

FastCC
Taming the Commons Category Tree

[[User:Dschwen | Daniel Schwen]]



Commons category tree?

Is it really a tree?

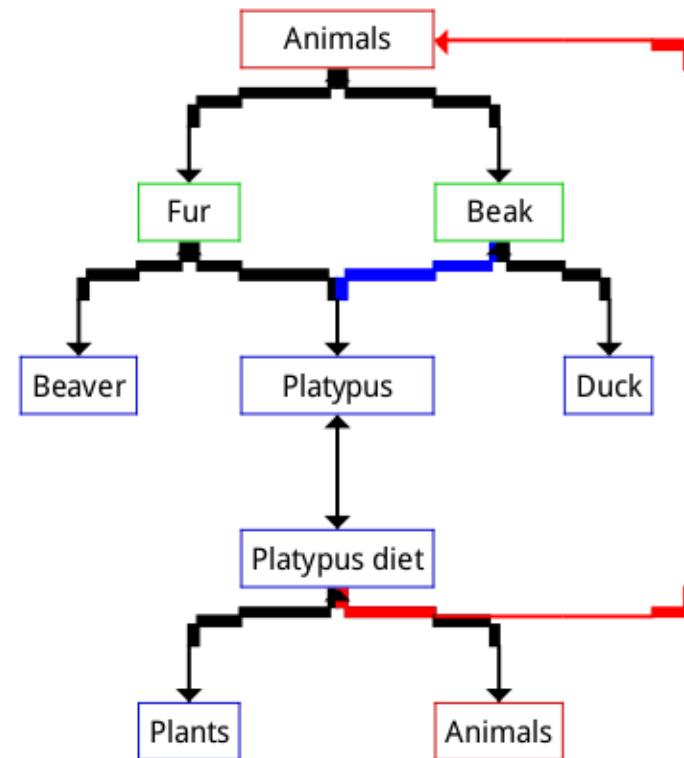
- Made-up example
- Diamond relations
- It gets worse...
- Loops!

It is a **directed graph**!

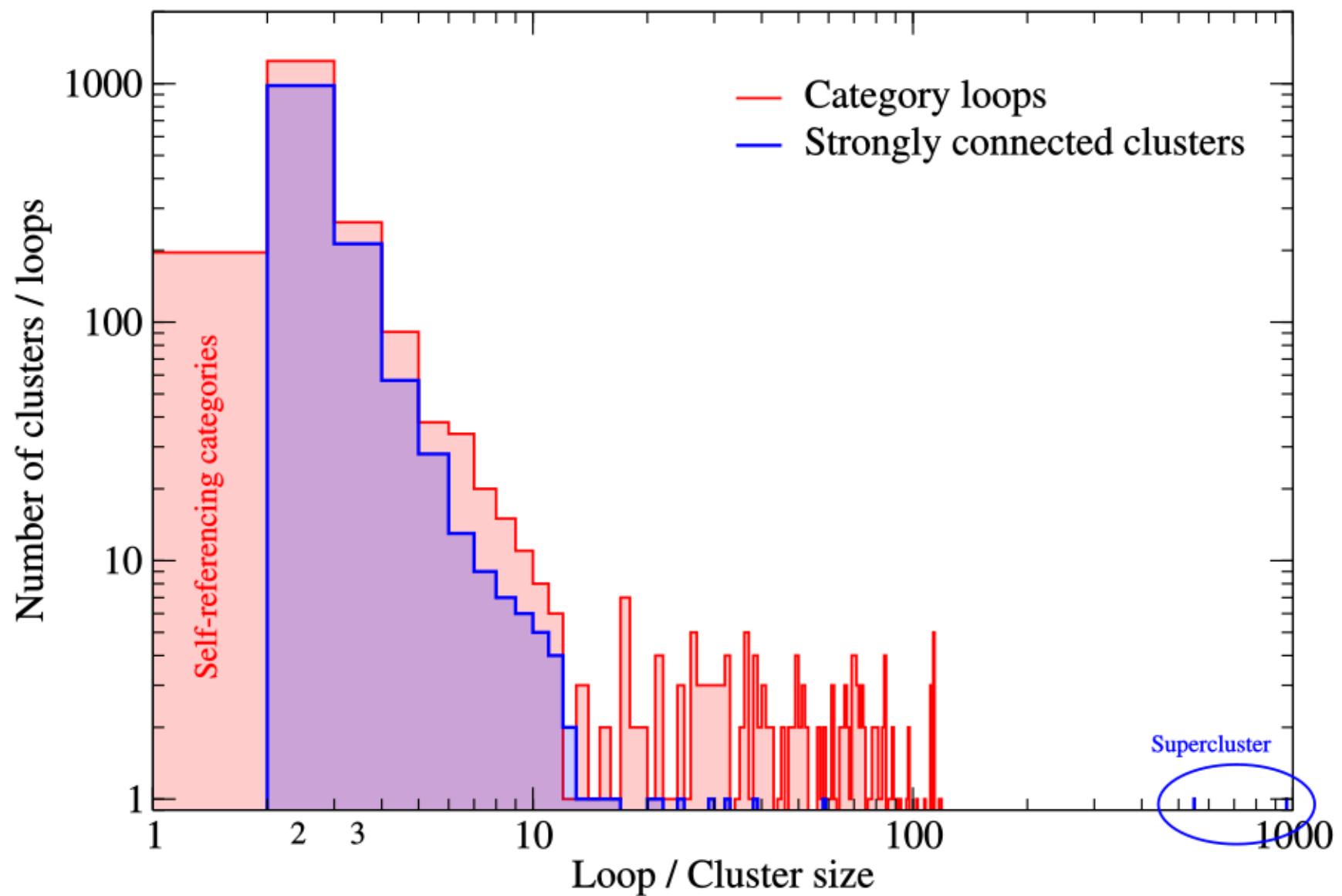
- Let's be careful with the arrow direction
- Well-defined element relation "*is subcategory of*"

