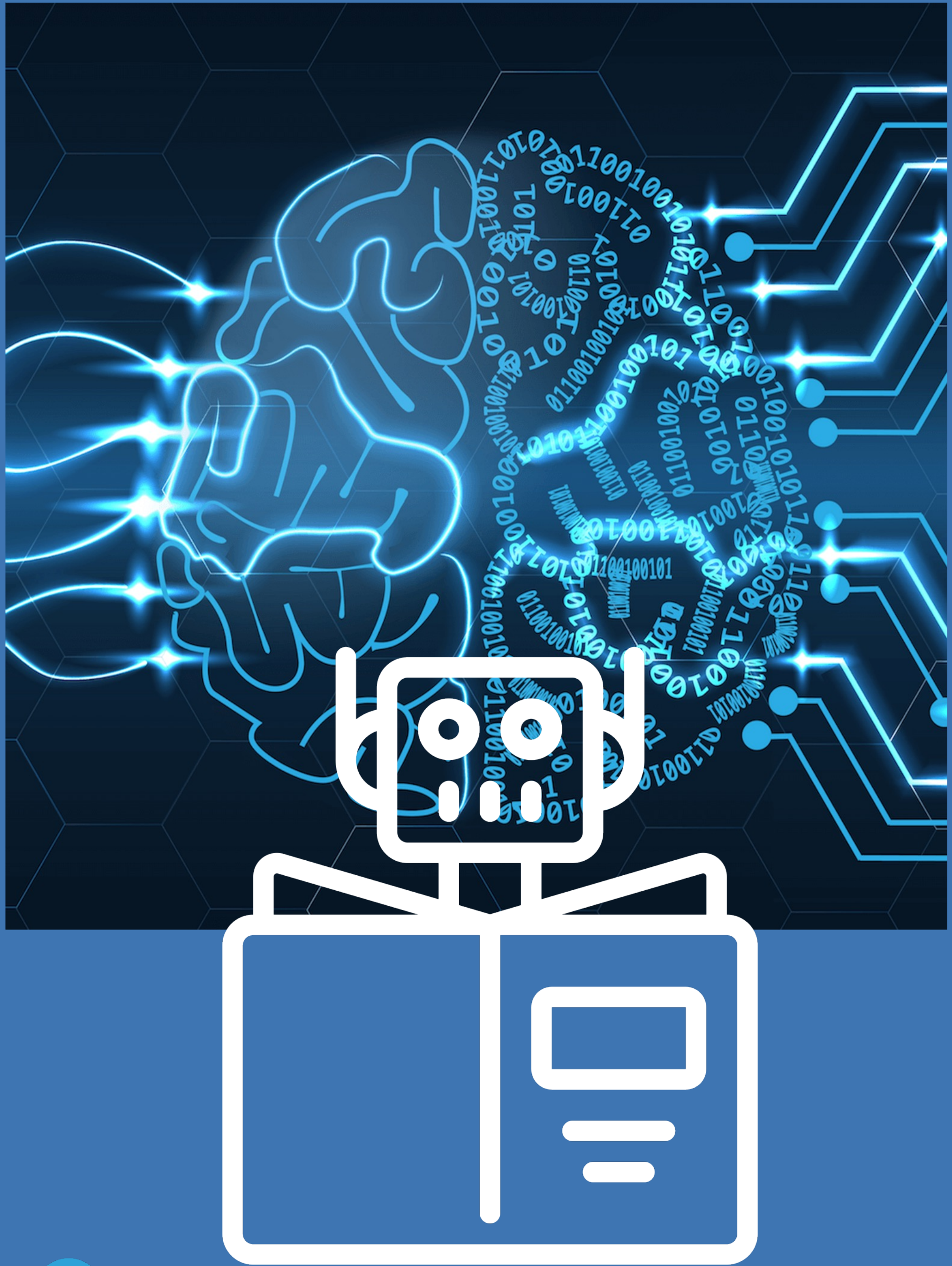


Opportunities for AI-enabled scientific knowledge exploration, analysis, and discovery

Karin Verspoor

Executive Dean, School of Computing Technologies, **RMIT University**

Fellow, **Australasian Institute of Digital Health**

Deputy Director, **ARC Training Centre in Cognitive Computing for Medical Technologies**

