



Dr. Eva Seidlmayer

**Towards (Automatic) Quality Control of Life Science Articles.  
Between Censorship and Good Research Practice**

© ZB MED / Sima Deghani, die Abbildung steht unter der Lizenz CC BY-ND 4.0



Session 071 Infodemic Management:  
**Strategies for Combatting Health Mis/Dis/Malinformation**  
July 26th 2022



# Outline

1. *Status quo* on mis-information at ZB MED holdings
2. Quality control at ZB MED: Providing information on
  - Good scientific practice
  - Automated classification of mis-information

# Status quo: ZB MED Discovery system LIVIVO

- Aggregating 50 databases (MEDLINE, AGRICOLA, BASE, library catalogues...)
- >70 M items
- No quality filtering

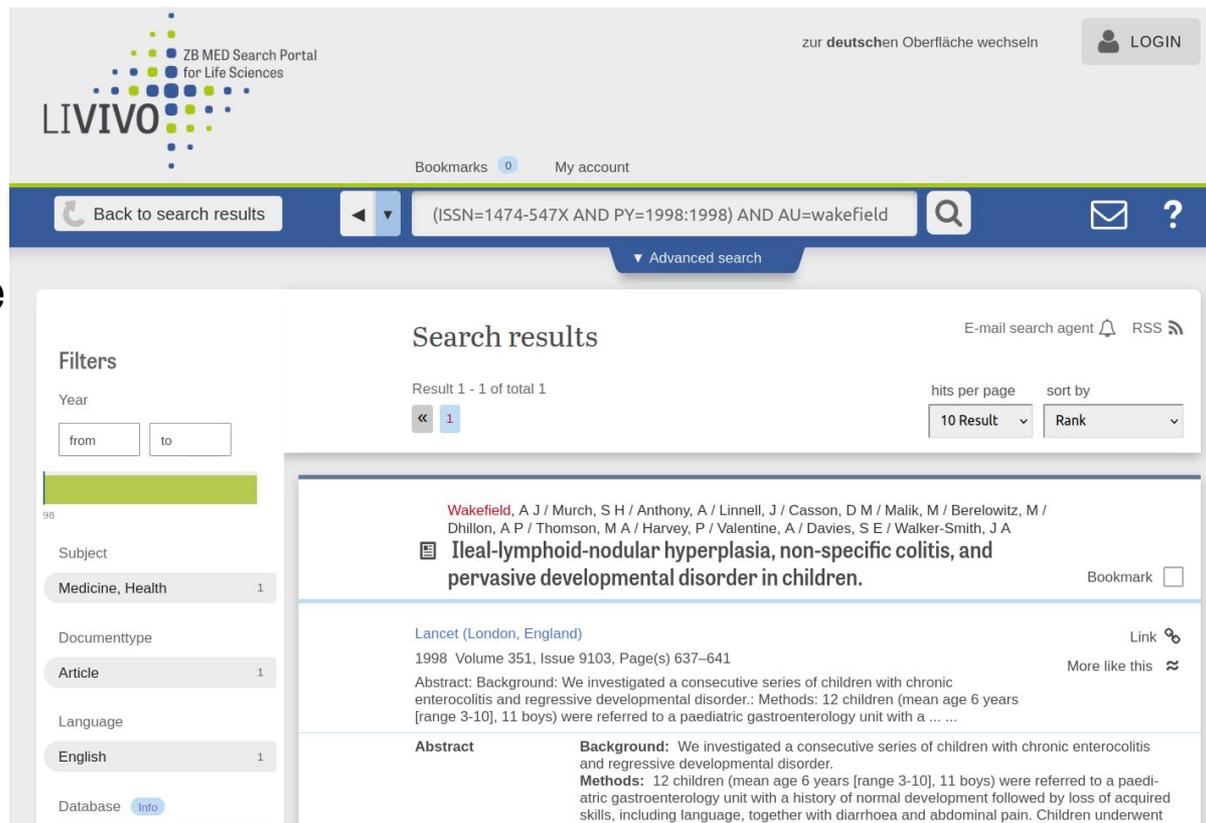


The screenshot shows the LIVIVO search portal interface. At the top right, there is a link to switch to the German interface ('zur deutschen Oberfläche wechseln') and a 'LOGIN' button. The main header features the LIVIVO logo and the text 'ZB MED Search Portal for Life Sciences'. Below the logo, there are links for 'Bookmarks' (0) and 'My account'. A search bar is prominently displayed with the placeholder text 'Enter your search terms here', a search icon, and a mail icon. Below the search bar is a link for 'Advanced search'. A