I 202: Information Organization & Retrieval Fall 2025

Class 27: Intellectual Property; Course Wrap up

Today's Outline

Intellectual Property

Course Wrap Up

Course Evaluations

"Let's Go Crazy" #1

https://www.youtube.com/watch?v=N1KfJHFWlhQ

A Landmark Legal Battle Over a Toddler Dancing to Prince Song Looks to Be Ending

After 11 years, and several brain-twisting takes, including one from the Trump administration, the lawsuit brought by Stephanie Lenz against Universal Music is settling.

BY ERIQ GARDNER JUNE 6, 2018 9:14AM

ELECTRONIC FRONTIER FOUNDATION

This Is What It Sounds Like When Fair Use Cries

'Dancing baby' YouTube copyright case settled after 11 years

28 June 2018

Last week the US Supreme Court refused to hear an appeal from Universal, which owns the rights to Prince's music, over a previous ruling by the Ninth Circuit.

This led to both parties coming to their own agreement to settle the case.

Mrs Lenz, represented in court by US digital rights group the Electronic Frontier Foundation, said **in a statement** that Universal's processes for handling such cases were now "much better".

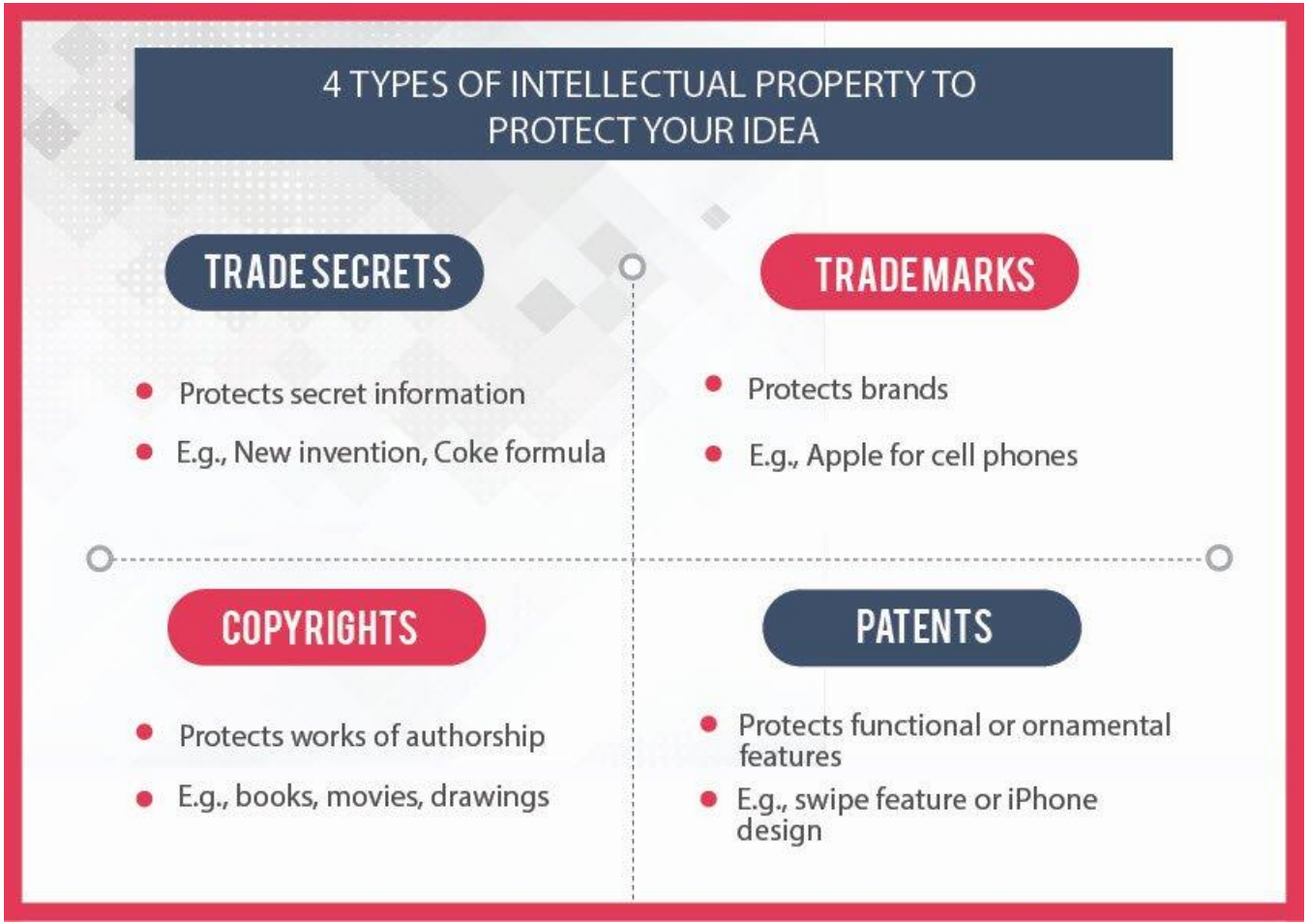
Universal's chief counsel, David Kokakis, said: "The Lenz case helped us to develop a fair and tempered process for evaluation of potential takedowns."

https://www.bbc.com/news/technology-44643419

WHAT IS INTELLECTUAL PROPERTY?

“A category of property that includes intangible creations of the human intellect”

wikipedia



https://ocpatentlawyer.com/four-types-intellectual-property-protect-idea/

Types of Intellectual Property

Trade marks

- NOKIA
- Product "208"
- Start-up tone

Patents & utility models

- Data-processing methods
- Operating system
- Operation of user interface

Design

- Form of overall phone
- Arrangement and shape of buttons
- Position and shape of screen

Trade secrets

- Some technical know-how kept "in-house" and not published

Copyright

- Software
- User manuals
- Ringtones
- Start-up tone
- Images

EUIPO

Patent basics

- > Patent process step-by-step
- > Search patents
- > Frequently asked questions (FAQs)
- > Procedures (MPEP) and guidance

Apply for a patent

- > Get started filing online
- > Register and easier filing resources
- > Check application status
- > Filing fees and payment
- > Forms

Application assistance

- > Contacts for application questions
- > Advanced application resolution
- > Petitions
- > Appeals and proceedings (PTAB)

Maintain your patent

- > Maintenance fees and payment
- > Calculate expiration date
- > Reinstate an expired patent
- > Ownership change and search

Helpful resources

- > CARES Act deadline waivers
- > International patent filing
- > Examiner interview requests
- > In-person help near you
- > Free assistance programs

Independent inventors

- > File a patent application on your own
- > Legal assistance and resources
- > Free resources in your state

Trademark basics

- > Do I need a U.S.-licensed attorney?
- > Trademark fee information
- > Timeline to process an application
- > Free resources in your state

Search trademarks

- > Develop a search strategy
- > How to use the Trademark Electronic Search System (TESS)

Apply for a trademark

- > How do I apply using the Trademark Electronic Application System (TEAS)?
- > Respond to an office action
- > Select goods & services in ID Manual

Protect trademarks

- > Scam awareness
- > Protect your trademark rights
- > Watch out for counterfeit goods

Maintain your registration

- > How do I renew my trademark?
- > What happens if my trademark registration is audited?

Laws and rules

- > Trademark Modernization Act
- > CARES Act FAQs
- > Rule making
- > Examination guides

What Can / Cannot be Protected by Copyright?

Let’s do an online poll:

pollev.com/i202

Who Owns a Recipe? A Plagiarism Claim Has Cookbook Authors Asking.

U.S. copyright law protects all kinds of creative material, but recipe creators are mostly powerless in an age and a business that are all about sharing.

Alan Richardson and Karen Tack, the authors of "Hello, Cupcake," saw their signature corn-on-the-cob cupcake on the cover of a women's magazine in 2011 — but the recipe didn't give them any credit. Timothy Mulcare for The New York Times

https://www.nytimes.com/2021/11/29/dining/recipe-theft-cookbook-plagiarism.html

Associated Press Settles Copyright Lawsuit Against Obama 'Hope' Artist

Street artist Shepard Fairey and The Associated Press are settling a copyright dispute over who owned the rights to the iconic Obama “Hope” poster, the two said in a joint statement Wednesday. The out-of-court settlement ends the closely watched, 2-year-old lawsuit without resolving the underlying legal issue: whether Fairey had a fair use right under [...]

https://www.wired.com/2011/01/hope-image-flap/

Fairey had long claimed he based his abstract graphic rendition on a photo of Obama seated next to actor George Clooney. But he later admitted he actually used a solo shot of Obama from the same event, and had destroyed and fabricated evidence to support his story.

Though both photos at issue were shot by the same AP photographer, the fact that the solo shot of Obama was the source was important because the more one transforms a photograph, the higher the chances that the resulting art constitutes a fair use of the original work.

https://www.wired.com/2011/01/hope-image-flap/

What Does Fair Use Mean?

- “Fair use … lets people use and adapt copyrighted works without getting the explicit permission of their owner”
- “Instead of setting out specific statutory exemptions to copyright, as many other countries do, U.S. law issues four broad factors which guide whether the permission-less use of a copyrighted work is fair.”
- “This means that fair use can evolve and change over time; it also means that the only real way to find out if something is “fair use” is to ask a federal court.”

Meyer, After 10 Years, Google Books Is Legal, *The Atlantic*, 2015

The Four Factors of Fair Use

Purpose and Character of Use

Was material transformed with new expression or meaning?
Was value added by creating new information, aesthetics, insights?

Nature of Copyrighted Work

Dissemination of facts benefits the public, so you have more leeway to copy from factual works than fiction. Fair use is narrower from unpublished work.

Just acknowledging source material is not enough to allow for copying for commercial use

Amount / Sustainability of Portion Taken

The less you take, the more likely copying can be considered fair use. But not if you copy the "heart" of the work.

Effect of Use Upon the Potential Market

Deprives copyright owner from income?
Undermines potential market for the work?

https://fairuse.stanford.edu/overview/fair-use/four-factors

Fair Use in Teaching

Single copying for teachers	
A single copy generally may be made of any of the following for teaching purposes:	A chapter from a book
	An article from a periodical or newspaper
	A short story, short essay or short poem, whether or not from a collective work
	A chart, graph, diagram, cartoon, or picture from a book, periodical, or newspaper

Some examples of activities that courts have regarded as fair use
Quotation of excerpts in a review for purposes of illustration, criticism or comment
Quotation of short passages in a scholarly or technical work, for illustration or clarification
Parody of the content of the work
A summary of an article, with brief quotations
Reproduction of a small part of a work by a teacher or student to illustrate a lesson

https://copyright.universityofcalifornia.edu/use/teaching.html#p=in_the_classroom

Transformative Use is Key

- Allows for search results snippets and thumbnails
  - An image to help people decide what web site to go to is different than being entertained or informed by it
- What about scanning library books (in cooperation with libraries) – both in and outside of copyright – and making searchable snippets available?

https://www.theatlantic.com/technology/archive/2015/10/fair-use-transformative-leval-google-books/411058/

Fast forward from 2007 to today…

How does copyright work on YouTube?

How To Use Copyrighted Music On YouTube

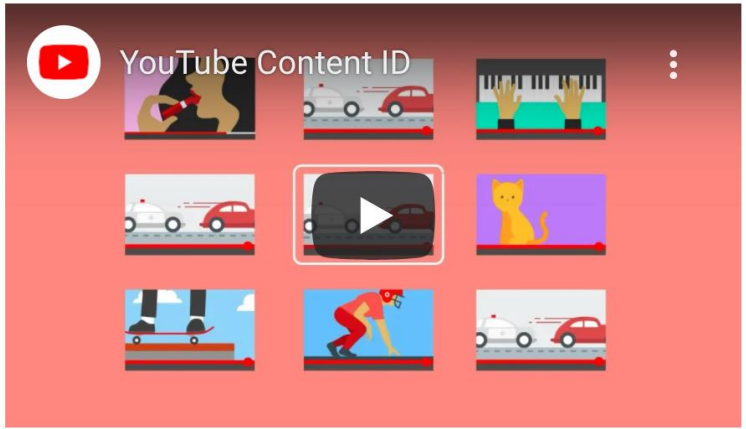
Updated: October 5, 2021 37 Comments

https://www.tunepocket.com/use-copyrighted-music-youtube/

How Content ID works

Copyright owners can use a system called Content ID to easily identify and manage their content on YouTube. Videos uploaded to YouTube are scanned against a database of files that have been submitted to us by content owners.

Copyright owners get to decide what happens when content in a video on YouTube matches a work they own. When a match is found, the video gets a Content ID claim.



YouTube Copyright Checks

Content ID claims

Unlike copyright removal requests, which are defined by law, Content ID is a tool created by YouTube. When Content ID finds a match, it applies a Content ID claim on the matching content.

What happens if my content gets a Content ID claim?

Depending on the copyright owner's Content ID settings, Content ID claims can:

- **Block** content from being viewed.
- **Monetize** content by running ads on it and sometimes sharing revenue with the uploader.
- **Track** the viewership statistics on the content.

Any of these actions can be geography-specific. For example, a video can be monetized in one country/region and blocked or tracked in a different country/region.

Keep in mind that when content is tracked or monetized, it **stays viewable on YouTube** with the active Content ID claim on it. Usually, copyright owners choose to track or monetize videos, not block them.

https://support.google.com/

YouTube Copyright Checks

Content ID claims in three-minute Shorts

Starting on October 15 2024, all new vertical videos that are 1-3 minutes in length will be categorized as Shorts on YouTube. Shorts longer than one minute that have an active Content ID claim, regardless of the policy, will be blocked on YouTube.

If a claim is found when you upload a 1-3-minute Short, you'll get a notification. You may remove claimed content from your videos. If you believe the claim was made in error, you can file a dispute. Once the claim is resolved, your Short will be viewable. Learn more about 1- to 3-minute Shorts.

https://support.google.com/youtube/

Beware uses of Music, Images, Video

- Warning: remember the statement earlier that only a court can really decide!
- That said, it seems that you should be very careful with uses of:
  - Music, including sampling: seems to lose in court if you do not obtain permission
  - Images (including tracing); see the Fairey case above; you need permission to use images that are copyrighted in most cases
- Even if you own a piece of artwork, you can't necessarily legally sell images of it on the web – the artist might own the rights!

The good news: lots of progress has been made on more reasonable approaches to online IP

LET'S TEST OUR UNDERSTANDING

Let's do an online poll:

pollev.com/i202

Web Search and Fair Use

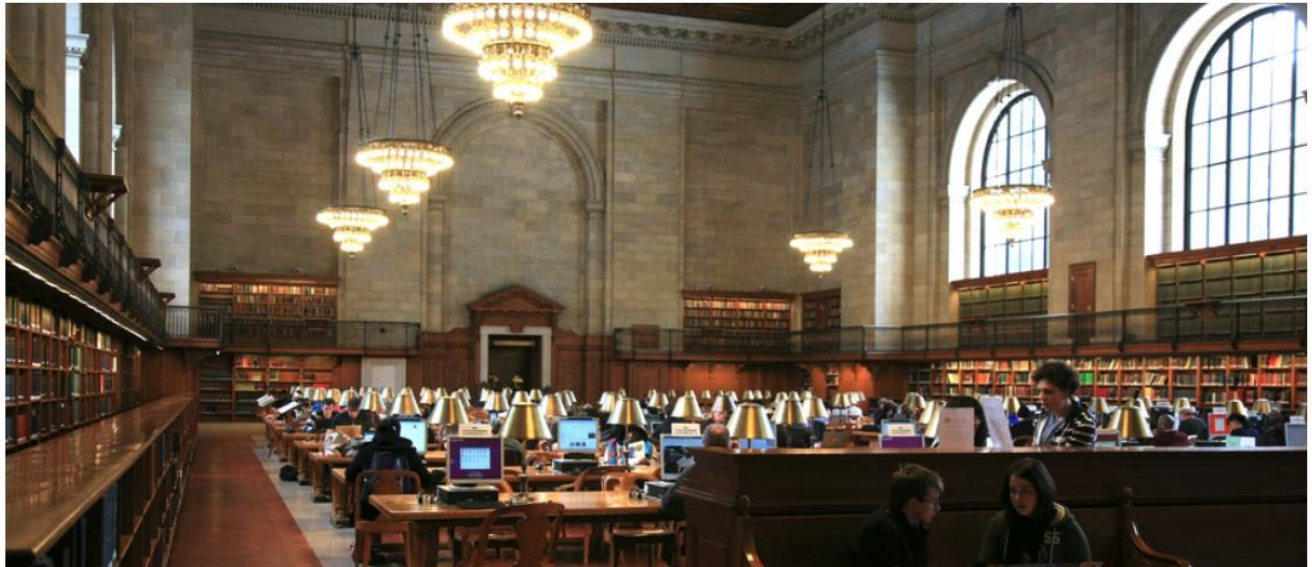
- **Google's caching of websites**
  - Field v Google 2006, Parker v Google 2007
  - Found that crawling + caching served a new and valuable function, this is highly transformative, and not a replacement for the original work
- **Image search and thumbnails**
  - Perfect 10 v Amazon / Google 2007
  - Found search-engine thumbnails and indexing are protected fair use, even when the underlying works are copyrighted images.
- **Web scraping**
  - Ticketmaster v RMG Technologies 2007
  - Found against the scraper—but for contract breach and circumvention, not copyright
  - Helps clarify that fair-use analysis applies when the issue is copying, not access violation.

TECHNOLOGY

After 10 Years, Google Books Is Legal

Thanks to a landmark ruling, information just got a little more free.

