# AUTOMATED DATA EXTRACTION IN SYSTEMATIC REVIEWS

Dream big.
Then call ICF.

## A Case Study from Regulatory Toxicology

Arun Varghese, Michelle Cawley, Joanne Trgovcich, and George Agyeman-Badu | ICF

Contact
Arun Varghese
arun.varghese@icf.com

## Systematic Reviews in Toxicology: The Traditional Approach

### What are Systematic Reviews?

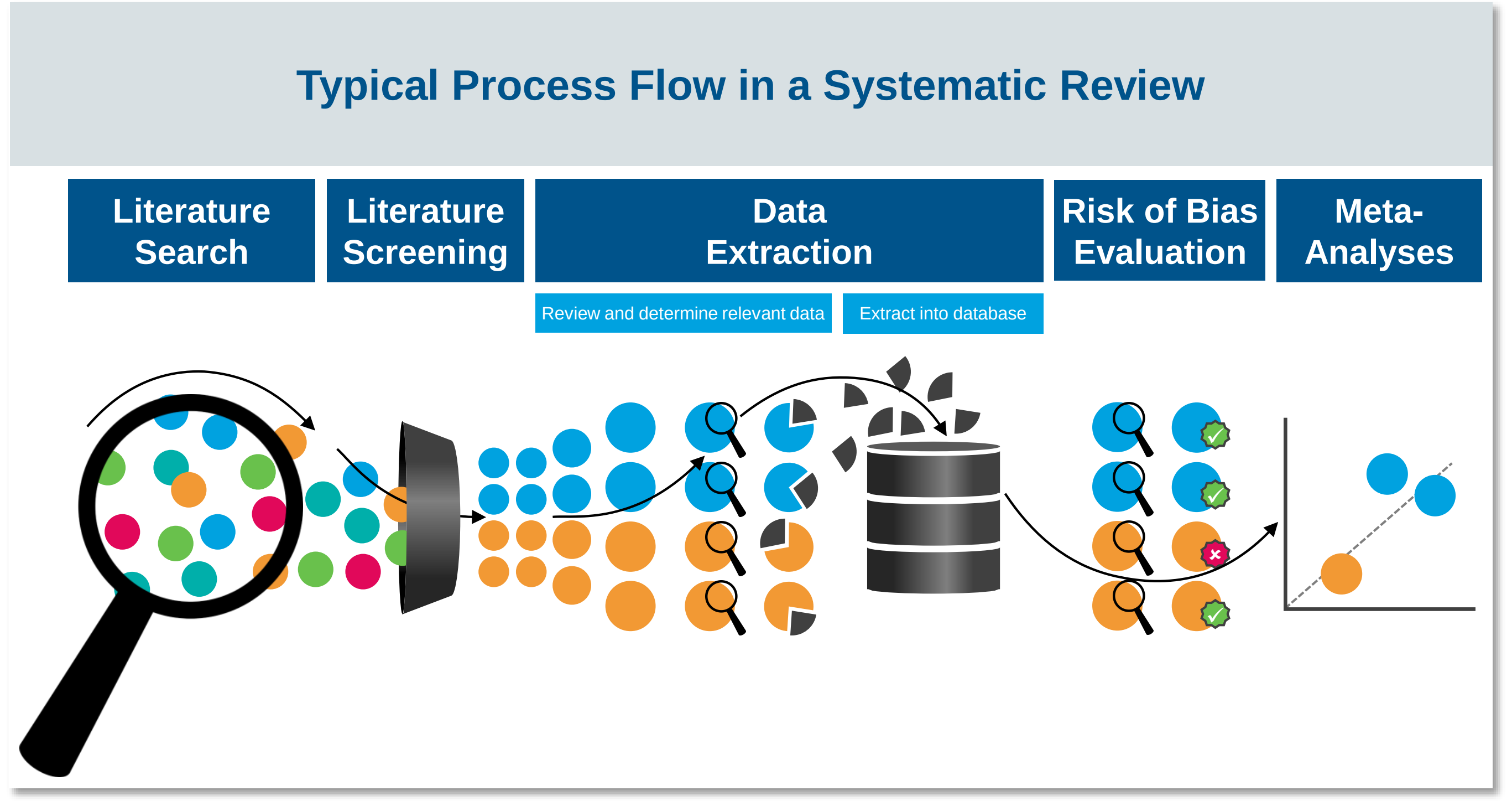
Systematic reviews identify, assess, synthesize, and interpret published evidence to support decision-making in regulatory and commercial applications of toxicology. The steps in a systematic review are illustrated in the Figure to the right.

