

TEXT MINING SCIENTIFIC PAPERS with R



Rocío Joo

UF UNIVERSITY of
FLORIDA

MIAMI
R-Ladies



rocio.joo@ufl.edu



@rocio_joo

Colaborators:



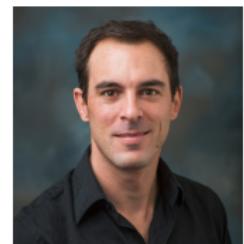
M. Boone



S. Picardi



V. Romero



M. Basille



T. Clay



S. Clusella-Trullas



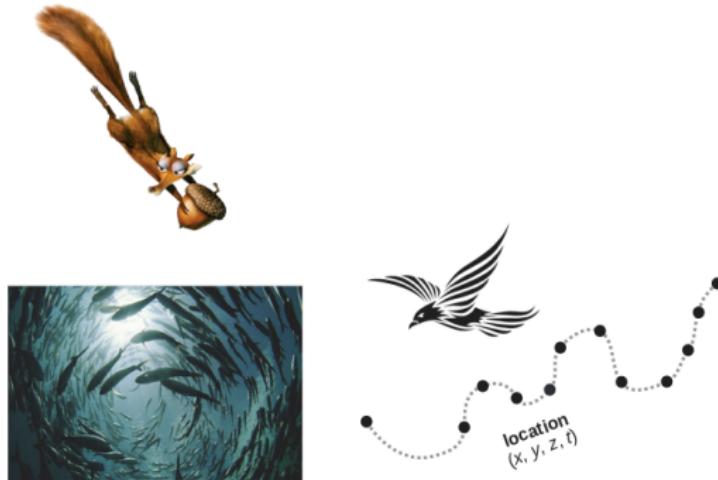
S. Patrick



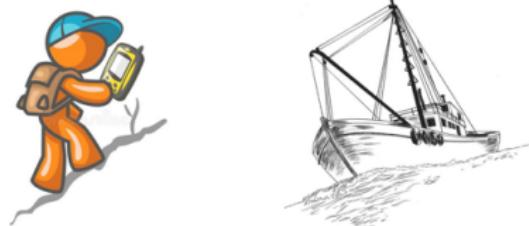
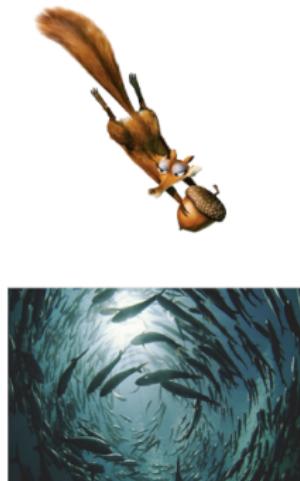
Context: movement ecology



