



Hoboken

Single Day Event

July 18th 2019

Stevens Institute of Technology, Hoboken

**Artificial Intelligence and
the Digital Transformation
of Healthcare**



Using NLP Techniques for Data Identification and Extraction from Pooled Research Sources

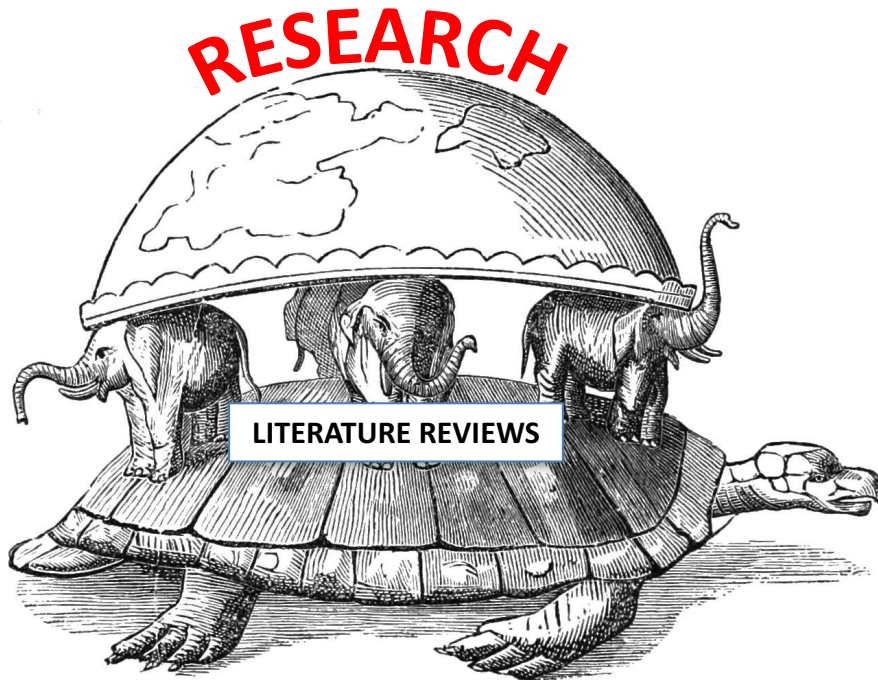
Genesis Research and Stevens School of Business

Dr. Rong Liu, Stevens School of Business

Dr. Marko Zivkovic, Genesis Research

Literature Review – Foundation of Proper Research

A critical analysis of a segment of a published body of knowledge through summary, classification, and comparison of prior research studies, reviews of literature, and theoretical articles¹



A comprehensive recapitulation on the given scholarship from past to present to:

- Explore existing information
- Ascertain relevant formative works
- Identify key researchers
- Establish current landscape of main ideas, conclusions, and theories
- Clarify research methodologies
- Expose gaps in existing research
- Detect relationships among studies
- Provide context for your own research

¹ University of Wisconsin Writing Center

Literature Reviews – Useful, but Labor-Intensive

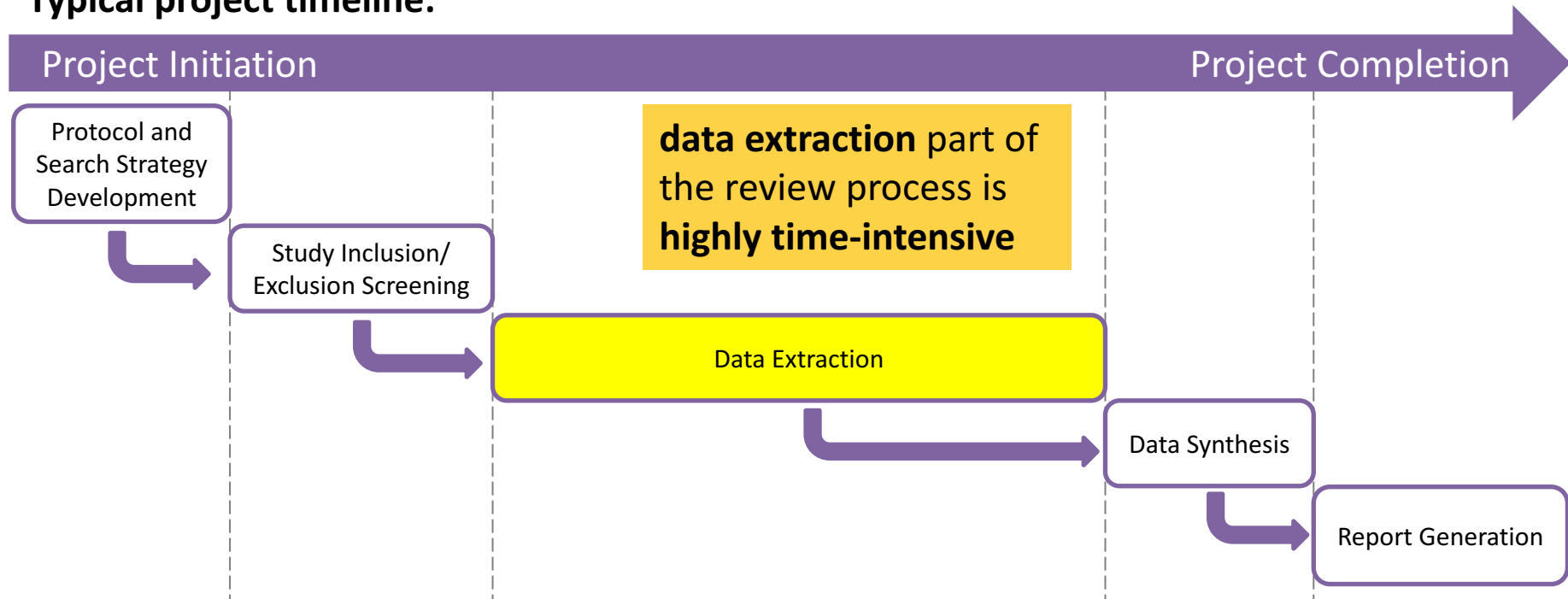
- High quality systematic reviews require two independent researchers to screen studies and extract data¹
- There is a high prevalence error in data extraction for literature reviews and meta-analyses.^{3,4}

