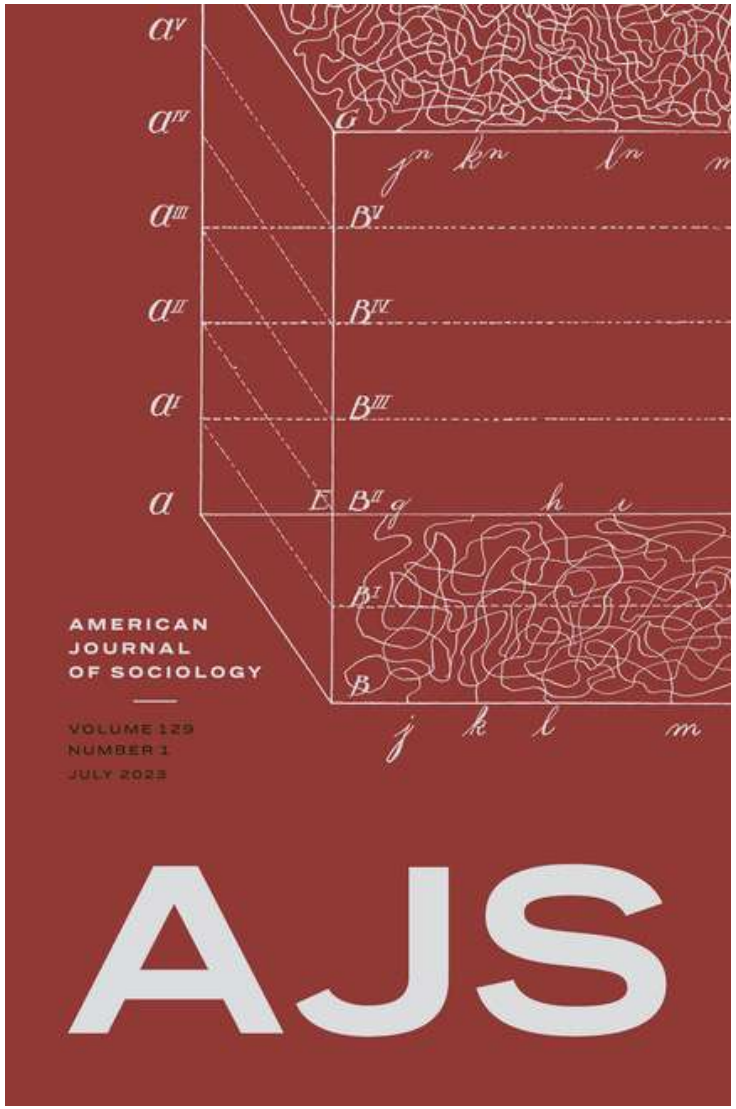
