UNIVERSITY OF
Southampton

UNIVERSITY OF Southampton

Search Engines

COMP3220 Web Infrastructure

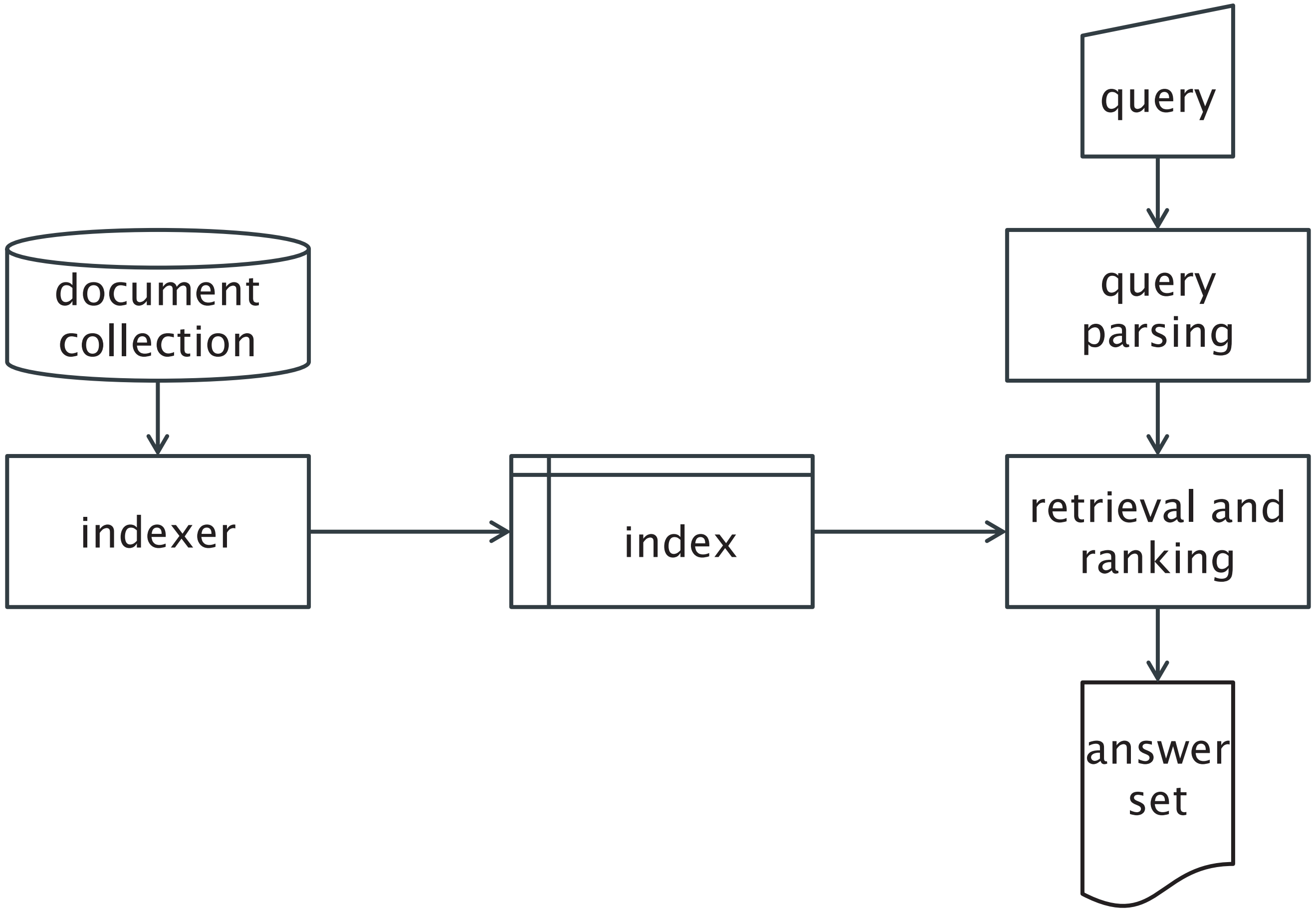
Dr Heather Packer – hp3@ecs.soton.ac.uk

Search Engines

Information Retrieval

- The primary goal of an information retrieval system is to retrieve all the documents that are relevant to a user query while retrieving as few as few non-relevant documents as possible
  - An **information need** is a topic which a user desires to know more about
  - A **query** is what the user conveys to the computer in an attempt to communicate their information need
  - A document is **relevant** if it is one that the user perceives as containing information of value with respect to their personal **information need**