The actual meaning is very loosely defined:

- **Hyponymy:** a sort/kind/type of...
- **Meronymy:** a part of..., a member of...
- **Attributes:** color, shape, size, nationality, technique, quality...
- **Location:** where, in..., from... (place/event, place/building, place/people, country/language...)
- **Timing:** when (time/event, time/depicted situation, time of birth, inception or construction...)
- **Agentive and influence relations:** creator/work, device/product, company/product...
- **Modification:** original/modified or modified/original (avoid cyclic structure)



Loops!



Bundle similar images



Media in category "Mister Softee trucks in Hong Kong"

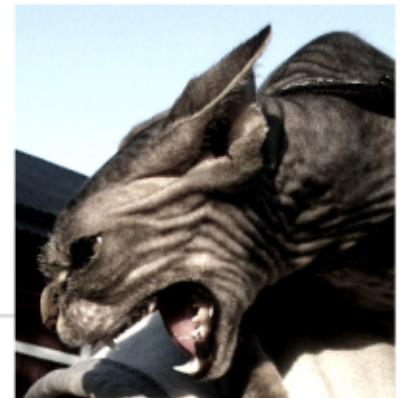
The following 14 files are in this category, out of 14 total.

						
HK MisterSoftee GoldenComputerArcade, MaOnShan.JPG [Copy Vio] 1.68 MB	HK MisterSoftee MaOnShan.JPG [Copy Vio] 285 KB	HK MisterSoftee Mongkok NullahRd.JPG [Copy Vio] 303 KB	HK MisterSoftee SaiKung.JPG [Copy Vio] 306 KB	HK MisterSoftee ShatinMTR.JPG [Copy Vio] 1.5 MB	HK MisterSoftee TST.jpg [Copy Vio] 768 KB	HK MisterSoftee VictoriaPark.JPG [Copy Vio] 1.72 MB
						
HK MobileSoftee CentralPiers.JPG [Copy Vio] 2.69 MB	HK MobileSoftee FoTanHeadquarters.JPG [Copy Vio] 1.31 MB	HK MobileSoftee SaiKung.JPG [Copy Vio] 2.58 MB	Mister Softee HK1.jpg [Copy Vio] 515 KB	Mister Softee HK2.jpg [Copy Vio] 142 KB	StarHouse.jpg [Copy Vio] 287 KB	Temple Street 9, Hong Kong, Mar 06.JPG [Copy Vio] 536 KB

Category:Mister Softee trucks in Hong Kong

"So happy smiling cat" by Yumi Kimura from Yokohama, JAPAN. Licensed under CC BY-SA 2.0 via Wikimedia Commons.

Categories hide content



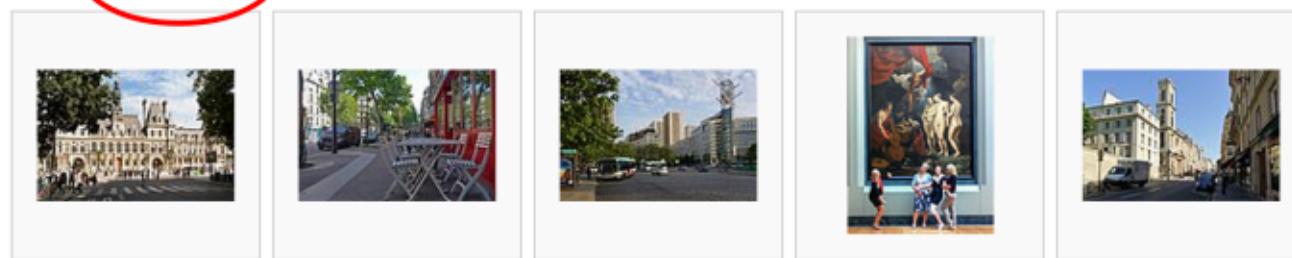
Pages in category "Paris"

This category contains only the following page.

- [Paris](#)

Media in category "Paris"

The following 5 files are in this category, out of 5 total.



[Avenue Victoria and Hotel de Ville, Paris 2014.jpg](#) [Copy/Via]

1.04 MB

[Dans les rues de Paris, 2014.jpg](#) [Copy/Via]

736 KB

[Place d'Italie, Paris 2014.jpg](#) [Copy/Via]

876 KB

[The Education of the Princess in the Louvre, Paris 2014.jpg](#) [Copy/Via]

1.87 MB

[Église Saint-Jacques-du-Haut-Pas, Rue Saint-Jacques, Paris 2014.jpg](#) [Copy/Via]

659 KB

[Categories \(+‡\): Cities and villages in France by department \(–\) \(±\) \(↓\) \(↑\)](#) | [Cities and villages in Île-de-France \(–\) \(±\) \(↓\) \(↑\)](#) | [Paris urban area \(–\) \(±\) \(↓\) \(↑\)](#)
Non-topical/index: [Categories with a gallery for a better choice of sub-categories \(–\) \(±\) \(↓\) \(↑\)](#)

Diffusion into countless sub-categories makes discovering images hard.