Mean time to
complete a review:



67.3 weeks²

Typical project timeline:



The Burden of Data Extraction

First-stage screening

Includes screening titles and abstracts.

The largest amount of material to go through but limited in number of sentences per article (abstracts are usually short).

Tasking, but not too demanding. No extraction needed!

Can be automated!!!

DistillerSR

Evidence Partners

The screenshot displays the DistillerSR web application. At the top, it shows the study title: 'Beverley Shea, Lex M Bouter, Jeremy M Grimshaw, Daniel Francis, Zulma Ortiz, George A Wells, Peter S Tugwell, Maarten Boers (2006). Scope for improvement in the quality of reporting of systematic reviews. From the Cochrane Musculoskeletal Group. The Journal of rheumatology, 33(1), 9'. Below this, there's a 'Reference Labels' section with an 'Add Labels here' button. The main content area shows the abstract and parts of the methods and results sections. On the right, there's a 'Submit Form' button and a 'This Form - Next Reference' dropdown. Below the form, there's a 'DistillerSR's Intuitive 5 Step Process' diagram with steps: Upload References, Create Forms, Assign Reviewers & Begin Screening, Monitor, and Export Data.

The difficult part

Each article must be read in full.

No useful automation... yet!!!

Title and abstracts

identified and screened

(n=6,585)

Excluded by screening title and/or abstracts (n=5,035):

- Indication not of interest (n=2,950)
- Animal study (n=104)
- Study design (case studies) (n=1,137)
- Non-English (n=638)
- English articles with no access (n=206)

Full text articles (n=1,550):

- Full text retrieved and assessed for eligibility (n=1,538)
- Awaiting access to full text (English articles) (n=12)

Excluded by full text screening (n=845):

- Indication not of interest (n=252)
- Animal/in-vitro study (n=65)
- Study design (n=431)
- Outcome not of interest (n=97)

Gray literature
(n= 21)

Total number of references
included for data extraction

(n=726)

Data Extraction Example

 **Article 1**
"Maintenance zinc therapy after initial penicillamine chelation to treat symptomatic hepatic Wilson's disease in resource constrained setting"

 Read the article in detail

Identify + highlight + extract values for aspects of interest



 **Article 2**
"A genetic study of Wilson's disease in the United Kingdom"

 Read the article in detail

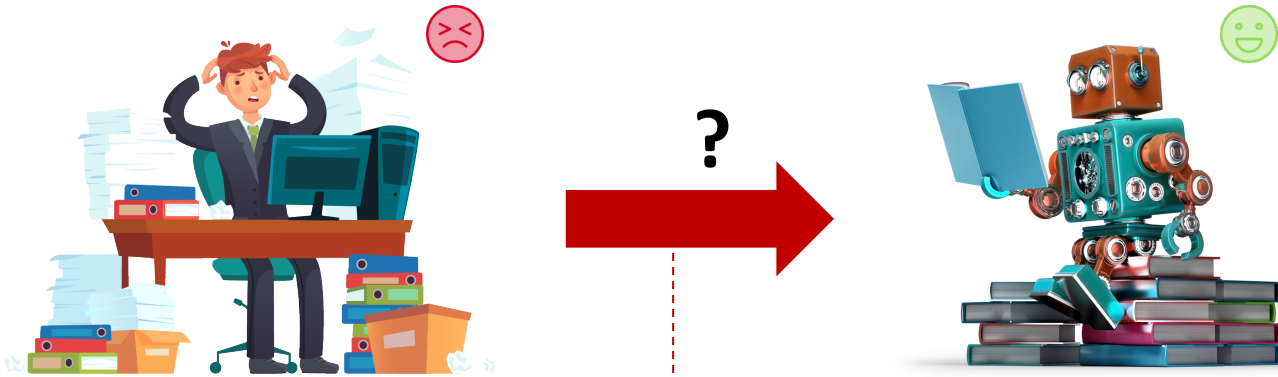
Identify + highlight + extract values for aspects of interest