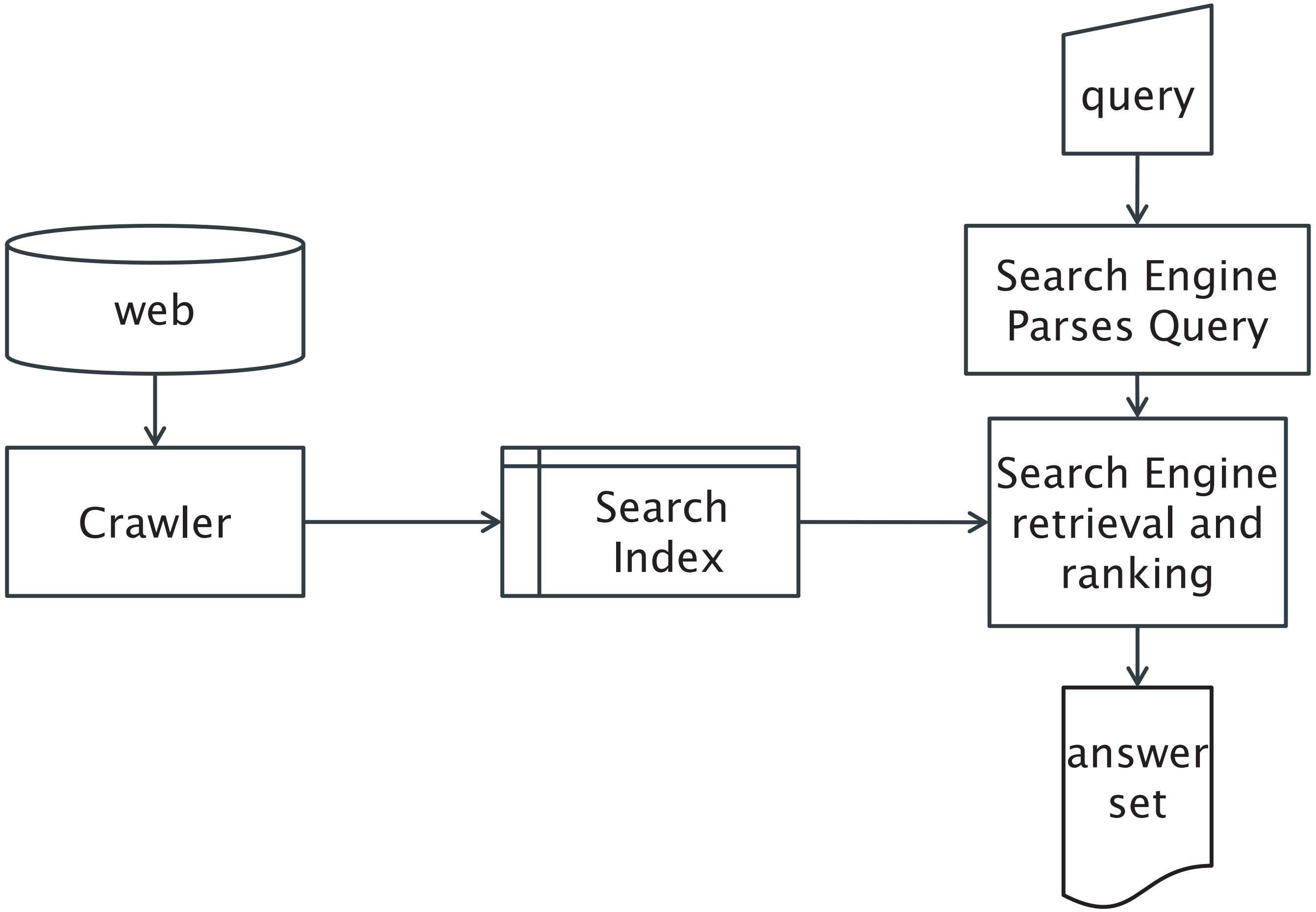
High-Level System Architecture – Information Retrieval