central banner with a green checkmark icon announces a special LIVIVO COVID-19 collection, stating: 'LIVIVO also offers a special LIVIVO COVID-19 collection. We have developed a special search for COVID-19 / SARS-CoV-2: the [LIVIVO COVID-19 collection](#).' At the bottom, the LIVIVO logo and tagline 'The Search Portal for Life Sciences' are on the left, and a horizontal list of search topics is on the right, including: Animal Protection, Biodiversity, Bioenergy, Breast Cancer, Carcinogenes, Climate Change, Colony Collapse Disorder, COVID-19, Crispr Cas, Dementia, Food Allergy, Glyphosate, Hospital Hygiene, HPV Vaccination, Influenza, Interleukin-33, Jumping Gene, Malaria, Microplastics, Nanomaterial, Nuclear Disaster, Nutrigenomics, Oncogene, Organ Donation, Organic Farming, Pesticide Residues, RuBisCo, SARS-CoV-2, Vegan Diet, Zika, and Zoonosis.

<https://www.livivo.de/>

# Status quo: Mis-information is widespread

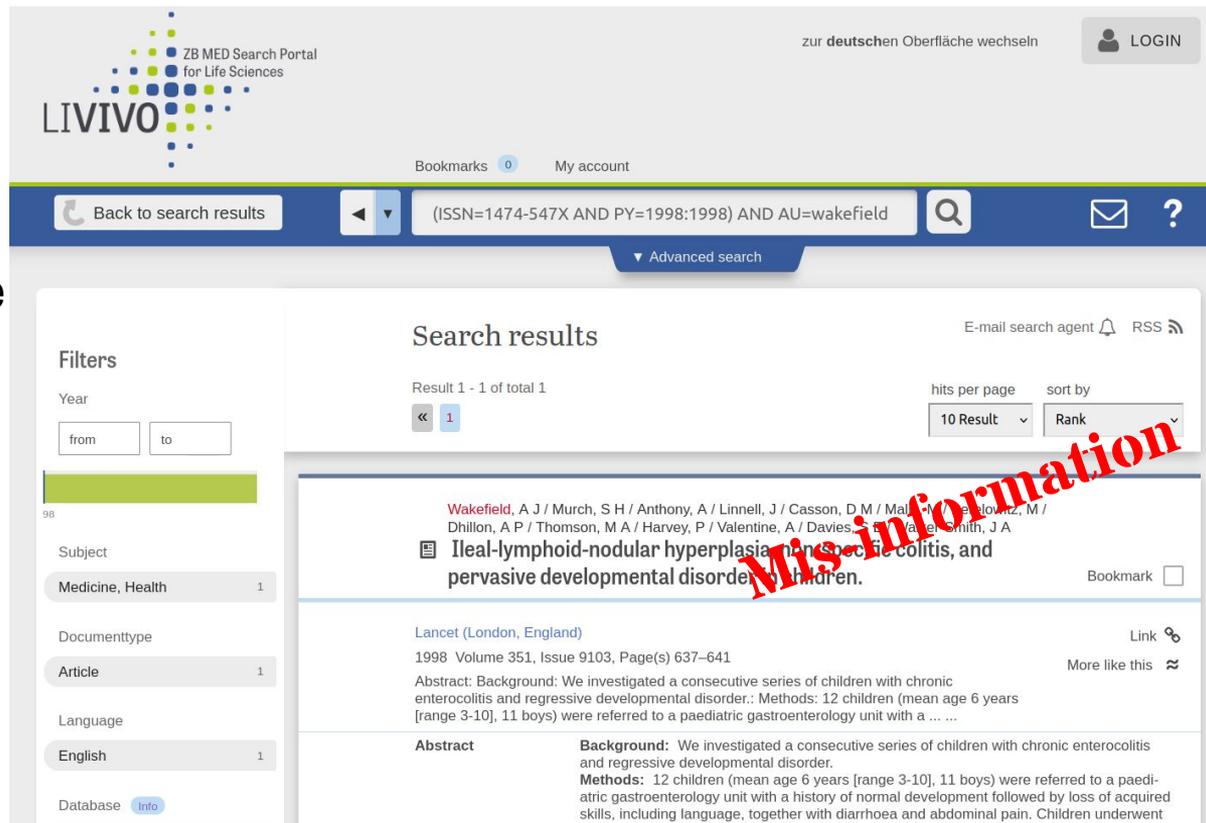
- Mis-information in LIVIVO
- Scientific literature portals are affected (*Holone 2016*)
- Mis-information is everywhere in public spaces, but also in (Life) Sciences
  - Homeopathy (*EU 2021a*)
  - WHO: One of the ten greatest health hazards worldwide: Vaccination refusal (measles...) (*WHO 2019*)