### How Long Do Systematic Reviews Take?

Systematic reviews are known to be time consuming. It typically takes 2.5 – 6.5 years for a primary study publication to be included and published in a systematic review. 23% of reviews are out of date within 2 years of publication, because of new evidence rapidly being published. (Jonnalagada, 2015)

### Which is the Rate Limiting Step?

The data extraction process in a systematic review is widely acknowledged to be the most time and resource intensive. Subject matter experts are required to review full text publications to extract data of interest, such as animal species, strain, administered dose, endpoint, amongst others, and to assess questions related to risk of bias.



Typical Process Flow in a Systematic Review

Literature Search → Literature Screening → Data Extraction (Review and determine relevant data; Extract into database) → Risk of Bias Evaluation → Meta-Analyses

## Can Artificial Intelligence Improve Efficiency in Systematic Reviews?

### Text Analytics Can Potentially Fully Automate the Information Retrieval Step

- Owing to the comprehensive evidence gathering nature of systematic reviews, the literature search step is often designed using low precision search strategies that retrieve many more publications than are relevant. Traditionally, literature screening is a manual process undertaken by subject matter experts.
- Text analytics algorithms, which apply machine learning and statistical algorithms to large volumes of text, are capable of partially or even fully automating the screening step by reliably identifying relevant literature and discarding background noise. This process is not the subject of this poster but is described in detail in ICF's adjacent poster.

### Natural Language Processing Can Partially Automate Information Extraction

- Natural language processing algorithms help understand and derive meaning from human language. They are typically based on a combination of statistical algorithms and encoding of the semantic structure of a particular language in ways that allow manipulation by computers.
- NLP algorithms can be trained to extract well defined data elements from unstructured sources such as text documents and convert them to a structured format, such as a data table or database. In the context of toxicology, NLP algorithms can be used to extract elements like animal species, strain, gender, chemical, dose, units, outcomes, etc.
- Owing to the more challenging nature of the information extraction problem compared to the information retrieval problem, data extraction algorithms are better positioned as an aid to a manual process driven by subject matter experts, rather than a fully automated solution. Even so, they can facilitate massive efficiencies in the data extraction process.

## Case Study Data and Objectives

### Scope and Purpose of Simulation

- We assess the potential of natural language processing algorithms in extracting data for a systematic review relating to the toxicological effects of PCBs. This is intended as an example application of methods that could be applied more widely, with minimal customization, to other toxicological systematic reviews.

### Approach

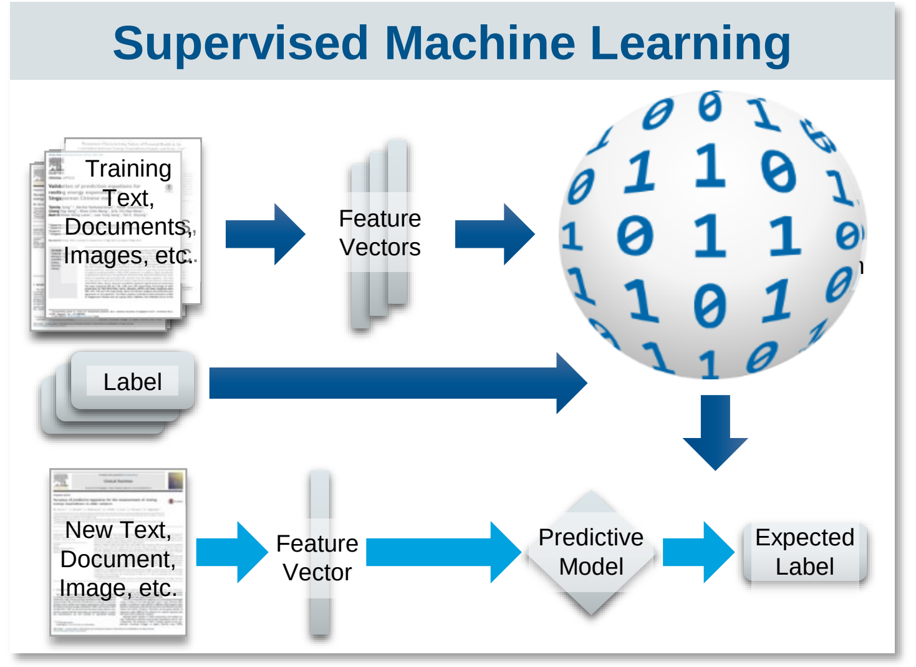
- We used subject matter expert extracted data tables relating to the developmental impacts of PCBs as our gold standard.
- The associated full text literature relating to approximately 270 studies was then used as the basis for which to run automated extraction algorithms to machine extract the identical data elements as in the gold standard extracted results.
- The extracted data elements included animal species, strain, number, gender, chemical, dose, units, exposure route, outcomes, and risk of bias.
- For some data elements, the extracted item was the complete entity (e.g. species = "Rat", strain="Wistar")
- For other elements, the extracted item was whole sentences assessed as informative to the extraction of the complete entity (e.g. number of subjects = "The treatment group contained 40 Wistar rats….." The 10 best sentences, ranked in order of estimated relevance, were considered.

