# Data-Extraction Automation in Systematic Reviews – A 2024 Living Review Update

Lena Schmidt

# Why do we need automation and living reviews?

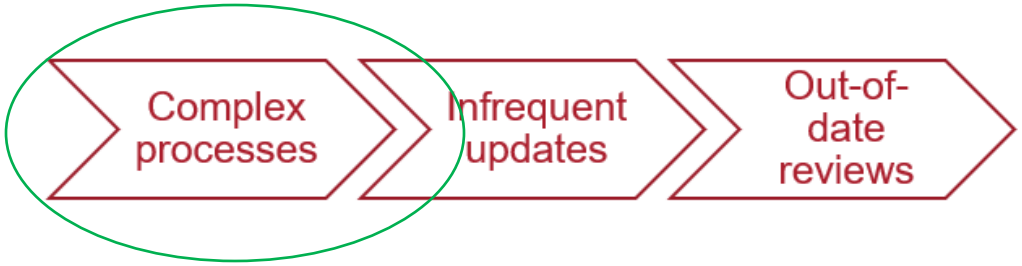
Expectation

Original vision for Cochrane:

*"include a library of trial overviews, which will be updated when new data become available"*[1]

**vs**

Reality



# Objective of living review of automated data extraction methods

- Increasing rate of published research
- LSR focuses on automating extraction of variables interesting for clinical SR and characteristics of studies (e.g. PICOs, N, design, ..)

1. Review published **methods and tools for data extraction** to (semi)automate the systematic reviewing process.
2. Review this evidence in the scope of a **living review**. To keep information **up to date and relevant** to the challenges faced by systematic reviewers at any time

Lena Schmidt, Ailbhe N. Finnerty Mutlu, Rebecca Elmore, Babatunde K. Olorisade, James Thomas, Julian P.T Higgins
Data extraction methods for systematic review (semi)automation: Update of a living systematic review [version 3; peer review: 3 approved]. *F1000Research* 2025, **10**:401 (https://doi.org/10.12688/f1000research.51117.3)

# Why Automate Data Extraction?

- Manual extraction:
  - Slow
  - Costly
  - error-prone

**→ automation looks tempting!**

# How this review works

- Repeated searches of up to 5 databases, changes across 5 years!
- 117 includes (to Aug 2024)
- ACL, ArXiv.cs, dblp, PubMed, EPPI-Reviewer+OpenAlex