The screenshot shows the LIVIVO search portal interface. At the top, there is a search bar with the query: (ISSN=1474-547X AND PY=1998:1998) AND AU=wakefield. The search results section displays one result from the Lancet journal, titled "Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children." The article is from 1998, Volume 351, Issue 9103, pages 637-641. The abstract and background information are visible, detailing the investigation of a consecutive series of children with chronic enterocolitis and regressive developmental disorder.

# Status quo: Mis-information is widespread

- Mis-information in LIVIVO
- Scientific literature portals are affected (*Holone 2016*)
- Mis-information is everywhere in public spaces, but also in (Life) Sciences
  - EU: Homeopathy (*EU 2021a*)
  - WHO: One of the ten greatest health hazards worldwide: Vaccination refusal (measles...) (*WHO 2019*)



The screenshot shows the LIVIVO search portal interface. At the top, there is a search bar with the query: (ISSN=1474-547X AND PY=1998:1998) AND AU=wakefield. Below the search bar, the search results are displayed. The first result is a paper by Wakefield, A J / Murch, S H / Anthony, A / Linnell, J / Casson, D M / Mallik, D / Lewis-Jones, M / Dhillo, A P / Thomson, M A / Harvey, P / Valentine, A / Davies, S / Walker, S / Smith, J A. The title of the paper is "Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children." The abstract is visible, and a large red watermark reading "Mis-information" is overlaid on it. The interface also shows filters for Year, Subject (Medicine, Health), Documenttype (Article), Language (English), and Database.

## Status quo: Mis-information cannot be completely filtered out

- Technical difficulties
- Often part of correct context (*Gensing 2020*)
- Complex character: intended disinformation, retracted incorrect information, simplified popular information...
- Difficult distinction: new / unpopular / critical work and mis-information
- Particular research question is unknown

→ No censorship of suspicious items but provision of information

→ Strengthening information literacy (*CILIP 2018*)



## What to do?

# Solution

Quality control at *ZB MED*: Providing information on

- Good scientific practice
- Automated classification of mis-information

Quality control at *ZB MED*: Providing information on

- **Good scientific practice**
- Automated classification of mis-information

# Solution Part 1: Aspects of good scientific practice

- *German National Research Alliance (DFG):*  
 „Guidelines for Safeguarding Good Research Practice. Code of Conduct” (DFG 2019)
- 19 Guidelines
- Guideline No. 7:

“**If researchers** have made their findings publicly available and subsequently **become aware of inconsistencies or errors** in them, **they make the necessary corrections.** [...]”