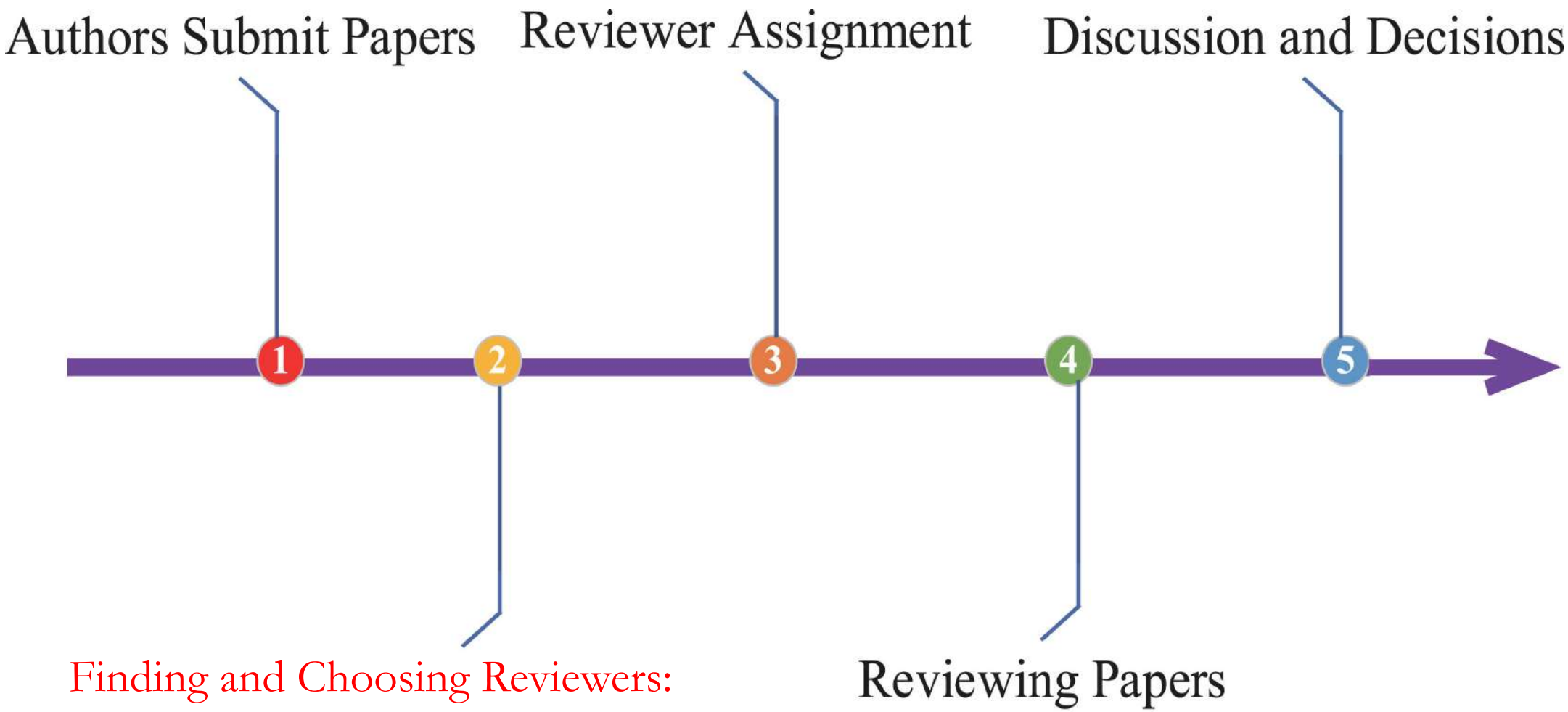
Data-Driven Reviewer Selection