Search Engines

- Search engines are a service
- They allow users to search for content using **keywords**
- A query returns a **ranked** set of results
- They **DO NOT** access the web directly
- They **USE** huge databases

High-Level System Architecture – Search Engine


query
web
Search Engine Parses Query
Crawler
Search Index
Search Engine retrieval and ranking
answer set

Specific Board and Vague Queries

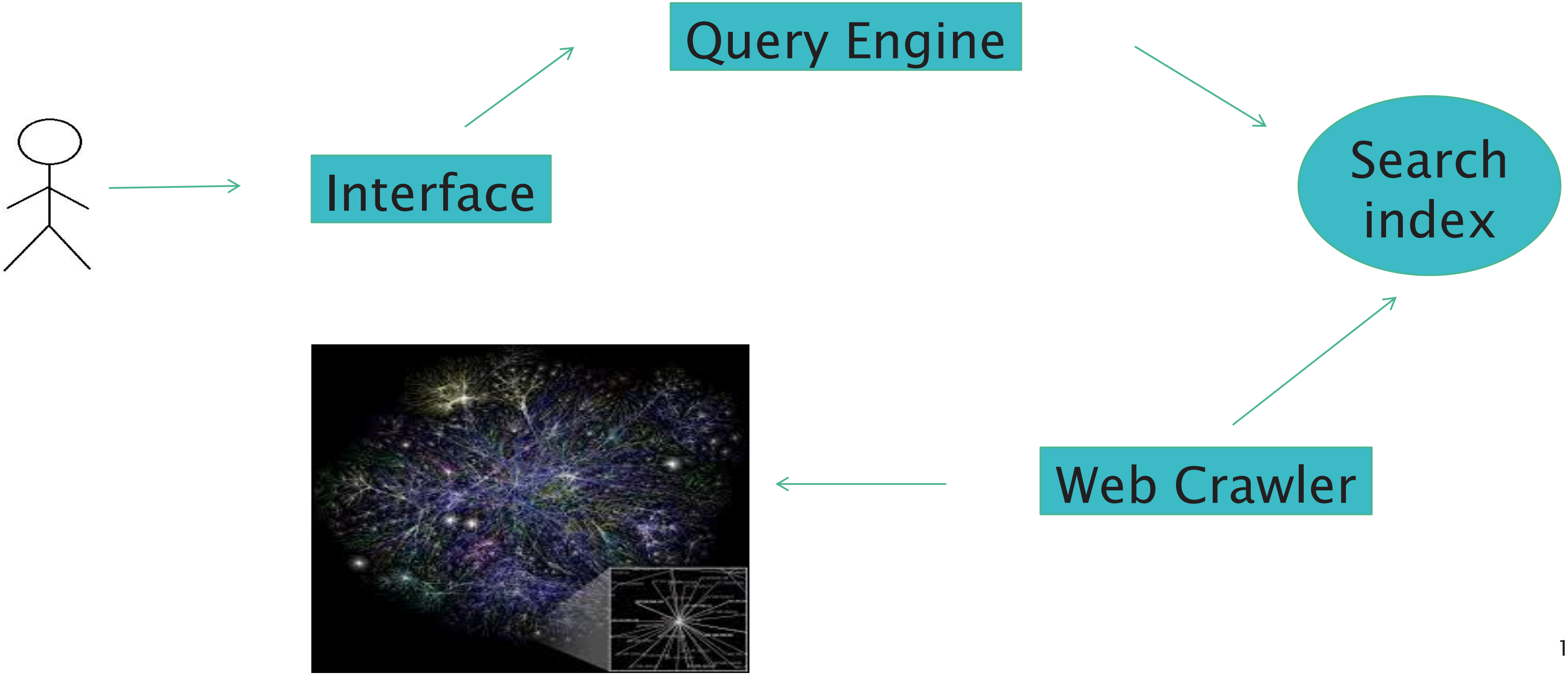
	Specific	Broad	Vague
I know specifically what I'm looking for	✓	✗	✗
I know where to look for what I'm looking for	✓	✓	✗
Example	Looking for a specific gene in the Gene Database	Looking for the manager of HR, in the company directory	Google