**The origin of the data, organisms, materials and software** used in the research process **is disclosed** and the reuse of data is clearly indicated; **original sources are cited.** “ (DFG 2019)

# Solution Part 1: Aspects of good scientific practice

Quality features:

- Peer-review?
- Scientific references?
- Already cited?
- Has it been retracted?

# Solution Part 1: Aspects of good scientific practice

Quality features:

- Peer-review?  
→ Metadata: data type, publishing journal
- Scientific references?
- Already cited?
- Has it been retracted?

# Solution Part 1: Aspects of good scientific practice

## Quality features:

- Peer-review?
  - Metadata: data type, publishing journal
- Scientific references?
  - Reference data: *Crossref* database
- Already cited?
  - Reference data: *Crossref* database
- Has it been retracted?



<https://www.crossref.org>

## Solution Part 1: Aspects of good scientific practice

Quality features:

- Peer-review?
  - Metadata: data type, publishing journal
- Scientific references?
  - Reference data: *Crossref* database
- Already cited?
  - Reference data: *Crossref* database
- Has it been retracted?
  - *Retraction Watch* database



<https://www.crossref.org>

<https://retractionwatch.com>

# Solution Part 1: Aspects of good scientific practice

5 **Wakefield, A J / Murch, S H / Anthony, A / Linnell, J / Casson, D M / Malik, M / Berelowitz, M / Dhillon, A P / Thomson, M A / Harvey, P / Valentine, A / Davies, S E / Walker-Smith, J A**

**Leal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children.** Bookmark

---

**Lancet** (London, England) Link

1998 Volume 351, Issue 9103, Page(s) 637–641 More like this

Abstract: Background: We investigated a consecutive series of children with chronic enterocolitis and regressive developmental disorder.: Methods: 12 children (mean age 6 years [range 3-10], 11 boys) were referred to a paediatric gastroenterology unit with a ... ..

More links ▶
 peer-reviewed
 scientifically referencing
 scientifically referenced
 not listed as retracted

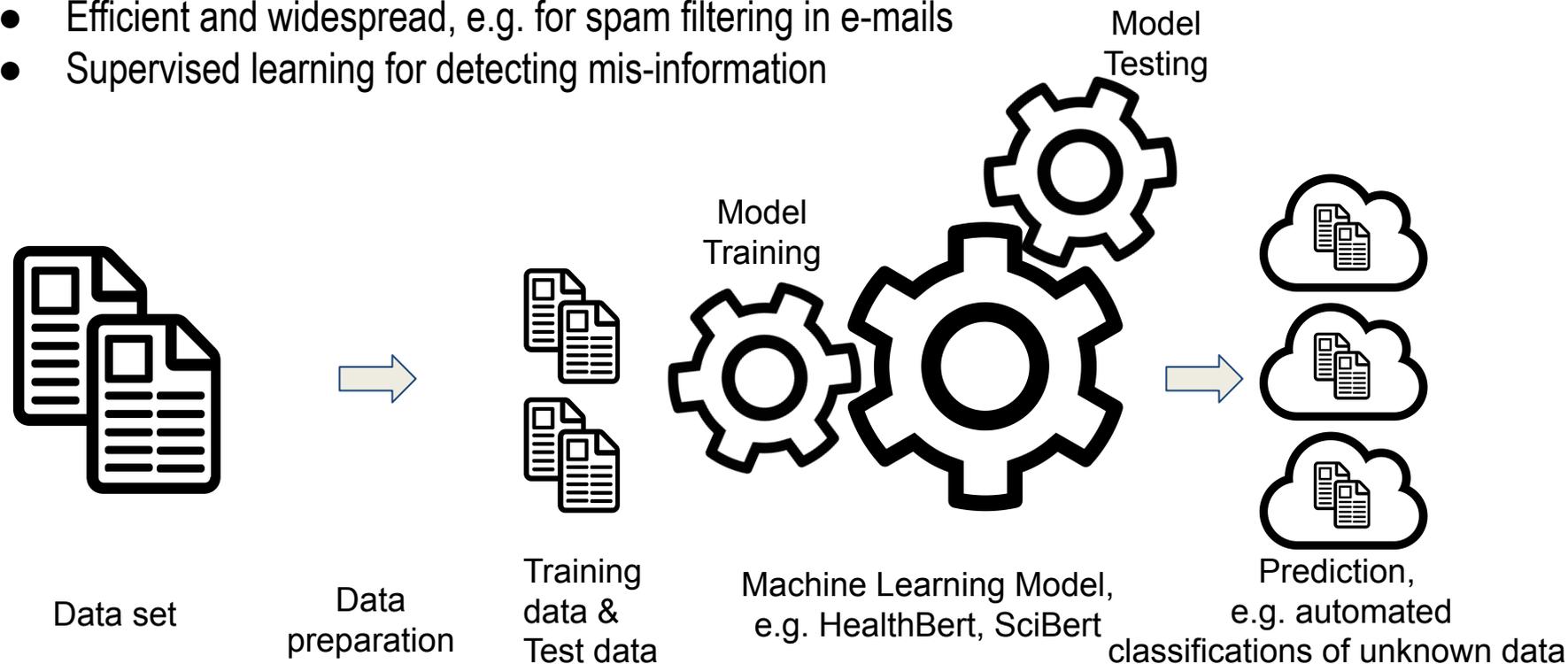
Details ▼
Full text online 
See ZB MED holdings
Order with fees