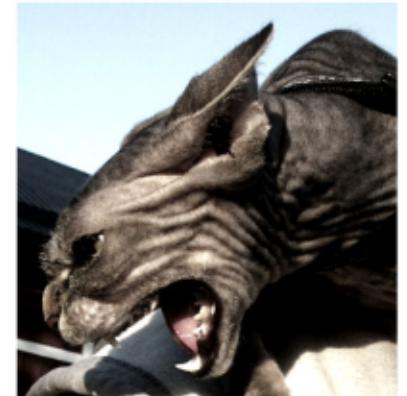
Categories can encode complex relations

- More than just a simple **and**:
 - Category:Star symbols **on top of** objects
 - Category:People **inside** flowers
Category:People **holding** flowers
Category:People **with** flowers
 - Category:Nandi sculptures **outside** India
6 levels below **Category:India**
 - Category:Sundials **without** sun
5 levels below **Category:Sun**



"So happy smiling cat" by Yumi Kimura from Yokohama, JAPAN. Licensed under CC BY-SA 2.0 via Wikimedia Commons.

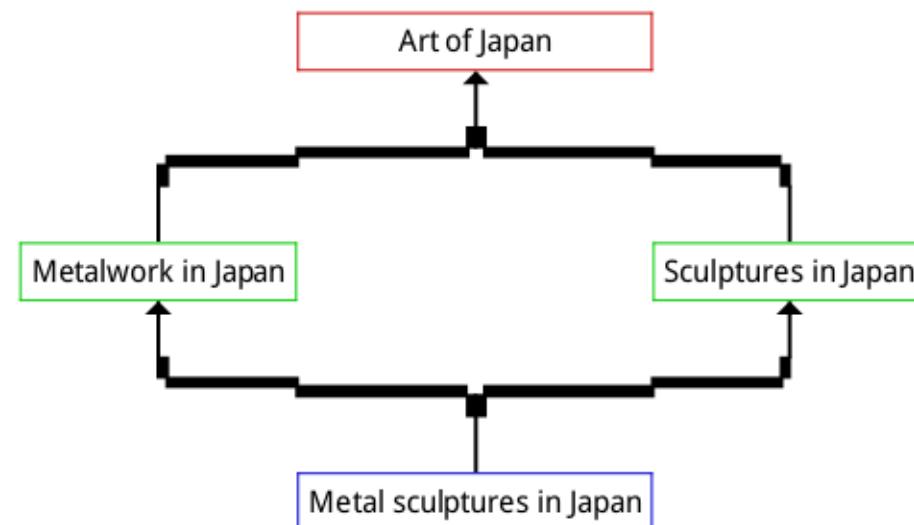
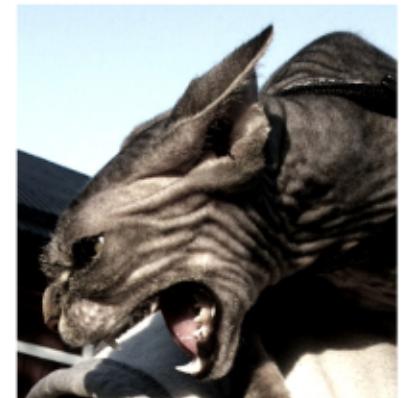
Categories *do* encode *complex* relations



Category: Demographic maps of 15-17 year old dependent children whose fathers speak another language and did not state proficiency in English and whose mothers speak another language and speak English not well or not at all in Australia by state or territory

And this monster has **sub-categories!**

Some categories are natural intersections



= Japan & Metalworks & Sculptures

2.850.563 category diamonds found (not all permutations!)

`./fastcci_diamond(<3s)`

"AngrySphynx" by User:Trilobite2. Licensed under CC BY-SA 3.0 via Wikimedia Commons.

What about using tags?

- Tags or "Topics" are commonly used in popular Photo communities (Flickr, 500px)
- No hierarchy
- We may get tags (through WikiData), however:
 - Don't hold your breath!
 - Community consensus?
 - Additional maintenance burden?
 - Can not replace all Category applications
 - Won't solve **today's** problems

Category intersection

- Treat categories as Tags
- Certainly works for higher level categories
Kittens, Wyoming, Bicycles, Green
- Consider all images in and below all sub-categories
- Meaning
 - All images which both labels apply to
- Problem
 - Higher up categories loose meaning
the further "down" in the graph we go