### Validation Approach

- We compared the manually extracted gold standard results with the automatically extracted suggestions for 50 randomly sampled studies.
- The automated suggestion was considered accurate if either (i) the complete entity was identical or (ii) if any one of the 10 suggested sentences contained the required data for manual extraction.

## 1 Machine Learning to Extract Sentences Containing Relevant Data

- **Approach:** To identify candidate sentences that contained information necessary to answer questions related to elements such as risk of bias, we created a training dataset of sentences drawn from an independent set of literature. Sentences related to risk of bias were labelled as positive and the majority of unrelated sentences were labelled as negative to that topic. We then experimented with the following machine learning based classification approaches that are available in ICF's DoCTER tool:
- **Supervised ML:** Supervised machine learning uses statistical algorithms to identify the features (words such as "randomized" or "blind") most predictive of relevant sentences and then builds a predictive model. This model was applied to the sentences in the development literature, which were then assigned a probability score reflecting their relevance to the risk of bias question being assessed. The top 10 sentences were returned.
- **Semi-Supervised ML:** The supervised clustering method is based on semi-supervised machine learning and is an effective alternative when the user does not have a large training set. It works by using a limited number as seeds within non-supervised clustering algorithms to create an effective sentence prioritization and classification system.
- **Active ML:** Active machine learning allows users to interactively train the machine learning algorithm to help the algorithm decide whether sentences are relevant to the topic of interest or not. In active learning, the algorithm identifies the sentences it most wants to be trained on. The promise of this method is that it can also reduce the training burden compared to passive machine learning and thus save time and resources.



Supervised Machine Learning

Training Text, Documents, Images, etc.
Feature Vectors
Label
New Text, Document, Image, etc.
Feature Vector
Predictive Model
Expected Label

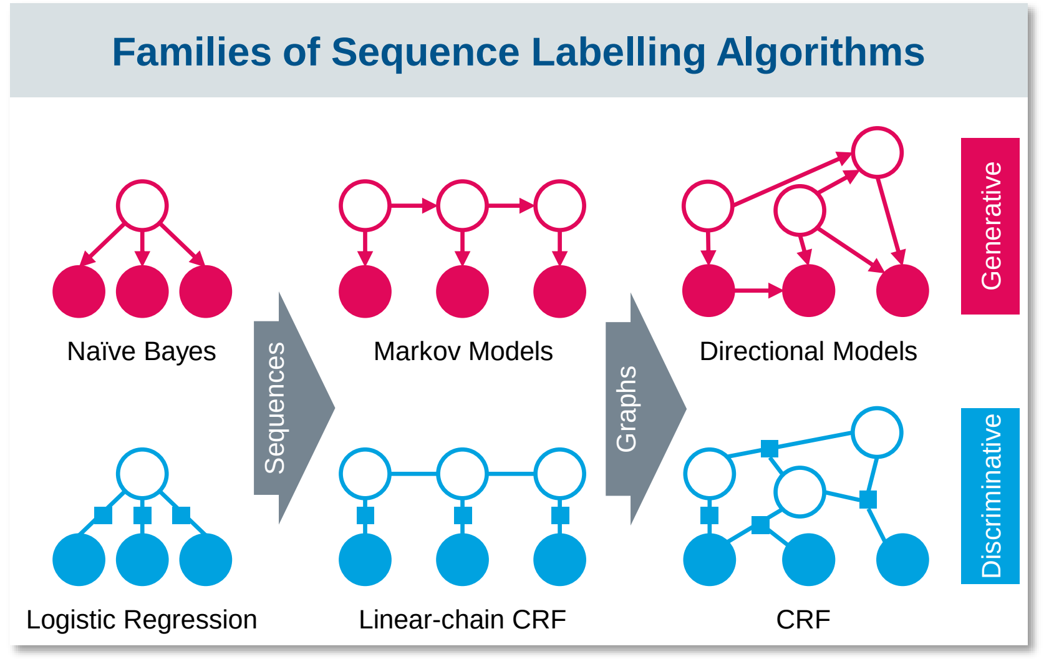
## 2 Hand-Crafted and Regular Expression Rules to Extract Entities