University of BRISTOL

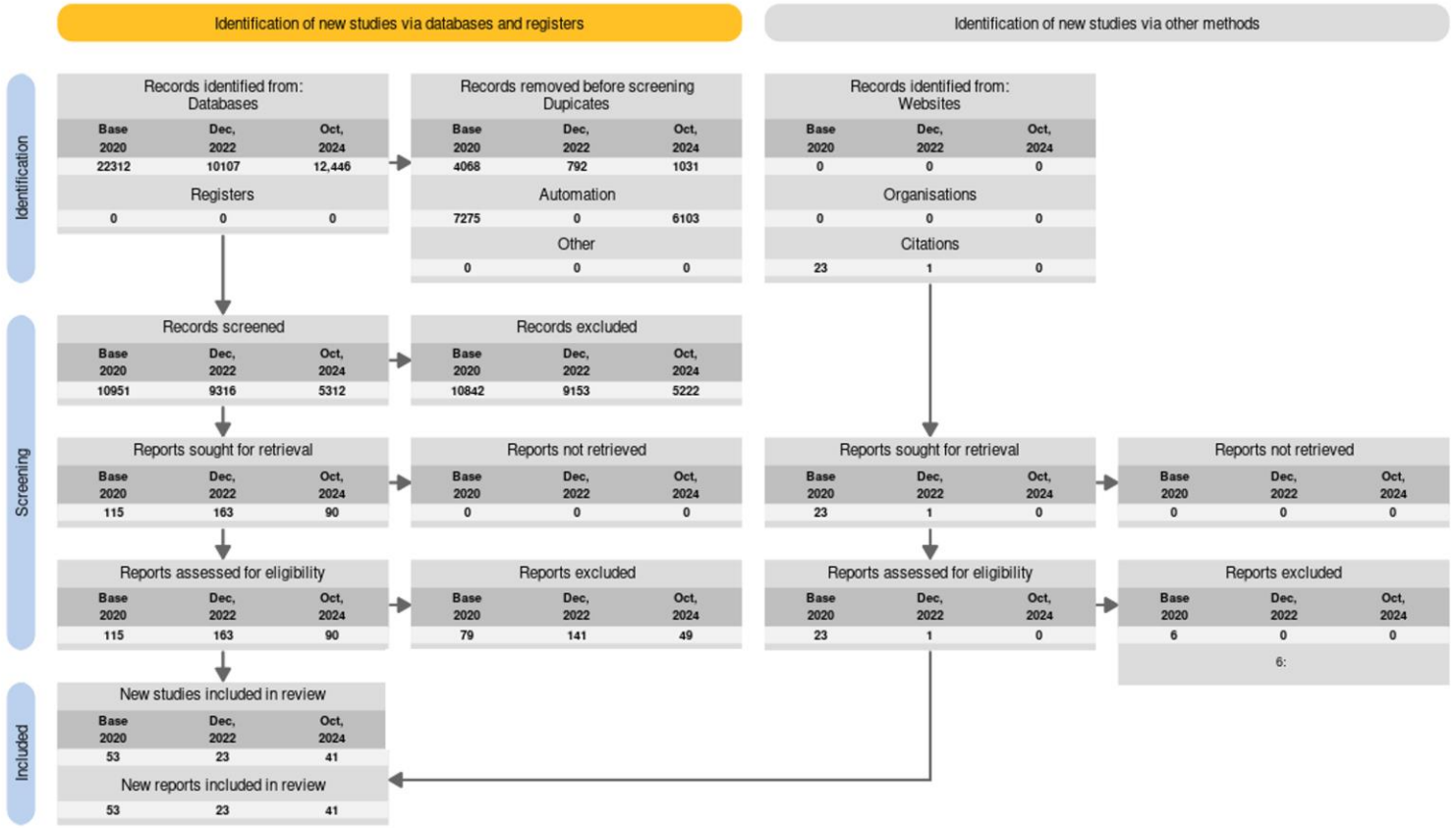
NHS University Hospitals Bristol NHS Foundation Trust

Databases for continuous searches (Data-extraction LSR)

| Traditional | Specialised computer science databases | Preprint/other |
|---|---|---|
| PubMed | Association for Computational Linguistics | arXiv.org |
| EPPI REVIEWER + OpenAlex | dblp computer science bibliography | SR TOOLBOX |

* *Included only in base-review:* WEB OF SCIENCE, IEEE Advancing Technology for Humanity

Funded by NIHR



# Assessment of included papers

## Data Extraction

- AI approaches used (e.g. machine-learning, LLMs)
- Metrics used for reporting results (e.g. Precision, Recall)
- Type of data
  - Study types (e.g. RCT)
  - Fields (e.g. Abstract vs. Full Text)
  - Model in- and Output (e.g. RIS file w. abstracts, JSON)
- Granularity of extraction (Named Entities vs. sentence-classification)
- Other outcomes defined by papers (e.g. time saved)

## Quality of Reporting

- Assessed the quality of reporting of all included papers WRT. 5 domains:
  - **Reproducibility:** Data sources and processing of data described?
  - **Transparency of methods:** Algorithms, data characteristics, hardware,source-code described or available?
  - **Testing:** Model assessment, basic metrics, precision/recall tradeoffs described?
  - **Availability of the final model or tool:** Is an actual tool (not scripts) for end-useds available, and is the dataset accessible for training own instances
  - **Internal/external validity of the model:**
    - Internal: overfitting avoided, description of separate train/test data, tested on multiple datasets
    - External: comparability to other tools using the same/similar datasets.