The longest Category chain

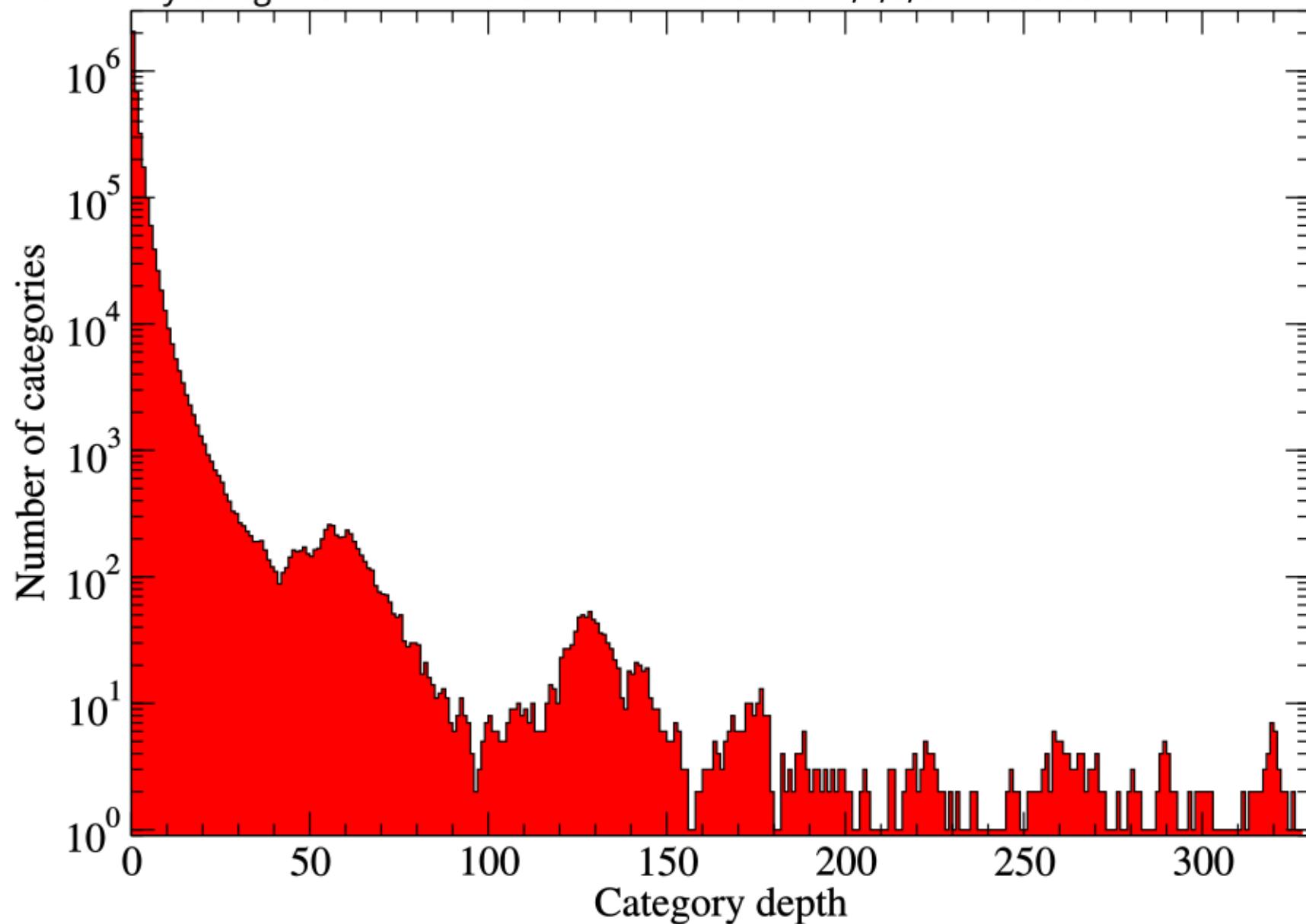
CommonsRoot → Commons → Commons maintenance content → Categories requiring diffusion → Categories requiring permanent diffusion → Science → Scientific disciplines →
 Interdisciplinary fields → Science studies → Knowledge → Academia → Academic disciplines → Engineering → Engineering by type → Mechanical engineering → Manufacturing → Resources
 → People → People by medium → Portraits → Portraits by model → Portraits of animals → Animal heads → Mammal heads → Primate heads → Hominidae heads → Human heads →
 Human faces → Facial expressions → Anger → Violence → Fighting → Fights → Martial arts → Martial arts by type → Archery → Arrows → Arrow symbols → Vectors → Vector mathematics
 → Direction → Velocity → Acceleration → Movement recording → Video and movie technology → Computer graphics → Computer graphics by type → Animations → Animations by subject
 → Medical animations → Computer generated images → Digital art → Computer art → Digital image processing → Algorithms → Numerical analysis → Finite difference method → Grids →
 Lattices → Square lattices → Magic squares → Collaborative work → VIAF inter project linking → Categories with authority control data → Electronic circuits → Circuit diagrams → Electronic
 diagrams → Waveforms → Transverse wave → Electromagnetic radiation → Optics → Optical technology → Photography → Photographs → Photographs by topic → Views → Views by
 subject → Landscapes → Geomorphology → Landforms → Mountains → Hills → Hillslopes → Hillslope processes → Mass movements → Fluid flows → Travel → Migration → Human
 migration → Humanitarian aid → Welfare → Health → Health problems → Environmental problems → Endangered species → Species by IUCN Red List category → IUCN Least Concern
 species → Homo sapiens → Human behavior → Body language → Bodily functions → Thinking → Abstraction → Abstract art → Art by genre → Architecture → Structures → Construction →
 Infrastructure → Infrastructure by function → Communication infrastructure → Communication technology → Digital technology → Computers → Computer code → Computer programs
 → Software → Software engineering → Data modeling → Ontology → Life → Biology → Behavior → Ethology → Learning → Critical thinking → Belief → Philosophy → Philosophy of science
 → Cognitive science → Mind → Consciousness studies → Philosophy of mind → Cognition → Intelligence → Creativity → Art → Humanities and art → Linguistics → Language → Vocabulary
 → Words → Verbs → Design → Modelling → Scientific modeling → Information visualization → Graphics → Symbols → Allegories → Personifications → Virtues → Allegories of virtues →
 Truth → Logical truth → Logical consequence → Deductive reasoning → Theories of deduction → Formalism (deductive) → Formal sciences → Classification systems → Taxonomy → Tree
 structures → Tree diagrams → Phylogenetic trees → Phylogenetic tree of life → WikiProject Tree of Life → Taxon categories → Neomura → Eukaryota → Opisthokonta → Animalia →
 Eumetazoa → Bilateria → Deuterostomia → Chordata → Vertebrata → Gnathostomata → Tetrapoda → Mammalia → Theria → Eutheria → Primates → Haplorrhini → Simiiformes →
 Catarrhini → Hominoidea → Hominidae → Homo → Homo in art → People in art → Figures → Body → Animal body → Zootomy → Zootomy by taxon → Mammal anatomy → Primate
 anatomy → Hominidae anatomy → Human anatomy → Organs of the human body → Human body organs in art → Human heads in art → Human faces in art → Faces of men in art →
