Machine Learning for Research Classification

Manolis Antonoyiannakis
Gokhan Izgi
Minyi Huang

The problem: Research assessment & classification

Research assessment of scientists, departments, countries, to:

- Allocate research funding (identify cutting-edge research, emerging fields)
- Facilitate hiring & promoting decisions

How? Analyze scientific publications (research papers)

- Number count: how many papers were published?
- Citation count: counting citations to papers from other papers

But: Different research fields have different publication & citation practices

➔ Need to allocate papers to their respective research fields: This is not trivial

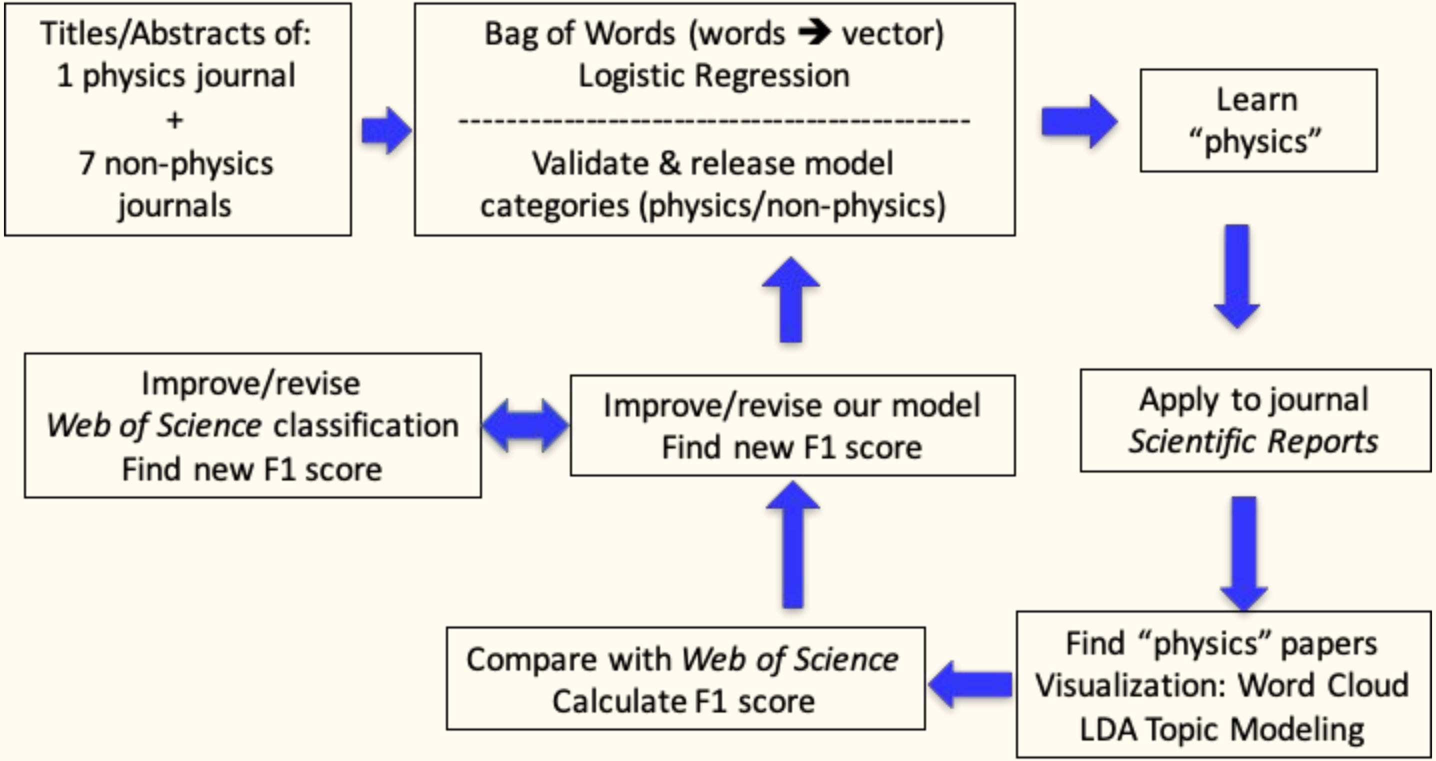
➔ Multidisciplinary journals cover all sciences (from astrophysics to sociology)!