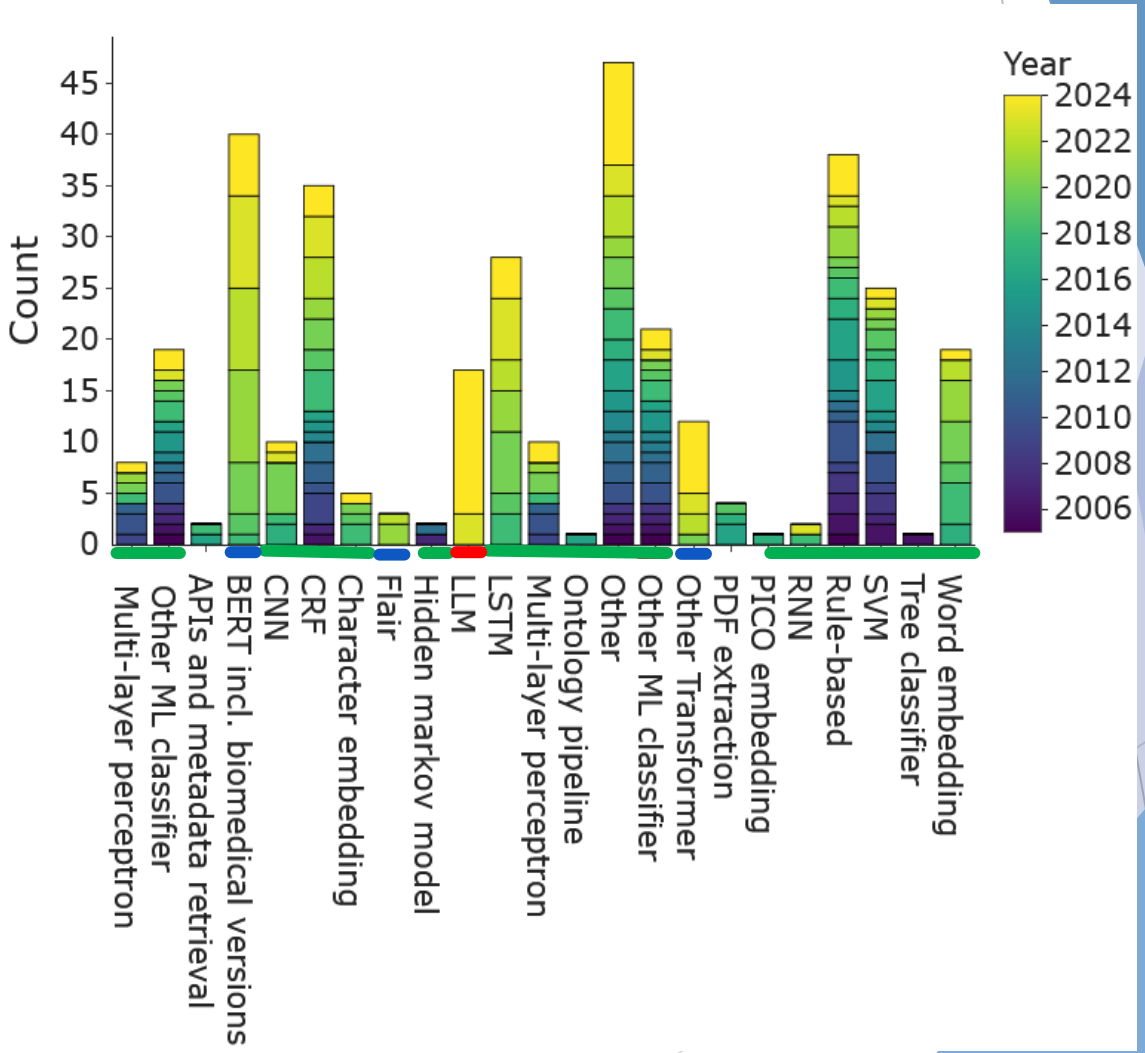
Additionally, we collected and summarized information on caveats described by report authors


All interactive maps are here: https://l-ena.github.io/living_review_data_extraction/LatestUpdates.html

# Research Output Keeps Climbing

- Review now cover three broadly distinct ‘eras’ of automation

1. Pre-2018, **array of methods**: rule-base, machine-learning, embedding, and first neural networks (e.g. LSTM)
2. 2018-2022, **‘discriminative’ Transformer models** based on BERT, classifying input texts after fine-tuning
3. 2022-now, **‘generative’ Transformers** such as GPT models, with zero-, few-shot classification and fine-tuning



# Datasets & Code availability: A Success Story ?

- Table 4 lists 76 unique datasets; almost half downloadable (EBM-NLP, PubMed-PICO, EvidenceInference 2.0 …). We have LOTS of data!
- Code sharing jumped from 15 % (pre-2021) to 42 % and all repositories are listed in Table 2. We have LOTS of code and implemented methods!
- This should enable comparability, re-usal and progress – at least on paper