 Beards in art → Saint Peter → Places named after Saint Peter → Saint Petersburg → People of Saint Petersburg → Alfred Nobel → Nobel Prize → Nobel Prize by discipline → Nobel Peace
 Prize → Nobel Peace Prize laureates → European Union → Members of the European Union → Denmark → Geography of Denmark → Nature of Denmark → Water in Denmark → Bodies of
 water in Denmark → Baltic Sea → Countries bordering the Baltic Sea → Middle Europe → Countries of Middle Europe → Slovenia → Geography of Slovenia → Nature of Slovenia → Water in
 Slovenia → Bodies of water in Slovenia → Adriatic Sea → Adriatic Sea Basin → Italian Peninsula → Italy → Geography of Italy → Seas of Italy → Mediterranean Sea → History of the
 Mediterranean → Classical antiquity → Classics → Ancient history → Ancient peoples → Semitic peoples → Arabs → Arab world → Greater Middle East → Algeria → Geography of Algeria →
 Geomorphology of Algeria → Landforms of Algeria → Mountains of Algeria → Mountain ranges of Algeria → Atlas Mountains → Habitat → Territories → States and territories by year of
 establishment → States and territories established in the 2nd millennium → States and territories established in the 18th century → States and territories established in the 1770s → States
 and territories established in 1776 → United States → Society of the United States → Education in the United States → Science and technology in the United States → Geology of the United
 States → Geomorphology of the United States → Landforms of the United States → Coasts of the United States → Coastal geomorphology of the United States → Islands of the United States
 → Islands of the United States by state → Islands of New York → Islands of New York City → Islands of Manhattan → Liberty Island → Statue of Liberty → People associated with the Statue of
 Liberty → François Mitterrand → Things named after François Mitterrand → Quai François-Mitterrand (Paris) → Palais du Louvre → Musée du Louvre → Collections of the Louvre →
 Antiquities in the Louvre → Greek antiquities in the Louvre → Greek antiquities in the Louvre - Nike of Samothrace stair → Nike of Samothrace → Copies of Nike of Samothrace → Prix Ars
 Electronica → Prix Ars Electronica winners → Wikipedia → Wikipedia community → Wikipedians → Wikipedians by continent → Wikipedians in Europe → Wikipedians in Poland →
 Wikimedia meetups in Poland → Polish Wikiexpeditions → Wikiekspedycja 2011 → Places visited during Wikiekspedycja 2011 → Sanok → Geography of Sanok → Places in Sanok → Visitor
 attractions in Sanok → Architecture of Sanok → Buildings in Sanok → Buildings in Sanok by function → Military buildings in Sanok → Barracks in Sanok → Former Austro-Hungarian Army
 barracks in Sanok → Old military barracks in Olchowce, Sanok 1876-1989 → Galizisches Ulanen Regiment "von Böhm-Ermolli" Nr. 13 → Konfederatka → November Uprising → People of
 the Polish-Russian War 1830-1831 → November Uprising participants → Juliusz Słowacki → Monuments and memorials to Juliusz Słowacki → Things named after Juliusz Słowacki → Juliusz
 Słowacki streets → Słowackiego streets in Poland → Juliusza Słowackiego Street in Gdańsk → Gdańsk Lech Wałęsa Airport → Aircraft at Gdańsk Lech Wałęsa Airport → Aircraft at Gdańsk Lech
 Wałęsa Airport by aircraft type → Airbus aircraft at Gdańsk Lech Wałęsa Airport → Airbus A320 at Gdańsk Lech Wałęsa Airport → Airbus A320 of Wizz Air at Gdańsk Lech Wałęsa Airport

