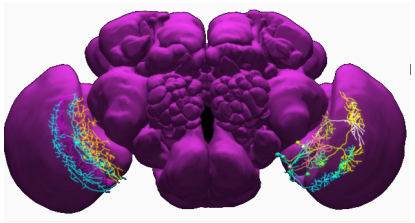
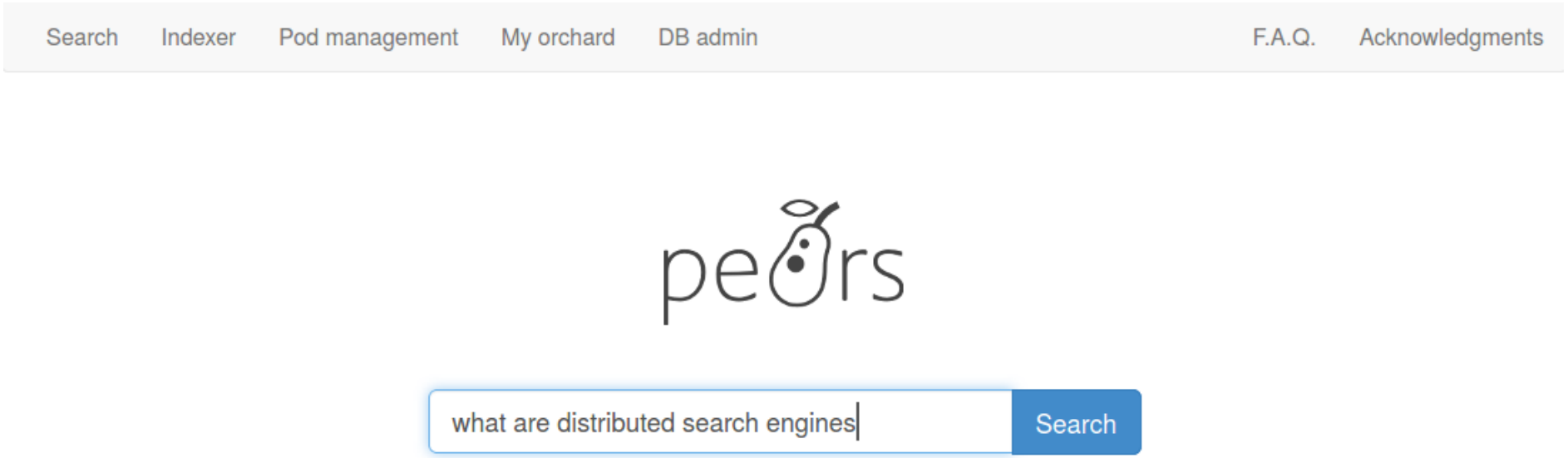
The PeARS project