- Key Problems:
  - Only very few of the benchmarking datasets are actively re-used
  - Those who are re-used often get adapted/improved/extended in new publications, limiting comparability
  - No consensus on validation scripts; different application of same metrics
  - New architectures such as LLMs require different datasets and validation metrics

# Corpora for training/evaluation

| Publication | Also used by | Name | Description | Classes | Size/type | Availability | Note |
|---|---|---|---|---|---|---|---|
| 96 | 39,54,87,95,98,136 Dataset adaptations: 60, 167 | PubMedPICO | Automatically labelled sentence labels from structured abstracts up to Aug'17 | P, IC, O, Method | 24,668 abstracts | Yes, https://github.com/jind11/PubMed-PICO-Detection | |
| 55 | 32,33,36,61,74,85,95,98,100,106,130,135,138,140,157,165,178,179, Via BLURB-Benchmark: 132, 169 Dataset adaptions: 34,37,50,67,134,139,145 | EBMNLP, EBM-PICO | Entities | P, IC, O + age, gender, and more entities | 5,000 abstracts | Yes, https://github.com/bepnye/EBM-NLP | |
| 97 | | | Entities | I and dosage-related | 694 abstract/full text | Yes, https://ii.nlm.nih.gov/DataSets/index.shtml | Domain drug-based interventions |
| 48 | | | Entities | P, O, Design, Exposure | 60 + 30 abstracts | Yes, http://gnteam.cs.manchester.ac.uk/old/epidemiology/data.html | Domain obesity |
| 75 | | | Sentence level 90,000 distant supervision annotations, 1000 manual. | Target condition, index test and reference standard | 90,000 + 1000 sentences | Yes (labels, not text), https://zenodo.org/record/1303259 | Domain diagnostic tests |
| 52 | 64 (includes classifiers from), 40,53,54,102,107–110,147,153 | NICTA-PIBOSO | Structured and unstructured abstracts, multi-label on sentences. | P, IC, O, Design | 1000 abstracts | Yes, https://drive.google.com/file/d/1M9QCgrRjERZnD9LM2FeK-3jjvXJbjRTI/view?usp=sharing | Multi-label sentences |
| 47 | | | Sentences | Drug intervention and comparative statements for each arm | 300 (500 in available data) sentences | Yes, https://dataverse.harvard.edu/file.xhtml?fileId=4171005&version=1.0 | Domain drug-based interventions |
| 98 | | | Sentences | P, IC, O | 5099 sentences from references included in SRs, labelled using active-learning | Yes, https://github.com/wds-seu/Aceso/tree/master/datasets | Domain heart disease |
| 62 based | 32,61,99,171, Extending/adapting dataset: 177,149 | Evidence-inference | Sentences | P, I, O | Fulltext: 12,616 prompts | Yes, http://evidence-inference.ebm-nlp.com/ | Triplets for relation |

**Table 4. Corpora used in the included publications.**

RCT, randomized controlled trials; IR, information retrieval; PICO, population, intervention, comparison, outcome; UMLS, unified medical language system.

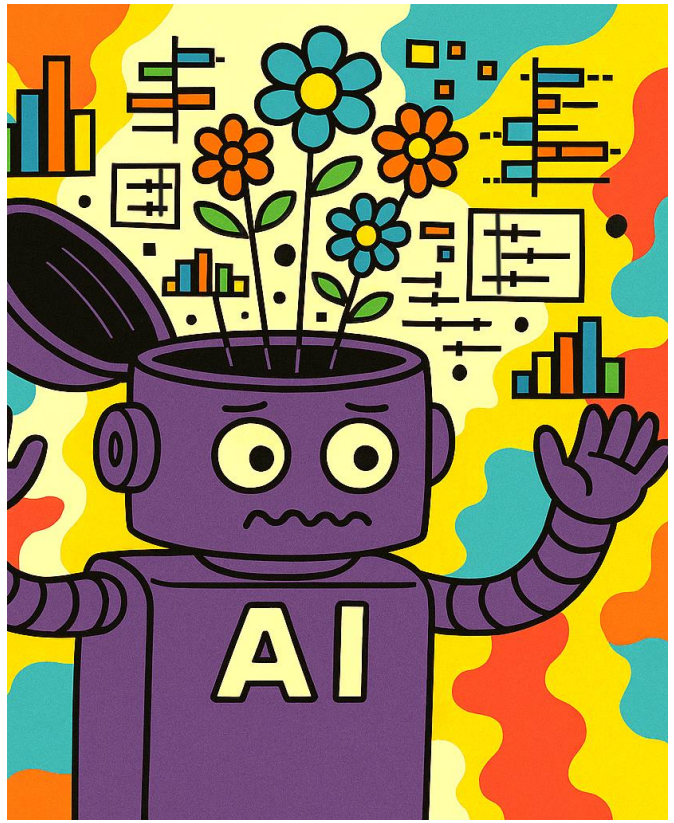
# Adaptation and evaluation problems with LLMs