By Robinson Meyer



Google Books and Fair Use

- 2004: Google starts scanning books; wants to make “orphaned” works available
- 2005: Authors Guild sued to stop it
- 2011: almost settled on a for-fee model – the judge ruled against -- that would be a monopoly!
- 2013: Judge Chin said the scans were fair use
  - ““It advances the progress of the arts and sciences, while maintaining respectful consideration for the rights of authors and other creative individuals, and without adversely impacting the rights of copyright holders”
- 2014: Authors Guild appealed – and lost!

https://www.theatlantic.com/technology/archive/2015/10/fair-use-transformative-leval-google-books/411058/

Google Books and Fair Use

Appeals Court Judge Leval wrote:

"For nearly 300 years, since shortly after the birth of copyright in England in 1710, courts have recognized that, in certain circumstances, giving authors absolute control over all copying from their works would tend in some circumstances to limit, rather than expand, public knowledge"

"Google's unauthorized digitizing of copyright-protected works, creation of a search functionality, and display of snippets from those works are non-infringing fair uses. The purpose of the copying is highly transformative, the public display of text is limited, and the revelations do not provide a significant market substitute for the protected aspects of the originals."

https://www.theatlantic.com/technology/archive/2015/10/fair-use-transformative-leval-google-books/411058/

Upshot of Google Books Case

Does this settlement mean you can write a web scraper to download all of the journals from a publisher and do text analysis on them?

Upshot of Google Books Case

Does this settlement mean you can write a web scraper to download all of the journals from a publisher and do text analysis on them?

No, it does not – if you did this using the UC Berkeley license to access online journals, that would violate the terms of use (and might shut down access for the entire campus!)

IP and LLMs: Training Data

Legally Speaking: Does Using In-Copyright Works as Training Data Infringe?

Communications of the ACM, forthcoming

UC Berkeley Public Law Research Paper

6 Pages • Posted: 15 Jul 2025

Pamela Samuelson

University of California, Berkeley - School of Law

Date Written: July 01, 2025

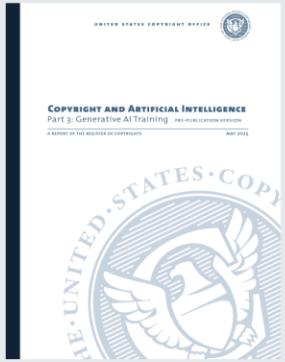
TWO LLM TRAINING DATA CASES

- *2 cases: Bartz v. Anthropic* and *Kadrey v. Meta*,
- Question is whether genAI developers made a fair use or not by collecting and copying in-copyright works as training data for their LLMs.
- The judges largely found the developers' training uses to be fair, although applying differing analyses and coming to different conclusions on particular issues.
- The judges' two biggest disagreements concerned:
  - The implications of using "pirated" books as training data for foundation models and
  - A novel theory of "market dilution" which speculates that the general, indirect market effects of genAI outputs capable of producing massive numbers of potentially competitive works is a market harm under copyright law.

LLM Training: Prediction

- It looks likely that future training will be done on licensed data and this issue will dissipate
- However, there are concerns that for information that is not licensed, there will be little incentive to produce it (especially web pages)

What is the US Copyright Office Saying about LLMs/AI?



Report on Copyright and Artificial Intelligence

Copyright and Artificial Intelligence analyzes copyright law and policy issues raised by artificial intelligence (AI). This Report is being issued in several Parts. Part 1 was published on July 31, 2024, and addresses the topic of digital replicas. Part 2 was published on January 29, 2025, and addresses the copyrightability of outputs created using generative AI. On May 9, 2025, the Office released a pre-publication version of Part 3 in response to congressional inquiries and expressions of interest from stakeholders. A final version of Part 3 will be published in the future, without any substantive changes expected in the analysis or conclusions.

Part 1: Digital Replicas

Part 2: Copyrightability

Part 3: Generative AI Training (Pre-publication)

https://www.copyright.gov/ai/

IP and LLMs: Copyright

Topic	USCO Position
Human Authorship	Copyright protects the **original expression created by a human author**. Copyright does not extend to purely AI-generated material or material where there is insufficient human control over the expressive elements.
AI as a Tool	Using AI as an **assisting tool** in a human creative process (e.g., for editing, ideation, or rendering) does **not** bar copyright protection for the final work.
AI Prompts	**Prompts alone do not provide sufficient human control** to make the user the author of the resulting output. The Office views prompts as a user's mental conception or idea, not the controlled expression required for authorship.
Modifications/Arrangement	A human can claim copyright for their contributions to an AI-generated work if they demonstrate creativity through the **selection, coordination, or arrangement** of the AI-generated material, or through **creative modifications** of the output.
Registration Requirement	Applicants must **disclose** the inclusion of AI-generated material when registering their works, and they can only claim copyright for the human-authored components.

Summary by Gemini

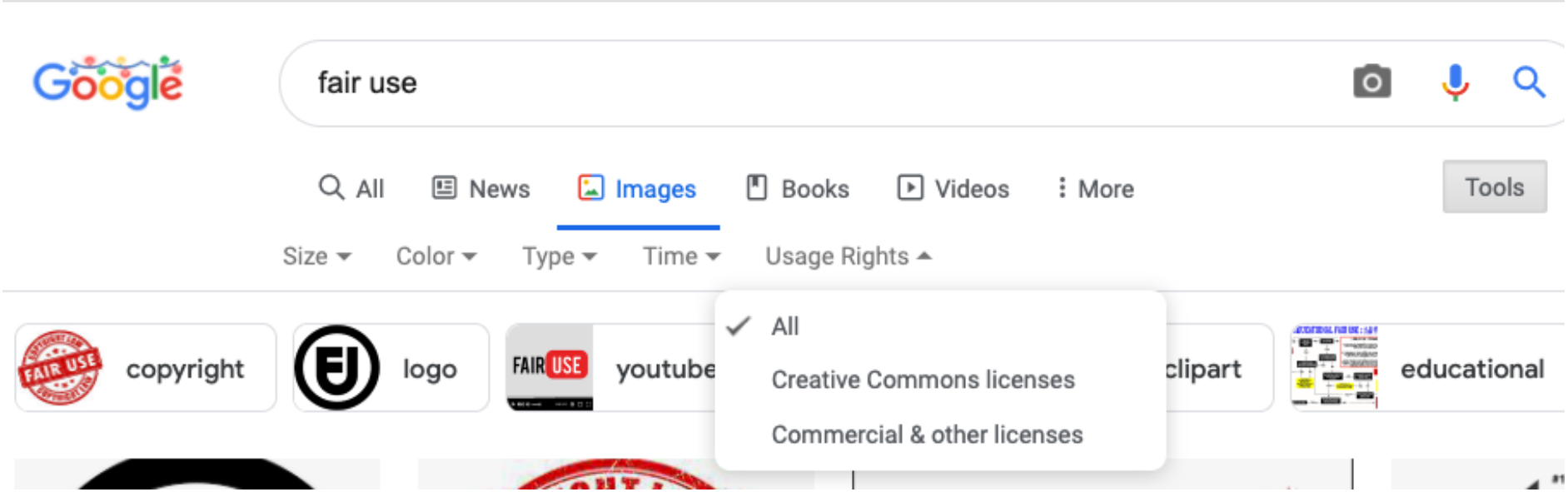
A Brief History of the Creative Commons


creative
commons

To Get Rights to Music / Images

- Make your own
- Get permission / pay for it
- Support / contribute to open access efforts
- Support efforts to reform copyright law

Google Images allows you to try to find images with sharable rights



Many Organizations Trying To Make Free Sharable Content Available

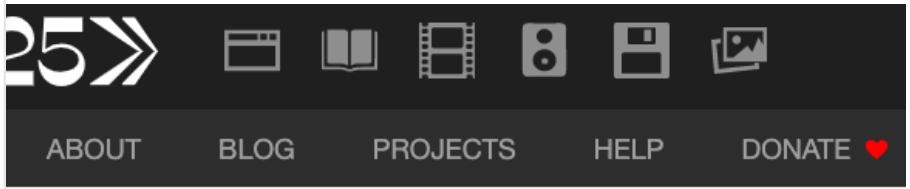
creative commons | Share your work | Use & remix

Creative Commons > What We Do > Program areas > Open Culture > Legal Music for Remixing and Sampling

Legal Music for Remixing and Sampling

Producers, DJs, and remixers: Looking for free sounds?

Many musicians choose to release their songs under Creative Commons licenses, which give you the legal right to do things like remix them and use samples from them in your tracks for free.

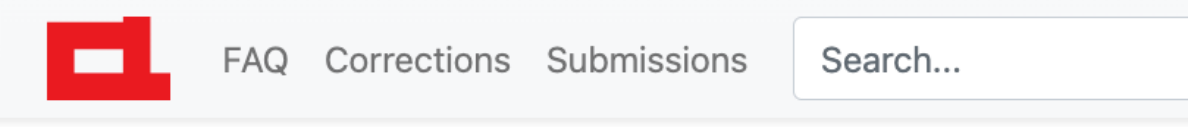


Prelinger Archives

View thousands of films from the Prelinger Archives!

More...

Academic Publishing is Slowly Winning the Battle for Open Access



Academic Publishing is Slowly Winning the Battle for Open Access

Open Access

ACM's Transition to Full Open Access

On January 1, 2026, ACM will become a fully Open Access Publisher. All ACM publications, including ACM journals, will be 100% Open Access. ACM believes a sustainable Open Access future best serves the global computing community. Open Access will enable ACM authors to gain a competitive edge in visibility, impact, and global reach. Open Access papers in the ACM Digital Library are:

Proceedings of the AAAI Conference on Artificial Intelligence

The proceedings are sponsored by the Association for the Advancement of Artificial Intelligence, which has chosen to provide them in open access to the entire worldwide scientific community. Copyright to individual papers as well as the proceedings as a whole is fully owned by the Association for the Advancement of Artificial Intelligence. Permission is required for republication. Please consult the AAAI copyright form for details.

Breaking boundaries. Empowering researchers. **Opening Science.**

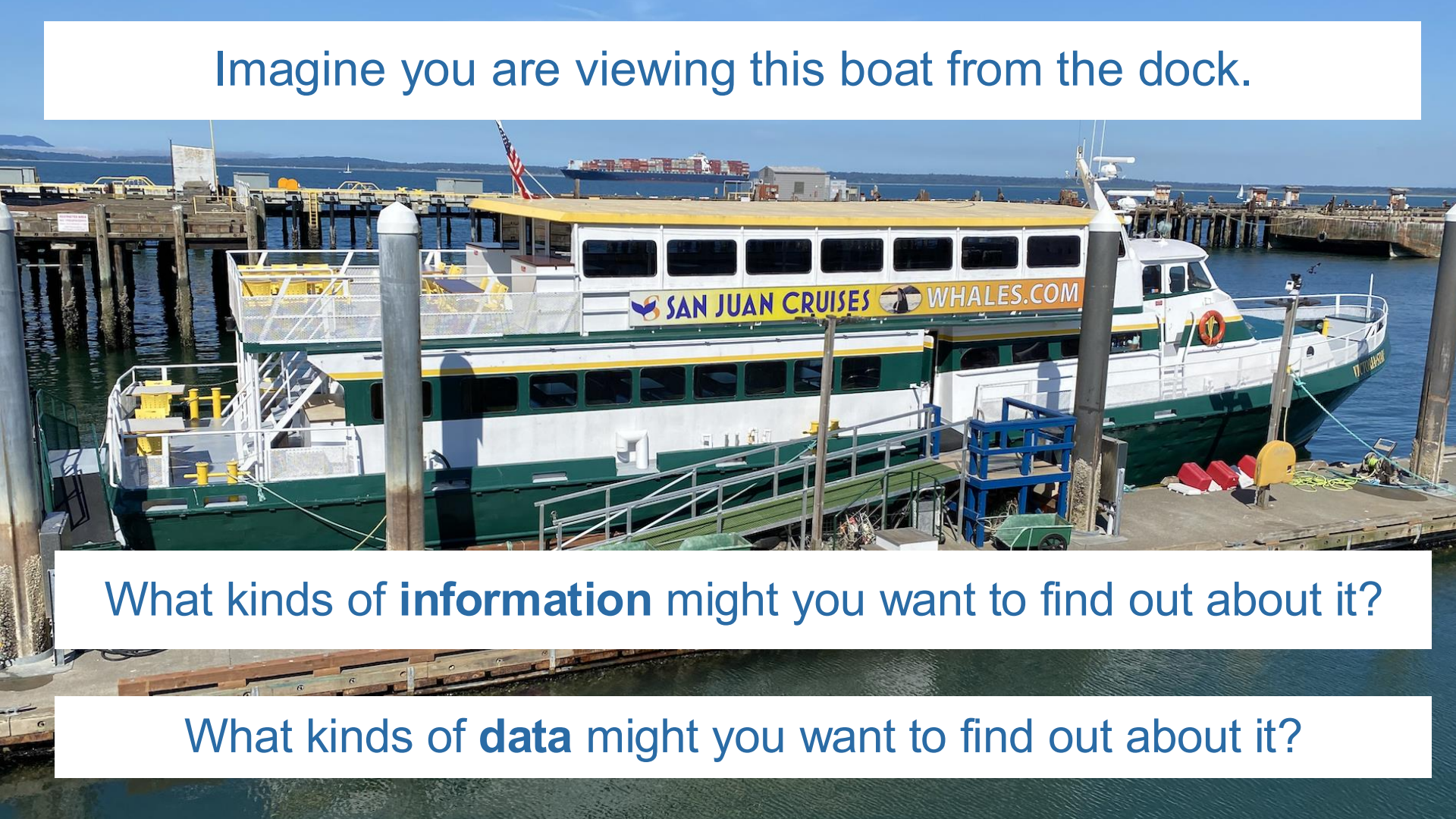
PLOS is a nonprofit, Open Access publisher empowering researchers to accelerate progress in science and medicine by leading a transformation in research communication.

About PLOS

What Intellectual Property Considerations Should You Consider For your Final Projects/Papers?

Course Wrap Up

Imagine you are viewing this boat from the dock.



What kinds of **information** might you want to find out about it?

What kinds of **data** might you want to find out about it?

Borges' Library of Babel

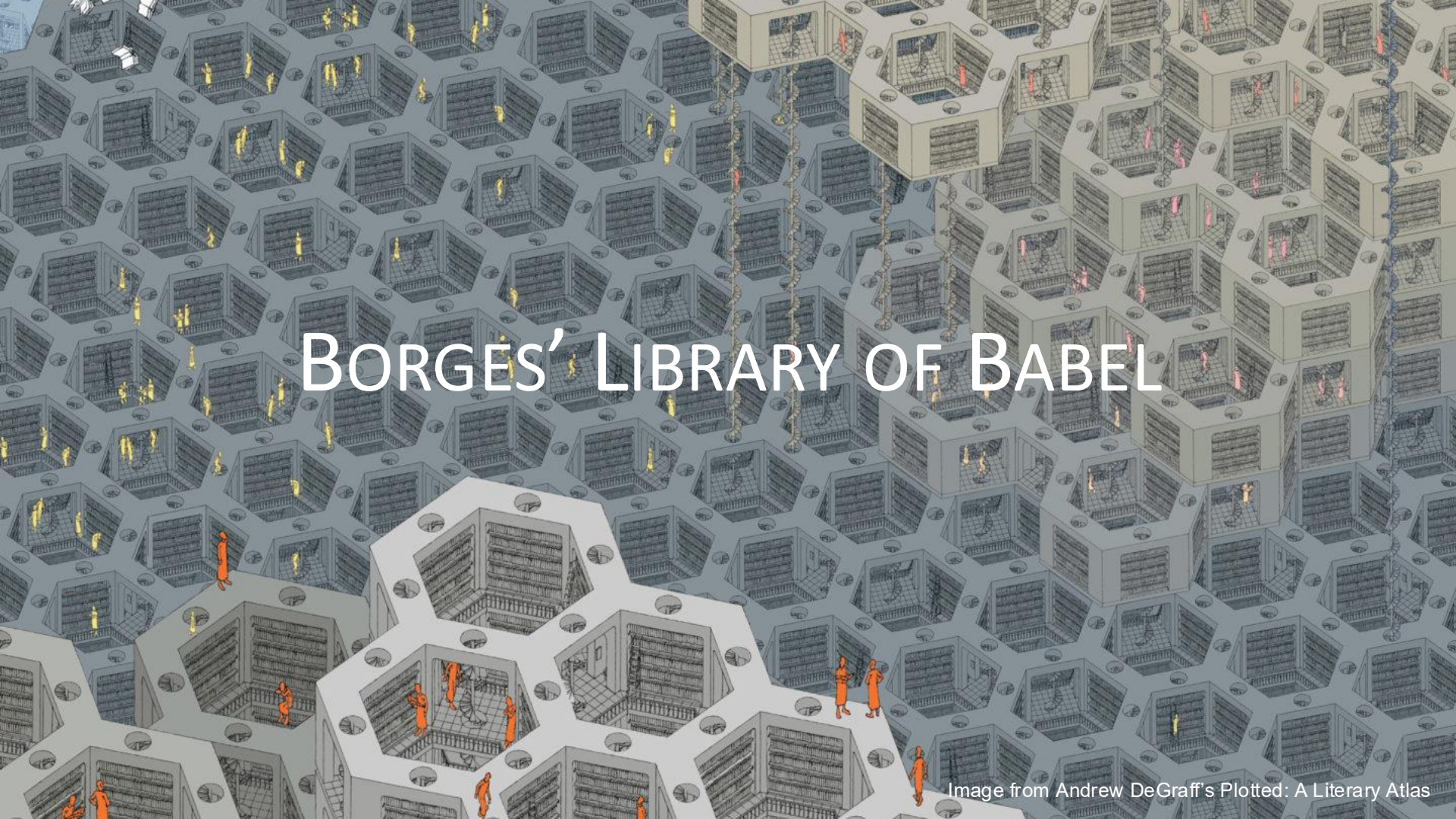


Image from Andrew DeGraff's Plotted: A Literary Atlas

Data, Descriptions, Metadata, Metadata Description

Phenomena in the World	Data: What Items are Collected	Data Representation	MetaData	Metadata Description Language
	Photos of birds, stored on hard drive	Pixels	Image size, Image format, Dat/time taken, Lat/Long, Camera settings	EXIF (automatic) IPTC (manual); now XMP is a new standard
	Real bird specimen	Physical bird	Measurements, type, date, instrument, etc	Specimens database Schema (e.g. EME)
Work done on a contract	PDF document of an invoice	Dollar amount embedded in PDF	Date, invoice number, payee, amount due, etc	Database Schema

JSON vs XML

```
{
  "student": [

     {
        "id":"01",
        "name": "Tom",
        "lastname": "Price"
     },

     {
        "id":"02",
        "name": "Nick",
        "lastname": "Thameson"
     }
  ]
}
```

```
<?xml version="1.0" encoding="UTF-8" ?>
<root>
        <student>
                <id>01</id>
                <name>Tom</name>
                <lastname>Price</lastname>
        </student>
        <student>
                <id>02</id>
                <name>Nick</name>
                <lastname>Thameson</lastname>
        </student>
</root>
```

https://www.guru99.com/json-vs-xml-difference.html

XML can be nested (hierarchical)

```
▼<Class>
  ▼<Order Name="TINAMIFORMES">
    ▼<Family Name="TINAMIDAE">
       <Species Scientific_Name="Tinamus major"> Great Tinamou.</Species>
       <Species Scientific_Name="Nothocercus">Highland Tinamou.</Species>
       <Species Scientific_Name="Crypturellus soui">Little Tinamou.</Species>
       <Species Scientific_Name="Crypturellus cinnamomeus">Thicket Tinamou.</Species>
       <Species Scientific_Name="Crypturellus boucardi">Slaty-breasted Tinamou.</Species>
       <Species Scientific_Name="Crypturellus kerriae">Choco Tinamou.</Species>
     </Family>
   </Order>
  ▼<Order Name="GAVIIFORMES">
    ▼<Family Name="GAVIIDAE">
       <Species Scientific_Name="Gavia stellata">Red-throated Loon.</Species>
       <Species Scientific_Name="Gavia arctica">Arctic Loon.</Species>
       <Species Scientific_Name="Gavia pacifica">Pacific Loon.</Species>
       <Species Scientific_Name="Gavia immer">Common Loon.</Species>
       <Species Scientific_Name="Gavia adamsii">Yellow-billed Loon.</Species>
     </Family>
   </Order>
```

http://webdam.inria.fr/Jorge/prog/birds.xml

Netflix

Resource Selection Strategy

What is the **Right** Way to Organize Spices/Herbs?