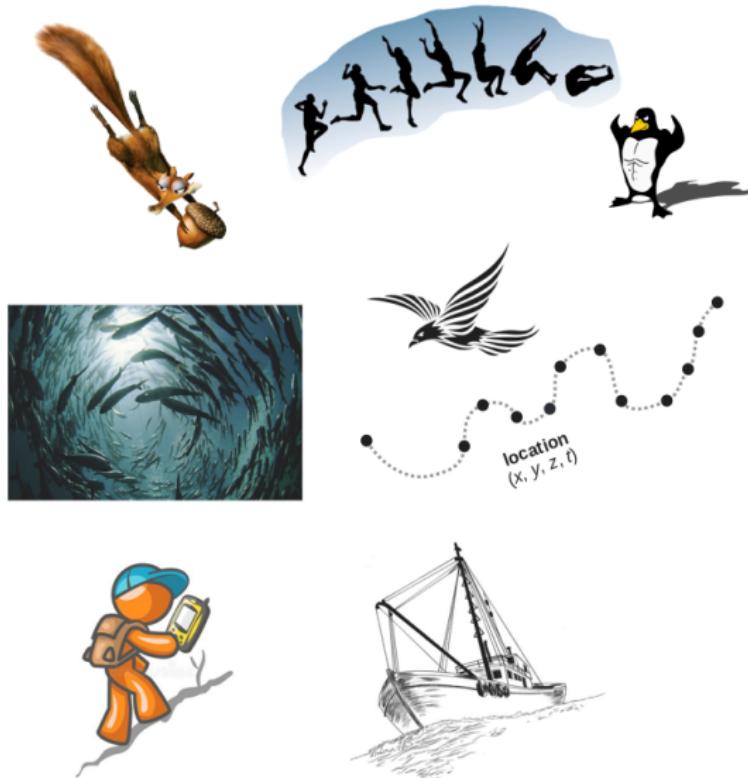
Context: movement ecology



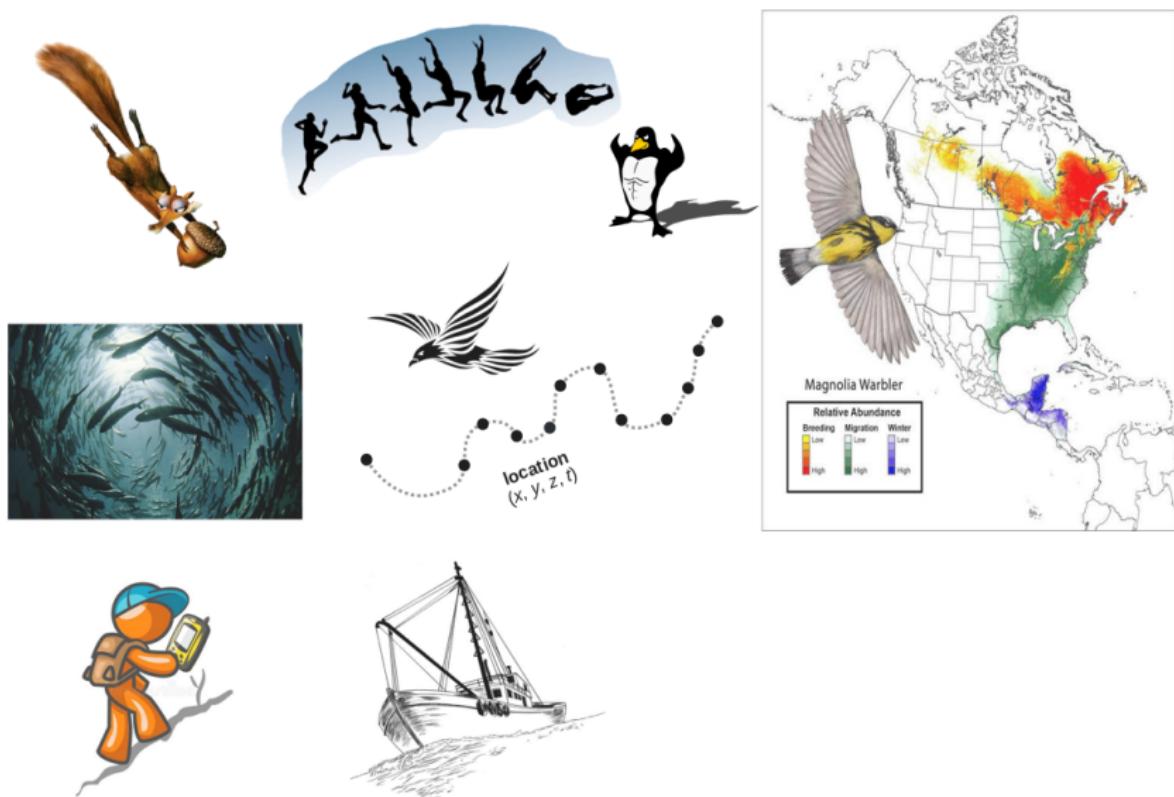
Context: movement ecology



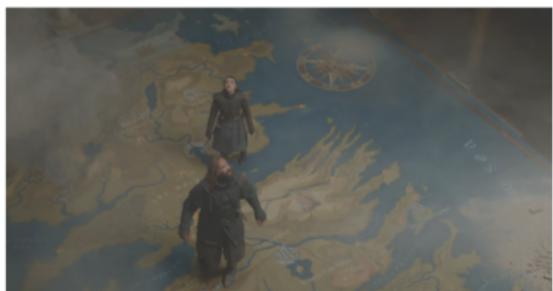
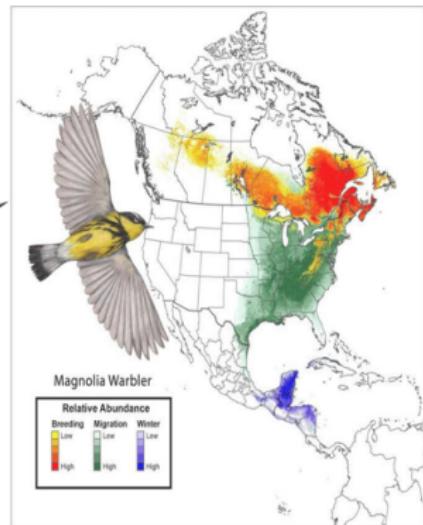
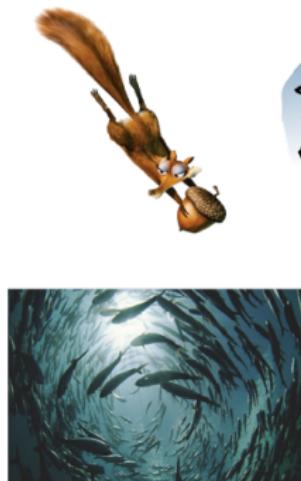
Context: movement ecology