Honorary Professor, **The University of Melbourne**

@karinv

RMIT UNIVERSITY

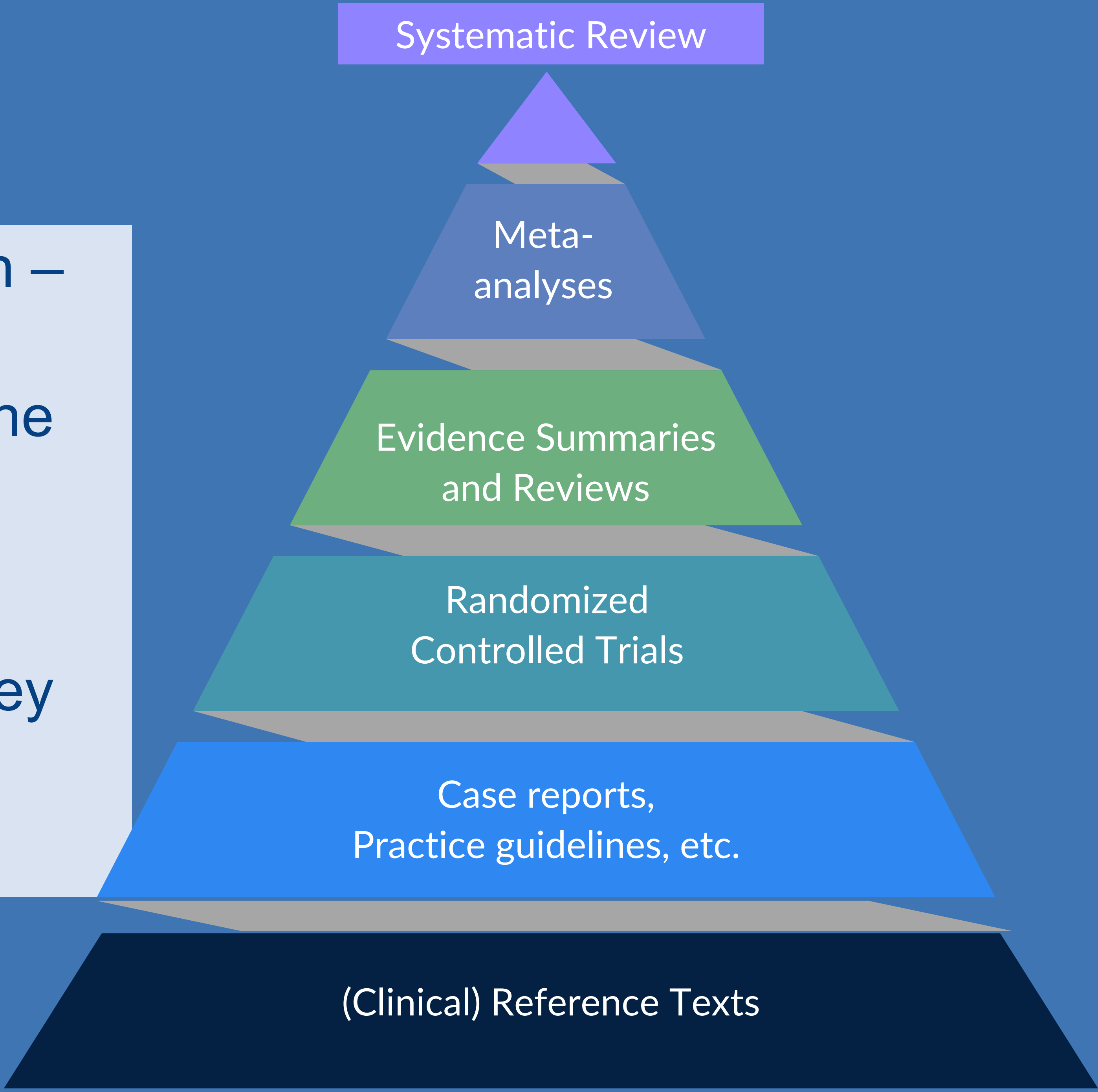
ARC TRAINING CENTRE IN COGNITIVE COMPUTING FOR MEDICAL TECHNOLOGIES

THE UNIVERSITY OF MELBOURNE

Research Literature

Evidence derived from research – and published in the scientific literature – is considered to be the **gold standard** for knowledge, particularly in medical practice.

Research literature is also the key source of **knowledge** driving scientific progress.



Clinical Randomised Control Trial Structure

P Patient, Problem, Population

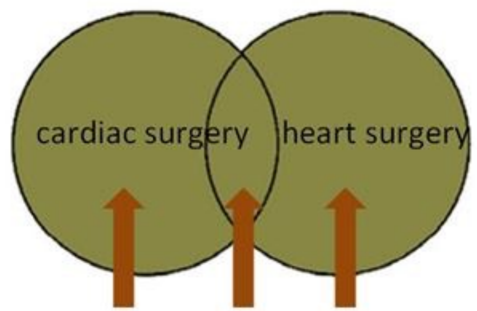
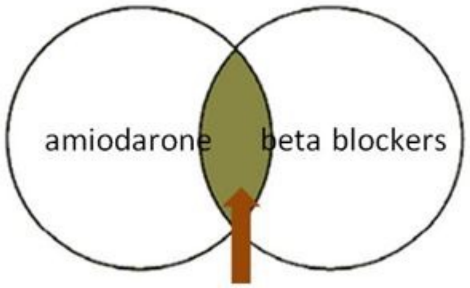
I Intervention

C Comparison or Control

O Outcome

"To assess the effects of [intervention] compared to [comparison/control] for [condition/problem] in [population] in [context] on [outcomes]."