# Solution

1. Quality control at *ZB MED*: Providing information on
  - Good scientific practice
  - **Automated classification of mis-information**

## Solution Part 2 : Automated classification

- Efficient and widespread, e.g. for spam filtering in e-mails
- Supervised learning for detecting mis-information



## Solution Part 2: Automated classification

- Classification for supervised learning approach is required
- Basic (e.g.):
  - Scientific information
  - Mis-information
- More differentiated (e.g.):
  - Scientific information
  - Mis-information
  - Popular information



## Solution Part 2: Automated classification

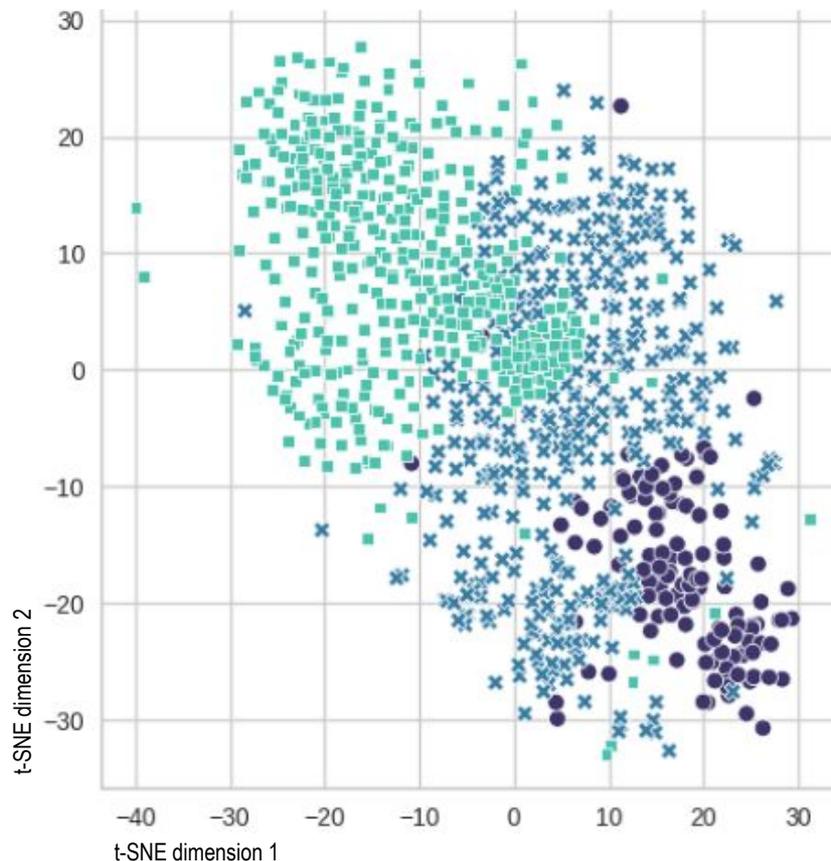
- Data set for training and testing
- Full texts for all classes
- Preprocessing to avoid overfitting (*Oshikawa/Qian/Wang 2020*)
- Reuse of data sets from Life Sciences
  - PUBHEALTH (*Kotonya/Toni 2020*)
  - Health&Well Being (HWB) Fake News Dataset (*Singh/Deepak/Anoop 2020*)
- Data set within classification:
  - Scientific texts: PubMed Central...
  - Popular science texts: MedlinePlus, Medhelp, Wikipedia, PUBHEALTH, HWB...
  - Mis-Information: Signs of the times, PUBHEALTH, HWB...