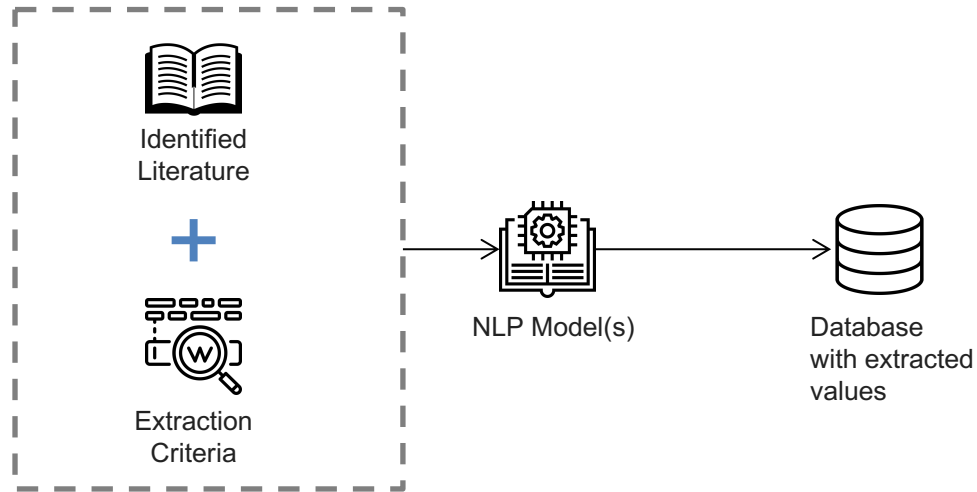
+ 724 other articles

	A	B	C	D	E	F	G	H	
	Author	Year	Title	Data Source Type	Data Source Name	Study Design	Study Period	Study Length	Country
1	Abdelgaffar	2011	Phenotypic and genetic characterization of Patient Charts	NR	NR	NR	1992 to 2009	17 years	Egypt
2	Abdelgaffar	2008	WILSON DISEASE (WD) PRESENTING TO THE Patient Charts	NR	Retrospective Cohort	NR	Mean follow-up: 6.4 years (range 1-14)	NR	Egypt
3	Abdelgaffar	2014	Zinc mono-therapy in pre-symptomatic Wilson's disease	NR	Children's Hospital of Fudan University	Retrospective Cohort	January 2006 to March 2013	Mean follow-up time was 1.54	China
4	Adnan	2011	Clinical profile of Wilson's disease in central Patient Charts	NR	Chaitam Hospital and Resea	Prospective Cohort	NR	3 year	India
5	Aggarwal	2017	Neurological worsening in patients under Patient Charts	NR	Wilson's Disease Clinic	Prospective Cohort	2005-2016	11 years	India
6	Aggarwal	2019	Complete neurological recovery in Wilson's disease Patient Charts	NR	NR	Prospective Cohort	2005-2013	NR	NR
7	Aggarwal	2014	A novel objective brain MRI assessment scale Patient Examinations	NR	Validation	NR	NR	NR	NR
8	Aggarwal	2009	A novel global assessment scale for Wilson's NR	NR	Jaslok Hospital and Research	Validation	April 2006-June 2007	NR	India
9	Alali	1991	The psychiatric presentations of Wilson's disease Patient Charts	NR	University of Michigan	Retrospective Cohort	NR	NR	United States
10	Ala	2015	Prospective pilot study of a single daily dose Clinical Trial Data	NR	NR	NR	November 2011-July 2012	12 months	India
11	Alam	2018	AARIC-AACF SCORE: BEST PREDICTOR OF QOL IN	NR	NR	Prospective Cohort	NR	90 days	India
12	Alam	2017	Hepatic encephalopathy and bilirubin prognostic Patient Charts	NR	Institute of Liver and Biliary Sciences	Retrospective Cohort	January 2011 to January 2017	6 years	India
13	Alam	2017	Predicting "poor outcome" in decompensated Patient Charts	NR	Retrospective Cohort	NR	2011-2016	NR	India
14	Alam	2014	Low albumin and high alanine transaminase Patient Charts	NR	Institute of Liver and Biliary Sciences	Retrospective Cohort	NR	3 years	India
15	Alam	2013	Prediction of poor outcome on medical therapy	NR	Institute of Liver and Biliary Sciences	Retrospective Cohort	NR	NR	India
16	Albrecht	2012	Retinal Neurodegeneration in Wilson's Disease Patient Examinations	NR	Heinrich-Heine University	Prospective Case-Control	NR	Mean follow-up time: 9.8 (SD: 3.0)	Germany
17	Al-Hilouy	2012	LONG TERM OUTCOMES OF LIVER TRANSPLANT Patient Charts	NR	University Hospitals Birmingham	Retrospective Cohort	1987-2011	NR	United Kingdom
18	Al-Hilouy	2012	Presentation, diagnosis and outcome of pre Patient Charts	NR	King Fahad National Guard Hospital	Retrospective Cohort	January 1994 - July 2008	mean (SD) follow up period: 6.5 (SD: 4.0)	Saudi Arabia
19	Antczak-Kowalska	2018	Gastroscopy findings in Wilson's disease patient Patient Examinations	NR	2nd department of Neurology, Prospective Cohort	NR	NR	NR	Poland
20	Antczak-Kowalska	2018	Myocardial integrated ultrasonography in Wilson's disease Patient Examinations	NR	NR	Prospective Cohort	NR	NR	Turkey
21	Arat	2011	Phenotypic and genetic characterization of Wilson's disease	NR	NR	NR	NR	NR	NR
22	Arat	2008	WILSON DISEASE (WD) PRESENTING TO THE Patient Charts	NR	NR	NR	NR	NR	NR
23	Arat	2014	Zinc mono-therapy in pre-symptomatic Wilson's disease	NR	NR	NR	NR	NR	NR
24	Arat	2017	Clinical profile of Wilson's disease in central Patient Charts	NR	NR	NR	NR	NR	NR