- **Approach:** To extract complete entities for fields such as animal and species, we used hand crafted regular expression rules that work off of traditional text search algorithms.
- **Search patterns:** We compiled search strategies based on comprehensive lists of species, strain, chemicals, exposure route, outcomes, and other fields of interest.
- **Search algorithm:** The Boyer Moore Horspool algorithm was deployed for exact and fuzzy searches.
- **Hand-crafted rules:** Rules for exact search, fuzzy search and proximity search were designed to extract entities and identify potential sentences of interest.
- **Regular Expression rules:** In future, we could use regular expression patterns to help extract data elements that follow a predictable format, like dose units (e.g. xx/xx/xx).
- **Example:** Return instances of gender ("male" or "female") in sentences also containing a species term AND a strain term.

| Regular Expression Syntax | Meaning |
|---|---|
| [0-9]{3}-[0-9]{2}-[0-9]{4} | Social security number |
| ([0-9]{3})[1-9]{3}-[0-9]{4} | Phone number (xxx)yyy-zzzz |
| (\d{4} [- ]?){3}\d{4} | Credit Card (three occurrences of four digits followed optionally by a space or dash, and one 4-digit series) |
| [1-9][A-Z]{3}[0-9]{3} | Car License Plate in California |
| [A-Z][a-z]+(\s+[A-Z][a-z]*)?\s+[A-Z][a-z]+ | First name, optional Middle Initial/name, and Last name |
| /[\w._%+-]+@[\w.-]+\.[a-zA-Z]{2,4}/ | E-mail |

## 3 Sequence Labelling to Refine Entity Extraction

- **Context:**. Grammatical constructs in sentences show distinct patterns. For instance, in the terminology of English language grammar, verbs often follow subjects in sentences and are followed by objects. Natural language processing algorithms have been used to parse sentences into their parts of speech. These algorithms can also be used to predict labels for other types of structures in sentences that relate to their role in the particular context of the extraction problem. For example, the label "dose" could be applied to the quantity indicating the dose of chemical supplied and the label "units" could be applied to the units term. Non-relevant words could be labeled as "other".
- **Sequence Labelling Algorithms:** Algorithms such as hidden Markov models and conditional random fields train themselves on a labelled set of sentences and build predictive models of label sequence structure. These predictive models may then be used to predict labels on unlabeled sentences.
- **Approach:** We manually labelled a small selection of sentences to include labels such as dose, units, and subject number. We then applied the CRF algorithm to build a predictive model and made prediction on the remaining sentences in the corpus. We applied hand crafted rules to extract the entities relating to dose and units.
- **Example:** To extract dose and units, we identified sentences in which both labels occurred based on the CRF algorithm predictions. We then extracted those fields from the relevant sentences and presented them as candidate suggestions.