Low-resource web search applications from cognitive models

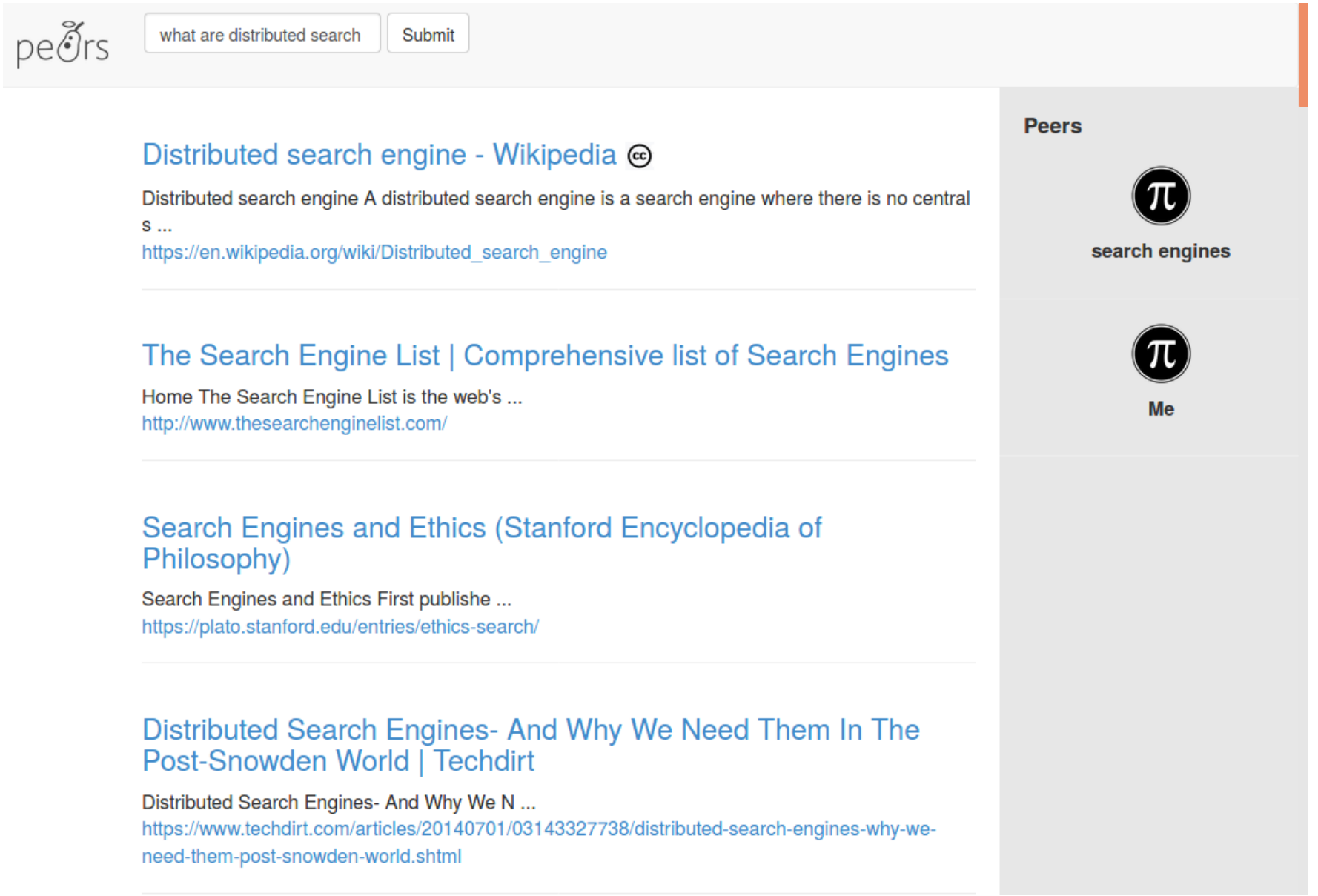
Aurelie Herbelot, Tanise Pagnan Ceron,
Le Minh Nhut Truong