A metric to weigh intersection results:

Path lengths in the graph

Subcategory depth

How many categories do we have on commons with 1,2,3,..n levels below?



Ye olde intersection (The SQL way) = recursive queries = painfully slow!

FastCCI

Category intersection needs...

- ...to be **fast**
- ...an easy interface
 - shallow learning curve
 - powerful options available

FastCCI's Killer Application

- Parallel category trees for ★ FP, ● QI, ○ VI
 - ★ Category:Featured picture
640 × Category:Featured pictures of...
 - ● Category:Quality image
1825 × Category:Quality images of...
 - ○ Category:Valued image
578 × Category:Valued Images of...
- Ideal target for intersection
- Easy interface → one button!

FastCCI Interface

Screenshot of the Wikimedia Commons "Category:Idaho" page, showing the FastCCI interface.

The page title is "Category:Idaho". The main content area displays the following information:

- States of the United States:** Alabama, Alaska, Arizona, Arkansas, California, Colorado, Connecticut, Delaware, Florida, Georgia, Hawaii, Idaho, Illinois, Indiana, Iowa, Kansas, Kentucky, Louisiana, Maine, Maryland, Massachusetts, Michigan, Minnesota, Mississippi, Missouri, Montana, Nebraska, Nevada, New Hampshire, New Jersey, New Mexico, New York, North Carolina, North Dakota, Ohio, Oklahoma, Oregon, Pennsylvania, Rhode Island, South Carolina, South Dakota, Tennessee, Texas, Utah, Vermont, Virginia, Washington, West Virginia, Wisconsin, Wyoming.
- District of Columbia:** Washington, D.C.
- Insular areas:** American Samoa, Guam, Northern Mariana Islands, Puerto Rico, Panama Canal Zone, United States Virgin Islands.

Below this text are two flags: the Idaho state flag and the United States national flag.

The text "English: Idaho is a state in the northwestern United States." is present.

A map of the United States highlights the state of Idaho in red.

Subcategories:

This category has the following 28 subcategories, out of 28 total.

A	G	P cont.
▶ Architecture of Idaho (7 C)	▶ Gardens and parks in Idaho (1 C)	▶ Photographs of Idaho (4 C, 1 F)
B	▶ Geography of Idaho (16 C, 38 F)	▶ Populated places in Idaho (4 C)
▶ Buildings in Idaho (32 C, 48 F)	▶ Geology of Idaho (9 C, 17 F)	▶ Postcards of Idaho (1 F)
C	▶ Government of Idaho (7 C, 3 F)	▶ Pronunciation of Idaho (2 F)
▶ Cities in Idaho (52 C, 2 P)	H	S
▶ Counties in Idaho (48 C)	▶ History of Idaho (21 C, 27 F)	▶ Sports in Idaho (7 C, 4 F)
▶ Culture of Idaho (10 C, 5 F)	M	▶ Stereo cards of Idaho (57 F)

FastCCI Interface

The screenshot shows a browser window displaying the 'Category:Idaho' page on Wikimedia Commons. The URL in the address bar is <https://commons.wikimedia.org/wiki/Category:Idaho>. The page title is 'Category:Idaho'. A sidebar on the left contains links for Main page, Welcome, Community portal, Village pump, Language select (set to English), Participate (Upload file, Recent changes, Latest files, Random file, Help, Contact us, Donate), Tools (In Wikipedia, Afrikaans, Aragonés, Ænglisc, العربية, Azərbaycanca, Беларуская, Беларуская (тарашкевіца), Гэл意志), and a 'Create account' and 'Log in' link at the top right.

The main content area displays a grid of nine images from the Idaho category. The images include a plane taking off, a mountain range, a lake, a landscape with mountains, a forest scene, and a river. A sidebar on the right, titled 'Good pictures', lists filters: All images, Featured pictures, Quality images, Valued images, In this category and in..., and In this category but not in... Each filter has a corresponding colored icon (star, green dot, yellow dot, red/purple dot, blue/purple dot).

FastCCCI Interface

The screenshot shows a web browser window displaying a Wikimedia Commons page. The URL in the address bar is https://commons.wikimedia.org/wiki/File:Glacier_Point_at_Sunset,_Yosemite_NP,_CA,_US_-_Diliff.jpg. The page title is "File:Glacier Point at Sunset, Yosemite NP, CA, US - Diliff.jpg". The breadcrumb trail indicates the file's path: Idaho → Geography of Idaho → Great Basin (USA) → Mountain ranges of the Great Basin → Sierra Nevada (United States) → Mountains of the Sierra Nevada (U.S.) → Mountains in Yosemite National Park → Half Dome. The page content includes a large image of Half Dome at sunset, navigation links for "Nominate this image for QI", "File", "File history", "File usage on Commons", "File usage on other wikis", and "Metadata". The left sidebar contains links for Main page, Welcome, Community portal, Village pump, Participate (Upload file, Recent changes, Latest files, Random file, Help, Contact us, Donate), and Tools (What links here, Related changes, Special pages, Printable version, Permanent link, Page information, Nominate for deletion).

"Glacier Point at Sunset, Yosemite NP, CA, US - Diliff" by Diliff - Own work. Licensed under CC BY-SA 3.0 via Wikimedia Commons.

FastCCI Advanced Interface

Screenshot of the Category:Wyoming page on Wikimedia Commons, showing the FastCCI interface.

The page title is "Category:Wyoming". The sidebar includes links for Main page, Welcome, Community portal, Village pump, Language select (English), Participate (Upload file, Recent changes, Latest files, Random file, Help, Contact us, Donate), Tools (In Wikipedia, Afrikaans, Aragonés, Ænglisc, آئلر, Azerbaiyacanca, Български, Brezhoneg, Bosanski, Català, Tssetséhestâhese).

The main content area shows the "Category:Wyoming" page with sections for States of the United States, District of Columbia, Insular areas, and flags for Wyoming and the USA. A summary states: "English: Wyoming is a state in the western United States." Subcategories listed include G cont., H, L, M, N cont., and P.

A sidebar on the right lists categories: Good pictures, All images, Featured pictures, Quality images, Valued images, In this category and in..., and In this category but not in... It also includes a map of the United States with a yellow dot over Wyoming.

FastCCI Advanced Interface

Screenshot of the Wikimedia Commons "Category:Wyoming" page. A search dropdown is open, showing results for "Bison".