Structured querying of biomedical literature

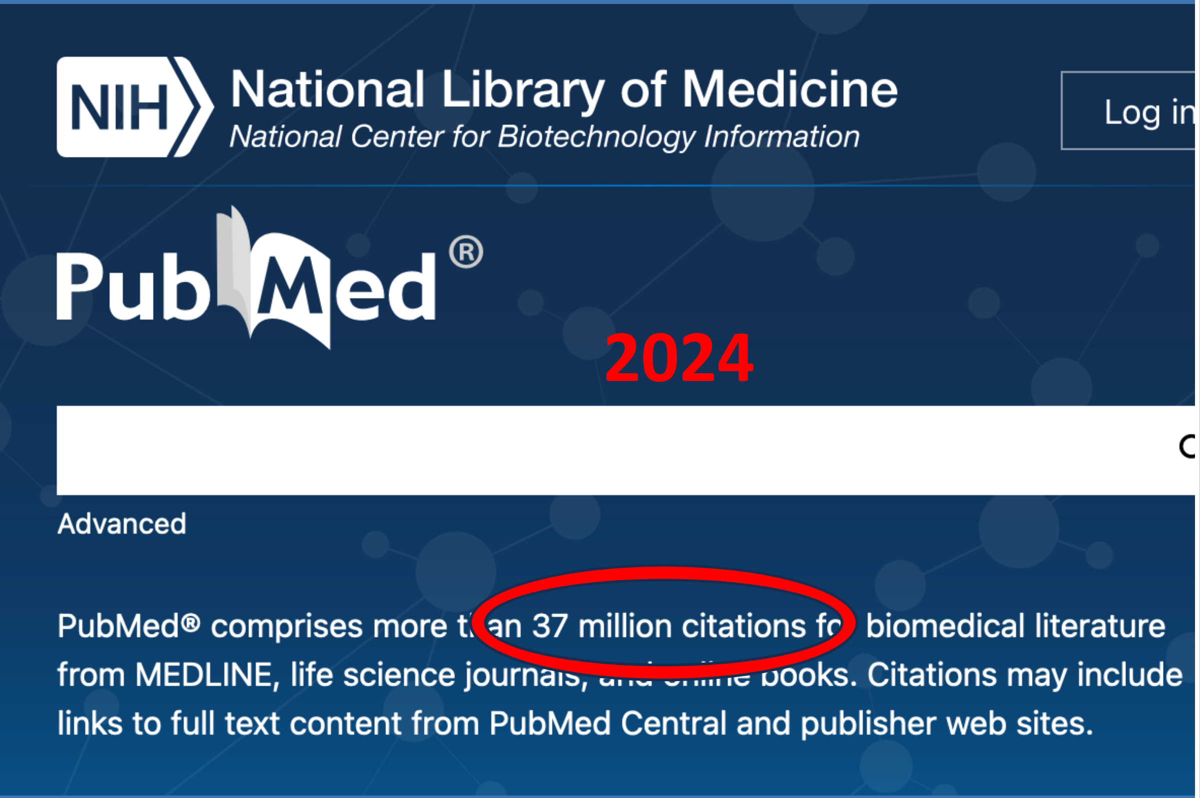
Using **OR** lets you broaden your search by combining synonyms to appropriately cover a concept. This search will retrieve articles containing each term separately, as well as both terms together.	
Using **AND** lets you narrow your search and is used to combine concepts. This search will retrieve articles containing both terms only.	