Specific Board and Vague Queries

	Specific	Broad	Vague
I know specifically what I'm looking for	✓	✗	✗
I know where to look for what I'm looking for	✓	✓	✗
Example	Looking for a specific gene in the Gene Database	Looking for the manager of HR, in the company directory	Google

Search Engines

Simple Framework – Altravista’s Framework 1994



Query Engine

Interface

Search index

Web Crawler

Web Crawler, Spider or bot

- An algorithm that systematically browses the web
- A basic algorithm
  1) Start at a webpage
  2) Follow the hyperlinks that webpage points to
  3) Then follows the links those webpages point to
- Each page it visits it collects metadata about it

- Stores a file for each resource, with its metadata in a search index
- Crawlers consume resources
- Can visit sites without approval

Web Crawler - robots.txt

- Block all web crawlers

```
User-agent: *

Disallow: /
```

- Allow all web crawlers

```
User-agent: *

Disallow:
```

- Block a specific web crawler from a specific folder

```
User-agent: Googlebot

Disallow: /example-subfolder/

Disallow: /index.html
```

Web Crawler Policies

The behaviour of a web crawler is based on:

1. Selection Policy
2. Re-visit Policy
3. Politeness Policy
4. Parallelisation Policy

Web Crawler Selection Policy

- Search engines only index part of the web
- It’s important to download the most relevant pages
- A selection policy states which pages to index
- Strategies:
  - Random
  - Breadth first
  - Back link count
- Only request pages that have searchable content (HTML, PDF etc)

Web Crawler Re-visit Policy

- Its worth revisiting web pages because they change over time
- An ideal search engine would have the most up-to-date version of every page in its index
- Strategies
  - Re-visit all pages equally frequent
  - Prioritise pages that change often (but not too often!)
- May take page quality into account

Web Crawler Politeness Policy

- Issues of schedule and load when large collections of pages are accessed
- Strategies
  - Do not make parallel calls to the same server
  - Spread out requests
  - Abide by Crawler delay in Robots.txt

Web Crawler Parallelisation Policy

- An efficient crawler needs to access many web servers at once
- Run multiple processes in parallel
- Could find the same URL on multiple pages
- Strategies:
  - Dynamically assign pages to crawler processes
  - Static mapping eg based on a hash function

Web Crawler - Crawlability

- Broken links
- Denied access
- Outdated URLs
- URL errors
- Blocked
- No out links
- Slow load speed
- Flash content
- JavaScript (Googlebot executed from 2014)
- HTML frames (outdated and thus poorly indexed)
- Data
  - Unstructured data - gifs, pdf, etc
  - Redundant data - 30% pages are near duplicates
  - Quality of data - False, poorly written, invalid, misspelt