25	Arat	2017	Neurological worsening in patients under Patient Charts	NR	NR	NR	NR	NR	NR
26	Arat	2019	Complete neurological recovery in Wilson's disease Patient Charts	NR	NR	NR	NR	NR	NR
27	Arat	2014	A novel objective brain MRI assessment scale Patient Examinations	NR	NR	NR	NR	NR	NR
28	Arat	2009	A novel global assessment scale for Wilson's disease	NR	NR	NR	NR	NR	NR
29	Arat	1991	The psychiatric presentations of Wilson's disease	NR	NR	NR	NR	NR	NR
30	Arat	2015	Prospective pilot study of a single daily dose Clinical Trial Data	NR	NR	NR	NR	NR	NR
31	Arat	2018	AARIC-AACF SCORE: BEST PREDICTOR OF QOL IN	NR	NR	NR	NR	NR	NR
32	Arat	2017	Hepatic encephalopathy and bilirubin prognostic Patient Charts	NR	NR	NR	NR	NR	NR
33	Arat	2017	Predicting "poor outcome" in decompensated Patient Charts	NR	NR	NR	NR	NR	NR
34	Arat	2014	Low albumin and high alanine transaminase Patient Charts	NR	NR	NR	NR	NR	NR
35	Arat	2013	Prediction of poor outcome on medical therapy	NR	NR	NR	NR	NR	NR
36	Arat	2012	Retinal Neurodegeneration in Wilson's Disease Patient Examinations	NR	NR	NR	NR	NR	NR
37	Arat	2012	LONG TERM OUTCOMES OF LIVER TRANSPLANT Patient Charts	NR	NR	NR	NR	NR	NR
38	Arat	2012	Presentation, diagnosis and outcome of pre Patient Charts	NR	NR	NR	NR	NR	NR
39	Arat	2018	Gastroscopy findings in Wilson's disease patient Patient Examinations	NR	NR	NR	NR	NR	NR
40	Arat	2018	Myocardial integrated ultrasonography in Wilson's disease Patient Examinations	NR	NR	NR	NR	NR	NR
41	Arat	2011	Phenotypic and genetic characterization of Wilson's disease	NR	NR	NR	NR	NR	NR
42	Arat	2008	WILSON DISEASE (WD) PRESENTING TO THE Patient Charts	NR	NR	NR	NR	NR	NR
43	Arat	2014	Zinc mono-therapy in pre-symptomatic Wilson's disease	NR	NR	NR	NR	NR	NR
44	Arat	2017	Clinical profile of Wilson's disease in central Patient Charts	NR	NR	NR	NR	NR	NR
45	Arat	2017	Neurological worsening in patients under Patient Charts	NR	NR	NR	NR	NR	NR
46	Arat	2019	Complete neurological recovery in Wilson's disease Patient Charts	NR	NR	NR	NR	NR	NR
47	Arat	2014	A novel objective brain MRI assessment scale Patient Examinations	NR	NR	NR	NR	NR	NR
48	Arat	2009	A novel global assessment scale for Wilson's disease	NR	NR	NR	NR	NR	NR
49	Arat	1991	The psychiatric presentations of Wilson's disease	NR	NR	NR	NR	NR	NR
50	Arat	2015	Prospective pilot study of a single daily dose Clinical Trial Data	NR	NR	NR	NR	NR	NR
51	Arat	2018	AARIC-AACF SCORE: BEST PREDICTOR OF QOL IN	NR	NR	NR	NR	NR	NR
52	Arat	2017	Hepatic encephalopathy and bilirubin prognostic Patient Charts	NR	NR	NR	NR	NR	NR
53	Arat	2017	Predicting "poor outcome" in decompensated Patient Charts	NR	NR	NR	NR	NR	NR
54	Arat	2014	Low albumin and high alanine transaminase Patient Charts	NR	NR	NR	NR	NR	NR
55	Arat	2013	Prediction of poor outcome on medical therapy	NR	NR	NR	NR	NR	NR
56	Arat	2012	Retinal Neurodegeneration in Wilson's Disease Patient Examinations	NR	NR	NR	NR	NR	NR
57	Arat	2012	LONG TERM OUTCOMES OF LIVER TRANSPLANT Patient Charts	NR	NR	NR	NR	NR	NR
58	Arat	2012	Presentation, diagnosis and outcome of pre Patient Charts	NR	NR	NR	NR	NR	NR
59	Arat	2018	Gastroscopy findings in Wilson's disease patient Patient Examinations	NR	NR	NR	NR	NR	NR
60	Arat	2018	Myocardial integrated ultrasonography in Wilson's disease Patient Examinations	NR	NR	NR	NR	NR	NR
61	Arat	2011	Phenotypic and genetic characterization of Wilson's disease	NR	NR	NR	NR	NR	NR
62	Arat	2008	WILSON DISEASE (WD) PRESENTING TO THE Patient Charts	NR	NR	NR	NR	NR	NR