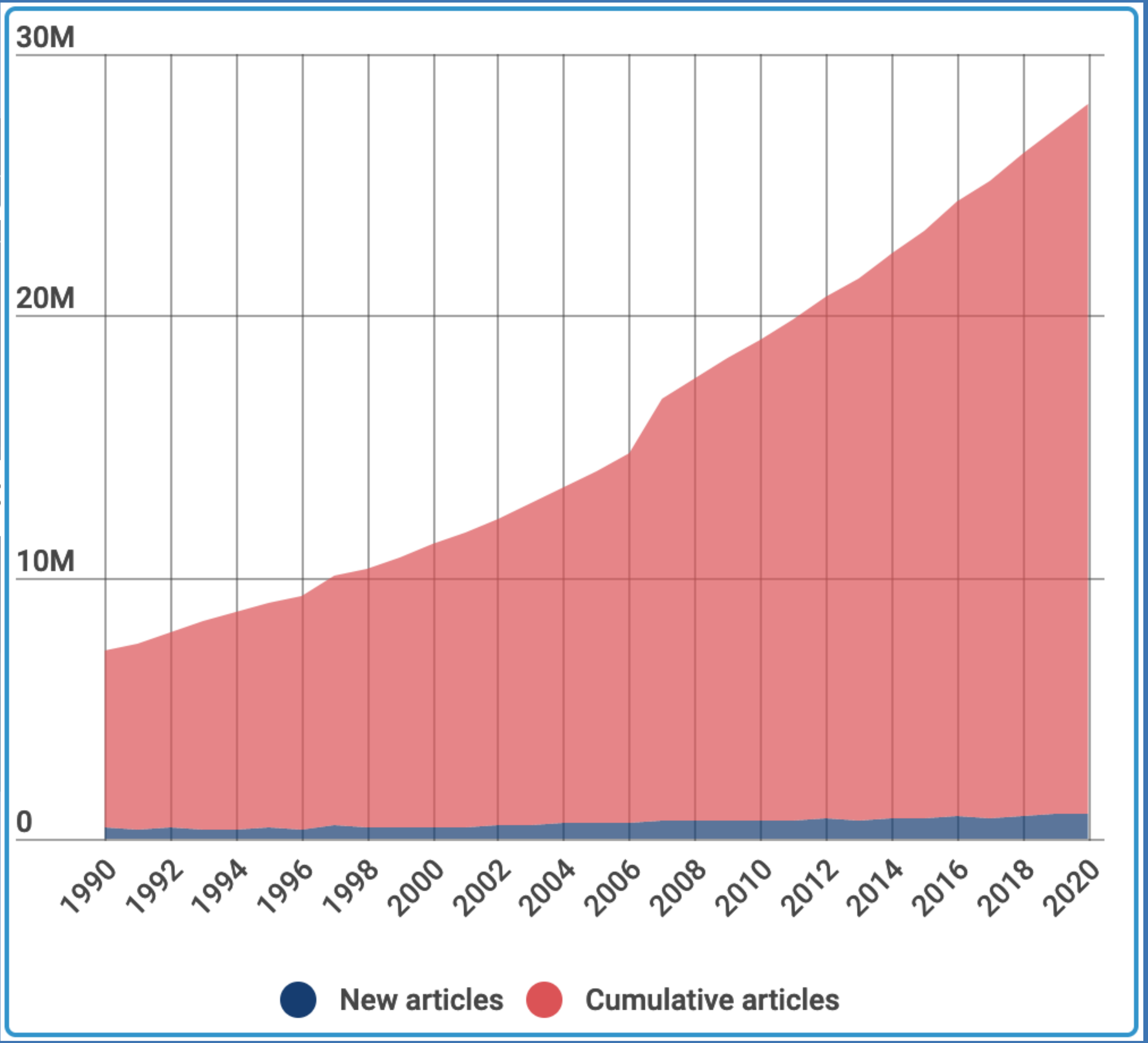
Search Process

Define main **concepts** in search topic.
Focus on key terms, phrases, synonyms or variants.
Add in **MeSH terms** to constrain results.
….
Select and read papers.
"Snowball search" to find papers citing relevant papers.

("Alzheimer Disease"[mh] OR "alzheimer's"[tiab] OR "alzheimer"[tiab] OR ad[tiab] OR "alzheimers"[tiab] OR "alzhiemer"[tiab] OR "alzhiemers"[tiab] OR "alzhiemer's"[tiab] OR "cognition disorders"[mh] OR cognitive[tiab] OR cognition[tiab] OR "dementia"[mh:noexp] OR dementia[tiab])

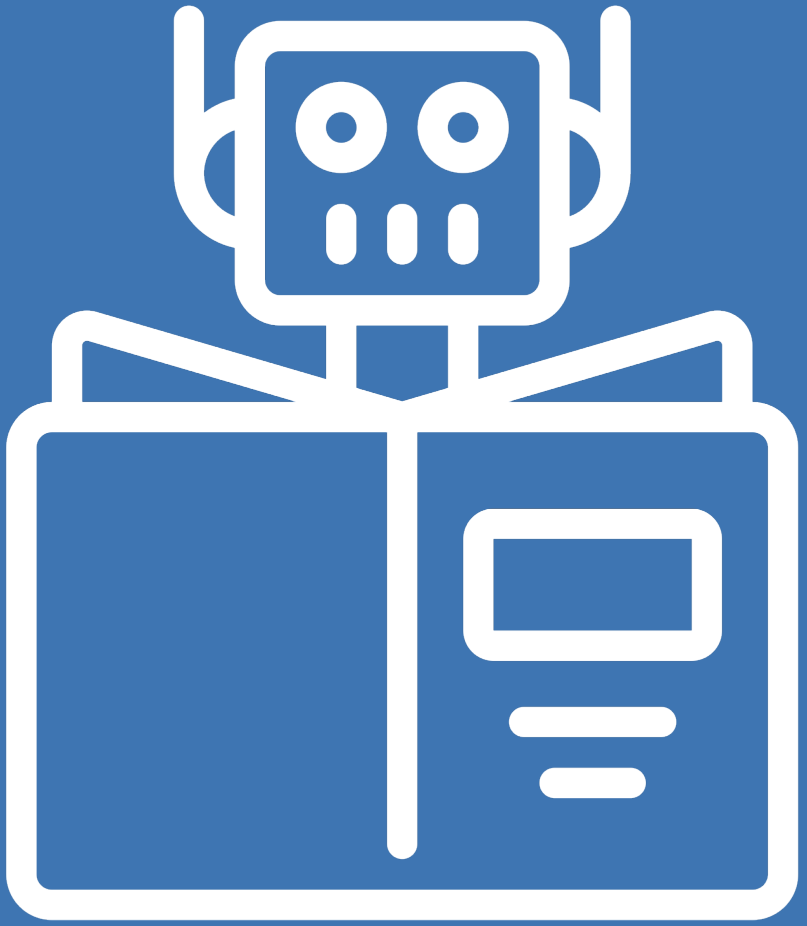
The research literature is huge and growing