The page includes the following text:

States of the United States: Alabama · Alaska · Arizona · Arkansas · California · Colorado · Connecticut · Delaware · Florida · Georgia · Hawaii · Idaho · Illinois · Indiana · Iowa · Kansas · Kentucky · Louisiana · Maine · Maryland · Massachusetts · Michigan · Minnesota · Mississippi · Missouri · Montana · Nebraska · Nevada · New Hampshire · New Jersey · New Mexico · New York · North Carolina · North Dakota · Ohio · Oklahoma · Oregon · Pennsylvania · Rhode Island · South Carolina · South Dakota · Tennessee · Texas · Utah · Vermont · Virginia · Washington · West Virginia · Wisconsin · Wyoming

District of Columbia: Washington, D.C.

Insular areas: American Samoa · Guam · Northern Mariana Islands · Puerto Rico · Panama Canal Zone · United States Virgin Islands

The search dropdown shows the following results:

- Bison
- Bison, South Dakota
- Bison (clothing)
- Bison (clothing) in art
- Bison (ship, 1975)
- Bison B.C.
- Bison Stematec
- Bison anatomy
- Bison antiquus
- Bison as symbol of Manitoba
- History of Wyoming (29 C, 18 F)
- Luke Reiner (1 F)
- M
- N
- Parks In Wyoming (9 C, 4 F)
- People of Wyoming (9 C, 19 F)
- Photographs of Wyoming (11 C)
- Populated places in Wyoming (4 C)
- Postcards of Wyoming (1 C, 5 F)
- Pronunciation of Wyoming (2 F)

FastCCI Advanced Interface

The screenshot shows a web browser displaying the Category:Wyoming page on Wikimedia Commons. The URL in the address bar is <https://commons.wikimedia.org/wiki/Category:Wyoming>. The page title is "Category:Wyoming". On the left, there is a sidebar with links to Main page, Welcome, Community portal, Village pump, Language select (set to English), Participate (Upload file, Recent changes, Latest files, Random file, Help, Contact us, Donate), Tools, and In Wikipedia (links to various languages). The main content area displays a grid of nine images. A "Strong match" filter is applied, and a "Weak match" button is visible. The images include: 1) A white pickup truck stopped on a road with several bison in the foreground. 2) Two bison grazing in a grassy field. 3) A herd of bison grazing in a large, open field. 4) A landscape with a cluster of tall, thin trees on a hillside under a cloudy sky. 5) A wide view of rolling green hills. 6) A brown bison standing in front of a weathered wooden building with a blue roof.

Work with Page IDs

- Unique across namespaces
- Just integers (small storage requirement)
 - 15 million files → 60Mb
 - 300 million **category relations** → 1.5GB
 - Fits into RAM
- Use the API to look up the names (client side)

...		
2729056	Category:	Nikola_Šubić_Zrinski
2729058	File:	BA_roof.jpg
2729059	Category:	Zrinski_family
2729060	File:	BlenderExtrusion3ErrCor.jpg
2729061	Category:	Nobility_of_Croatia
2729062	File:	Hadik_Janos.jpg
2729063	File:	Ordine_di_consegna_a_d'Annunzio.JPG
2729064	File:	Teherannosign.jpg
2729065	File:	Perperikon_Bulgaria_1994_CIA_map.jpg
2729066	File:	Emblema_SKVO.jpg
...		

Servers

FastCCI uses two labs instances

- Master to re-build the FastCCI database
- Worker to perform the queries

Master

- DB Format
- Two tables
 - `fastcci.tree` (category relations)

0	<code>cend</code>	Pointer to end of Subcategories	Category
1	<code>cfile</code>	Pointer to end of File list	
2	<code>page_id</code>	1st Subcategory	
2	<code>page_id</code>	2nd Subcategory	
...			
cend-1	<code>page_id</code>	Last Subcategory	
cend	<code>page_id</code>	1st File	
cend+1	<code>page_id</code>	2nd File	
...			
cfile	<code>page_id</code>	Last Subcategory	

- `fastcci.cat` (index into the tree)