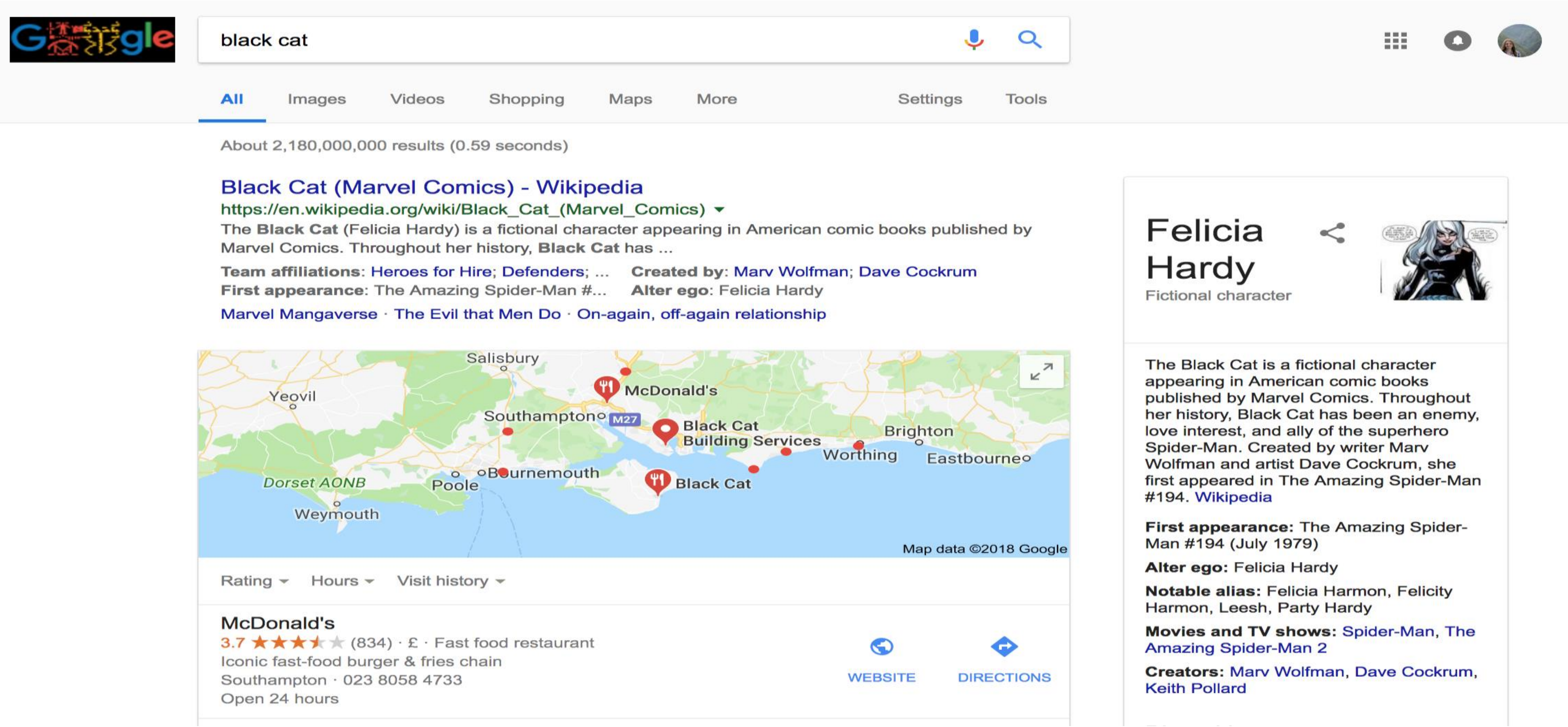
Search Index Ranking

- Query results can be ranked using many features:
  - How many times does the page contain the keywords
  - Do keywords appear in the title or url
  - Does it contain synonyms for your keywords
  - Is it from a quality source
  - What is it’s importance
  - How often a page is updated
  - Freshness of information
  - Page load time

User Search Problems

- Users may get unexpected answers because they are not aware of the input requirement of the search engine.
  - For example, some search engines are case sensitive.
- Users have problems understanding Boolean logic
- Around 85% of users only look at the first page of the result, so relevant answers might be skipped.
- Users do not understand how to provide a sequence of words for searches

User Search Problems: Ordering of Terms



black cat

About 2,180,000,000 results (0.59 seconds)

Black Cat (Marvel Comics) - Wikipedia
https://en.wikipedia.org/wiki/Black_Cat_(Marvel_Comics)
The **Black Cat** (Felicia Hardy) is a fictional character appearing in American comic books published by Marvel Comics. Throughout her history, **Black Cat** has ...