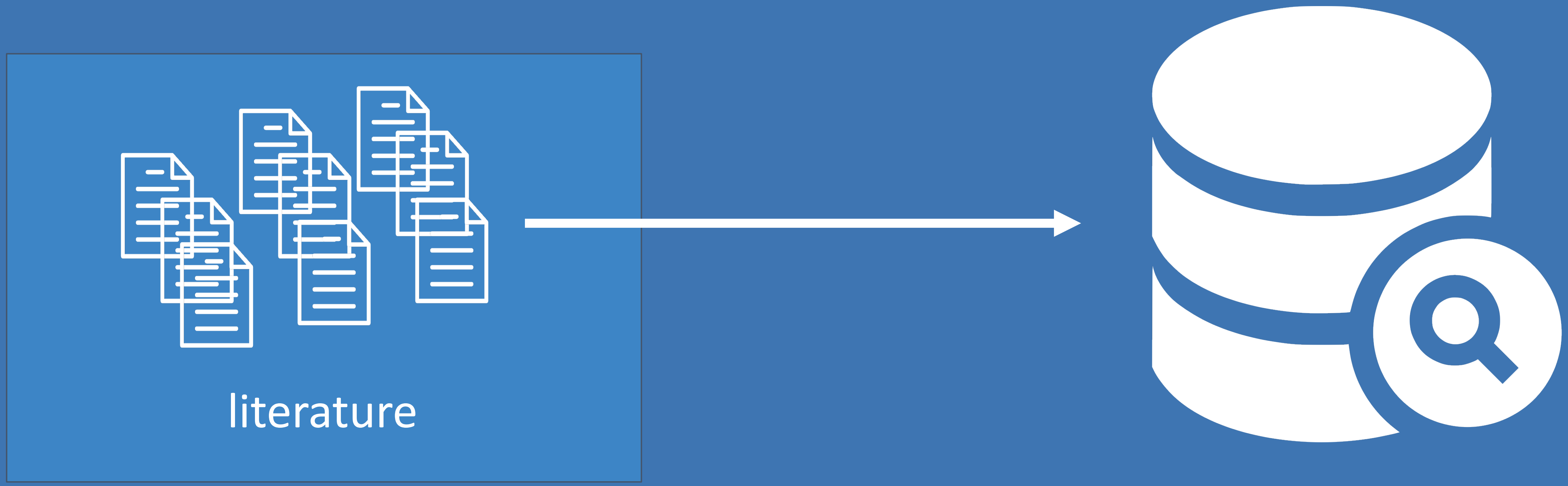
AI supporting science

- **Evidence detection** enabling *concept* search
- **Evidence exploration** tools allowing more open-ended literature navigation
- **Evidence summarization** and **synthesis**
- **Evidence discovery**



Evidence Detection

Organising knowledge



- Find key concepts, entities and events
- Map to standard identifiers and/or ontology terms
- Support indexing and retrieval

Recognising biological ontology concepts

Previous in vitro experiments using renal

GO:0005623 – “cell”

CL:0000000 – “cell”

PR:000004182 – “aquaporin-2”

EG:359 – “Aqp2”

cell lines suggest recessive Aqp2

SO:0001059 – “sequence_alteration”

GO:0006810 – “transport”

mutations result in improper trafficking

SO:0001059 – “sequence_alteration”

GO:0015250 – “water channel activity”

of the mutant water pore.

CHEBI:15377 – “water”

Funk C, Baumgartner WA, Garcia B, Roeder C, Bada M, Cohen KB, Hunter LE, Verspoor K. (2014)
Large-scale biomedical concept recognition: An evaluation of current automatic annotators and their parameters. BMC Bioinformatics 15(1):59.

Concept Recognition ≉ Named Entity Recognition

GO_cellular_component

Previous in vitro experiments using renal cell lines

protein mutation

suggest recessive Aqp2 mutations result in

GO_biological_process mutation GO_cellular_component

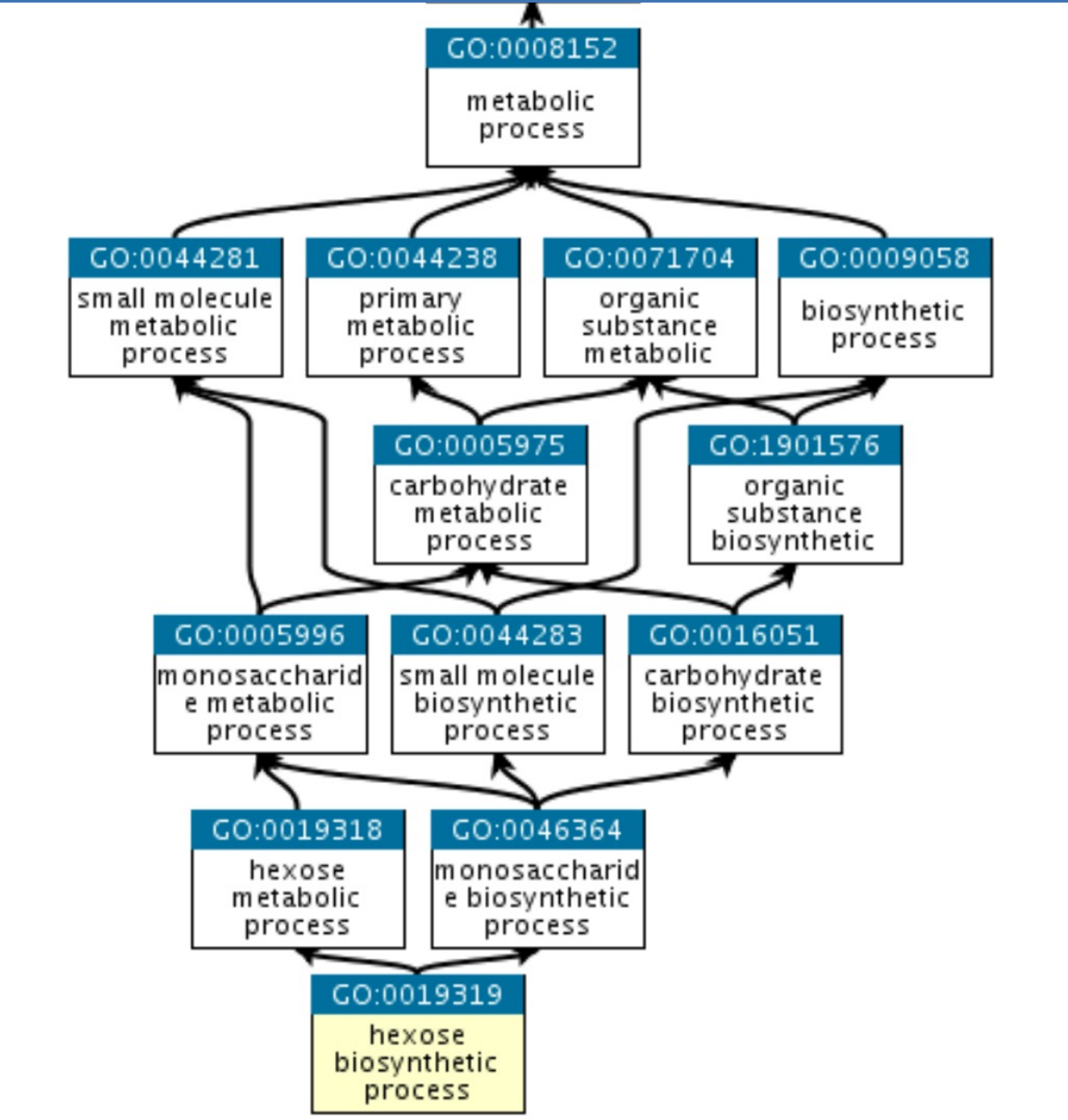
improper trafficking of the mutant water pore.