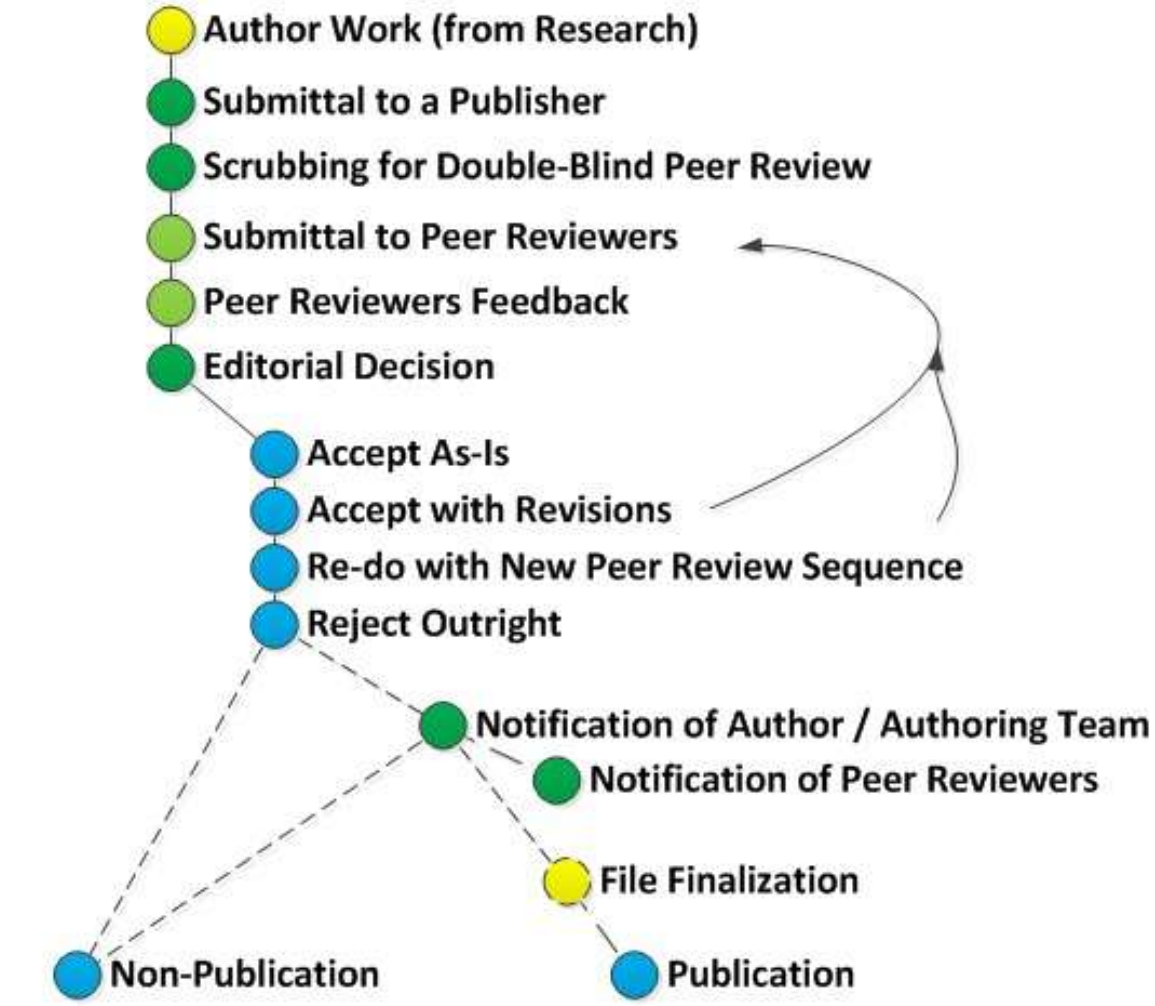
Using data science to improve scientific peer review