Glushko, TDO

by cuisine

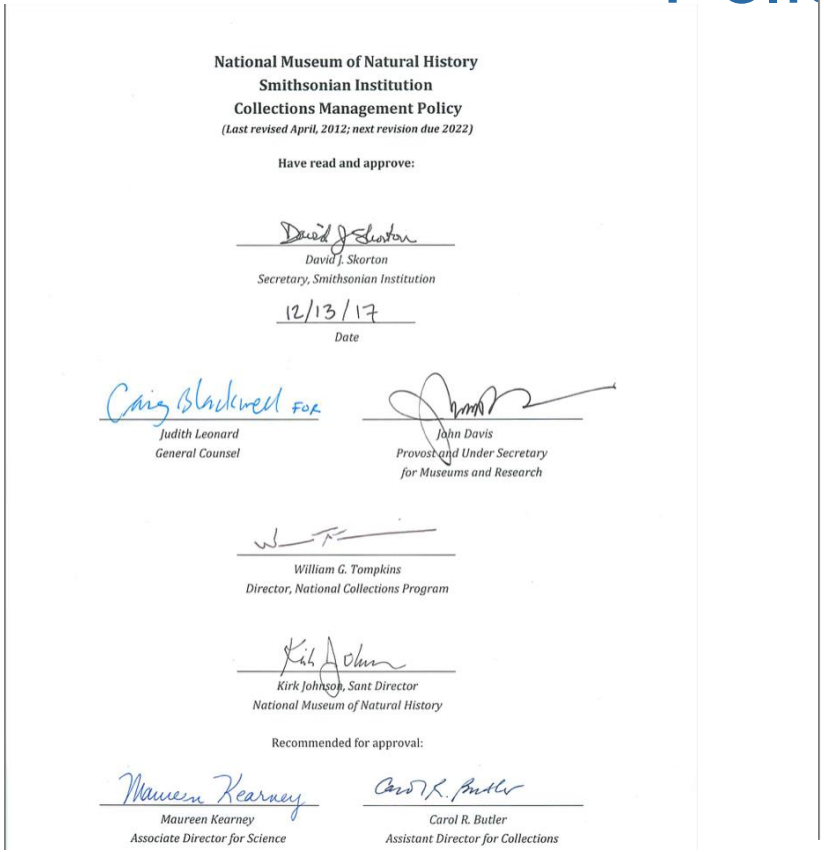
wikipedia

by price, eye appeal

alphabetically

Example:
Smithsonian Natural History Collection

Smithsonian Collections Management Policy



National Museum of Natural History
Smithsonian Institution
Collections Management Policy
(Last revised April, 2012; next revision due 2022)

Have read and approve:

David J. Skorton
David J. Skorton
Secretary, Smithsonian Institution

12/13/17
Date

Craig Blackwell FOR
Judith Leonard
General Counsel

John Davis
Provost and Under Secretary
for Museums and Research

William G. Tompkins
Director, National Collections Program

Kirk Johnson, Sant Director
National Museum of Natural History

Recommended for approval:

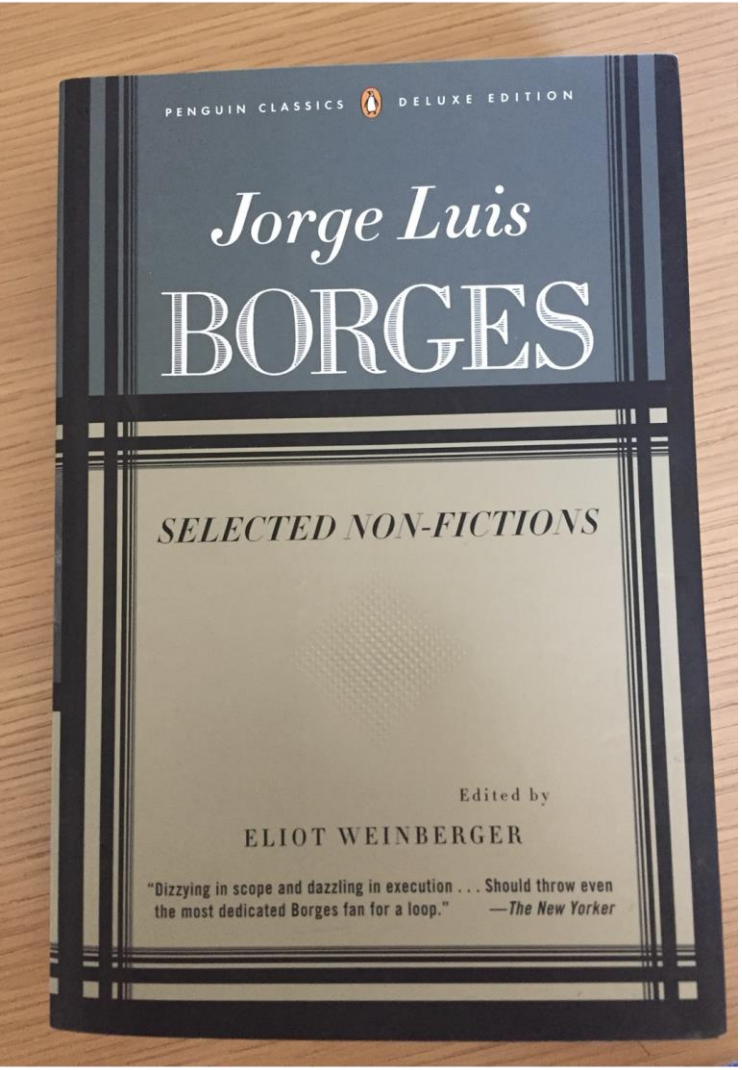
Maureen Kearney
Maureen Kearney
Associate Director for Science

Carol R. Butler
Carol R. Butler
Assistant Director for Collections

National Museum of Natural History
Smithsonian Institution
Collections Management Policy
Rev. December 13th, 2017

Table of Contents

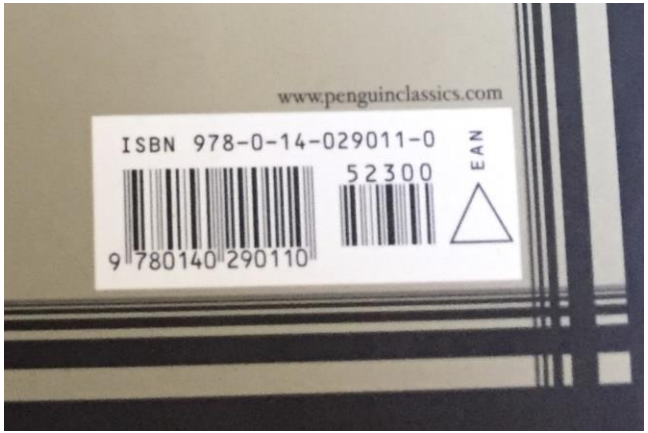
https://naturalhistory.si.edu/research/nmnh-collections/museum-collections-policies



There are many copies of this boo

Many books have the same title

How to identify them?



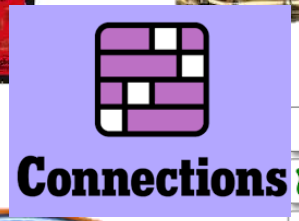
- Prefix: currently either 978 or 979
- Registration group: country, geo, language
- Registrant: publisher or imprint
- Publication: edition and format of the item
- Check digit: mathematically validates the rest of the number (sometimes called a check sum)

www.isbn-international.org/content/what-isbn

https://www.youtube.com/watch?v=UJMx_YfZiW0

Which Games / Game Shows are about Categories?

Categories



Individual information / facts / data

(Some) Principles for Intentionally Creating Categories

- Enumeration
- Single Property
- Multiple Properties
- Goal-based
- Theory-based

Car Brands

Climbable Things

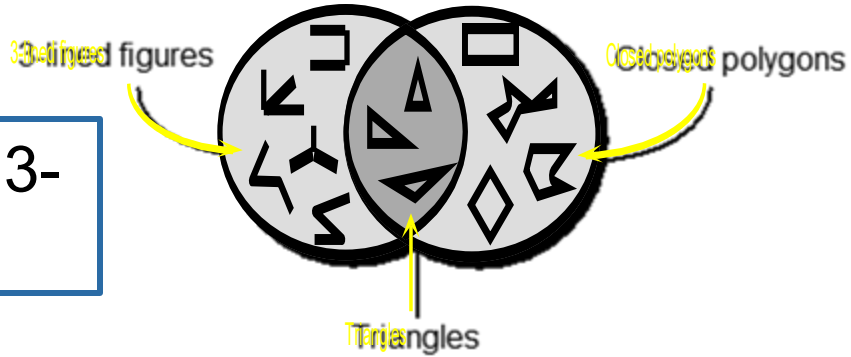
Expensive & Brittle Things

Things you take on a camping trip

Definition of a Triangle

CLASSIC VIEW OF CATEGORIES

1. Categories are defined by a list of properties shared by **all elements** in a category (necessary & sufficient)
2. Category membership is **binary** (in or out)
3. Because membership is defined by rules, **every** member in the category is **equally** a member



Example: triangles are 3-sided closed polygons

PROTOTYPES

- Which dog breed is central?
- Which are “better” or “worse” examples?

https://articles.hepper.com/great-dane-chihuahua-mix/

https://www.pageant.dog/breeds/dog/1120-Labrador-Retriever

Evidence For Prototypes

Typicality ratings

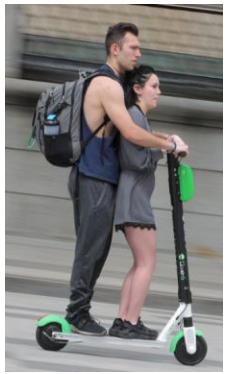
Order in which members are named

Time needed to verify category membership

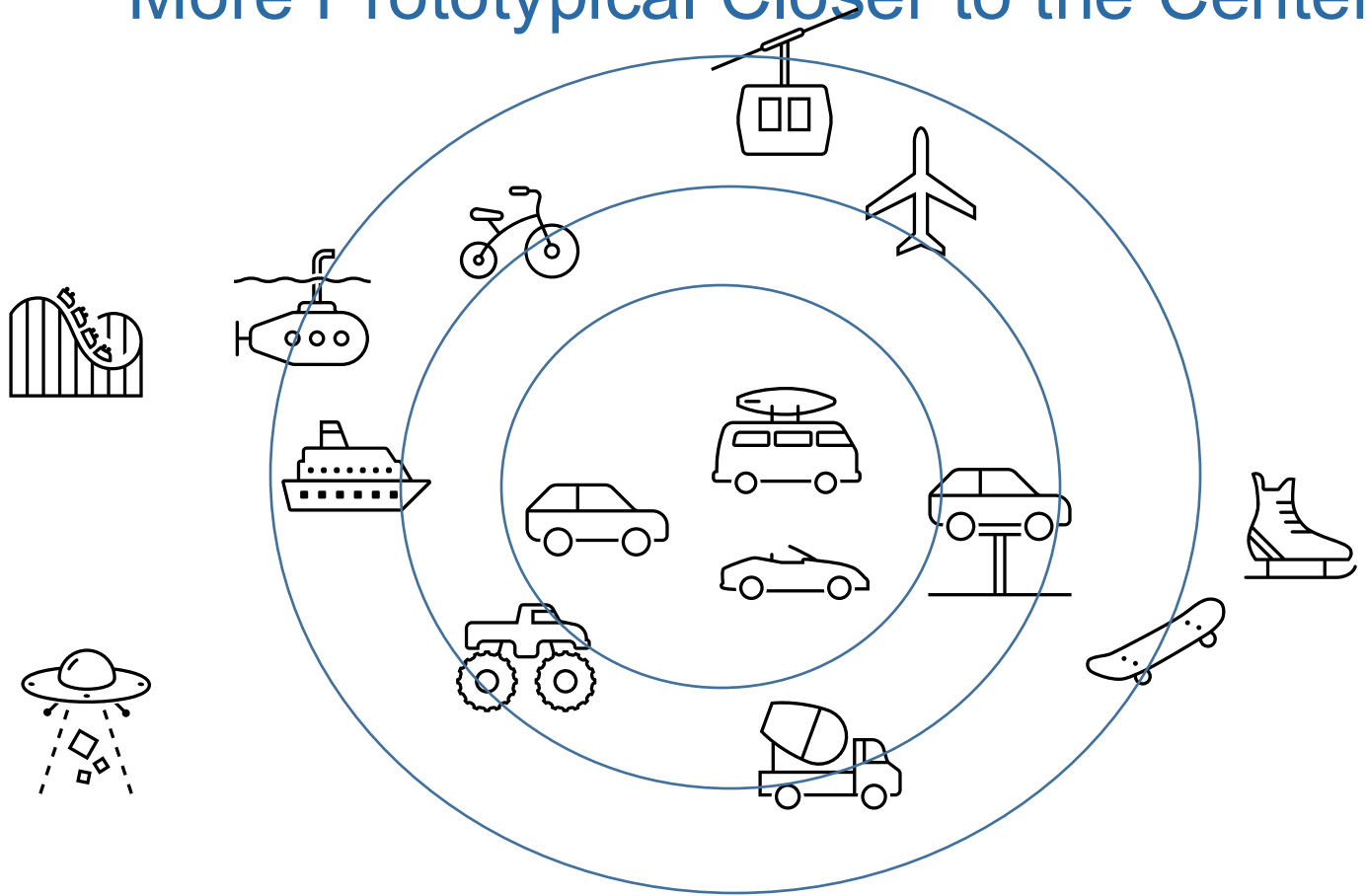
Istockphoto.com

Prototype vs. Classical In Practice

- Although we think more naturally with fuzzy boundaries, we sometimes are forced to make sharp distinctions
- Example: Dept of Motor Vehicles has to classify which vehicles require licenses
- Answers questions like:
  - *Licenses?*
  - *Helmets?*
  - *Sidewalk?*



Example: Vehicle
More Prototypical Closer to the Center



Superordinate and Subordinate Levels

SUPERORDINATE	animal	furniture	emotion
BASIC LEVEL	dog	chair	happy
SUBORDINATE	terrier	rocker	joy

What are other examples?