chemical

- NER: recognising terms of specific general *types*
- Concept recognition: NER + normalize term to ontology ID

Ontologies

- Comprehensive
  - Gene Ontology: 42k terms
  - SNOMED CT “Disease”: 49k terms
  - ICD 11: 17k codes;
    120k terms (5+ languages)
- Systematic (cf. natural)
- Hierarchical



Gene Ontology vs Natural Language

- Variation in PMID: 12925238

[Term]

id: GO:0006900

name: membrane budding

…

def: "The evagination of a membrane, resulting in formation of a vesicle."

…

synonym: "membrane evagination"

synonym: "nonselective vesicle assembly"

synonym: "vesicle biosynthesis"

synonym: "vesicle formation"

…

- Lipid rafts play a key role in **membrane budding**…
- …involvement of annexin A7 in **budding of vesicles**…
- …Ca2+-mediated **vesiculation process** was not impared.
- Red blood cells which lack the ability to **vesiculate** cause…
- Having excluded a direct role in **vesicle formation**…

Machine learning for concept recognition?

NER

- a handful of target classes
- annotated training data with many examples of each class

→ **Leverage annotated data**

→ **Supervised Learning**

Concept Recognition

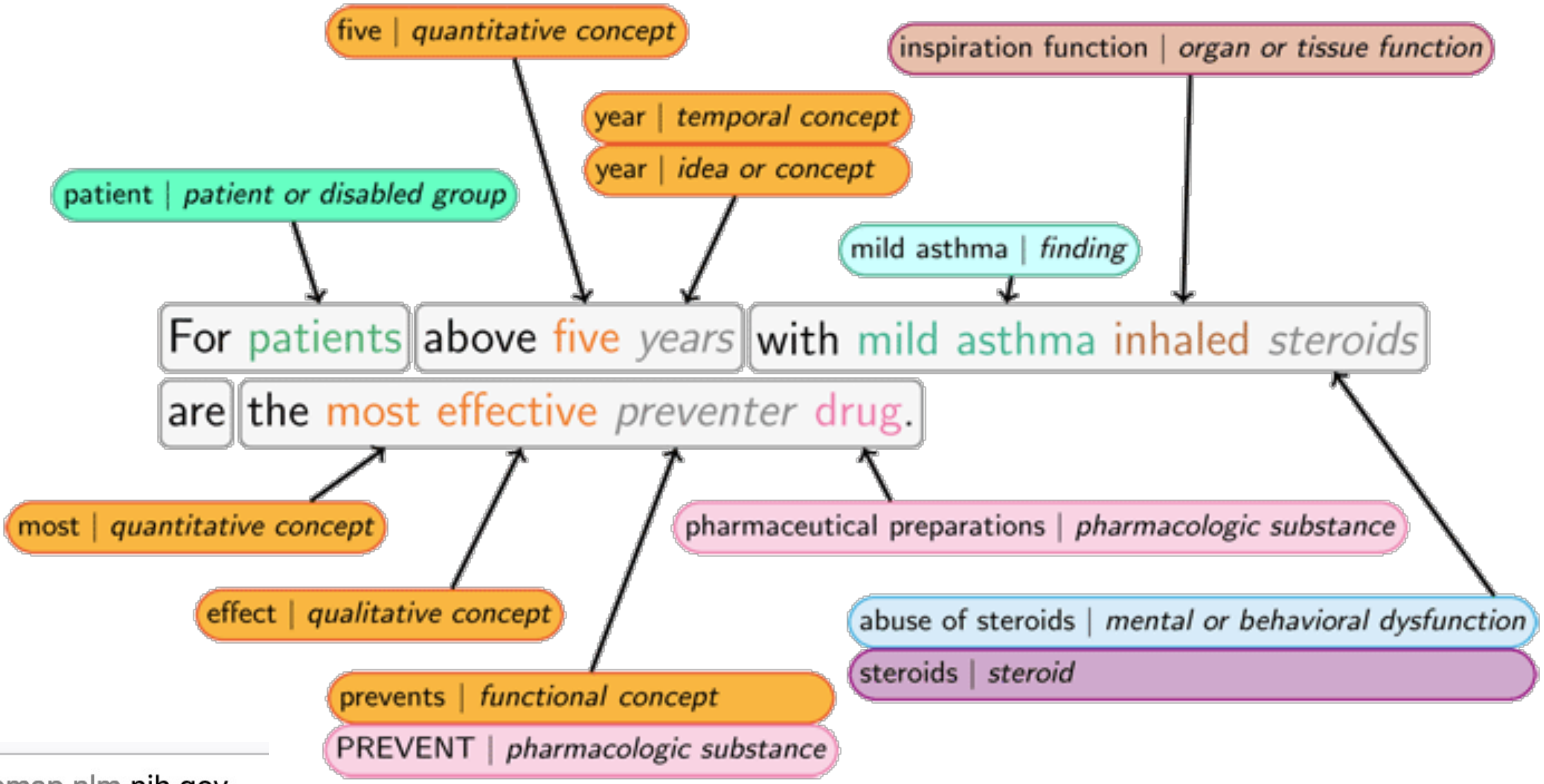
- 10s of thousands of target 'labels'
- difficult to produce enough training data to enable supervision

