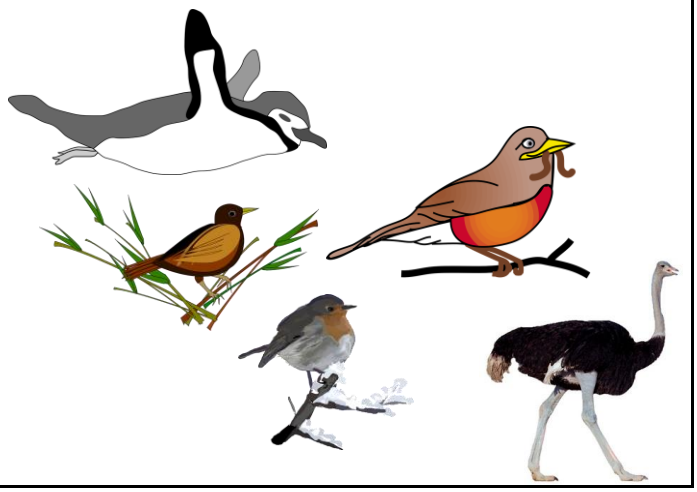
Does it have feathers?
Can it fly?

- clearly-defined
- mutually exclusive
- collectively exhaustive

Defined by necessary and jointly sufficient conditions

There are other ways to look at categories than the way Aristotle does.

Prototype theory of categories



For example, prototype theory says that categories don't have hard boundaries. Things aren't just in or out, but rather have degrees of belonging to a category based on similarity to an ideal exemplar. The more something is like the stereotypical bird, the more it belongs to the bird category.

Machine Learning

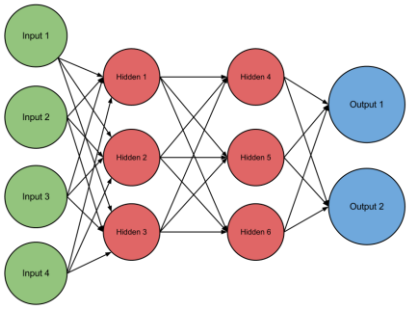
- Google Translate improvements
- Defeat of highly-ranked go players (10^{360} possible positions vs. 10^{123} for chess)

- Identification of cats in pictures

There are also other approaches to artificial intelligence. Machine learning is one that has had some significant successes recently. Machine learning has powered major improvements in automated translation and image identification. Go is a strategy game like chess where all the information is known by both players, but there are many, many more possible moves in go than in chess. Due to the complexity of go, experts had expected that it would be years before a computer would be good enough to beat a top-level human player. However, in March 2016 Google's AlphaGo defeated one of the world's best go players.

Machine Learning

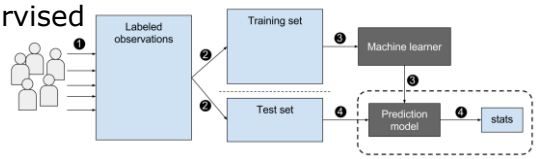
Give raw data to an algorithm and let the algorithm make its own rules



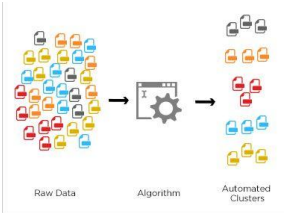
Linked data needs structured data as input. Machine learning works with raw data, even free text. It takes this raw data and weights it in various ways to produce its output. The algorithm may adjust these weights based on feedback.

Machine Learning

- Supervised



- Semi-supervised
- Unsupervised



This is one way to categorize machine learning. In supervised learning you start with a set of human-created data and the corresponding desired output. You feed a subset of this raw data with its corresponding answers to an algorithm that iteratively works out how to get from the input to the desired output. You then test the process that the program has developed on the remaining prepared data. When you are satisfied with the quality of output, you can use the algorithm on novel data.

In unsupervised learning, the answers are not

specified in advance. Instead, an algorithm works with the raw data to identify patterns.

Semi-supervised learning is a combination of these approaches.

Machine Learning

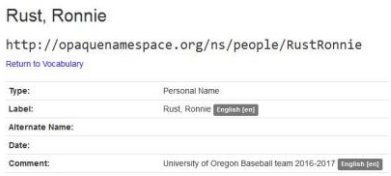
- Iterative, trial & error
- Statistical
- Fuzzy
- Black box

Machine learning is an iterative, statistics-based approach rather than being built on logic. With linked data, everything is transparent. You can see all the input and follow all the steps to see how a conclusion is reached. In comparison, machine learning is a black box and you have no idea how the algorithm reached its conclusions.

There are many ways to extract value from library data and we should beware of putting all our eggs in one basket.

Machine Learning

"directed by Clint Eastwood"


Rust, Ronnie

http://opaquenamespace.org/ns/people/RustRonnie

Return to Vocabulary

Type:	Personal Name
Label:	Rust, Ronnie English [en]
Alternate Name:	
Date:	
Comment:	University of Oregon Baseball team 2016-2017 English [en]

Current cataloging thought devalues transcription, but perhaps in the future this will be reconsidered. One function of transcription that I think is under-appreciated is its role in verification and reproducibility. Transcribed data serves the same kind of function as the requirement for citing sources in Wikipedia. It is also key in matching descriptions from multiple sources.

Perhaps we could get better results by feeding a machine the raw statements of responsibility and having the machine categorize and label them? Another idea that has been suggested to me is that perhaps this kind of artificial intelligence could be used to extract more value from comments or other free text associated with linked data.

Related Links

Operationalizing relationship designators
http://www.tandfonline.com/doi/abs/10.1080/01639374.2016.1200169?journalCode=wccq20

Machine learning
https://medium.com/@ageitgey/machine-learning-is-fun-80ea3ec3c471

https://www.nytimes.com/2016/12/14/magazine/the-great-ai-awakening.html

I hope I have given you some food for thought. Here are some references if you're interested.

http://www.tandfonline.com/doi/abs/10.1080/01639374.2016.1200169?journalCode=wccq20

https://medium.com/@ageitgey/machine-learning-is-fun-80ea3ec3c471

https://www.nytimes.com/2016/12/14/magazine/the-great-ai-awakening.html