- Children take longer to learn superordinate
- Superordinate not associated with mental images or motor actions

DIFFERENT LEXICAL FORM	SAME LEXICAL FORM
hypernyms (superordinate)	hypernyms (superordinate)
synonyms	polysemes
sibling terms	homographs
hyponyms (subordinate)	hyponyms (subordinate)

WORD

DIFFERENT LEXICAL FORM

SAME LEXICAL FORM

hypernyms (superordinate)

cooking utensil

Created by Gan Khoon Lay from Noun Project

polysemes

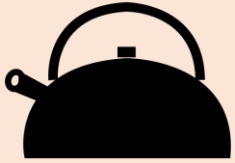
pot (betting)

Created by Parallel Digital Studio from Noun Project

pot (flower)

Created by Matthieu Mercier from Noun Project

pot



kettle

Created by Lim Qian Fang from Noun Project

Created by Aris Sunjayay from Noun Project

coffee pot

hyponyms (subordinate)

Created by ProSymbols from Noun Project

pot

homographs

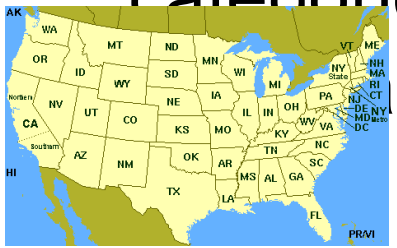
Hierarchy Composed of a Mix of Concepts Is Difficult to Navigate

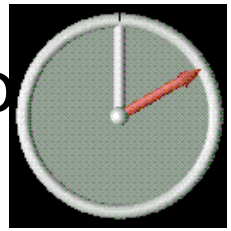
- The problem: **different** attributes in **one** hierarchy:
  - *Sound type > location type > beak type*
- But it is **not** a good solution to combine these either
  - *Ground-dwelling bird that sings with stubby beak*
- Solution: faceted metadata! **Separate** out the attributes and then assign **multiple** attributes to each information item.
  - *Sound type*
  - *Location type*
  - *Beak type*

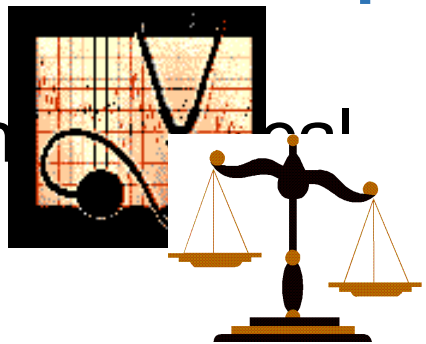
SOLUTION: FACETED CATEGORIES

- A **set** of different categories
  - Identify **different** aspects, attributes, or features
  - Resources are labeled with **multiple** categories

tego be h







GeoRegion + Time/Date + Topic + Role


Dish > main > tacos
Cuisine > Mexican
Occasion > Party
Occasion > Tailgate
Ingredients > Meat > Fish > Cod
Ingredients > Veg > Onion > Red Onion
Ingredients > Bread > Tortilla
Preparation > Saute
https://www.epicurious.com/recipes/food/views/fish-tacos-352976

Five Hierarchical Faceted Categories

Nature
Animal
Mammal
Horse

Is-a

Occupations
Cowboy

Is-a-type-of

Clothing
Hats
Cowboy Hat

Is-a

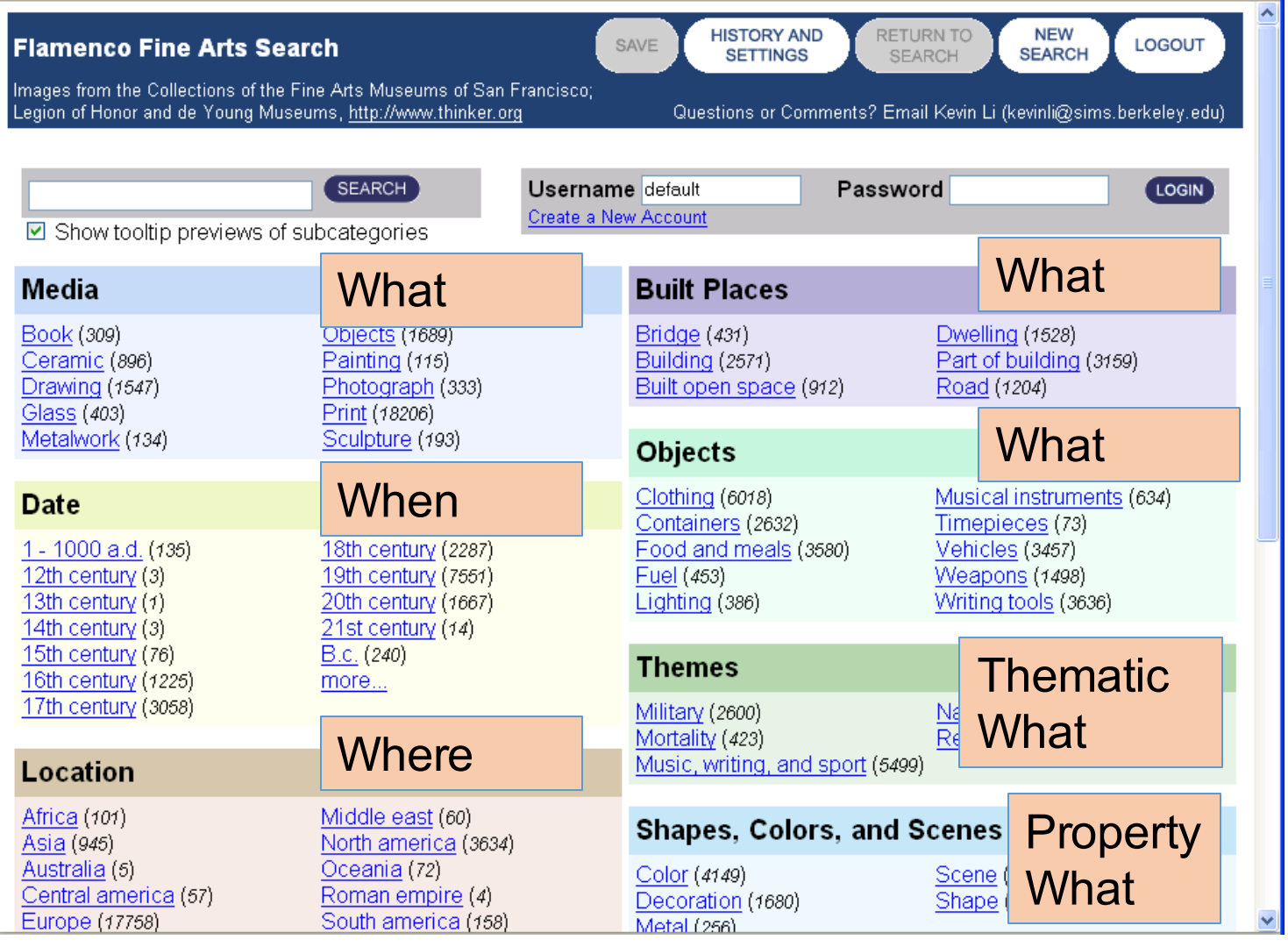
Media
Engraving
Wood Eng.

Is-a-type-of

Location
North America
America

Is-in or is-part-of

A Bucking Bronco by Henry Wolf, https://art.famsf.org/henry-wolf/bucking-bronco-19633026024



Notice for art objects, the "what" refers to both the art and what it depicts.

Let's drill down into this thematic category

Refine Your Search

Author

James Baldwin (56)
Grant Blackwood (9)
United States (9)
Carolyn G Hart (8)
Brad Taylor (8)
Show more ...

Year

2018 (144)
2017 (155)
2016 (138)
2015 (159)
2012 (143)
Show more ...

Language

English (2304)
Undetermined (29)
German (13)
French (9)
Chinese (5)
Show more ...

Content

Fiction (406)
Non-Fiction (2486)
Biography (184)

WorldCat

the fire next time

Advanced Search | Find a Library

Who (author)

When (published)

What (language)

What (genre)

Select All Clear All Save to: [New List] Save Sort by:

1. 

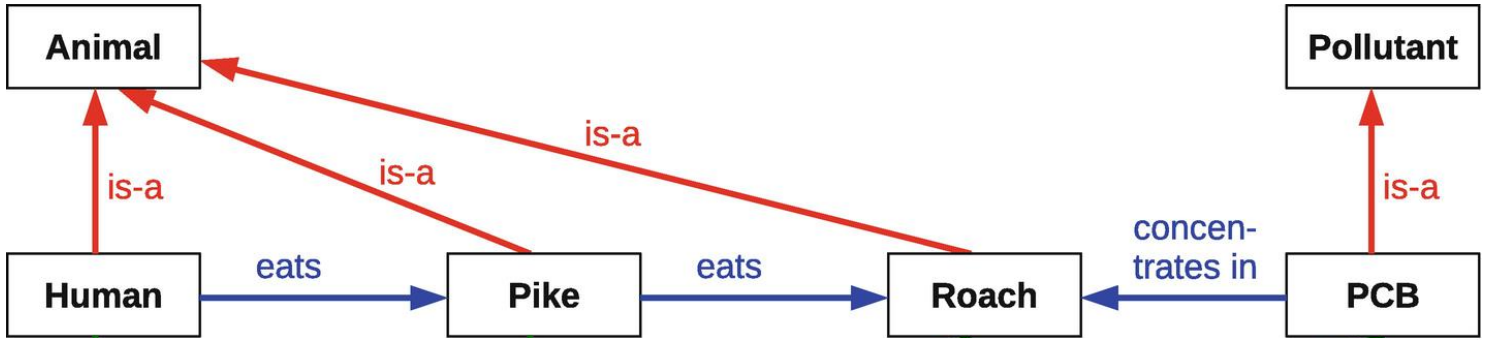
The fire next time
by James Baldwin
Print book View all formats and languages »
Language: English
York : Modern Library, 2021.
s »

2. Nach der Flut das Feuer = The fire next time
by James Baldwin; Miriam Mandelkow; Deutscher Taschenbuch-Verlag
View all formats and languages »
rman
chen dtv 2020
View all editions »

3. 

No fire next time : Black-Korean conflicts and the future of America's cities
Joyce
View all formats and languages »
glish
Publisher: Ithaca : Cornell University Press, 2003.
View all editions »

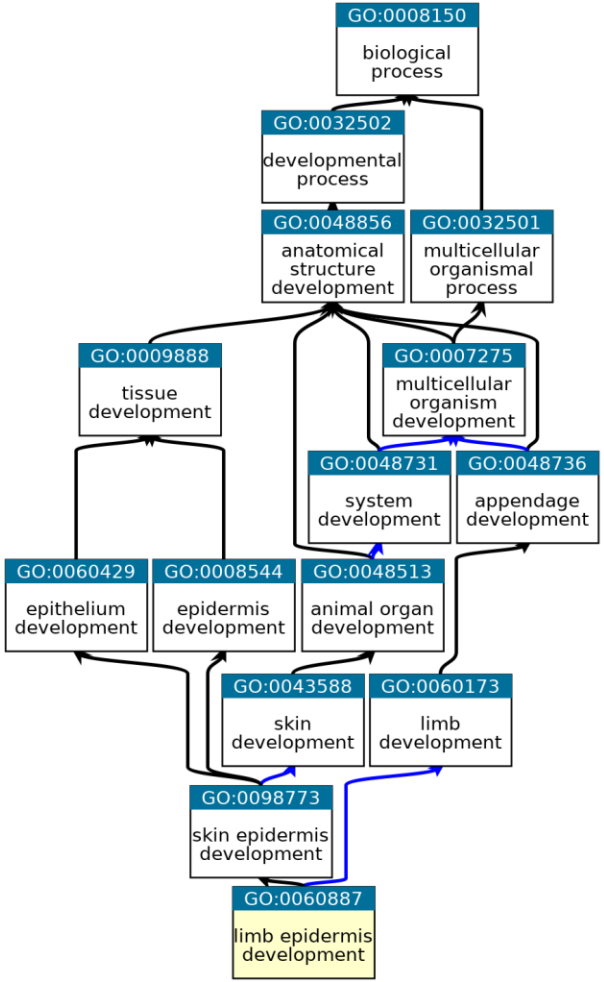
ONTOLOGY EXAMPLE: ECOLOGY

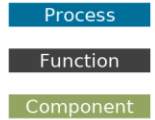


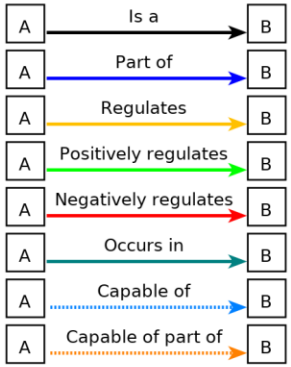
More rules, more inferences
Concentrated-in(p,a) and Eat(b,a) -> Intoxicated-by(b,p)
Eat(a,b) and Eat(b,c) -> Eat(a,c)

Can conclude:
Pike intoxicated-by PCB
Human intoxicated-by PCB

Jean-Baptiste, Ontologies with Python, O'Reilly, 2020







Go Ontology Example

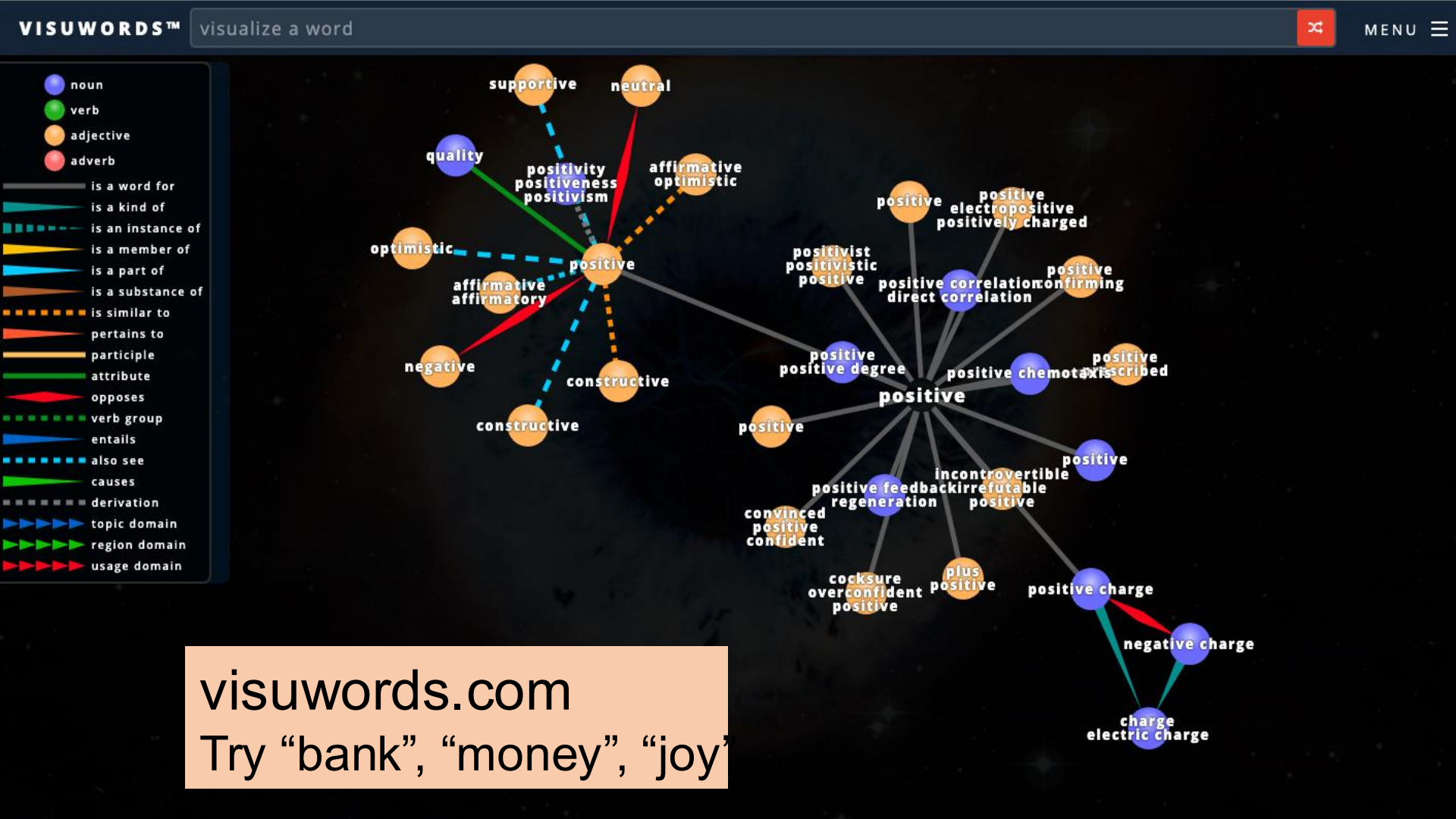
Taxon Constraints

The use of this term should conform to the following taxon constraints:

Ancestor GO ID	Ancestor GO Term Name	Relationship	Taxon ID	Taxon
GO:0008544	epidermis development	Never in Taxon	33090	Viridiplantae
GO:0009888	tissue development	Never in Taxon	147554	Schizosaccharomycetes
GO:0009888	tissue development	Never in Taxon	33630	Alveolata
GO:0009888	tissue development	Never in Taxon	33682	Euglenozoa
GO:0009888	tissue development	Never in Taxon	38254	Glaucocystophyceae
GO:0009888	tissue development	Never in Taxon	4891	Saccharomycetes

QuickGO - https://www.ebi.ac.uk/QuickGO

https://www.ebi.ac.uk/QuickGO/term/GO:0060887



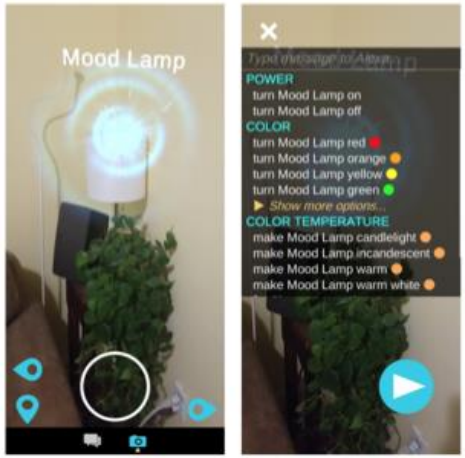
visuwords.com

Try “bank”, “money”, “joy”

Vocabulary Problem Example: Naming for Smart Rooms

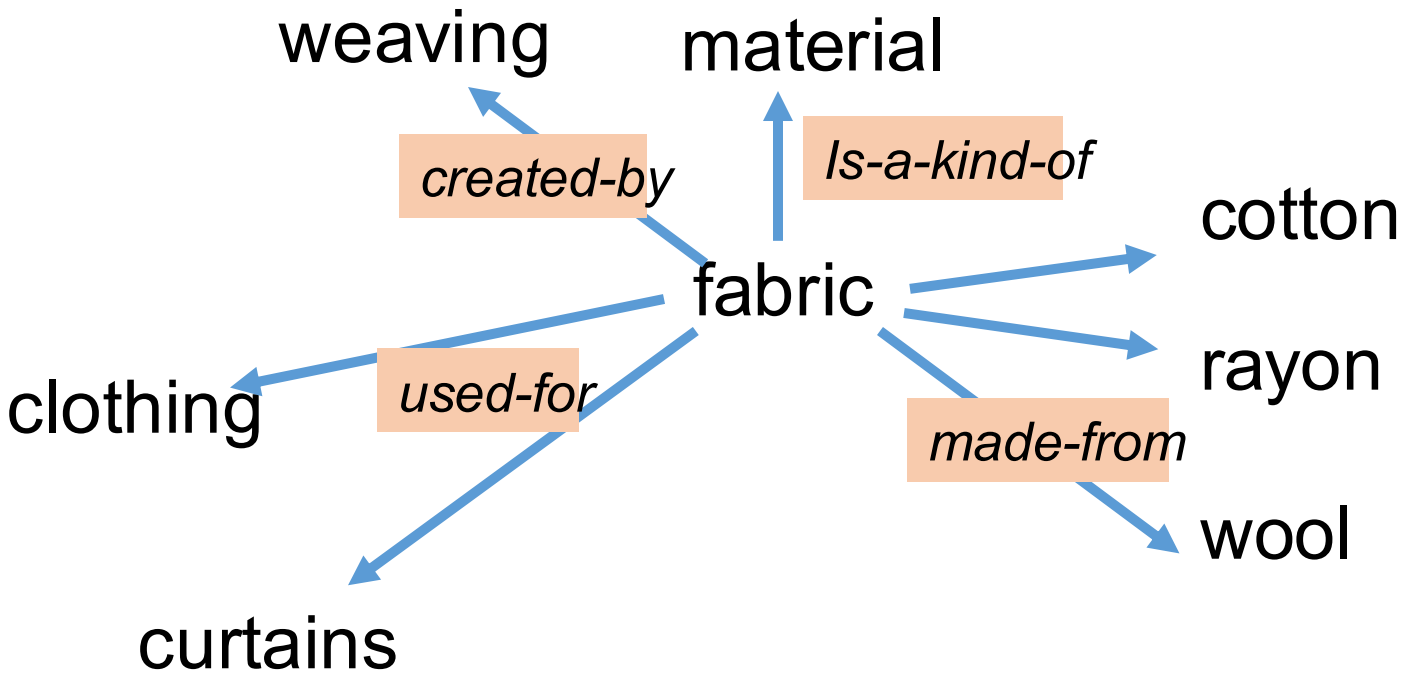
How did she solve this?