Authors Submit Papers
Reviewer Assignment
Discussion and Decisions
1
2
3
4
5
Finding and Choosing Reviewers:
the **Reviewer Selection Problem**
Reviewing Papers

The **Reviewer** Selection Problem

- Peer Review is crucial for scientific publishing…
- …but finding reviewers is
  - slow,
  - not accurate,
  - and biased



A Simplified Overview of the Academic Peer Review Process

PROPOSITION

- We will build an automated reviewer selection system that is:
  - Fast,
  - More or equally accurate than current solutions,
  - Less biased,
  - Actually useful!

- We will learn
  - Natural Language Processing
  - Machine Learning for graph data

Final result:

Input: a manuscript pdf

Output: a list of suitable reviewers

Data:

- A large database of papers from the OpenAlex API

OpenAlex is a fully open catalog of the global research system. It's named after the ancient Library of Alexandria and made by the nonprofit OurResearch.

This is the technical documentation for the **OpenAlex API.** Here, you can learn how to set up your

Data, details:

- We have
  - References = network of papers citing other papers
  - Abstract and keywords = textual data
  - Author's name = link author's other publications

For data set **MAG papers** and **AMiner papers**, each paper is a JSON object. Its data schema is:

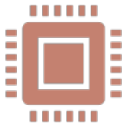
Field Name	Field Type	Description	Example
id	string	MAG or AMiner ID	53e9ab9eb7602d970354a97e
title	string	paper title	Data mining: concepts and techniques
authors.name	string	author name	Jiawei Han
author.org	string	author affiliation	department of computer science university of illinois at urbana champaign
venue	string	paper venue	Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial
year	int	published year	2000
keywords	list of strings	keywords	["data mining", "structured data", "world wide web", "social network", "relational data"]
fos	list of strings	fields of study	["relational database", "data model", "social network"]
n_citation	int	number of citation	29790
references	list of strings	citing papers' ID	["53e99ef4b7602d97027c2346", "53e9aa23b7602d970338fb5e", "53e99cf5b7602d97025aac75"]
page_stat	string	start of page	11
page_end	string	end of page	18
doc_type	string	paper type: journal, book title...	book
lang	string	detected language	en
publisher	string	publisher	Elsevier
volume	string	volume	10
issue	string	issue	29
issn	string	issn	0020-7136
isbn	string	isbn	1-55860-489-8
doi	string	doi	10.4114/ia.v10i29.873
pdf	string	pdf URL	//static.aminer.org/upload/pdf/1254/370/239/53e9ab9eb7602d970354a97e.pdf
url	list	external links	["http://dx.doi.org/10.4114/ia.v10i29.873", "http://polar.lsi.uned.es/revista/index.php/ia/ article/view/479"]
abstract	string	abstract	Our ability to generate...

Strategy for each manuscript:

(1) Constructing the candidate reviewer database.

Data and metric: closeness to the manuscript in a citation graph

Intuition: author's who cite the same papers are likely to be know enough to review each other's work



(2) Computing the matching degree between every paper–reviewer pair.

Data and metric: similarity based on natural language processing of keywords and abstracts

Intuition: authors who write about the same topics can review each other's work

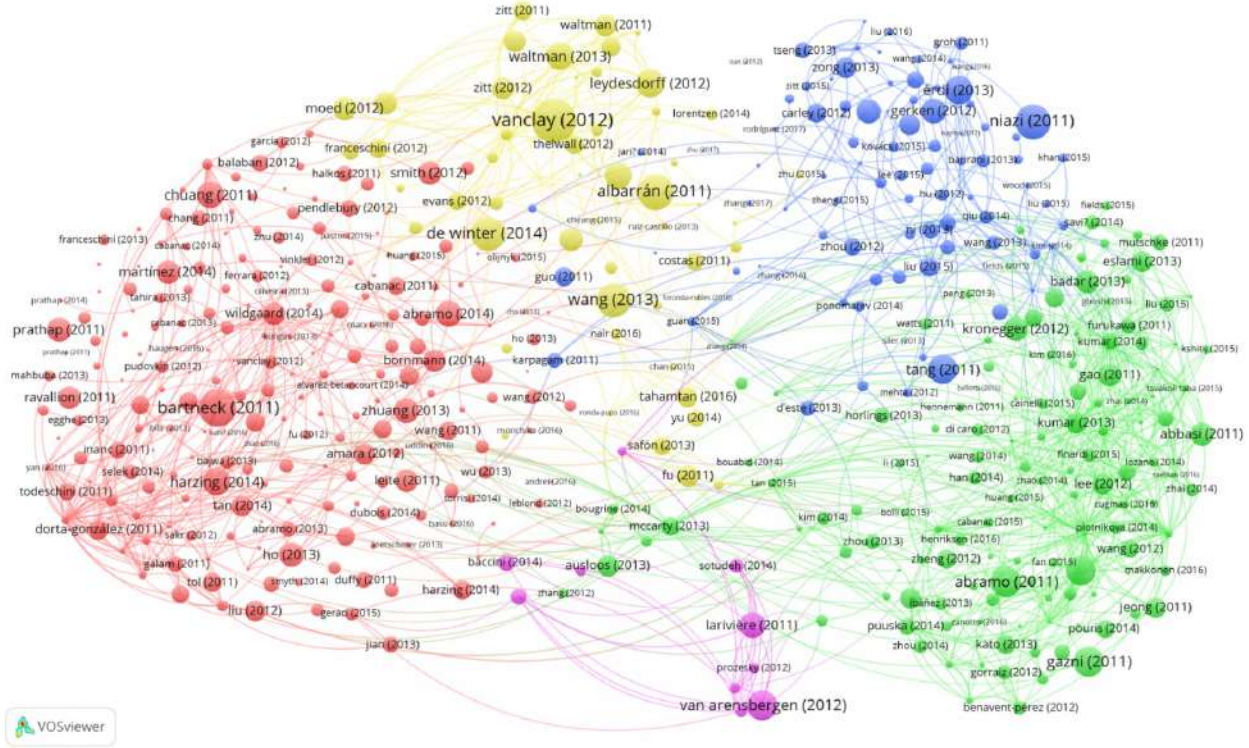
(3) Based on the matching degree matrix, choosing the best reviewer from the candidates