Team affiliations: Heroes for Hire; Defenders; ... **Created by**: Marv Wolfman; Dave Cockrum
First appearance: The Amazing Spider-Man #... **Alter ego**: Felicia Hardy

Marvel Mangaverse · The Evil that Men Do · On-again, off-again relationship

McDonald's
3.7 ★★★★☆ (834) · £ · Fast food restaurant
Iconic fast-food burger & fries chain
Southampton · 023 8058 4733
Open 24 hours

Felicia Hardy
Fictional character

The Black Cat is a fictional character appearing in American comic books published by Marvel Comics. Throughout her history, Black Cat has been an enemy, love interest, and ally of the superhero Spider-Man. Created by writer Marv Wolfman and artist Dave Cockrum, she first appeared in The Amazing Spider-Man #194. Wikipedia

First appearance: The Amazing Spider-Man #194 (July 1979)

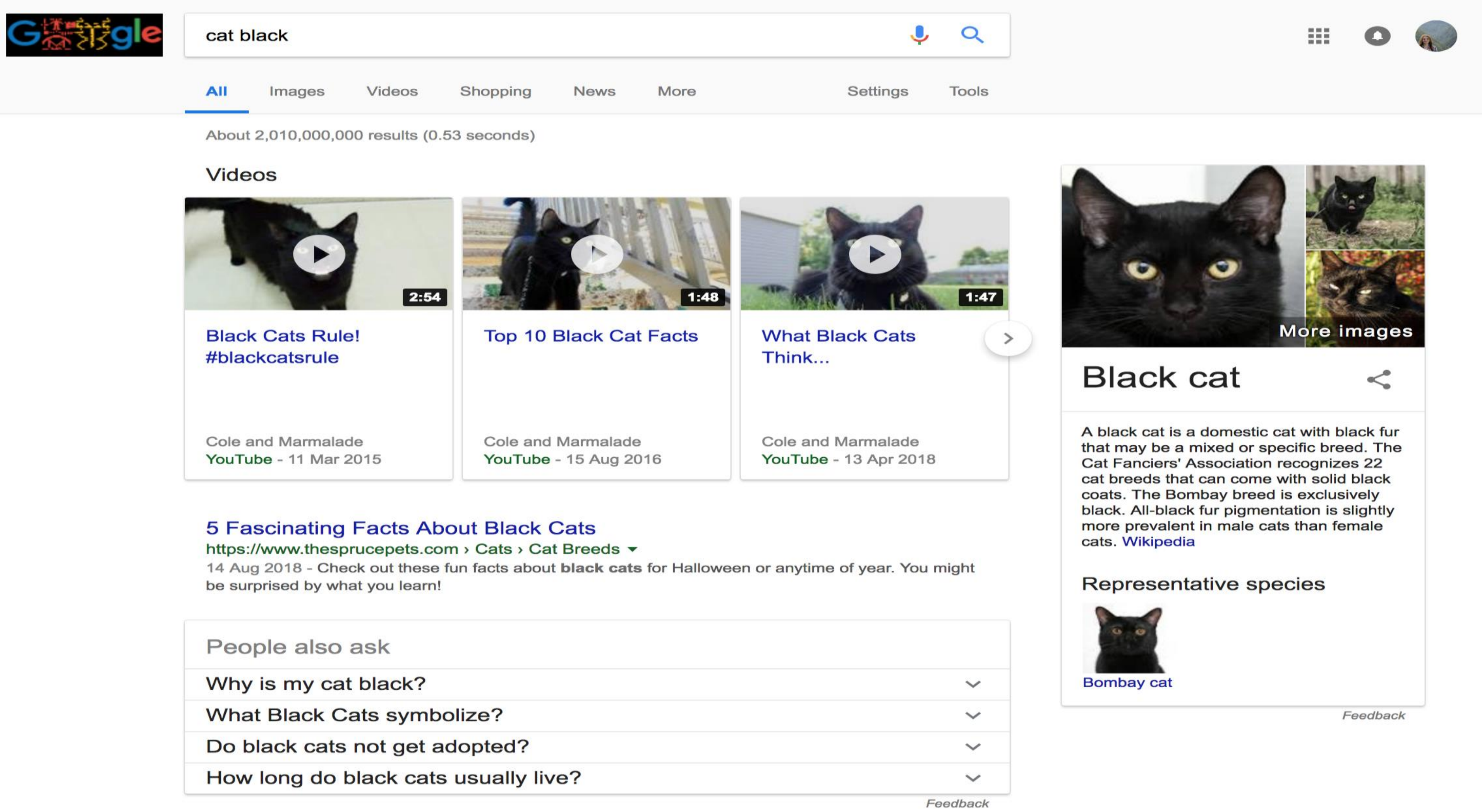
Alter ego: Felicia Hardy

Notable alias: Felicia Harmon, Felicity Harmon, Leesh, Party Hardy

Movies and TV shows: Spider-Man, The Amazing Spider-Man 2

Creators: Marv Wolfman, Dave Cockrum, Keith Pollard

User Search Problems: Ordering of Terms



cat black

All Images Videos Shopping News More Settings Tools

About 2,010,000,000 results (0.53 seconds)

Videos

Black Cats Rule! #blackcatsrule
Cole and Marmalade
YouTube - 11 Mar 2015
2:54

Top 10 Black Cat Facts
Cole and Marmalade
YouTube - 15 Aug 2016
1:48

What Black Cats Think...
Cole and Marmalade
YouTube - 13 Apr 2018
1:47

5 Fascinating Facts About Black Cats
https://www.thesprucepets.com › Cats › Cat Breeds
14 Aug 2018 - Check out these fun facts about **black cats** for Halloween or anytime of year. You might be surprised by what you learn!

People also ask

- Why is my cat black?
- What Black Cats symbolize?
- Do black cats not get adopted?
- How long do black cats usually live?