**Predictors of postdischarge outcomes from information acquired shortly after admission for acute heart failure: a report from the Placebo-Controlled Randomized Study of the Selective A1 Adenosine Receptor Antagonist Rolofylline for Patients Hospitalized With Acute Decompensated Heart Failure and Volume Overload to Assess Treatment Effect on Congestion and Renal Function ( PROTECT ) Study.**

BACKGROUND Acute heart failure is a common reason for admission, and outcome is often poor. Improved prognostic risk stratification may assist in the design of future trials and in patient management. Using data from a large randomized trial, we explored the prognostic value of clinical variables, measured at hospital admission for acute heart failure, to determine whether a few selected variables were inferior to an extended data set. METHODS AND RESULTS The prognostic model included 37 clinical characteristics collected at baseline in PROTECT, a study comparing rolofylline and placebo in 2033 patients admitted with acute heart failure. Prespecified outcomes at 30 days were death or rehospitalization for any reason ; death or rehospitalization for cardiovascular or renal reasons ; and, at both 30 and 180 days, all-cause mortality. No variable had a c-index > 0.70, and few had values > 0.60 ; c-indices were lower for composite outcomes than for mortality. Blood urea was generally the strongest single predictor. Eighteen variables contributed independent prognostic information, but a reduced model using only 8 items ( age, previous heart failure hospitalization, peripheral edema, systolic blood pressure, serum sodium, urea, creatinine, and albumin ) performed similarly. For prediction of all-cause mortality at 180 days, the model c-index using all variables was 0.72 and for the simplified model, also 0.72. CONCLUSIONS A few simple clinical variables measured on admission in patients with acute heart failure predict a variety of adverse outcomes with accuracy similar to more complex models. However, predictive models were of only moderate accuracy, especially for outcomes that included nonfatal events. Better methods of risk stratification are required. CLINICAL TRIAL REGISTRATION URL: http: //www.clinicaltrials.gov. Unique identifiers: NCT00328692 and NCT00354458.

| | EBM-NLP Gold Standard | LLM prediction |
|---|---|---|
| P | 'acute heart failure :', 'Patients Hospitalized With Acute Decompensated Heart Failure and Volume Overload', 'Acute heart failure', '2033 patients admitted with acute heart failure .', 'patients with acute heart failure | 2033 patients admitted with acute heart failure |
| IC | 'Placebo-Controlled', 'Selective A1 Adenosine Receptor Antagonist Rolofylline', 'rolofylline and placebo' | comparing rolofylline and placebo |
| O | ['death or rehospitalization for any reason ; death or rehospitalization for cardiovascular or renal reasons ;', 'Blood urea', 'all-cause mortality', 'variety of adverse outcomes', 'accuracy', 'moderate accuracy', 'outcomes that included nonfatal events .' | Prespecified outcomes at 30 days were death or rehospitalization for any reason; death or rehospitalization for cardiovascular or renal reasons; and, at both 30 and 180 days, all-cause mortality |

# Note on LLM evaluations




- **Challenging to directly compare** discriminative BERT models vs. generative LLM models on data extraction tasks
- Comparisons are **more straightforward** on other tasks in the SR process, e.g. with screening decisions or automated tagging of concepts for evidence-maps: predictions are binary.
- In SR automation, **exhaustive identification** and highlighting of all relevant text (verbatum!) is useful. In many of our included LLM papers **this wasn't assessed**. If it was assessed, older models outperformed LLMs (in 2024). **Note: this comparison isn't always fair!**
- Datasets for LLM validation were **very small,** because a human needs to assess generative output. This means that **while LLMs don't need training datasets, workload is shifted towards the evaluation phase**. They were also convenience datasets and splitting into prompt development and validation data wasn't described well.
- New LLM versions of the same model, in addition **to lack of reproducible** output, **are reducing reliability** of evaluations.
- **Reporting can be poor**, for example sharing prompts, or reporting on results from prompt-development vs. independent test sets is lacking