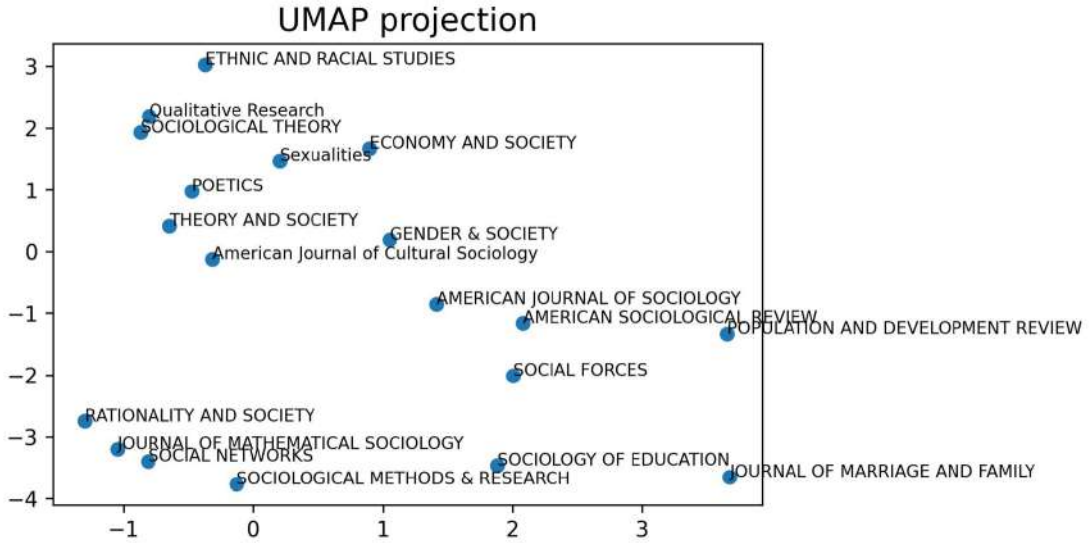
Evaluating and tuning of the model based on a "ground truth" dataset

Minimal model: a simple classification task

		FEATURE 1	FEATURE 2	BINARY TARGET
		citation_distance	textual_similarity	is_suitable_reviewer
OBSERVATION 1	reviewer_1			0
OBSERVATION 2	reviewer_2			0
OBSERVATION 3	reviewer_3			0
OBSERVATION 4	reviewer_4			1
OBSERVATION 5	reviewer_5			0
OBSERVATION 6	reviewer_6			0
OBSERVATION 7	reviewer_7			1
OBSERVATION 8	reviewer_8			0
OBSERVATION 9	reviewer_9			0
OBSERVATION 10	reviewer_10			0

Maximal model: a full citation graph and embeddings of all papers





But how will we know if it works?

- We use an existing dataset of suitable reviewers to evaluate and tune our model!

- This is courtesy of the open-source journal Frontiers