Augmented reality and autosuggest!



Meghan Clark, One-Shot Interactions with Intelligent Assistants in Unfamiliar Smart Spaces, UCB Dissertation, 2021

A Concept Has Many Relation Types (Associations)



Taxonomic and Thematic Similarity

High Taxonomic, Low Thematic

Word 1	Word 2
Breakfast	Dinner
Helmet	Crown
Salt	Sugar

Low Taxonomic, High Thematic

Word 1	Word 2
Helicopter	Pilot
Floss	Teeth
Pillow	Head

High Taxonomic, High Thematic

Word 1	Word 2
Ring	Bracelet
Shingle	Brick
Tape	Staple

Low Taxonomic, Low Thematic

Word 1	Word 2
Portrait	Report
Prisoner	Pupil
Bird	Lamb

Landrigan & Mirman, Taxonomic and Thematic Relatedness Ratings for 659 Word Pairs, JOPD 4, 2016

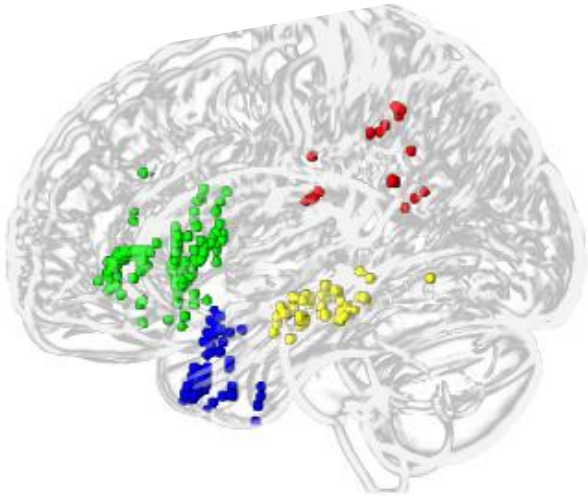
Taxonomic Vs Thematic Associations Evidence for Different Specialization in Regions in the Brain

Intracranial EEG readings suggest:

ATL specialized for taxonomic relations

IPL specialized for thematic relations

Close coordination is also suggested between the two regions.



Anterior Temporal Lobe (blue)
Inferior Parietal Lobule (red)

Thye et al, Intracranial EEG evidence of functional specialization for taxonomic and thematic relations, Cortex 140, 2021

Example: Grounded Coding of Tweets

Goal: better understand how people are writing outside the classroom

Approach: use tweets to analyze the writing practices of fans of Bruce Springsteen

Data: tweets before, during, and after a concert in 20

> **Example Corpus Tweet**
> I will never forget this night. I am officially the girl who danced on stage with Bruce Springsteen during dancing in the dark. So. Amazing.

wikipedia

Baby, We Were Born to Tweet: Springsteen Fans, The Writing Practices of *In Situ* Tweeting, and the Research Possibilities for Twitter, **Wolff, Kairos journal, 19.3, 2015** https://kairos.technorhetoric.net/19.3/topoi/wolff/

Measuring Inter-Annotator Agreement

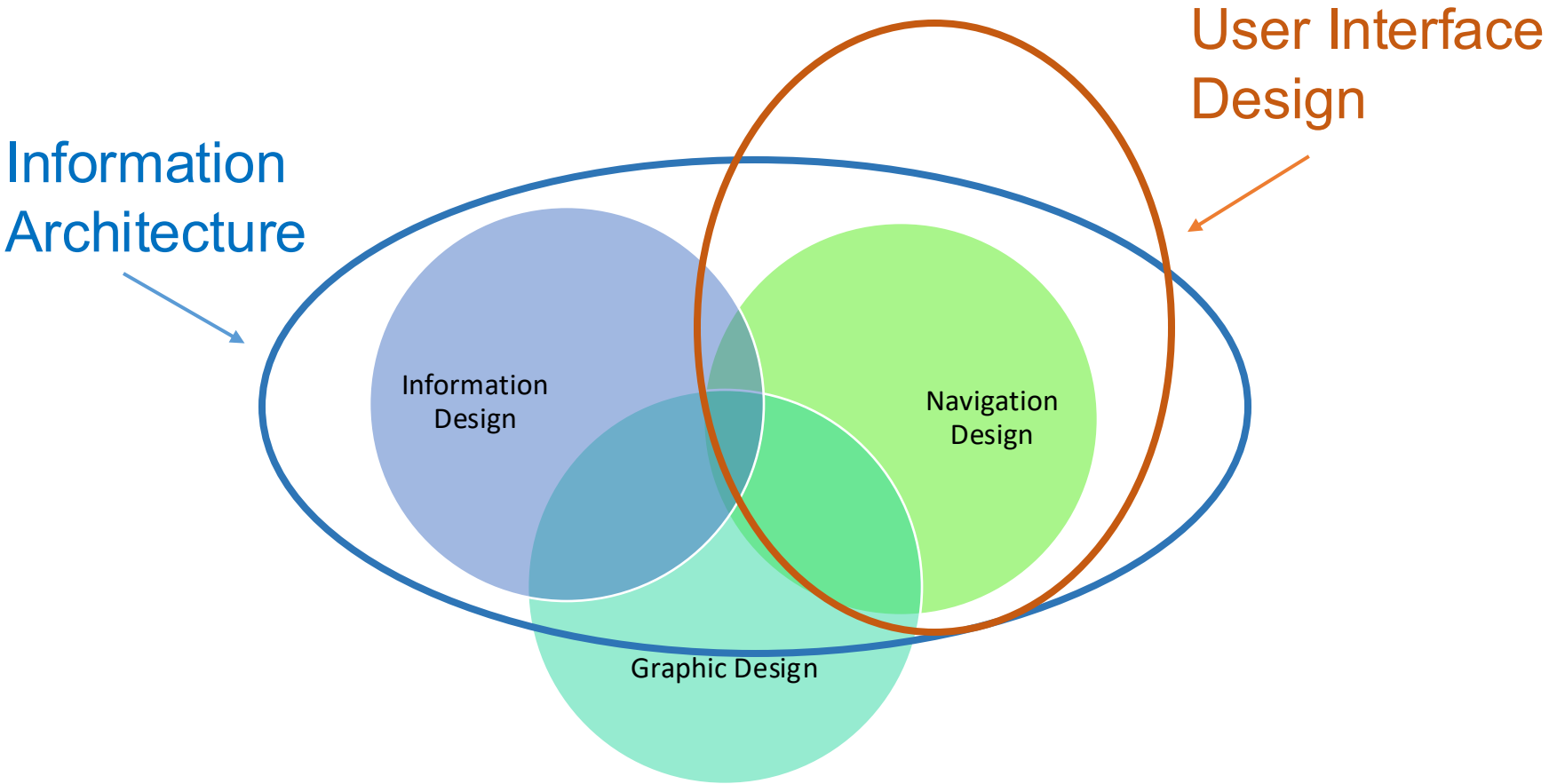
Simple method: take the proportion of agreement

Coder A	Coder B	Agree?
porpoise	dolphin	disagree
porpoise	dolphin	disagree
dolphin	dolphin	agree
dolphin	porpoise	disagree
porpoise	dolphin	disagree
porpoise	dolphin	disagree
dolphin	dolphin	agree
dolphin	porpoise	disagree
dog	dog	agree
dog	cat	disagree

3/10
vs
7/10

Coder A	Coder B	Agree?
cat	cat	agree
cat	cat	agree
cat	cat	agree
cat	cat	agree
porpoise	dolphin	disagree
porpoise	dolphin	disagree
dolphin	dolphin	agree
dolphin	porpoise	disagree
dog	dog	agree
dog	cat	disagree

Are these equivalent? Dolphin/Porpoise harder to distinguish than cat/dog!



Newman & Landay, DIS '00

Housing

Get information and services to help with finding and keeping a home.

How do I ...

- Change my address
- Find affordable housing
- Get help repairing my home

Affordable Rental Housing
Get help finding rental assistance or public housing.

Foreclosure
Learn some of the basics about avoiding and handling foreclosures.

Help Buying a Home
Learn about government programs that make it easier to purchase a home.

Housing Help
Find housing resources targeted to certain audience groups.

Housing Scams
Beware of these frauds and scams when buying or foreclosing on a home.

Housing-Related Complaints
Find out what to do if you have one of these complaints when buying or renting a home.

Mortgages
Learn some of the basics about mortgages.

Moving
Find resources to help you when you're moving.

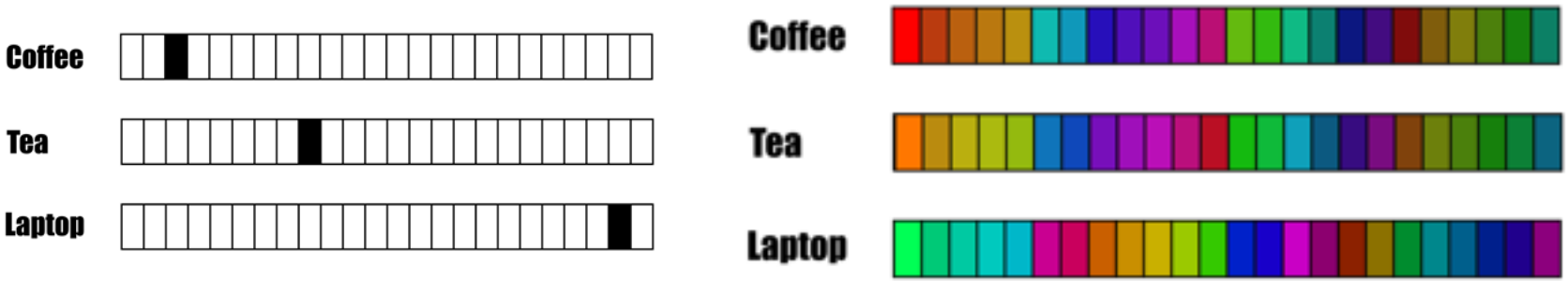
Repairing and Improving a Home
Look for help with repairing or making improvements to your home.

Intuition: Words with Similar Context Neighborhoods Have Similar Meaning

A cup of **tea**
A cup of **coffee**
Tea or **coffee**?
Coffee and **tea** have caffeine
Let's go for a **coffee**
Let's get a **tea**
Coffee vs **Tea**: Which is Best?
I avoid adding sugar to my **tea**
I drink **coffee** with two spoons of sugar

https://predictivehacks.com/a-high-level-introduction-to-word-embeddings/

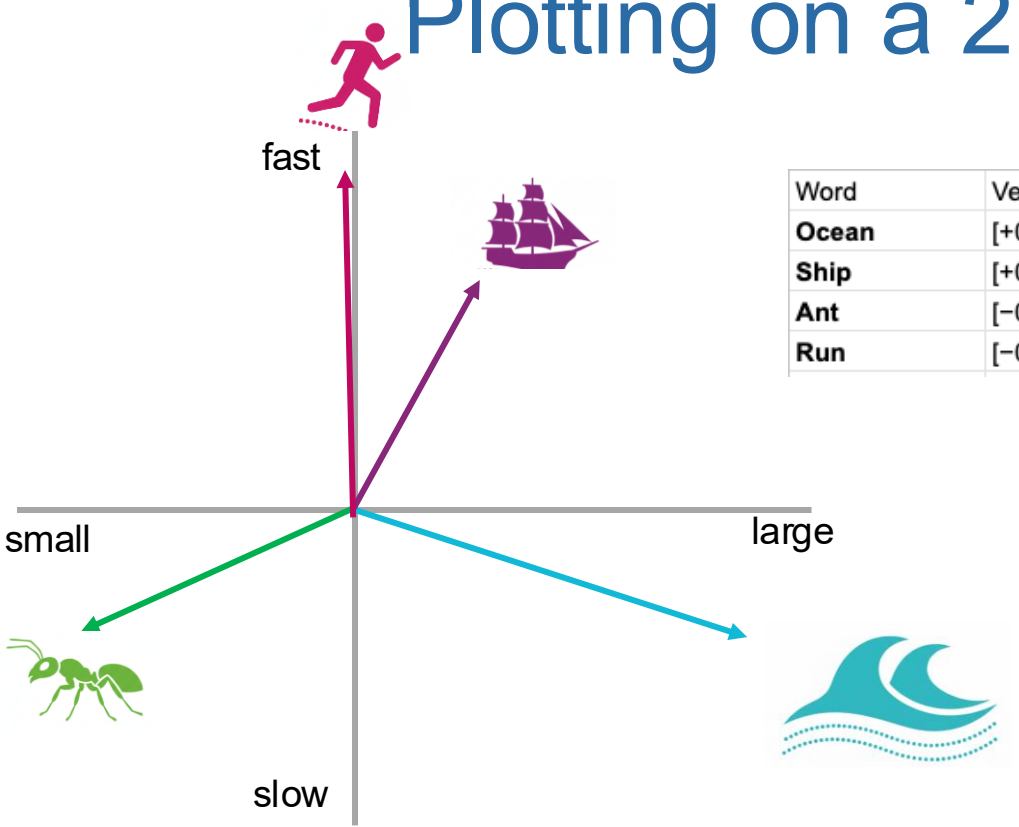
Computing Word Embeddings with Distributions Makes a Richer Representation



Think of the colors as showing complex nuance about which words have appeared in the same context

These are real numbers instead of frequency

Plotting on a 2D Graph



Word	Vector [X, Y]	Intuition
Ocean	[+0.9,−0.2]	Very Larget(high X), Very slow (low/negative Y).
Ship	[+0.5,+0.6]	Medium-large (medium X), Fast (high Y).
Ant	[−0.8,−0.4]	Very small (negative X), Slow (negative Y).
Run	[−0.1,+0.9]	Not big or small (near 0 X), Very fast (high Y).

A vector is not just a point on a graph; it's the **path** from the center to that point, defined by its components.

X-axis: a scale from Large (+X) to Slow (-X)
Y-axis: a scale from Fast (+Y) to Slow (-Y)
(0,0) is the “average” word

Computing Word Similarity with Word Embeddings

```
compareWords(wordPairs)
```

```
Similarity: cat          dog          : 0.8017
Similarity: cat          siamese cat  : 0.8670
Similarity: cat          calico cat   : 0.8437
Similarity: cat          free cat     : 0.7873
Similarity: cat          lion         : 0.5265
Similarity: cat          feline       : 0.6990
Similarity: cat          scratch      : 0.3427
Similarity: cat          whiskers     : 0.3962
Similarity: cat          bark         : 0.3596
```

Which relation types are scored as most similar?