63	Arat	2014	Zinc mono-therapy in pre-symptomatic Wilson's disease	NR	NR	NR	NR	NR	NR
64	Arat	2017	Clinical profile of Wilson's disease in central Patient Charts	NR	NR	NR	NR	NR	NR
65	Arat	2017	Neurological worsening in patients under Patient Charts	NR	NR	NR	NR	NR	NR
66	Arat	2019	Complete neurological recovery in Wilson's disease Patient Charts	NR	NR	NR	NR	NR	NR
67	Arat	2014	A novel objective brain MRI assessment scale Patient Examinations	NR	NR	NR	NR	NR	NR
68	Arat	2009	A novel global assessment scale for Wilson's disease	NR	NR	NR	NR	NR	NR
69	Arat	1991	The psychiatric presentations of Wilson's disease	NR	NR	NR	NR	NR	NR
70	Arat	2015	Prospective pilot study of a single daily dose Clinical Trial Data	NR	NR	NR	NR	NR	NR
71	Arat	2018	AARIC-AACF SCORE: BEST PREDICTOR OF QOL IN	NR	NR	NR	NR	NR	NR
72	Arat	2017	Hepatic encephalopathy and bilirubin prognostic Patient Charts	NR	NR	NR	NR	NR	NR
73	Arat	2017	Predicting "poor outcome" in decompensated Patient Charts	NR	NR	NR	NR	NR	NR
74	Arat	2014	Low albumin and high alanine transaminase Patient Charts	NR	NR	NR	NR	NR	NR
75	Arat	2013	Prediction of poor outcome on medical therapy	NR	NR	NR	NR	NR	NR
76	Arat	2012	Retinal Neurodegeneration in Wilson's Disease Patient Examinations	NR	NR	NR	NR	NR	NR
77	Arat	2012	LONG TERM OUTCOMES OF LIVER TRANSPLANT Patient Charts	NR	NR	NR	NR	NR	NR
78	Arat	2012	Presentation, diagnosis and outcome of pre Patient Charts	NR	NR	NR	NR	NR	NR
79	Arat	2018	Gastroscopy findings in Wilson's disease patient Patient Examinations	NR	NR	NR	NR	NR	NR
80	Arat	2018	Myocardial integrated ultrasonography in Wilson's disease Patient Examinations	NR	NR	NR	NR	NR	NR
81	Arat	2011	Phenotypic and genetic characterization of Wilson's disease	NR	NR	NR	NR	NR	NR
82	Arat	2008	WILSON DISEASE (WD) PRESENTING TO THE Patient Charts	NR	NR	NR	NR	NR	NR
83	Arat	2014	Zinc mono-therapy in pre-symptomatic Wilson's disease	NR	NR	NR	NR	NR	NR
84	Arat	2017	Clinical profile of Wilson's disease in central Patient Charts	NR	NR	NR	NR	NR	NR
85	Arat	2017	Neurological worsening in patients under Patient Charts	NR	NR	NR	NR	NR	NR
86	Arat	2019	Complete neurological recovery in Wilson's disease Patient Charts	NR	NR	NR	NR	NR	NR
87	Arat	2014	A novel objective brain MRI assessment scale Patient Examinations	NR	NR	NR	NR	NR	NR
88	Arat	2009	A novel global assessment scale for Wilson's disease	NR	NR	NR	NR	NR	NR
89	Arat	1991	The psychiatric presentations of Wilson's disease	NR	NR	NR	NR	NR	NR
90	Arat	2015	Prospective pilot study of a single daily dose Clinical Trial Data	NR	NR	NR	NR	NR	NR
91	Arat	2018	AARIC-AACF SCORE: BEST PREDICTOR OF QOL IN	NR	NR	NR	NR	NR	NR
92	Arat	2017	Hepatic encephalopathy and bilirubin prognostic Patient Charts	NR	NR	NR	NR	NR	NR
93	Arat	2017	Predicting "poor outcome" in decompensated Patient Charts	NR	NR	NR	NR	NR	NR
94	Arat	2014	Low albumin and high alanine transaminase Patient Charts	NR	NR	NR	NR	NR	NR
95	Arat	2013	Prediction of poor outcome on medical therapy	NR	NR	NR	NR	NR	NR
96	Arat	2012	Retinal Neurodegeneration in Wilson's Disease Patient Examinations	NR	NR	NR	NR	NR	NR
97	Arat	2012	LONG TERM OUTCOMES OF LIVER TRANSPLANT Patient Charts	NR	NR	NR	NR	NR	NR
98	Arat	2012	Presentation, diagnosis and outcome of pre Patient Charts	NR	NR	NR	NR	NR	NR
99	Arat	2018	Gastroscopy findings in Wilson's disease patient Patient Examinations	NR	NR	NR	NR	NR	NR
100	Arat	2018	Myocardial integrated ultrasonography in Wilson's disease Patient Examinations	NR	NR	NR	NR	NR	NR