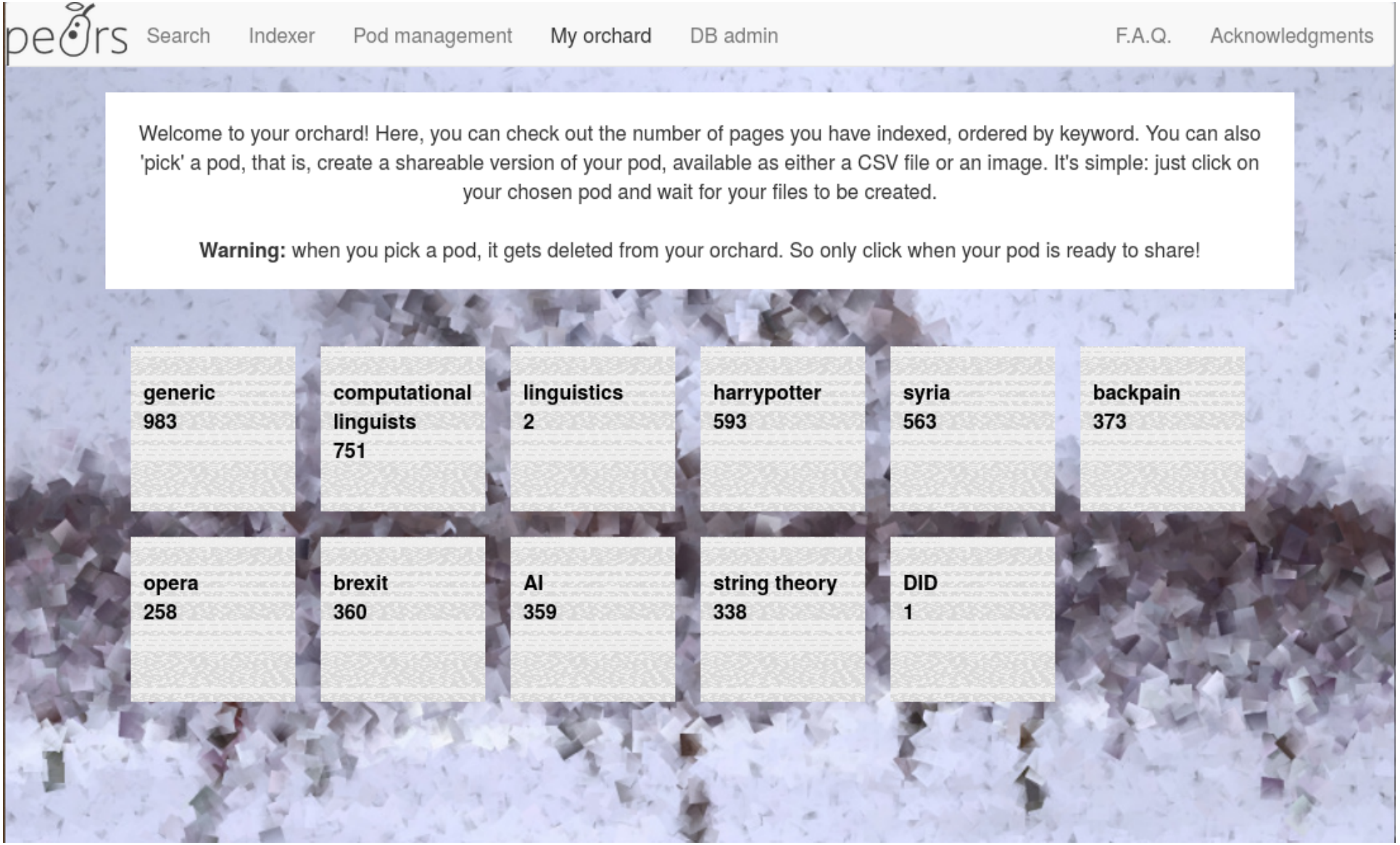
What is PeARS?



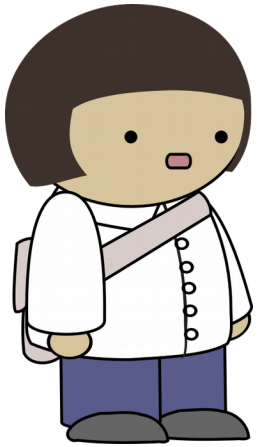
A local search engine



Index your Web

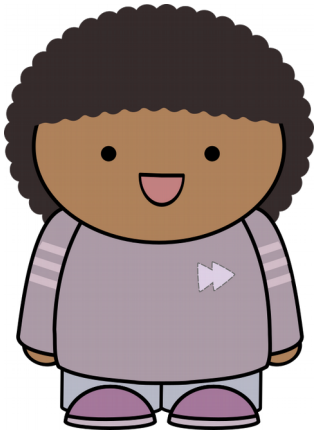


Share your Web



What is string theory?

Check out my pods!