➔ *Web of Science*, a major research database, does this on the fly… but is it accurate?

➢ Task: Use ML to find physics papers in the journal "Scientific Reports" and compare with classification by *Web of Science*.

Results: 2,286
(from Web of Science Core Collection)

The method



Titles/Abstracts of:
1 physics journal
+
7 non-physics journals

Bag of Words (words ➔ vector)
Logistic Regression

Validate & release model categories (physics/non-physics)

Learn "physics"

Apply to journal *Scientific Reports*

Find "physics" papers
Visualization: Word Cloud
LDA Topic Modeling

Compare with *Web of Science*
Calculate F1 score

Improve/revise our model
Find new F1 score

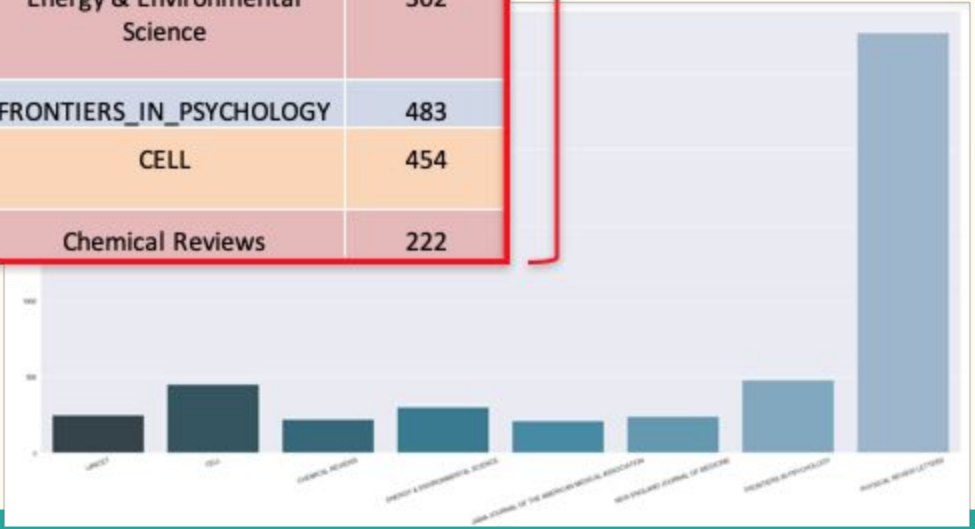
Improve/revise *Web of Science* classification
Find new F1 score

Input data

Field name	Journal	Papers
MULTIDISCIPLINARY SCIENCES	Scientific Reports	17158
PHYSICS	**Physical Review Letters**	**2773**
General & Internal Medicine	LANCET	248
General & Internal Medicine	JAMA - Journal of the American Medical Association	212
General & Internal Medicine	New England Journal of Medicine	239
Chemistry Energy & Fuels Engineering Environmental Sciences & Ecology	Energy & Environmental Science	302
Psychology	FRONTIERS_IN_PSYCHOLOGY	483
Biochemistry & Molecular Biology Cell Biology	CELL	454
Chemistry	Chemical Reviews	222