## Solution Part 2: Machine Learning classification

Unsupervised clustering of the German test data set with Doc2Vec (t-SNE projection)

- Specialized texts
- ✕ Popular-science texts
- Mis-informative texts



## Solution Part 2: Automated classification

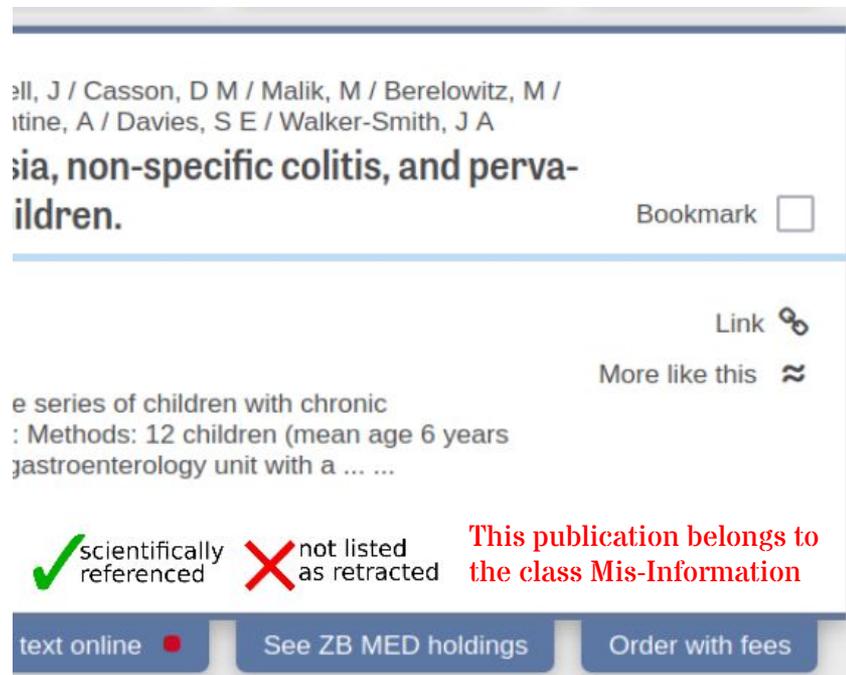
### Result representation

- Display of all similarity values, without harmonization, e.g. by using a bar chart
- Always risk of wrong classification
- Explainability of the workflow (*EU 2021b*)

## Solution Part 2: Automated classification

### Result representation

- Display of all similarity values, no harmonization, e.g. by using a bar chart
- Always risk of wrong classification
- Explainability of the workflow (EU 2021b)