Feedback

More images

Black cat

A black cat is a domestic cat with black fur that may be a mixed or specific breed. The Cat Fanciers' Association recognizes 22 cat breeds that can come with solid black coats. The Bombay breed is exclusively black. All-black fur pigmentation is slightly more prevalent in male cats than female cats. Wikipedia

Representative species

Bombay cat

Feedback

Search Engine Optimisation

- Search engines don’t host content
- 85% of people don’t look at the second page
- People try to optimise their site so that it ranks highly on Search Engines
- Could be fundamental to a website’s business model

Search Engine Optimisation

- Whole industry exists trying to boost search ranking to ensure pages are indexed by search engines

- Leads to arms race between SEO and search engines

- Legitimate SEO (White Hat)
  - Good Design
  - Valid metadata, alt tags on images

- Illegitimate SEO (Black Hat)
  - Often gaming search ranking algorithms
  - Deception

Combatting SEO

- Most search engines have rules against:
  - Invisible text
  - Meta tag abuse
  - Heavy repetition
  - “domain spam”
    - Overtly submission of “mirror” sites in an attempt to dominate the listings for particular terms

Google and other SEs are a Business

- Search Engines record tracking information
  - Google saves every voice search
  - IP addresses
  - Location
  - Saves your searches
- Google’s revenue is from adverts
  - Improve their revenue with targeted advertising
- Google has a large research department
  - Improve their technology

Overview

- What types of queries can be answered on the web
- Search engines and their basic framework:
  - Web crawler
  - Search index
  - Query engine
  - Interface
- Issues with Search Engines