Can We Automate Parts of Data Extraction?



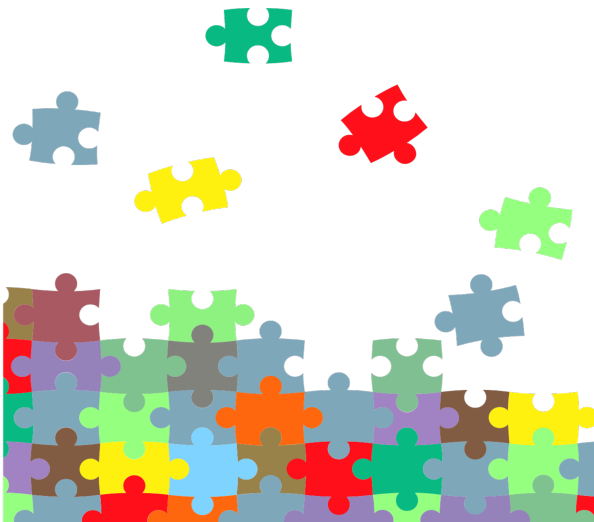
Basic Idea



Overview of Current Research

To date there are **no comprehensive, suitable and intuitive solutions** for automated data extraction of healthcare-related information from pooled research resources.

- Kiritchenko et al. developed ExaCT, tool that assists users with locating and extracting key trial characteristics such as eligibility criteria, sample size, drug dosage, and primary outcomes from full-text journal articles.¹
- Wei et al. developed deep learning approaches for medication and adverse drug event extraction from clinical text (2019).²
- Basu et al. proposed a novel framework to expedite systematic reviews by automatically building information extraction training corpora (2019).³

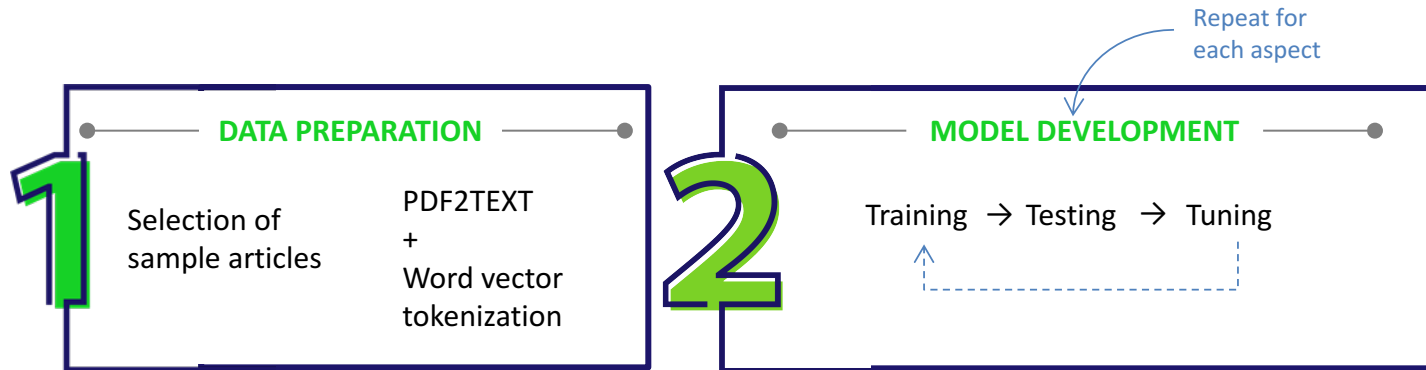


There is other promising research on the horizon, but each addresses a different aspect of the data extraction process.