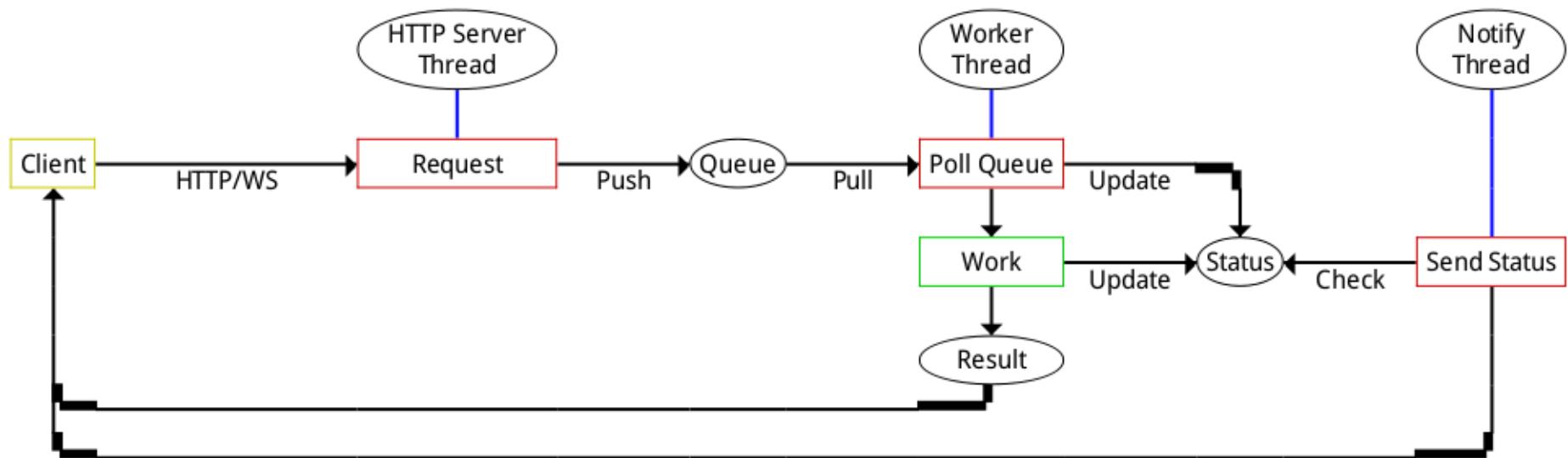
Queries on the category graph

- Generating a category subgraph
 - Walk the graph breadth first
 - Maintain a visitation map
(one byte per page_id, avoids loops, set to depth)
 - Maintain a list of categories in the subgraph
(and store depth with each category)
- Intesect by:
 - Intersect by taking two subgraphs

Queries on the category graph

- Pregenerate a union subgraph for FP/QI/VI at server start
- Store bitmask for awards (FP/QI/VI) for each image

FastCCI Server Internals



- <https://github.com/Commonists/fastcci>
- Stream updates and results via websockets
(immediate feedback: waiting, working, result)
- <http://fastccil.wmflabs.org/status>
- <https://github.com/davidmoreno/onion>

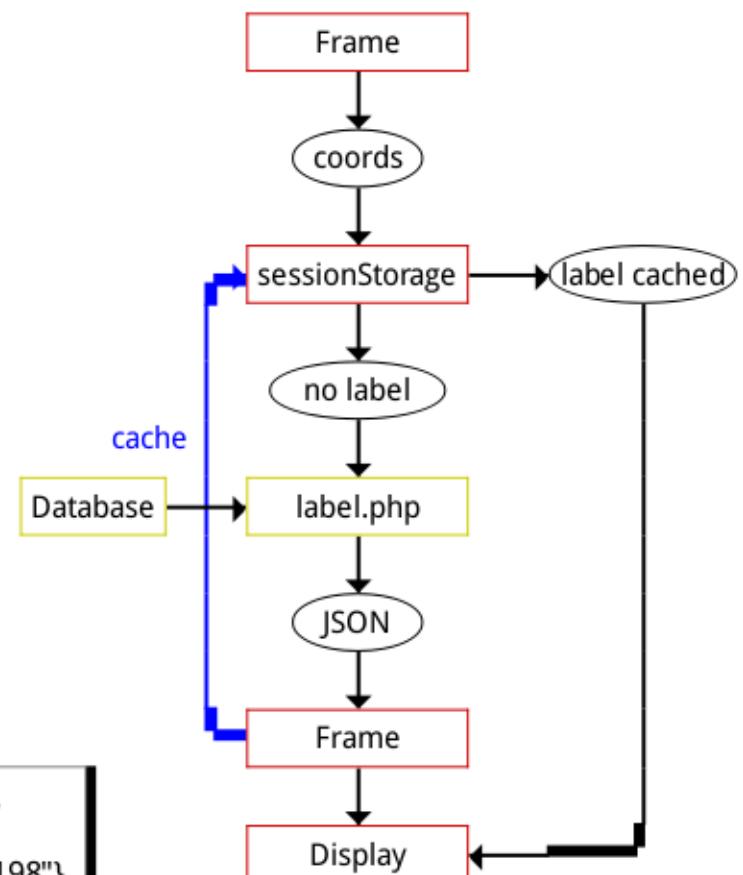
Strength of match slider

Getting the Label Data

- F When a new tile comes into view the *Frame script* looks in the sessionStorage for cached label data.
- F If found it is immediately displayed.
- TS Otherwise an Ajax request is made to the label server script on the toolserver.
- TS A local database is queried for the label data.
- TS The label data is encoded as a JSON string.
- F Label data is received, cached in sessionStorage, and displayed.

```
{"label": [{"style": "7", "lang": "en", "page": "Alexandria%2C_Virginia", "tx": 8, "ty": 93, "name": "Alexandria", "dx": "4829", "dy": "2198"}, {"style": "0", "lang": "en", "page": "Fort_Greble", "tx": 78, "ty": 44, "name": "Fort Greble", "dx": "4829", "dy": "2198"}, {"style": "5", "lang": "en", "page": "Forest_Heights%2C_Maryland", "tx": 117, "ty": 78, "name": "Forest Heights", "dx": "4829", "dy": "2198"}, {"style": "0", "lang": "en", "page": "United_States_Naval_Research_Laboratory", "tx": 62, "ty": 51, "name": "United States Naval Research Laboratory", "dx": "4829", "dy": "2198}], "z": 10}
```

```
{
  "style": "7", "lang": "en", "page": "Alexandria%2C_Virginia",
  "tx": 8, "ty": 93, "name": "Alexandria", "dx": "4829", "dy": "2198"
}
```



Thank you!

Contact [\[\[User:Dschwen\]\]](#) with suggestions,
<https://github.com/Commonists/fastcci>