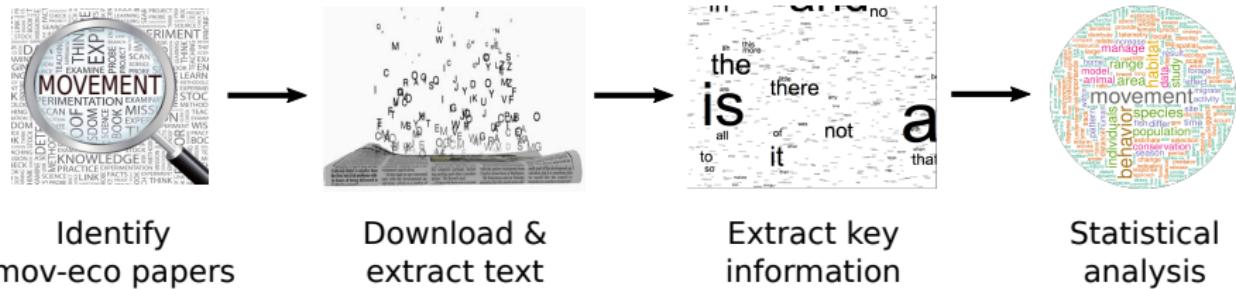
Context: movement ecology



Context: movement ecology



Workflow



Searching for papers

Web of Science [v.5.32] - Web of Science Core Collection Basic Search - Mozilla Firefox

E Statistical Modelling in E | S universidad nacional de | My Drive - Google Drive | Web of Science [v.5.32] +

https://apps.webofknowledge.com/WOS_GeneralSearch_input.do?product=WOS&search_mode=GeneralSearch&SID=6DeUcwSEcyIDsFAUKn&prefe...

Web of Science InCites Journal Citation Reports Essential Science Indicators EndNote Publons Kopernio

Rocio ▾ Help ▾ English ▾

Web of Science

Clarivate Analytics

Select a database Web of Science Core Collection

Tools ▾ Searches and alerts ▾ Search History Marked List

Basic Search Cited Reference Search Advanced Search Author Search

Example: oil spill* mediterranean Topic Search Search tips

+ Add row | Reset

Timespan Custom year range 1900 to 2019

More settings ▾

University of Florida

Clarivate Accelerating innovation

© 2019 Clarivate Copyright notice Terms of use Privacy statement Cookie policy

Sign up for the Web of Science newsletter Follow us  

Final keywords

Movement ecology papers: Papers that studied the voluntary movement of one or more living individuals

Final keywords

Movement ecology papers: Papers that studied the voluntary movement of one or more living individuals

4 groups of keywords were established for the search in WoK:

- ① **Behavior:** behavio
- ② **Movement:** moveme, moving, motion, spatiotemporal, kinematics, spatio-temporal
- ③ **Biologging:** telemetry, geolocat, biologg, accelerom, gps, geo-locat, bio-logg, reorient, vhf, argos, radar, sonar, gls, vms, animal-borne
- ④ **Individuals:** animal, individual, human, person, people, player, wildlife, fishermen

Papers had to have words from **at least 3 groups** to be selected.