Politics and society

Brexit

624 pages about the Brexit process

Anarchism

1175 pages about anarchist movements

The Black Panthers

506 pages about the political organisation founded in 1966

Science and culture

AI

359 pages about AI and its relation to society

String theory

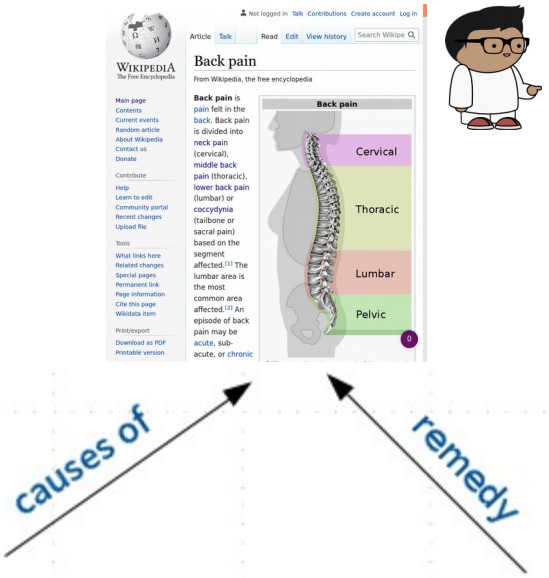
307 pages about string theory

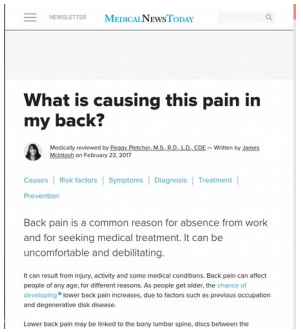
Opera

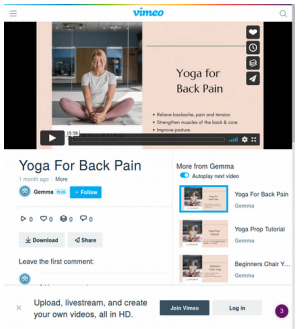
258 pages about opera

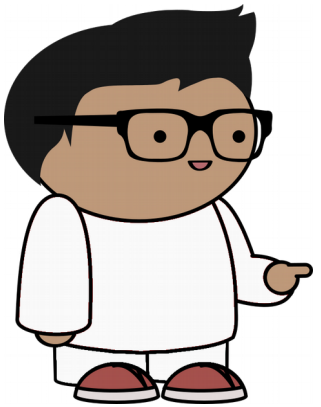
Benefits

- Decentralised search
- Curate Web data locally, share your annotation!









Can we help the curator?

https://www.medicalnewstoday.com/articles/172943

(Text: Causes of back pain)

https://praxisinstitute.org/wp-content/uploads/2019/10/Anatomy-of-SC.png

(Image: Anatomy of the spinal cord)

https://vimeo.com/553448612

(Video: Yoga for back pain)

Curation: lots of manual work if resources do not already contain structured data. The user might have a few thousands of URLs in their local pods. Many of those URLs won't be very informative with respect to the actual content of the resource, and they might span various formats, from text to images to videos.

Can we help the curator?

7196759210defdc0 = ***back pain***, cause, injury, *strain*, **spine**

(Text: Causes of back pain)

2638458674fhsgi4 = spinal cord, **spine**, anatomy, description

(Image: Anatomy of the spinal cord)

5782946738shsgd2 = ***back pain***, *strain*, remedy, yoga, exercise

(Video: Yoga for back pain)