# Accessible end-user tools

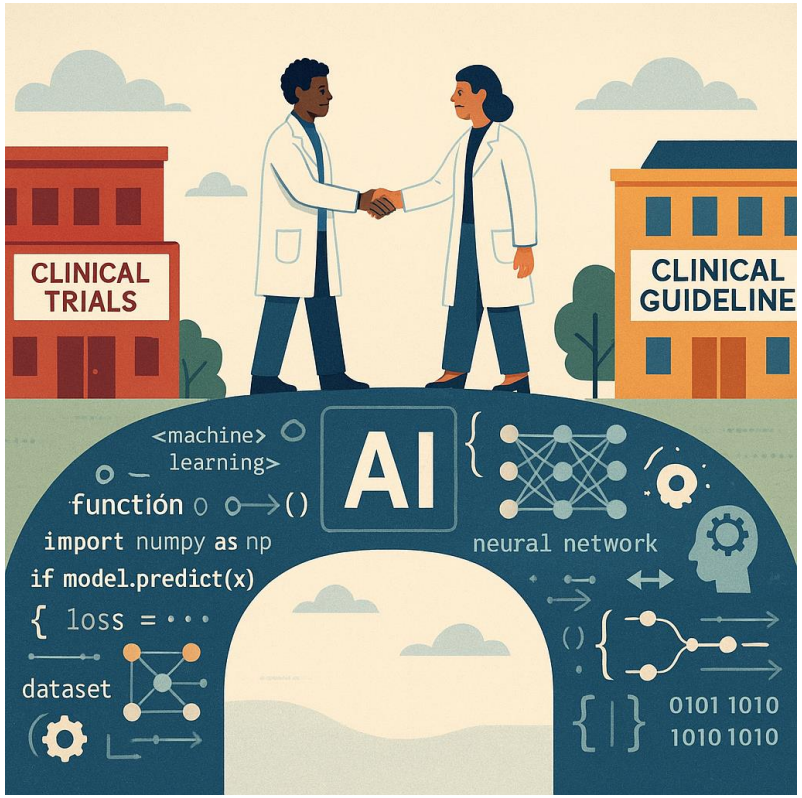
- More papers, more data, more code ≠ more practical automation
- Only 7 active web-apps (≈ 9 % of studies)
  - RobotReviewer, Trialstreamer, Trip PICO search – great, but mostly limited to PubMed abstracts, single-sentence outputs, and databases of pre-extracted information.
  - Very few functioning SR tools (*with published validations*) offer on-demand extraction for users' own PDFs. Those with non-existing broad validations aren't covered here.
  - Strong imbalance between clever prototypes and production-ready software.

# Implication for Practice

The SR seems to be willing to embrace automation approaches.

Barriers include:

- Scarcity of **deployed** user-friendly tools, and **limited interoperability** of tools
- Low re-usage of available code and data resources → **Culture to re-invent the wheel** (due to fears of not getting automation paper published?)
- No putting imperfect scores **into context with human imperfection** and problems on imperfect/unfitting gold standard that probably set an **invisible ceiling** for automation evaluation scores




# How do we encourage uptake of automation to alleviate the ‘SR bottleneck’?

# Conclusion?

- SR automation needs to be tackled on multiple levels:
  - ‘Micro’ level encouraging open, rigorous, reproducible data science, leveraging latest AI developments. Build on each other’s methods rather than feeling forced to re-invent the wheel in order to get published
  - ‘Intermediate’ level: availability of DEST
  - ‘Macro’ level through awareness and clear communication of standards set out by funders, publishers, producers of evidence synthesis to reassure reviewers and guide them
- Resources, collaboration, guidance and funding will be available over the next 5 years (see DESTINY, Hackathons, Cochrane AI group, ICASR, ESIC, ..) for those who are interested in contributing to solving these problems!