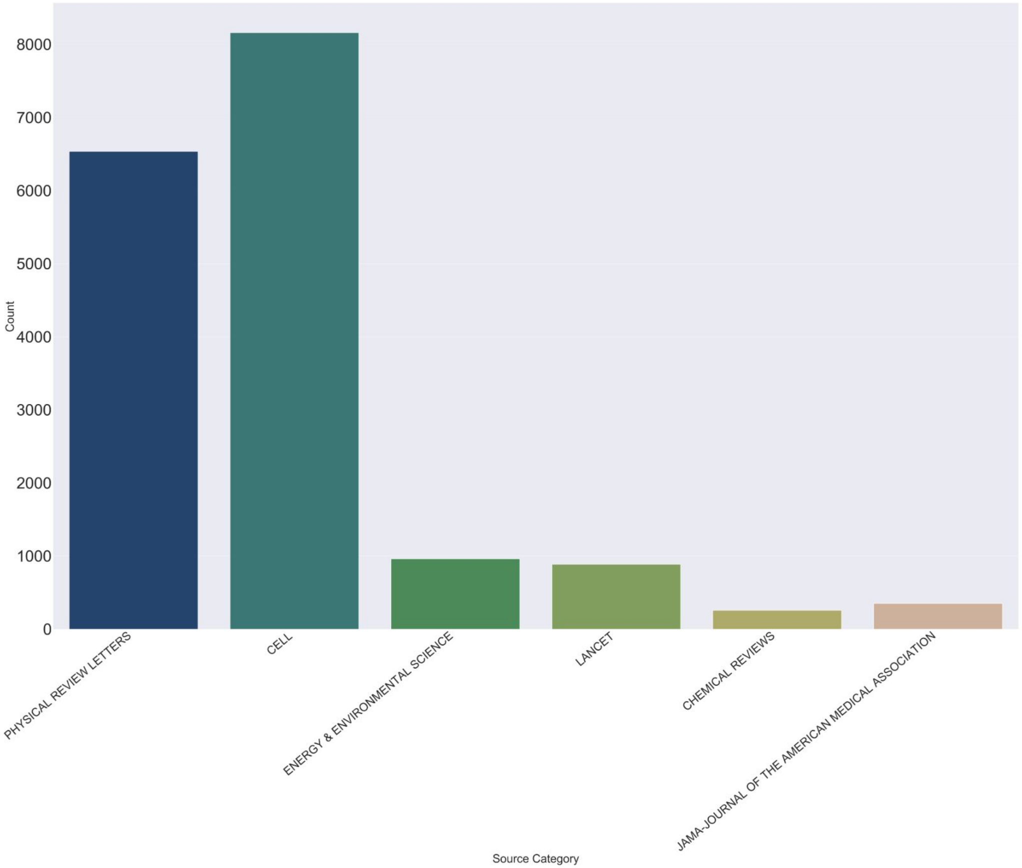
Predicting

Training/splitting/testing



Input journals for training

The predicted source Category based on the content of the research paper abstracts



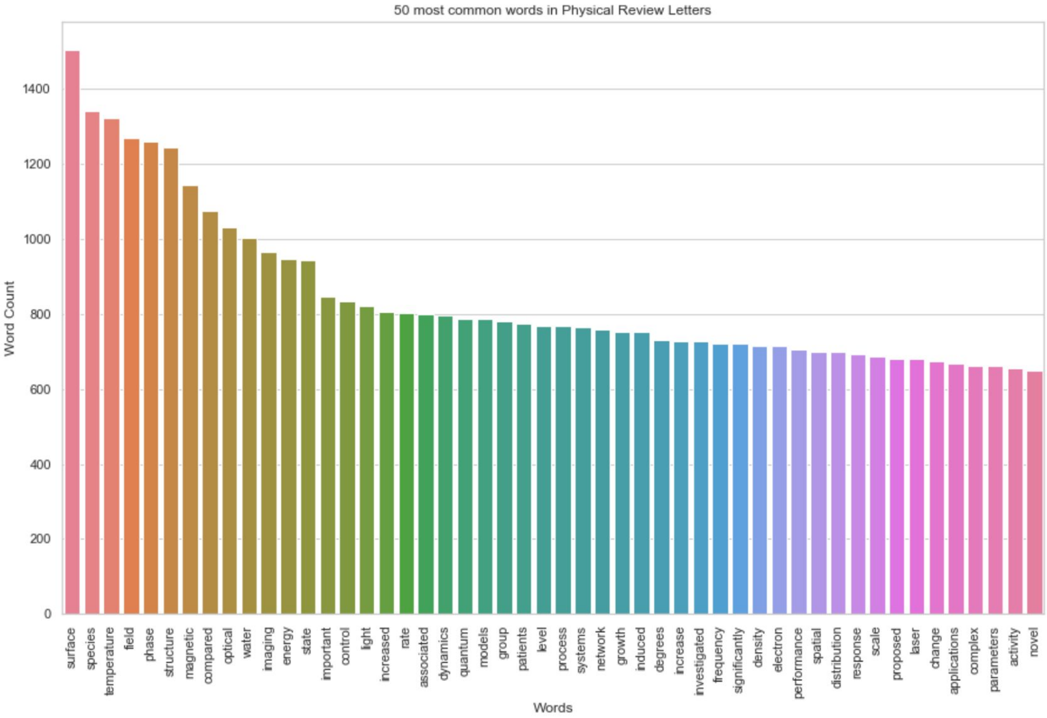
Prediction of the research area of each paper:

Step1: LEARN
Logistic Regression Algorithm along with the bag of words method, scans through the abstracts of each document (research paper) in train data and learns the word occurrence frequencies for each research area.

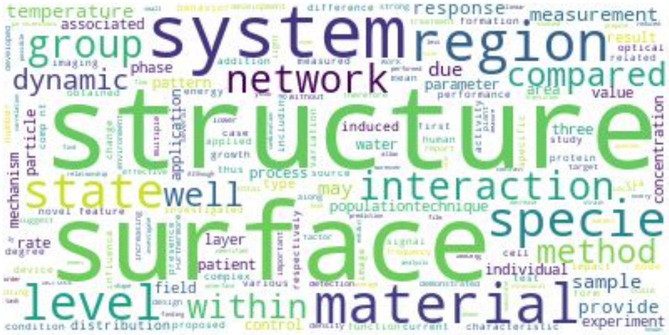
Step 2: PREDICT
In test data, where the specific research area of the papers is unknown, the regression algorithm predicts a research area for each document (research paper) named after the corresponding training journal.

The chart on the left shows how many predictions were generated for each research area in total.

Word frequency in Source Category *Physical Review Letters* (=physics)

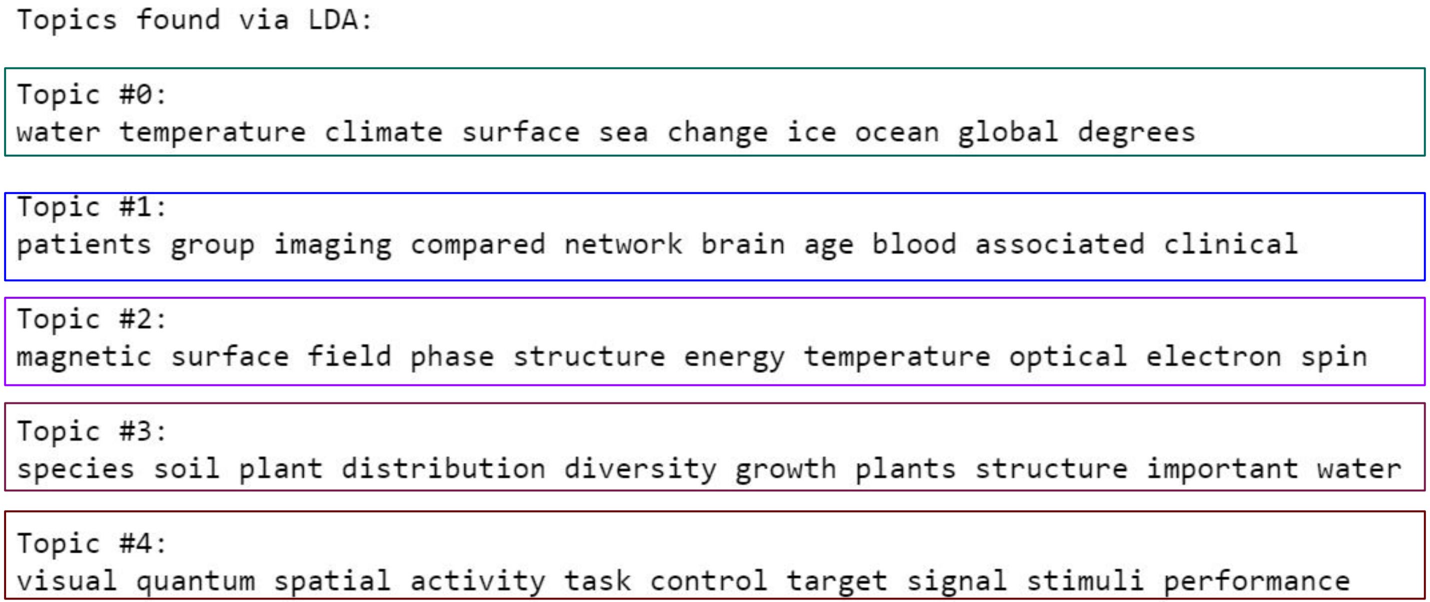


Wordcloud based on raw frequency of words



- ➢ The content of the entire abstract column for the sources (Physics),