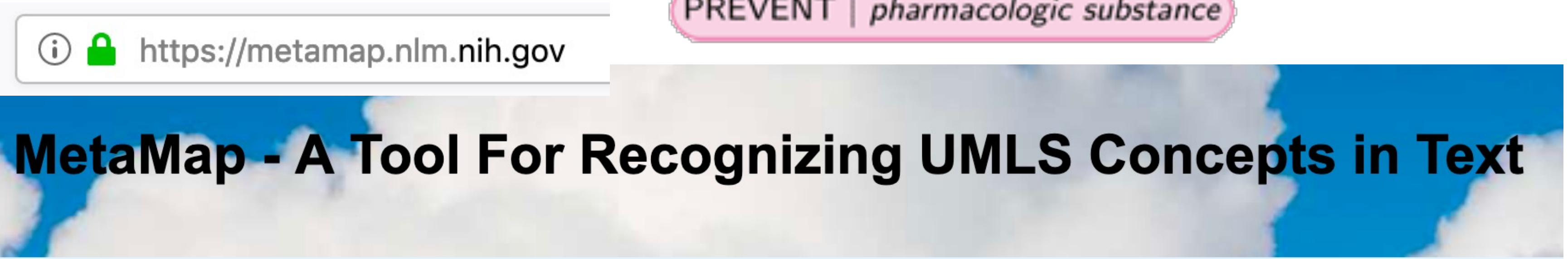
→ **Leverage Ontology itself**

→ **Match terms** (synonyms, etc.)

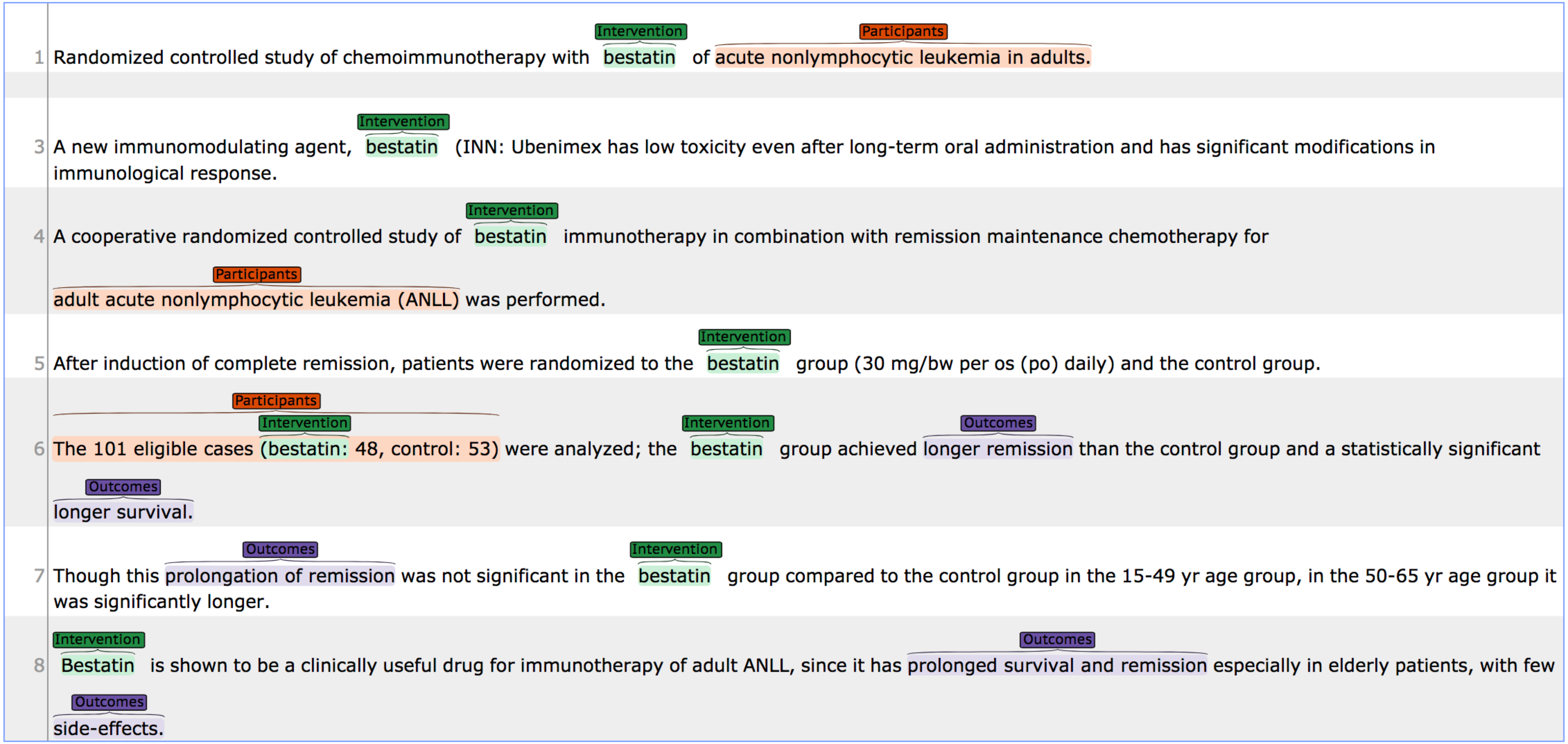
New opportunities with LLMs?