IMAGENET AND WORDNET

WordNet inspired the creation of ImageNet and provides its structure

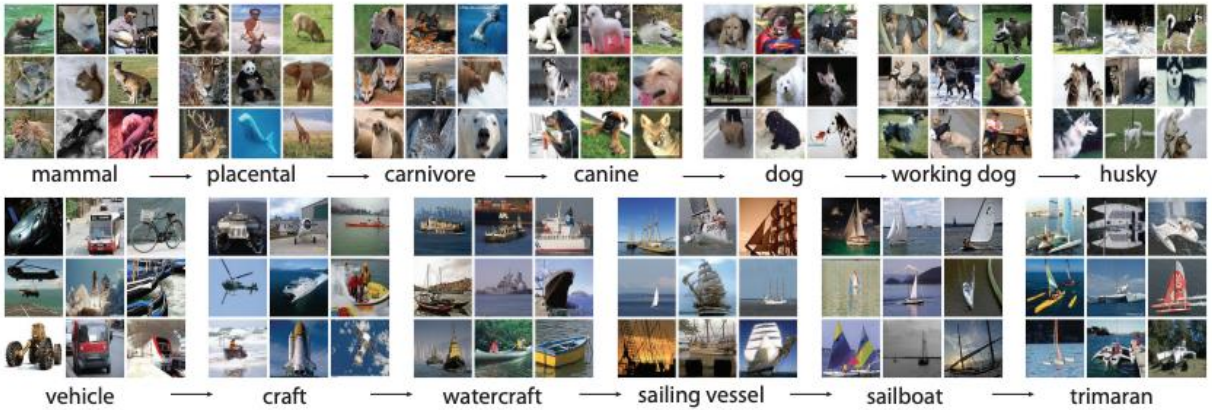
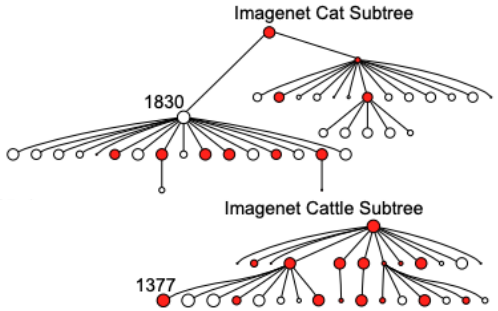
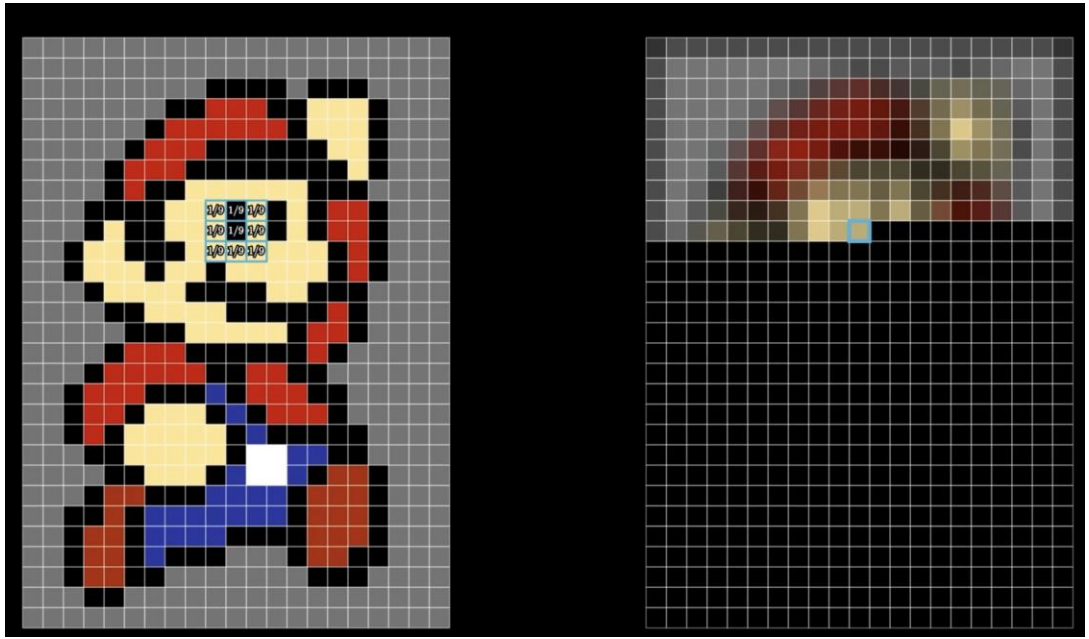


Figure 1: A snapshot of two root-to-leaf branches of ImageNet: the **top** row is from the mammal subtree; the **bottom** row is from the vehicle subtree. For each synset, 9 randomly sampled images are presented.



Deng, et al. "ImageNet: A large-scale hierarchical image database." *IEEE CVPR* 2009.

CONVOLUTIONS TO "AVERAGE" OVER AN IMAGE

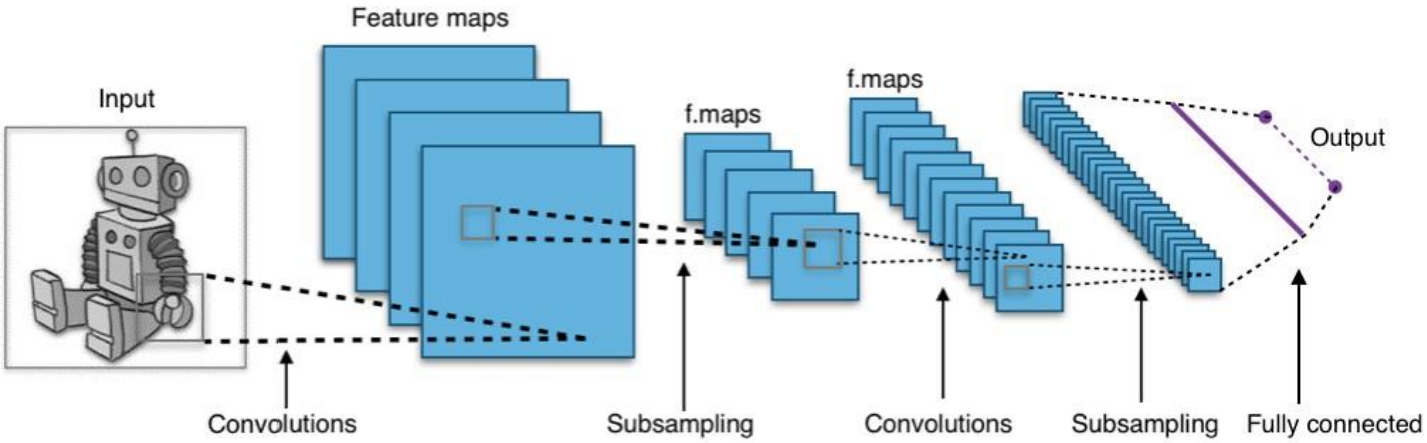


But What is a Convolution? 3Blue1Brown https://www.youtube.com/watch?v=KuXjwB4LzSA

Convolutional Neural Networks (CNNs)

Network Architecture

Convolutional Layer, Pooling Layer, Fully Connected Layer



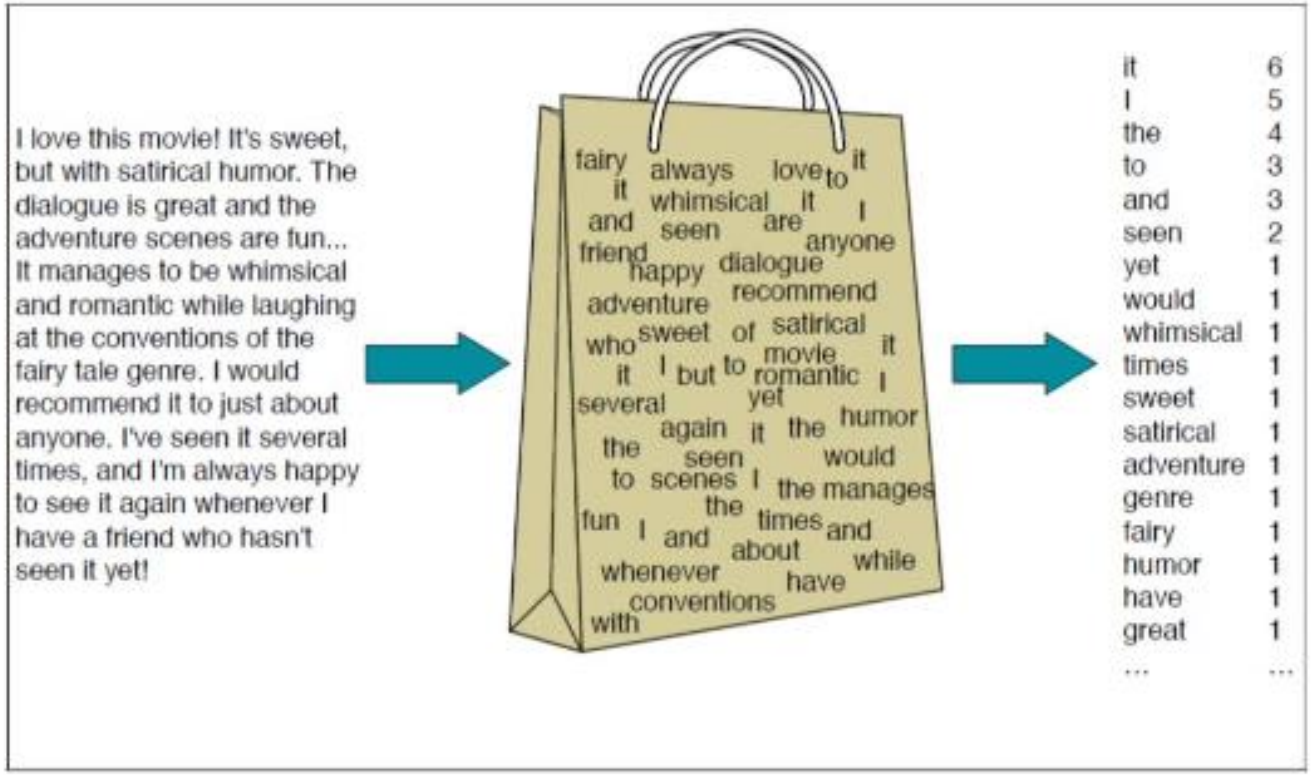
CLASSIFICATION

Let h(x) be the “true” mapping. We never know it. How do we find the best $\hat{h}(x)$ to approximate it?

One option: rule based

if x has characters in unicode point range 0370-03FF:
$\hat{h}(x)$ = greek

Bag of Words Representation



Jurafsky & Martin slides

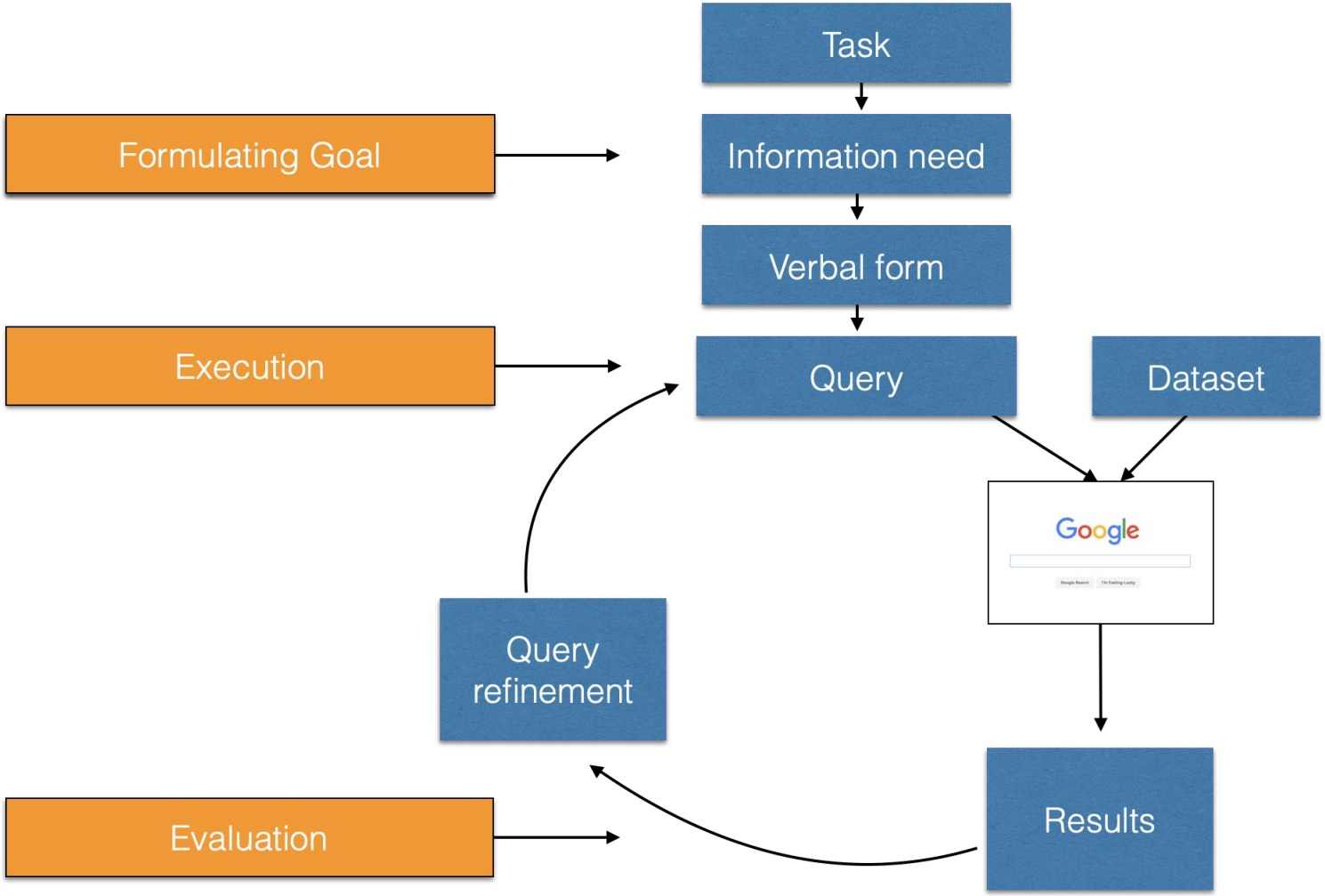
How Search Engines Work

Three main parts:

i. Gather the contents of all web pages (using a program called a **crawler** or **spider**)

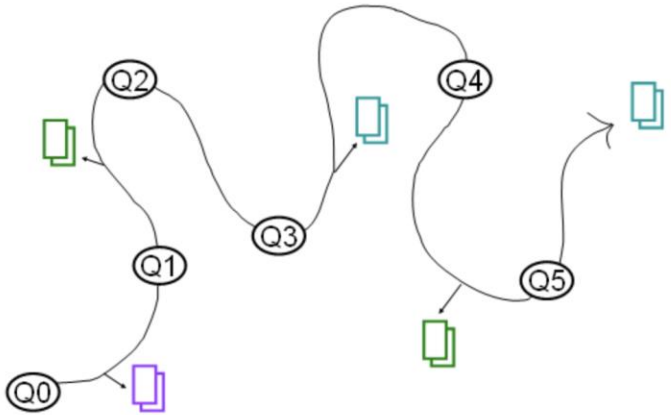
ii. Organize the contents of the pages in a way that allows efficient retrieval (**indexing**)

iii. Take in a query, determine which pages match, and show the results (**ranking** and **display** of results)


Task
Formulating Goal
Information need
Verbal form
Execution
Query
Dataset
Google
Query refinement
Results
Evaluation

Berry Picking Model

- Users' needs (and queries) change as a result of the search process
- Goals change in priority
- Information needs are not satisfied by finding single document; the trail of information is what's important.



Bates 1989

Main value is accumulated over the course of the search

How Does the Web Work?

sylvain kalache at wikicommons

Say a user named Oski using his computer at home (or in, say, Seoul) wants to find information about i202?

What happens when he:

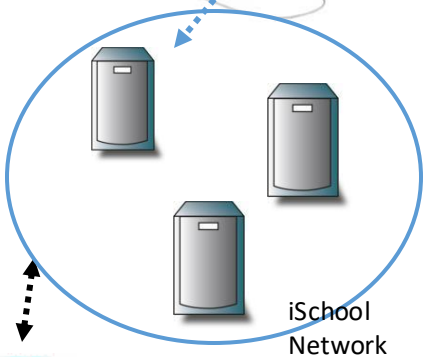
Brings up a search engine home page?
Types his query?

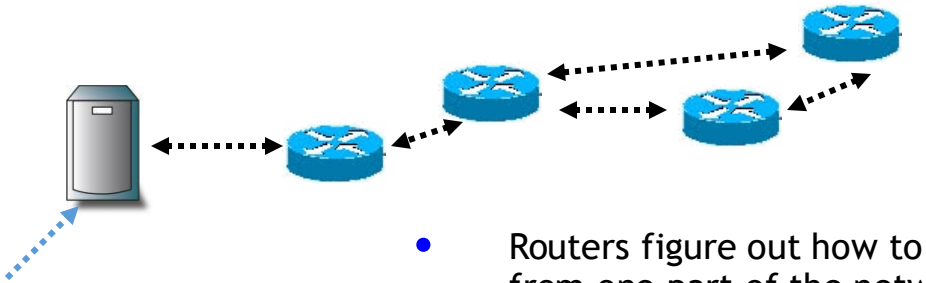
First, we have to understand how the WWW works!

Then we can understand search engines.