...ll, J / Casson, D M / Malik, M / Berelowitz, M /  
 ...tine, A / Davies, S E / Walker-Smith, J A  
**...ia, non-specific colitis, and perva-**  
**...ildren.**

Bookmark

Link 

More like this 

e series of children with chronic  
 : Methods: 12 children (mean age 6 years  
 gastroenterology unit with a ... ..

 scientifically referenced
  not listed as retracted
 **This publication belongs to the class Mis-Information**

text online 
 See ZB MED holdings
 Order with fees

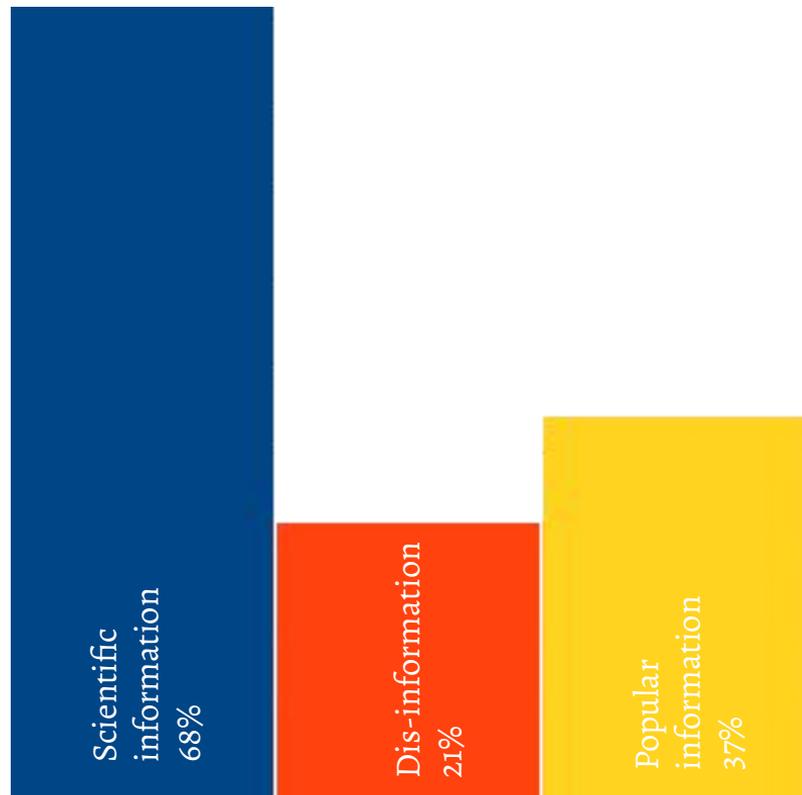
## Solution Part 2: Automated classification

### Result representation

- Display of all similarity values, no harmonization, e.g. by using a bar chart
- Always risk of wrong classification
- Explainability of the workflow (*EU 2021b*)

This publication has a high similarity to titles from the field “scientific information”.

What does this mean?



# Solution Part 2: Automated classification

5 **Wakefield, A J / Murch, S H / Anthony, A / Linnell, J / Casson, D M / Malik, M / Berelowitz, M / Dhillon, A P / Thomson, M A / Harvey, P / Valentine, A / Davies, S E / Walker-Smith, J A**

**Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children.** Bookmark

---

**Lancet** (London, England) Link

1998 Volume 351, Issue 9103, Page(s) 637-641 More like this

Abstract: Background: We investigated a consecutive series of children with chronic enterocolitis and regressive developmental disorder.: Methods: 12 children (mean age 6 years [range 3-10], 11 boys) were referred to a paediatric gastroenterology unit with a ... ..