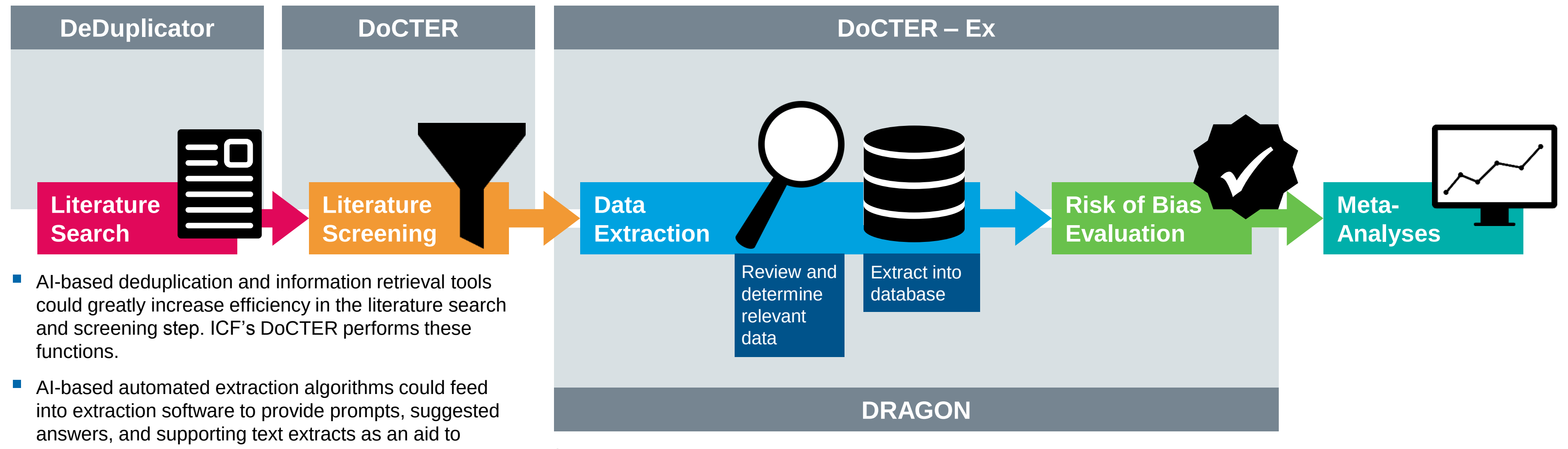
Families of Sequence Labelling Algorithms

## 4 Results

| Data Element --> | Chemical | Species | Strain | Gender | # Subjects | Dose | Units | Exposure Route | Outcome | Risk of Bias |
|---|---|---|---|---|---|---|---|---|---|---|
| Extracted Item --> | Entity | Entity | Entity | Sentence | Sentence | Entity | Entity | Entity | Sentence | Sentence |
| Method --> | HCR | HCR | HCR | HCR | CRF | CRF | CRF | HCR | ML | ML |
| Accuracy --> | 84% | 100% | 72% | 89% | 82% | 73% | 85% | 72% | 79% | 69% |

- **Approach:** 50 studies on which automated extraction strategies were applied were randomly selected for comparison to the gold standard. For entities, exact agreement was considered a match. For sentences, if any one of the 10 sentences proposed contained complete or partial information, it was considered a match.
- **Context and caveats:** Many of the chemicals in the gold standard field were mixtures of different chemicals. We considered it a match if any one of the chemicals in the mixture was captured. Similarly, the dose column in the gold standard often contained more than 1 dose regimen. We reported a match if any one of the extracted doses matched. Because of the complexity of the data structure in the gold standard, we did not compute F1 scores, precision, and recall.

## A Vision for Greater Automation in Systematic Reviews



- AI-based deduplication and information retrieval tools could greatly increase efficiency in the literature search and screening step. ICF's DoCTER performs these functions.
- AI-based automated extraction algorithms could feed into extraction software to provide prompts, suggested answers, and supporting text extracts as an aid to reviewers as they manually populate structured databases. ICF's DoCTER-Ex and DRAGON Online tools could work together to achieve this goal.

## Conclusions, Caveats, and Future Directions

- The machine learning and natural language processing-based data extraction algorithms developed in this prototype application could significantly increase efficiencies in the systematic review process, realizing savings in time and labor.
- Results achieved here would be improved by increasing investment in training data and further customizing the extraction algorithms.
- A graphical user interface that zooms in on the text extracts or sentences in the full text document as supporting evidence for the machine generated suggested entities for specific data elements would serve to increase the reviewer's confidence.
- The DoCTER-Ex data extraction tool could be integrated with the APIs of extraction software such as DRAGON Online to facilitate the seamless integration of AI algorithms into the manually mediated data extraction step.

DOCTER
Document Classification and Topic Extraction Resource

DRAGON

*Acknowledgements*

*ICF thanks U.S. EPA for permission to use their data in this prototype application*

*Logo Design and Graphics: **Courtney Skuce***