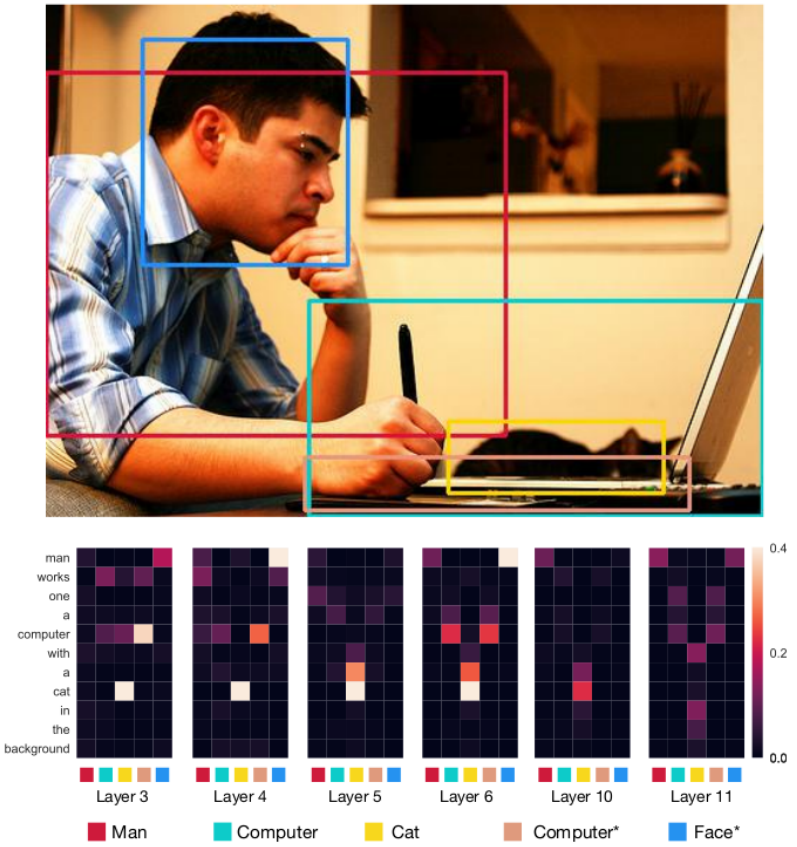
What if we could hash the content of those resources into comparable semantic representations?

→ Cluster documents into subtopics, automatically suggest keywords, associate multimodal content with text fragments… and of course, search better!

Semantics is expensive

Deep learning models can extract semantic representations from text and images, but they are too expensive to train for the average user.

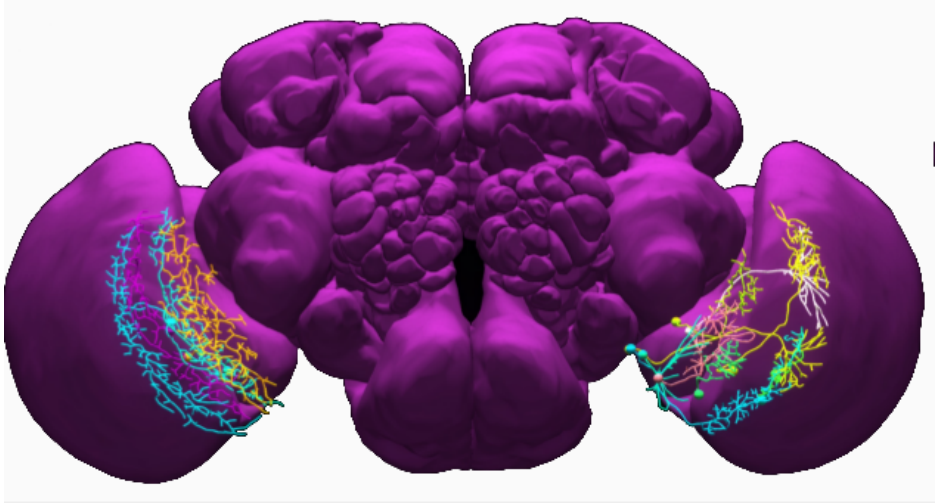
- High levels of CO2 emissions (training pollutes!)
- Pretrained models are provided by large companies as ‘black boxes’. No transparency.



Li et al (2019), https://arxiv.org/abs/1908.03557

The Fruit Fly Algorithm

- The olfactory system of the fruit fly: a powerful hashing algorithm, a simple feedforward architecture.
- Already used for hashing word and image vectors.
- Can it encode the semantics of *any* URI, from scratch?