[More links ▶](#) peer-reviewed scientifically referencing scientifically referenced not listed as retracted

This publication has a high similarity to titles from the field "scientific information". [What does this mean?](#)

[Details ▼](#) [Full text online](#)  [See ZB MED holdings](#) [Order with fees](#)

## Solution Part 2: Automated classification

### Potential

- Powerful capacity for classifying full texts in the future

### Challenges

- Careful selection of dataset for training
- Overfitting in training
- Risk of wrong classification by the machine learning model
- In use: Lack of full texts for publications in library holdings
- Results presentation of all similarity values
- Risk of loss of confidence in machine learning

# Summary

- Mis-information widespread - also in science
- Data literacy is better than censorship
- Provision on information compliance to good scientific practice, such as:
  - Peer-review status
  - References to scientific literature
  - Cited by scientific literature
  - Retraction status
- Machine learning can be used for automated classification, but:
  - Risk of overfitting and wrong classification
  - Recommendation: Display similarity values to all classes
  - Explainability: Transparent machine learning workflow

# Acknowledgements to:

YOU!

Prof. Dr. Konrad U. Förstner  
and my dear colleagues  
in the Data Sciences and Services Unit  
at ZB MED - Information Centre for Life  
Sciences



## References:

- Crossref, online: <https://www.crossref.org/> (2022-07-15).
- Chartered Institute of Library Information Professionals (CILIP) (2018). **Definitions of Information Literacy 2018**, <https://infolit.org.uk/ILdefinitionCILIP2018.pdf> (2022-07-16).
- DFG (2019). **Guidelines for Safeguarding Good Research Practice**, doi: [10.5281/zenodo.6472827](https://zenodo.org/record/6472827), online: [https://www.dfg.de/en/research\\_funding/principles\\_dfg\\_funding/good\\_scientific\\_practice/](https://www.dfg.de/en/research_funding/principles_dfg_funding/good_scientific_practice/) (2022-07-15).
- European Commission, Directorate-General for Communications Networks, Content and Technology (2021a). **European Commission Guidance on Strengthening the Code of Practice on Disinformation**. <https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX%3A52021DC0262>.
- European Commission (2021b). **Proposal for a Regulation of the European Parliament and of the council laying down harmonised rules on artificial intelligence artificial intelligence act and amending certain union legislative acts**. 2021/0106 (COD).
- Gensing, Patrick (2020). **Fakten gegen Fake News oder Der Kampf um die Demokratie**. Schriftenreihe Band 10500. Bonn: Bundeszentrale für politische Bildung.
- Holone, Harald (2016). **The filter bubble and its effect on online personal health information**. In: Croatian Medical Journal 57.3. doi: 10.3325/cmj.2016.57.298, p. 298–301.
- Kotonya, Neema und Francesca Toni (Okt. 2020). **Explainable Automated Fact-Checking for Public Health Claims**. In: arXiv:2010.09926 [cs]. arXiv: 2010.09926.
- LIVIVO, online: [www.livivo.de](http://www.livivo.de) (2022-07-15).
- Oshikawa, Ray, Jing Qian und William Yang Wang (2020). **A Survey on Natural Language Processing for Fake News Detection**. In: arXiv:1811.00770 [cs]. arXiv: 1811.00770.
- Retraction Watch, online: <https://retractionwatch.com/> (2022-07-15).
- Singh, Iknor, P. Deepak und K. Anoop (2020). **On the Coherence of Fake News Articles**. In: ECML PKDD 2020 Workshops. Ed. by Irena Koprinska et al. Vol. 1323. Cham, p. 591–607.
- WHO (2019). online: **Ten threats to global health in 2019** <https://www.who.int/news-room/spotlight/ten-threats-to-global-health-in-2019> (2022-07-15).
- ZB MED, online: [www.zbmed.de](http://www.zbmed.de) (2022-07-15).

Eva Seidlmayer, Dr. phil., M.LIS  
Data Sciences and Services, Research Fellow  
ORCID: 0000-0001-7258-0532  
Twitter: @kivilih

ZB MED - Information Centre for Life Sciences  
Gleueler Straße 60  
50931 Cologne  
Germany

seidlmayer@zbmed.de  
<http://www.zbmed.de/>

INFORMATION. KNOWLEDGE. LIFE.