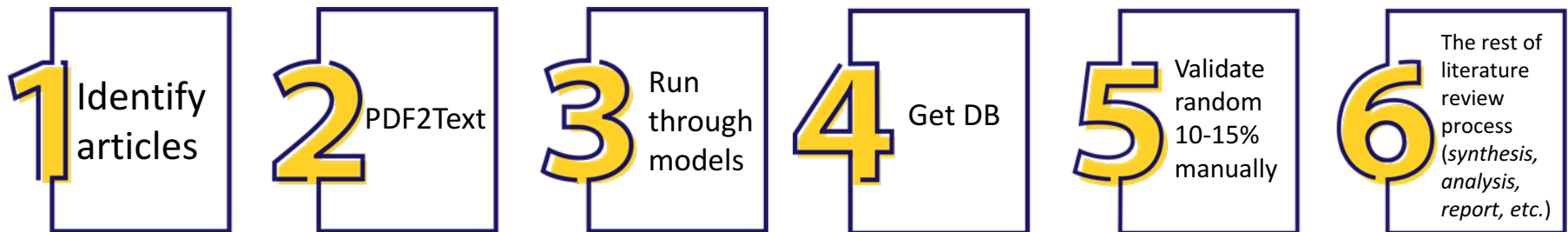
- Researchers in Denmark compared results of text mining full-text articles versus only their abstracts. Results indicated mining full-text articles consistently outperformed using abstracts only. While the study demonstrated the usefulness of mining different corpora, the Named Recognition Model depended heavily on the dictionaries and stop word lists and was very sensitive to ambiguous words.⁴
- The non-profit eLife Sciences is exploring whether computer vision can be used to provide a high-accuracy method to convert PDF to XML in order to overcome the limitations of unstructured text in PDFs.⁵
- NotesAid is a tool developed to assist patients in understanding their medical health records and physician notes through the use of NLP to link medical terms to lay definitions.⁶

Overview of Our Approach

Development Phase



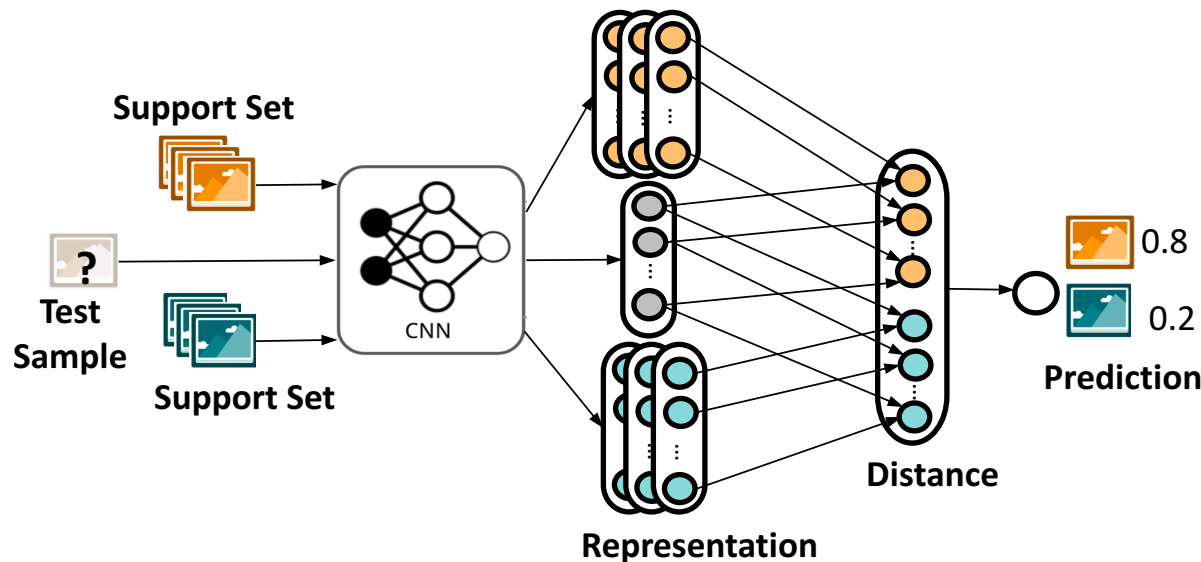
Use Phase



No training needed!

Neural Few Shot Learning Model

- Background of Few Shot Learning
 - Deep learning models usually require sufficient high-quality data to achieve good performance
 - However, it is expensive, and sometimes unrealistic, to obtain sufficient data
 - Few-shot learning discovers patterns in such challenging scenarios where data is limited
- Each class has a **prototype representation**, which is learned from labeled samples
- Train a neural network to **differentiate** the prototype representations of classes
- Given a **test sample** and **support sets** (i.e. labeled samples)
 - Use a **convolutional neural network** (CNN) to extract features (i.e. words or phrases)
 - Calculate the **distance** between the test sample and each support set
 - Compute the **weight** of each sample in support sets based on the distance
 - Produce the prediction as the sum of weights in each class



N-way, K-shot Learning Model

- **n-way**: # of classes
- **k-shot**: # of samples in each support set, k usually is small (e.g. < 20)