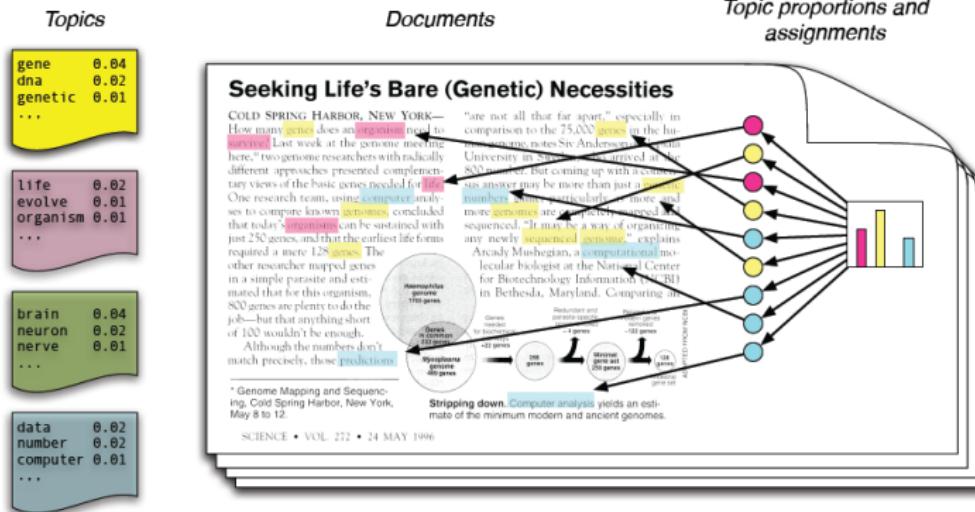
Cleaning query results

- Query results are processed again in R
- Mostly because of this exception:
 - **Unless** they had words from **group 3**, they shouldn't contain:
 - **Missleading words:** cell, DNA, enzyme, strain, neuro, atom, molecule, lymph, cortex, cortic, receptor, patient, prosthese, eye, particle, tectonic, counsel, cognit, market, spine, questionnaire, sedentary, insulin, peristal, muscle, amput, nervous, retinal, psychiatric, disease, virus, remotely-guided, tissue, polymer, mri

Example in R

Topic analysis

Latent Dirichlet Allocation (LDA)



Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare its own genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, these predictions

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

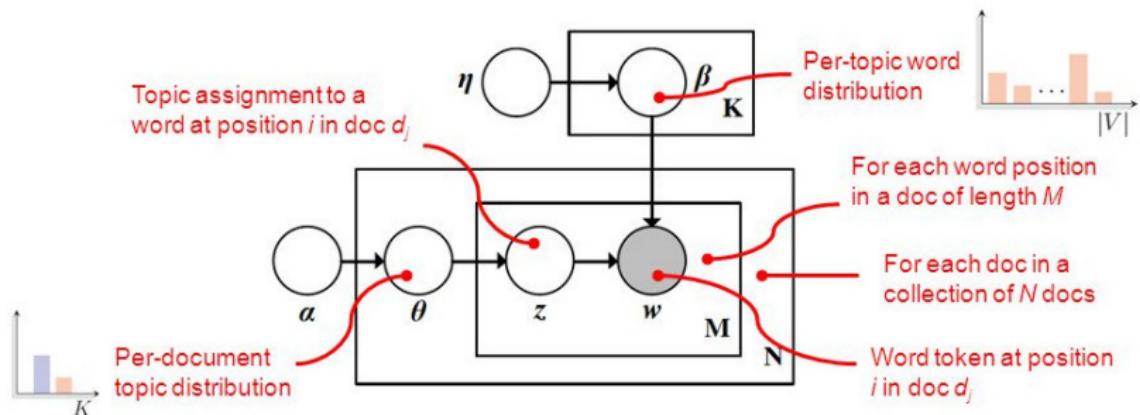
"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Sir Andersson of Uppsala University in Sweden, who arrived at the 800 number, but coming up with a consensus answer may be more than just a matter of numbers. As more and more genomes are sequenced, it will be sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing all

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

- Each **topic** is modeled as a distribution of words
- Each **document** is a mixture of topics
- Each **word** comes out of one of these topics

Topic analysis

Latent Dirichlet Allocation (LDA)



[Moens and Vulic, Tutorial @WSDM 2014]

$$p(\beta, \theta, z, \omega | \alpha, \eta) = \prod_{i=1}^K \left\{ p(\beta_i | \eta) \prod_{d=1}^N \left[p(\theta_d | \alpha) \left(\prod_{n=1}^N p(z_d | \theta_d) p(\omega_{d,n} | \beta_{1:K}, z_{d,n}) \right) \right] \right\}$$

Topic analysis

Steps to follow in practice:

- Define what a **document** is: an abstract
- **Filter out** non informative words (e.g. prepositions) and lemmatize
- Prepare input for LDA: list of words per document and their frequency
- Fit **LDA** with fixed number of topics
- **Interpret** topics

Example in R



✉ rocio.joo@ufl.edu

🐦 [@rocio_joo](https://twitter.com/rocio_joo)