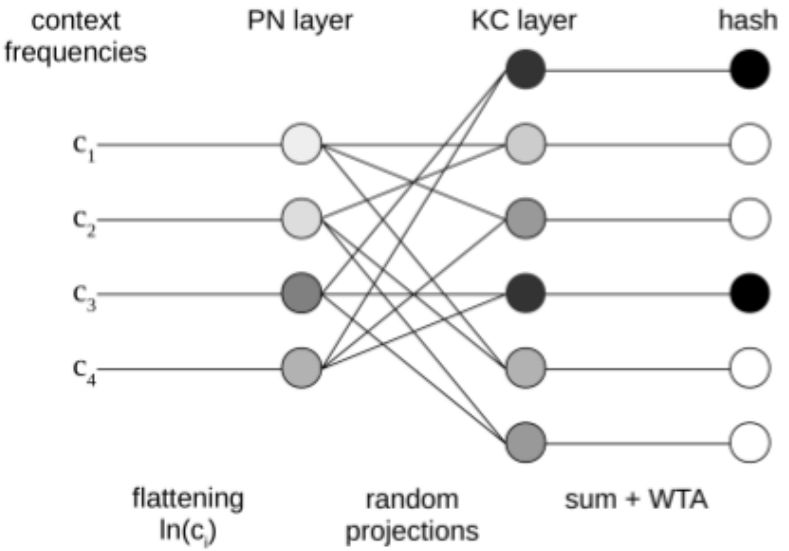
Ukani et al (2019), https://doi.org/10.1101/580290

Spawning fruit flies

We are spawning thousands of fruit flies, with the idea to produce the best possible Web document presentations.

The most intelligent specimen will be integrated into the PeARS framework to support search and local annotation.

A typical fruit fly model can be stored in 100-200KB. The algorithm is highly efficient. The process can be reproduced by anybody on their personal laptop.



https://www.flickr.com/photos/nasamarshall/15842201129, CC BY-NC 2.0

Read more about the project!

- Our website:

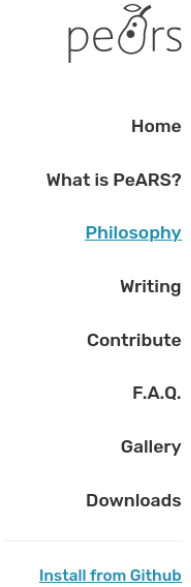
  https://pearsproject.org/index.html

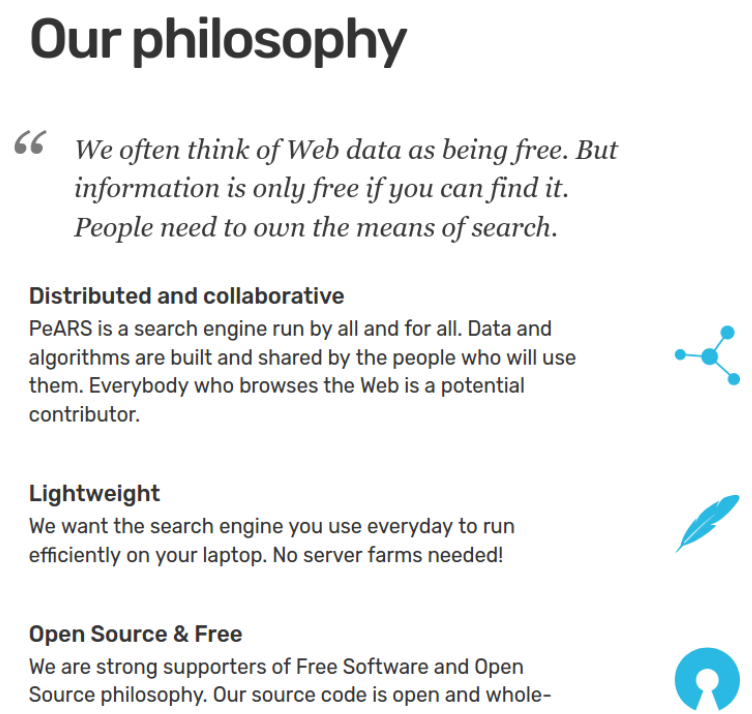
- Our GitHub:

  https://github.com/PeARSearch

- The Fruit Fly project will be evolving at:

  https://github.com/PeARSearch/PeARS-fruit-fly





Acknowledgements

https://openclipart.org/artist/anarres for the cartoon characters in this presentation.

The Center for Mind/Brain Sciences, for hosting the project