- ➢ Algorithm to extract the Top 10 most frequent words in the physical letters abstracts column.
- ➢ As seen, common words are dominant in the frequency chart.
- ➢ That is why, we apply topic modeling in later steps, to cluster the text based on the content of the subjects.

LDA Topic Modeling Results that extracts topics from Physics Research Papers

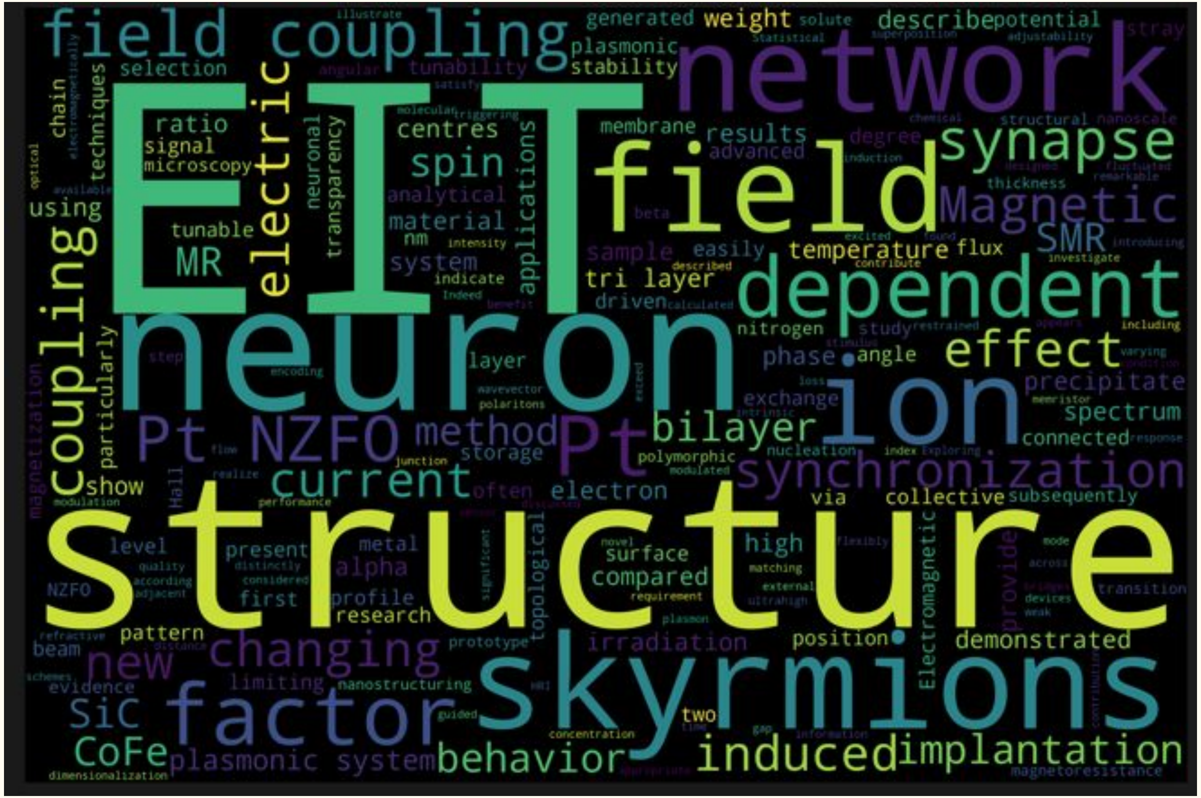


```
Topics found via LDA:

Topic #0:
water temperature climate surface sea change ice ocean global degrees

Topic #1:
patients group imaging compared network brain age blood associated clinical

Topic #2:
magnetic surface field phase structure energy temperature optical electron spin

Topic #3:
species soil plant distribution diversity growth plants structure important water

Topic #4:
visual quantum spatial activity task control target signal stimuli performance
```