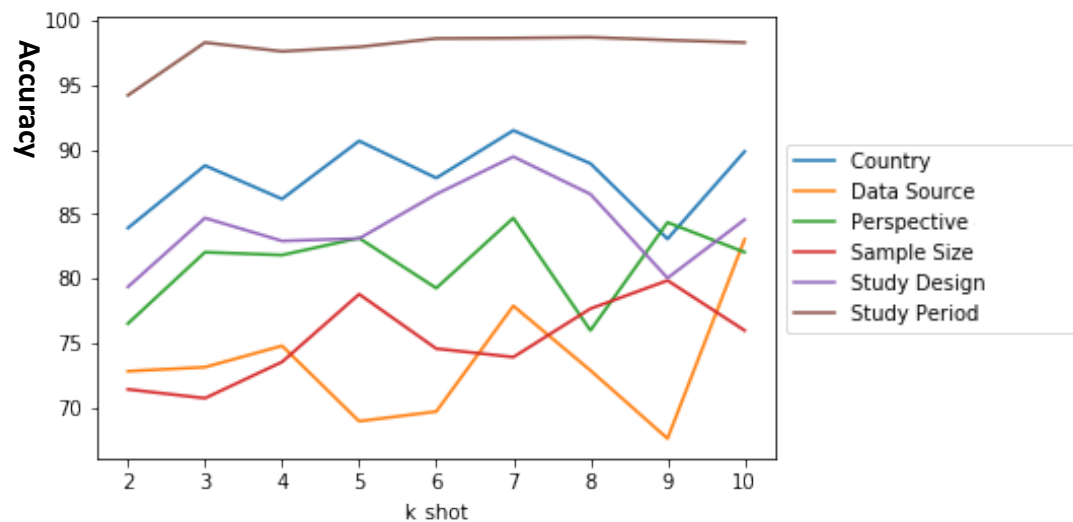
Model Design

- **Objective:** Identify sentences containing aspects of interest (i.e. **classes**) in research literature
Perspective Study Design Data Source ...
Country Study Period Sample Size ...
- **Preprocessing:** each sentence in a research article is tokenized into words, and words are mapped to pre-trained Glove word vectors in order to capture meaning
- Classes are **not mutually exclusive**. A sentence may describe multiple aspects
- **One-vs-the-Rest approach:** a binary model with two classes:
 - Positive class: one target class (e.g. study design)
 - Negative class: all the rest classes
- So our model is **2-way k-shot learning**, where **k** will be tuned, **k** = 1, 2, ...
- Each training sample is a 3-element tuple: (support set 1, support set 2, test sample), e.g.

Support Sets (Training Dataset)		Test Sample (Testing Dataset)
<u>Positive Class</u> Sentences randomly sampled from Study Design class	S_1 S_2 ... S_k	S_t randomly sampled from either Study Design or any of the rest classes
<u>Negative Class</u> Sentences randomly sampled from the rest of classes	S_1 (Study Period) S_2 (Country) ... S_k (Sample Size)	

Model Performance

- Our model is 2-way, k -shot learning, $k = 2, 3, \dots$
- Model training and validation:
 - ~40 labeled samples for each class, 2/3 (~27) for training, the rest for testing
 - The best k varies by model
 - Study Period has high accuracy, but Data Source and Sample Size are relatively difficult to identify
 - e.g. The study observation period was from 2000 to 2016 → Study period
 - Data Source examples
 - ... can be found at the national institutes of health **clinical trials registry** ...
 - ... **hospitalization data** were verified by linking the encrypted personal identification number ...



Model Analysis

- Performance is reasonably good considering small sample size
- **Capturing semantics** instead of just keyword matching
 - e.g. Cross sectional **retrospective** study was done → Perspective
 - e.g. ... analysis of an **ongoing** real-life study in CNB → Perspective
- **Multiple-aspect** detection: from a single sentence, the model can extract information relevant to multiple aspects
 - e.g. This is **retrospective** analysis of patients ... Paulbrousse Hospital Villejuif **France** → Perspective, Country
- However, there are a few challenges:
 - **Overfitting** due to small sample set
 - A sentence with **unseen patterns** can be easily misidentified, due to equal distance to both support sets
 - e.g. Poster presentations s630 journal of hepatology 2018 vol. → mislabeled as Study Design, Data Source
 - A sentence alone may not have the **right context** to make prediction
 - e.g. Thirty four patients (32%) were transplanted at mean age of years. → labeled as Sample size; however, we need to identify the total WD sample size

Future Work

- Model testing and improvement
 - Continuously testing our model with a wide variety of articles
 - Improve the model to be more robust
 - Regularization to overcome overfitting
 - How to deal with unseen patterns
 - Incorporate broader context of each sentence into learning
- Generate summary reports based on identified aspects
 - Extract specific aspect values from identified sentences: e.g. sample size, study period
 - Provide summary statistics
- Transfer learning: apply the model to different literature domains
- Implement the entire machine learning pipeline

The Brave Team

A lot of hard work done behind the scenes by an awesome team...

THANK YOU



Business vector created by coolvector
<https://www.freepik.com/free-photos-vectors/business>



CHEN LIAO



SIWEI WANG



YU HONG



COLIN NAVICKAS

... and they all are current or former Stevens students!