Routing Between Computers

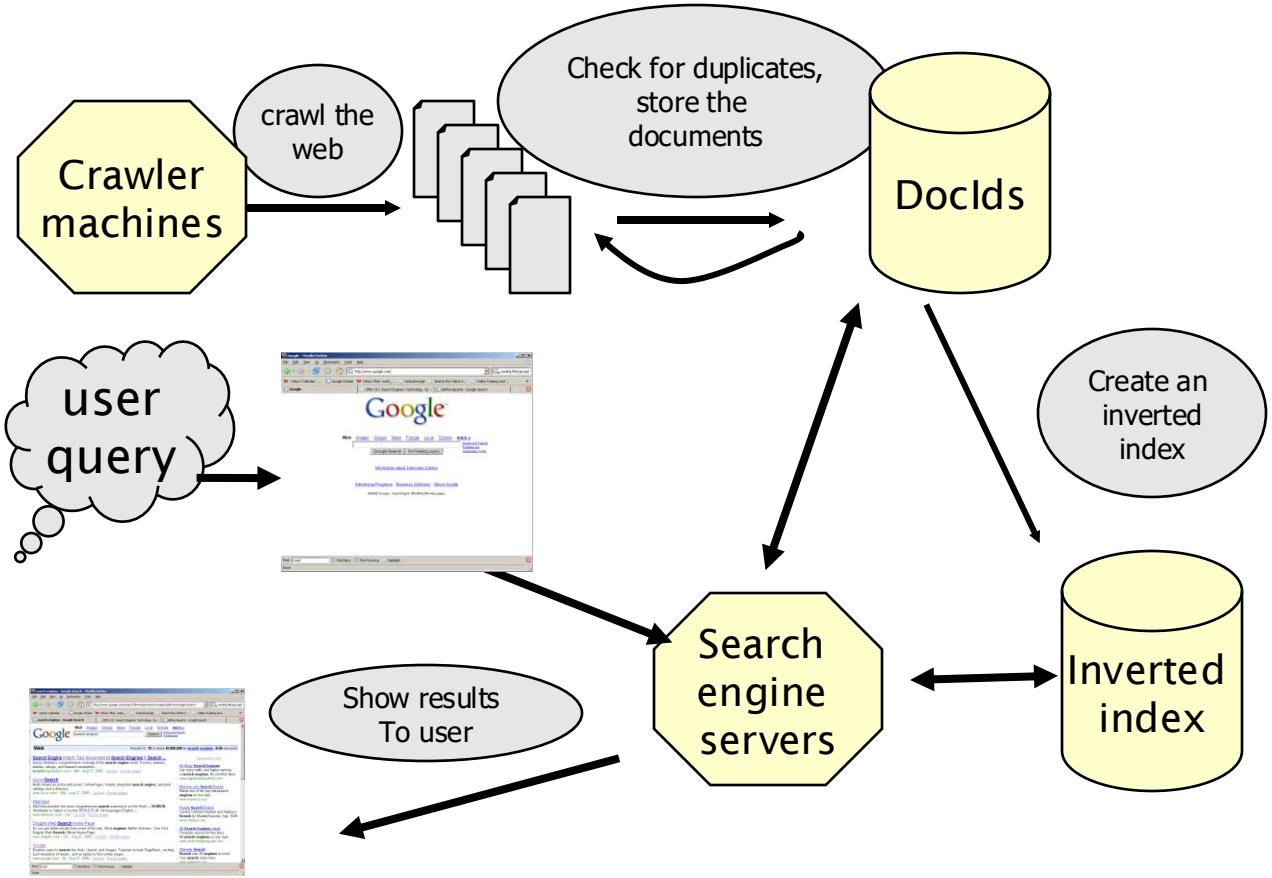
- How does the computer at Oski's desk figure out where the i202 web pages are?
- In order for him to use the WWW, Oski's computer must be connected to another machine acting as a web server (via his ISP).
- This machine is in turn connected to other computers, some of which are **routers**.





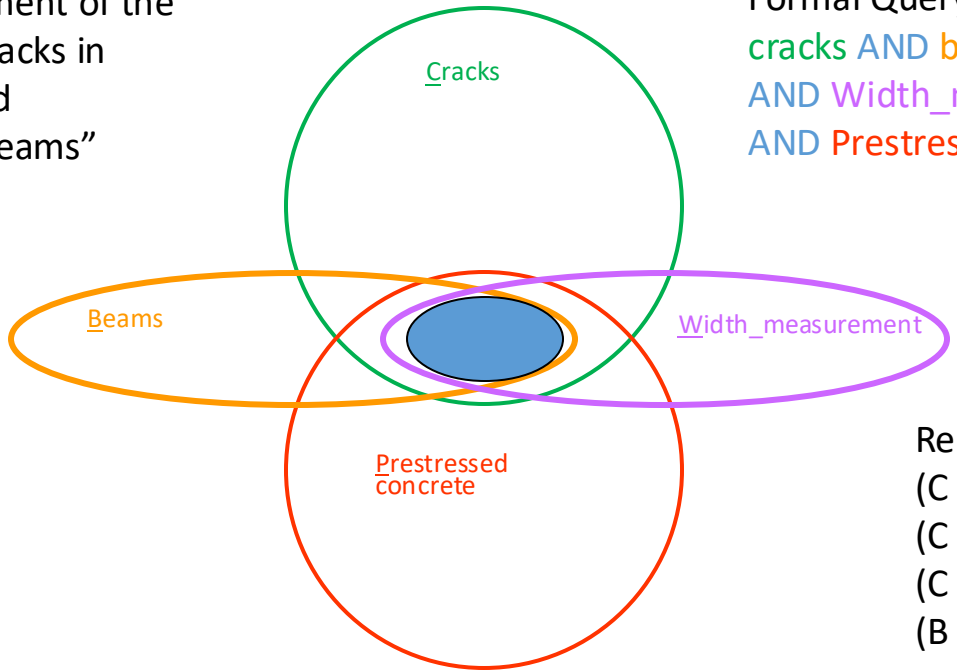
- Routers figure out how to move information from one part of the network to another.
- There are many different possible routes.

STANDARD WEB SEARCH ENGINE ARCHITECTURE



Converting Natural Language to Boolean

"Measurement of the width of cracks in prestressed concrete beams"



Formal Query:
cracks AND beams
AND Width_measurement
AND Prestressed_concrete

Relaxed Query:
(C AND B AND P) OR
(C AND B AND W) OR
(C AND W AND P) OR
(B AND W AND P)

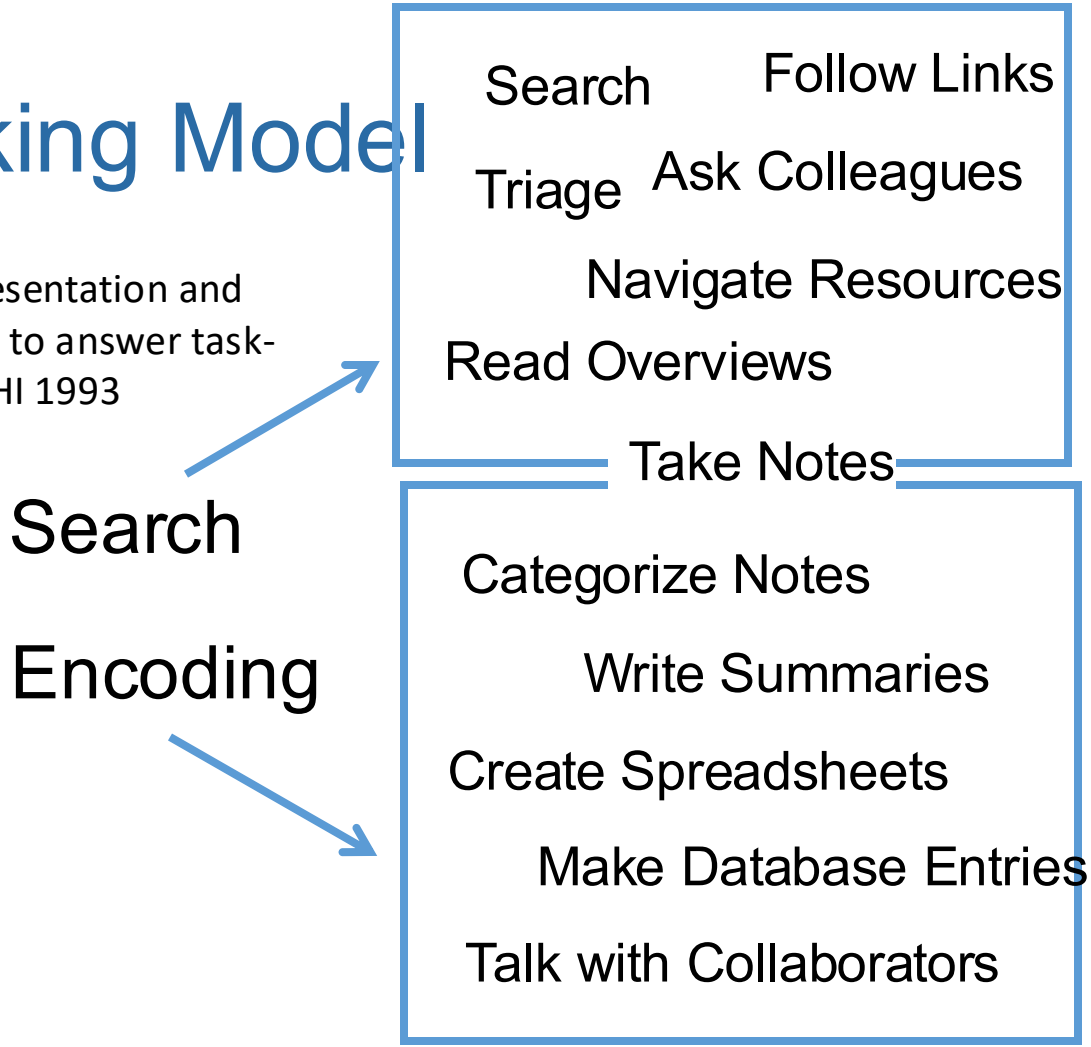
Foraging Theory

ANIMAL FORAGING		INFORMATION FORAGING
Food	**Goal**	Information
A site containing one or more potential sources of food	**Patch**	A website (or other source of information)
Search for food	**Forage**	Search for information
The animal's assessment of how likely it is that a given patch will provide food	**Scent**	How promising a potential source of information appears to the user
The totality of food types that an animal may consider in order to satisfy hunger	**Diet**	The totality of the information sources that a user may consider in order to satisfy an information need

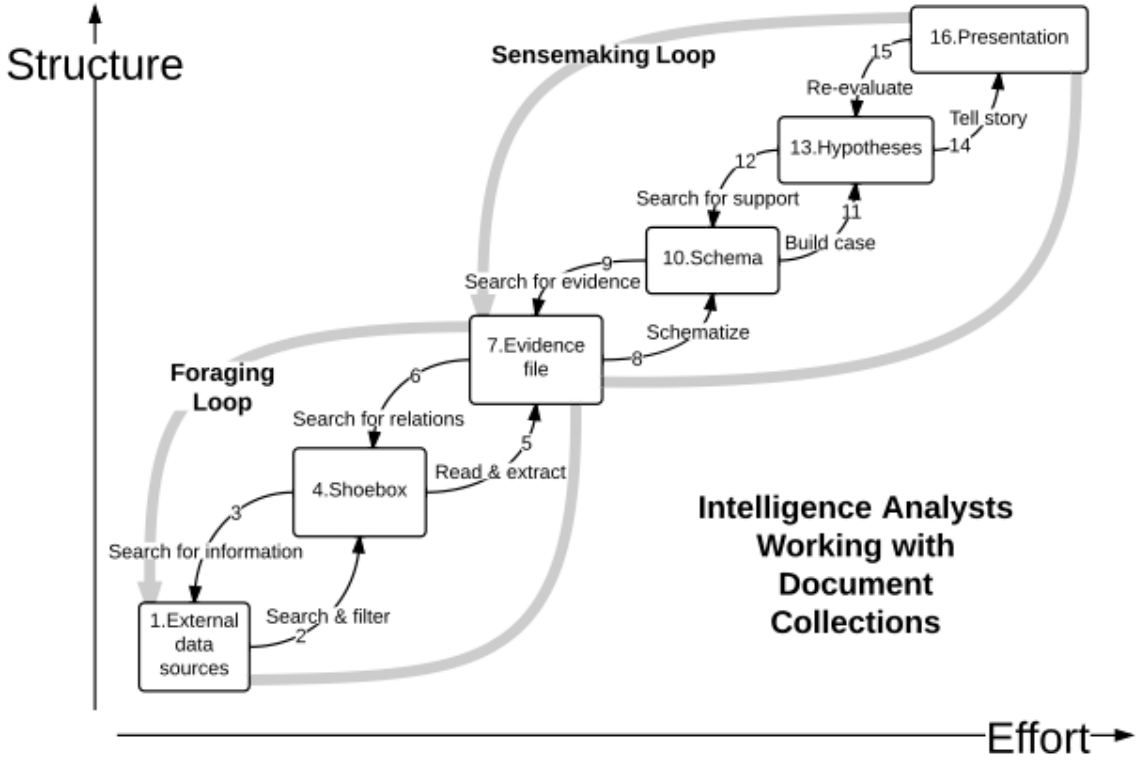
https://www.nngroup.com/articles/information-foraging/

The Sensemaking Model

"The process of searching for a representation and encoding data in that representation to answer task-specific questions." – Russell et al. CHI 1993



Intelligence Analysts' SenseMaking Loop



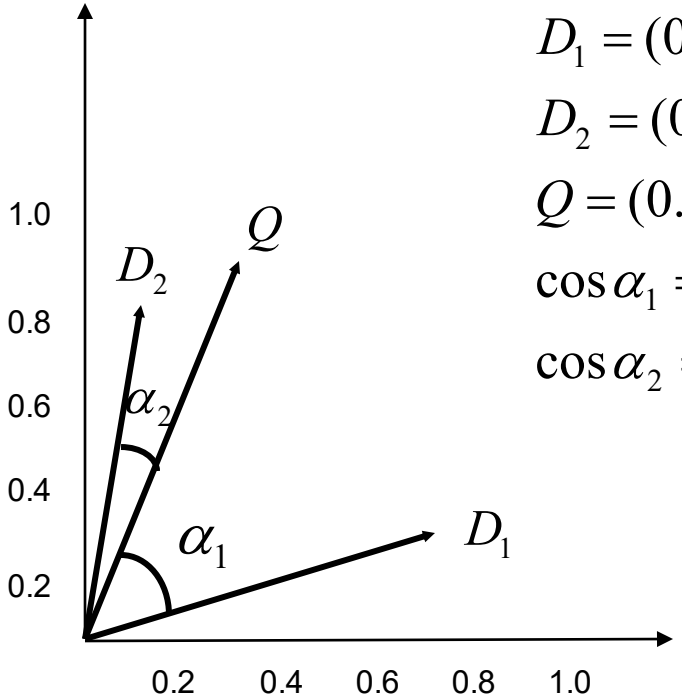
Pirolli & Card, The Sensemaking Process and Leverage Points for Analyst Technology as Identified Through Cognitive Task Analysis ,PICIA, 2005

Word Frequencies have Long Tails

Zipf's Law

Zdeněk Macháček on Unsplash

Computing Similarity Scores to Determine Which Documents are Close to the Query



$$D_1 = (0.8, 0.3)$$

$$D_2 = (0.2, 0.7)$$

$$Q = (0.4, 0.8)$$

$$\cos\alpha_1 = 0.74$$

$$\cos\alpha_2 = 0.98$$

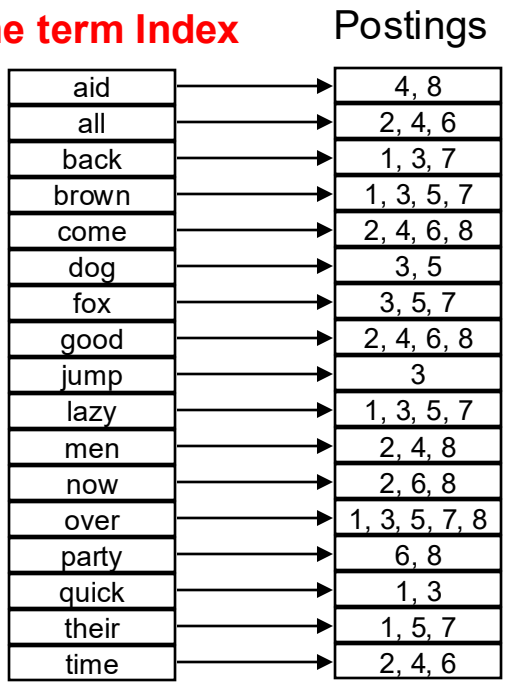
TF.IDF EXAMPLE

	tf				*idf*	$W_{i,j}$			
	1	2	3	4		1	2	3	4
complicated			5	2	0.301			1.51	0.60
contaminated	4	1	3		0.125	0.50	0.13	0.38	
fallout	5		4	3	0.125	0.63		0.50	0.38
information	6	3	3	2	0.000				
interesting		1			0.602		0.60		
nuclear	3		7		0.301	0.90		2.11	
retrieval		6	1	4	0.125		0.75	0.13	0.50
siberia	2				0.602	1.20			

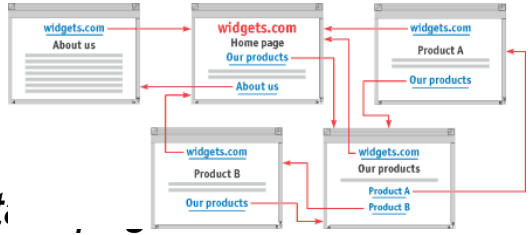
$$w_{i,j} = \mathrm{tf}_{i,j} \cdot \log \frac{N}{n_i}$$

Columns are document numbers, N=4; log base 10

Decoupling the Inverted Index



Measuring the Importance of Linking



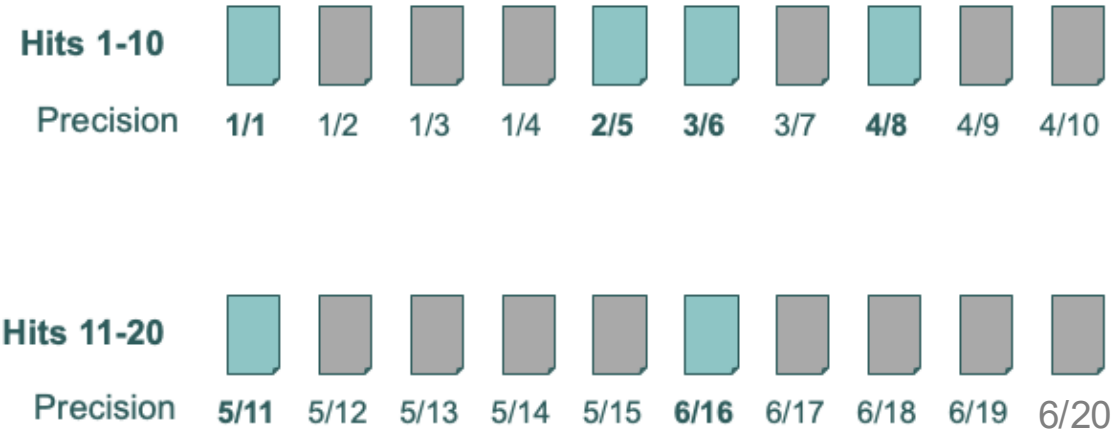
PageRank Algorithm

- *Idea: important pages are pointed to by other import*
- *Method:*
  - Each link from one page to another is counted as a “vote” for the destination page
  - But the importance of the starting page also influences the importance of the destination page.
  - And those pages scores, in turn, depend on those linking to them.