LDA (Latent Dirichlet Allocation): Topic modeling is the process of identifying topics in a set of documents. There are a few ways of doing topic modeling. One is LDA works by first making a key assumption: the way a document was generated was by picking a set of topics and then for each topic picking a set of words. Physical Review Letters Topics and the top terms that have high association with those topics. Unsupervised learning clusters words and terms that occurs together and assigns weights accordingly. 5 different topics are extracted here. Topics itself found by LDA. However It doesn't name the topics.

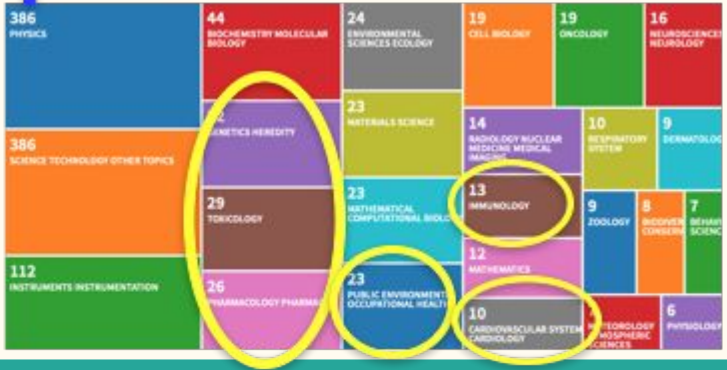
Validated model (train-test-split) improves results: **F1=0.47**

(3128 *physics* papers predicted in *Scientific Reports*)



Summary of results

Procedure	score	No. physics papers identified	Web of Science	Overlap
Fitting / predicting, no training, 8 categories	F1 = 0.13	6204	1479	524
Train/test/split on known papers, 1 category	model score = 0.98 F1 = 0.98	98% in test set	N/A	N/A
Testing on unknown papers; released model	F1 = 0.47	3128	1479	1091
Assess *Web of Science*: Remove 116 papers in clearly non-physics research areas that are also classified as physics in *Web of Science*	F1 = 0.48	3128	1363	1091
Assess *Web of Science:* 2037 "false positive" papers. Random sample of 100: About 50% of these ARE physics (misclassified in *Web of Science*) ➔ Assume 50% of 2037 "false positives" (1018 papers) are "true positive"	F1 = 0.76	3128	2381	2109



Genetics Heredity
Toxicology
Pharmacology Pharmacy
Public Env. Ocupp. Health
Immunology
Cardiology

Conclusions

- Machine Learning model to classify research papers in category “physics”
- Compare results with Web of Science classification
- Validated model works: F=0.47
- Model can be improved
- Aided by our model, we can improve the *Web of Science* classification

REFERENCES

- https://towardsdatascience.com/topic-modeling-and-latent-dirichlet-allocation-in-python-9bf156893c24
- https://towardsdatascience.com/end-to-end-topic-modeling-in-python-latent-dirichlet-allocation-lda-35ce4ed6b3e0
- https://www.mikulskibartosz.name/word-cloud-from-a-pandas-data-frame/
- https://clarivate.com/products/web-of-science/
- https://en.wikipedia.org/wiki/Web_of_Science