Image and explanation from http://www.economist.com/science/tq/displayStory.cfm?story_id=3172188

Measuring Precision and Recall

Assume there are a total of 14 relevant documents



Language Modeling Training in Action

TheUpshot

Watch an A.I. Learn to Write by Reading Nothing but <&$"<-]{,^\&

By Aatish Bhatia April 27, 2023

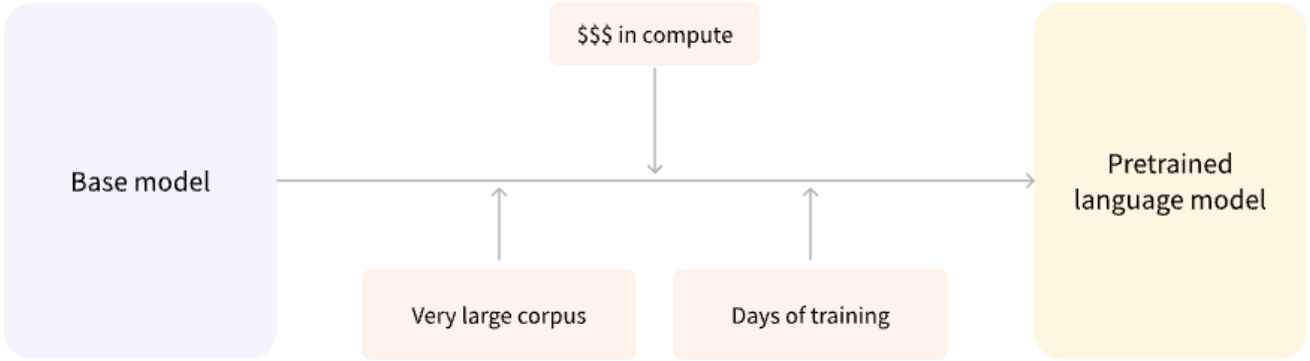
What this neural network actually generates is not letters but probabilities. (These probabilities are why you get a different answer each time you generate a new response.)

For example, when given the letters stai, it'll predict that the next letter is n, r or maybe d, with probabilities that depend on how often it has encountered each word in its training.

https://www.nytimes.com/interactive/2023/04/26/upshot/gpt-from-scratch.html

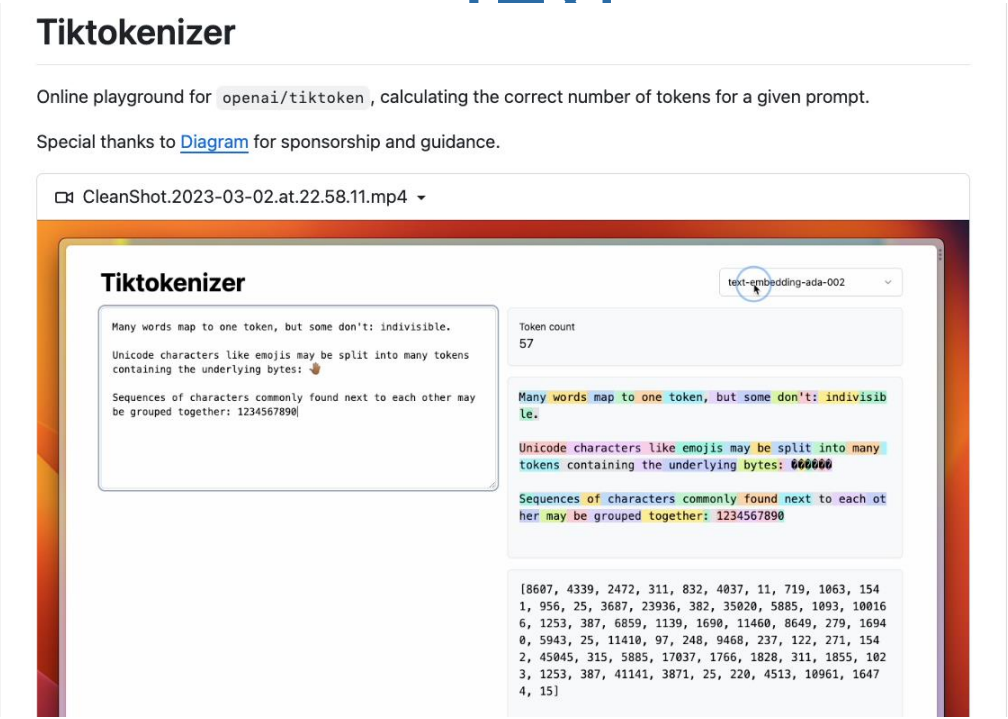
LLM Pretraining

Pretraining is the act of training a model from scratch: the weights are randomly initialized, and the training starts without any prior knowledge.



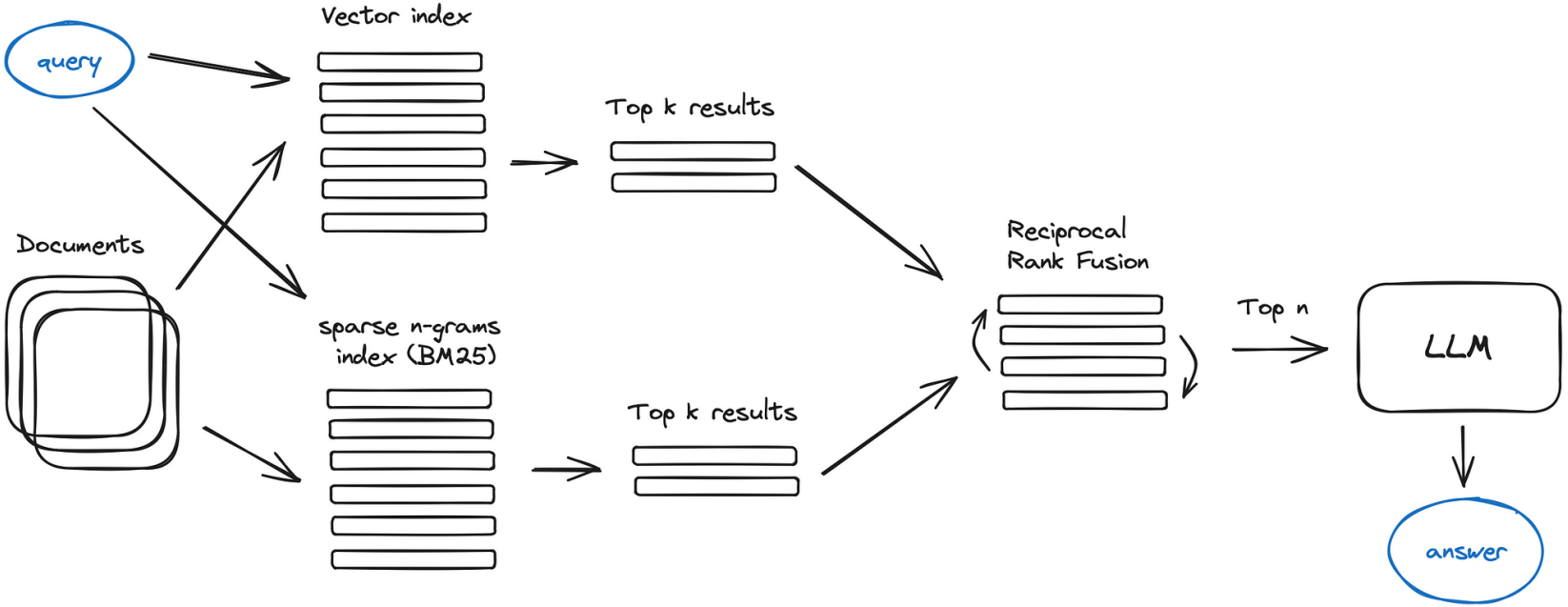
https://huggingface.co/learn/llm-course/en/chapter1/4

Exercise: See How LLMs Tokenize Text



https://tiktokenizer.vercel.app/

Hybrid Search Uses both Dense and Sparse (BM-25) Vectors



https://pub.towardsai.net/advanced-rag-techniques-an-illustrated-overview-04d193d8fec6

This article is more than **2 years old**

I thought I was immune to being fooled online. Then I saw the pope in a coat

Joel Golby

An encounter with an AI-generated image of his holiness has changed me: I now have sympathy for credulous baby boomers

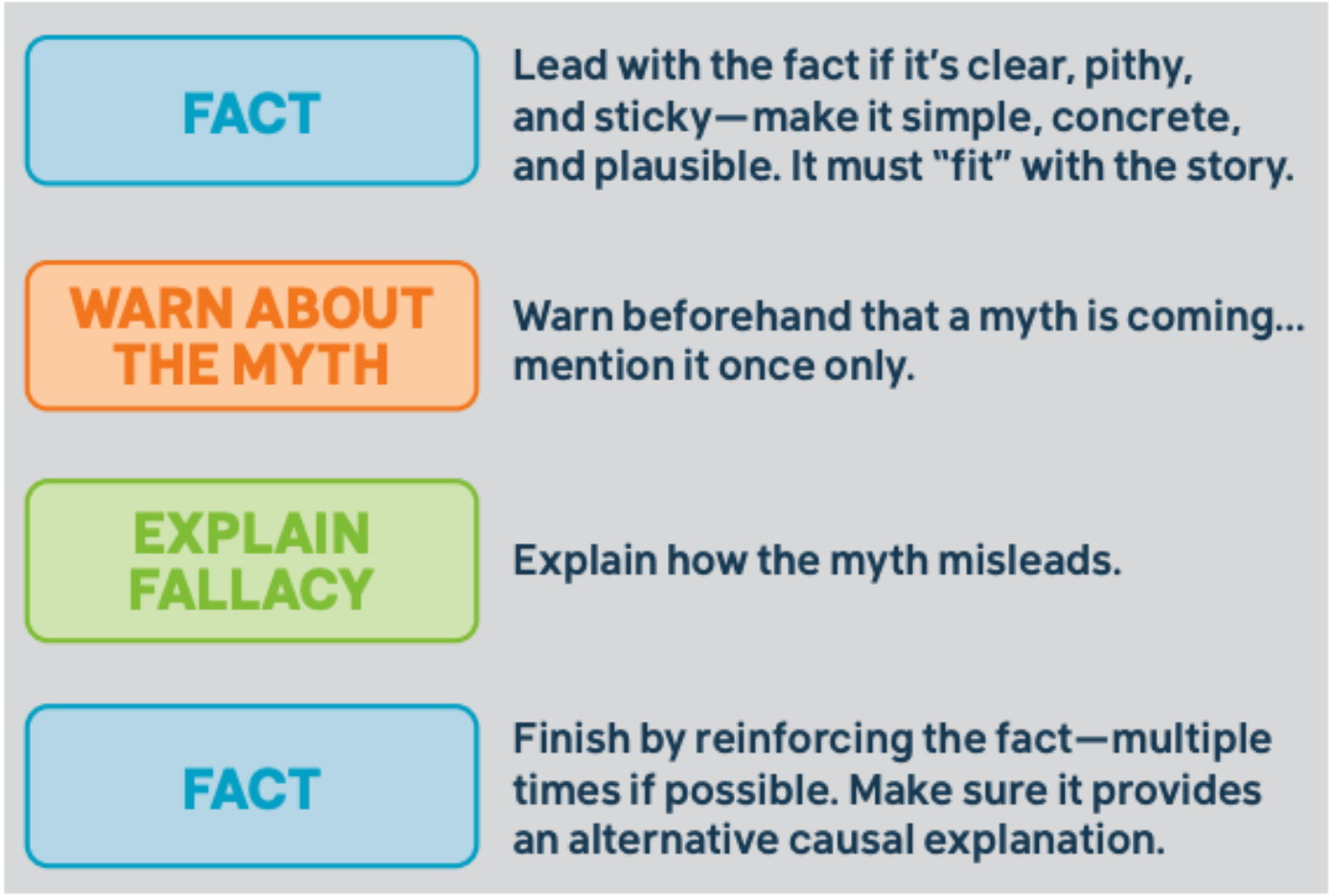
'I thought wearing a really big coat and looking like a Metal Gear Solid 2 boss battle might have been part of his ongoing cool guy shtick. Lord, forgive me.' Photograph: Reddit

https://www.theguardian.com/commentisfree/2023/mar/27/pope-coat-ai-image-baby-boomers

Strategies For Combating Misinfromation

- Standard approach: debunking
  - Can help, but there are barriers
  - Can backfire, but there is increasing evidence that this is not as much of an issue as once thought
- Newer, more successful approach: **inoculation**
  - Expose people to weakened versions of the misinformation
  - Slowly build up cognitive resistance
  - Prophylactic vs Therapeutic (for pre-existing views)

Lewandowsky et al., The Debunking Handbook 2020


FACT
Lead with the fact if it's clear, pithy, and sticky—make it simple, concrete, and plausible. It must "fit" with the story.
WARN ABOUT THE MYTH
Warn beforehand that a myth is coming... mention it once only.
EXPLAIN FALLACY
Explain how the myth misleads.
FACT
Finish by reinforcing the fact—multiple times if possible. Make sure it provides an alternative causal explanation.

What is Fairness?

What does it mean for an algorithm/system to be “fair” or “unfair”?

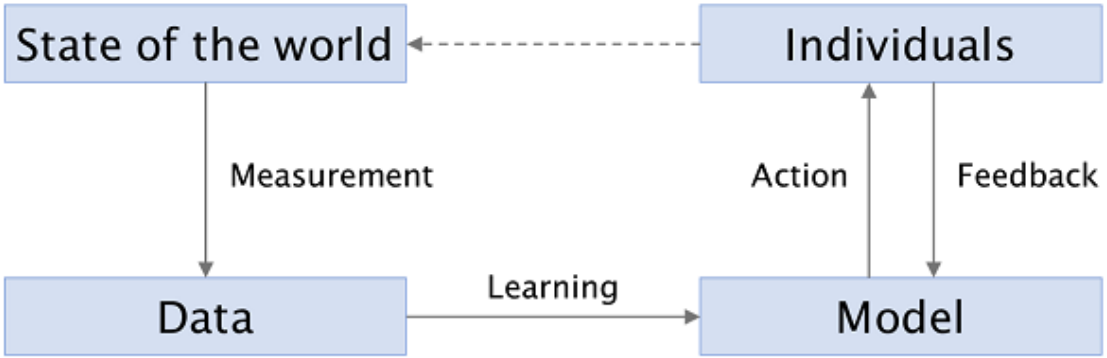
Keywords in Titles of Accepted Full Papers

(SIGIR is a Traditional Technical Conference, Core IR Research)

- SIGIR'17: Fairness: 0 Bias: 0
- SIGIR'18: Fairness: 1 Bias: 2
- SIGIR'19: Fairness: 0 Bias: 1
- SIGIR'20: Fair(ness): 4 Bias: 1
- SIGIR'21: Fair(ness): 7 Bias: 4

*excluding the statistical sense of "bias"

Machine Learning Feedback Loop



Do we have to learn from data? Why do we?

Do we have to learn from human's reactions to models? Why do we?

Fix Bias in the Data Stage of the Pipeline

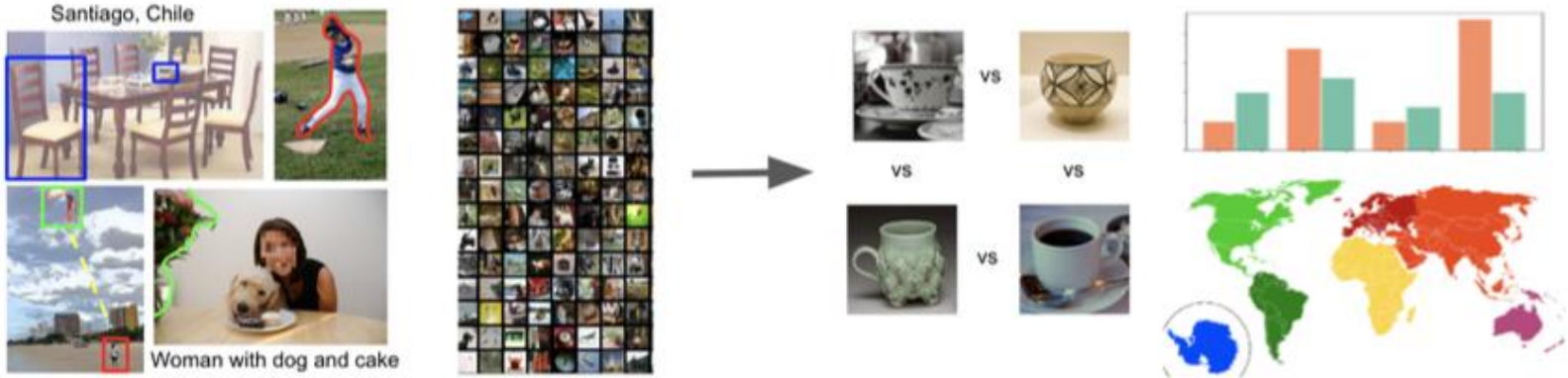


Fig. 1: Our tool takes in as input a visual dataset and its annotations, and outputs metrics, seeking to produce insights and possible actions.

Wang et al., "REVISE: A tool for measuring and mitigating bias in visual datasets." *European Conference on Computer Vision*. 2020.

That's a Wrap!

Thank you, Sunny and Sarah!

Course Evaluations

Please fill them out asap!