Recognising clinical concepts: UMLS





PICO spans



- Evidence Based Medicine (EBM-NLP) corpus
- BiLSTM-CRF model trained to recognise PICO elements

(EBM-NLP: Nye et al 2018)

PICO + clinical concepts (MeSH/UMLS)

OUTCOME

Hospitalization

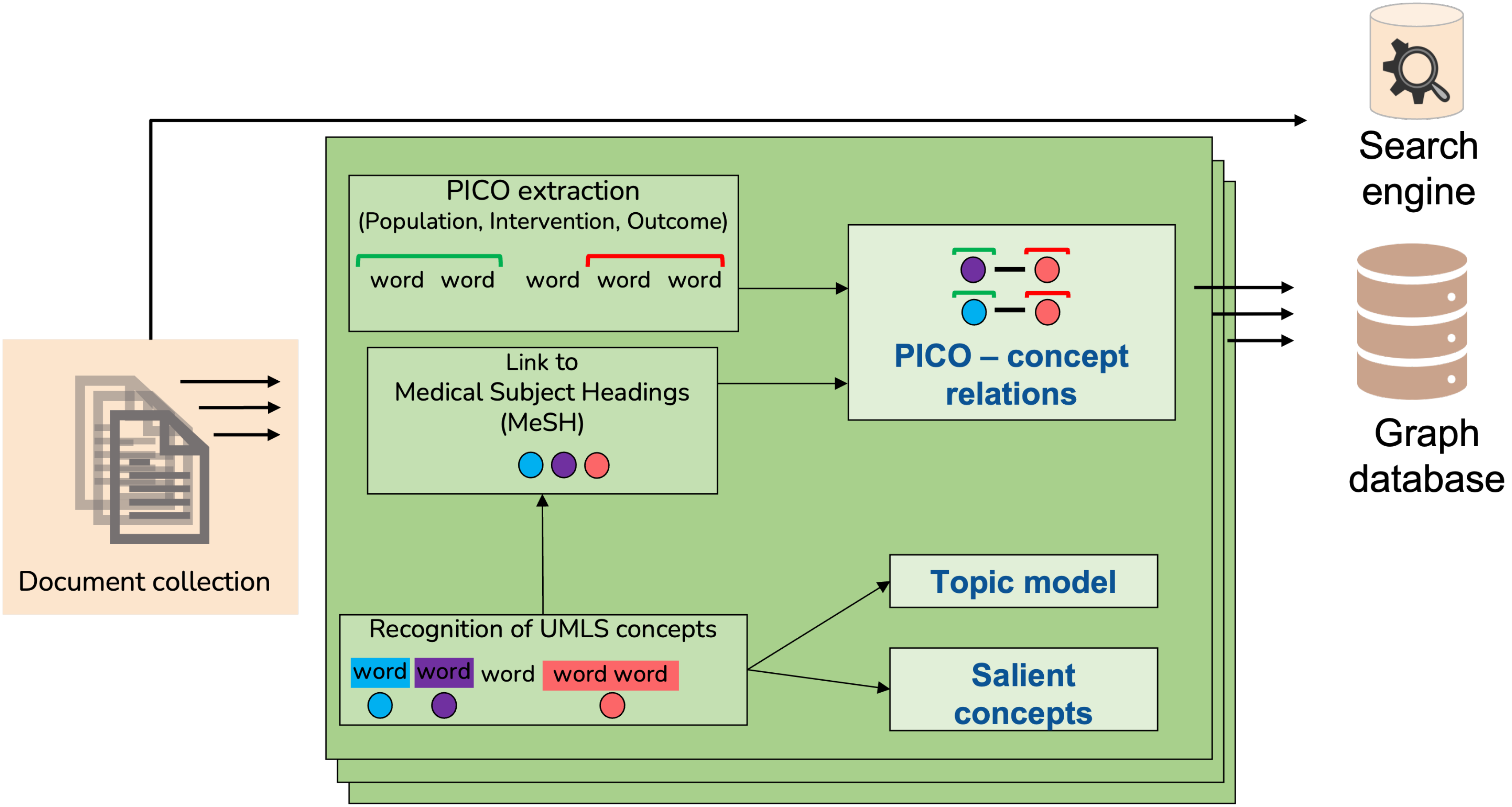
Mortality

“cumulative COVID-19-related hospitalization and death rates”

PICO concepts:

hospitalization$_{OUTCOME}$

mortality$_{OUTCOME}$



Structuring relations

- Capturing entities and relations
  - "PROTEIN interacts with PROTEIN"
  - "CHEMICAL treats DISEASE"
  - "MUTATION causes DISEASE"

- Incorporating knowledge
  - cf. "ACE inhibitor treats hypertension"
  - + benazepril –*isa*- ACE inhibitor

Cyclin E2 interacts with Cdk2 in a functional kinase complex.

protein protein interaction:
    interactor1: cyclin E2
    interactor2: cdk2

id: GO:0009358
name: polyphosphate kinase complex

Chemical-induced disease

Title	**Propylthiouracil**-induced **hepatic damage**
Abstract	Two cases of **propylthiouracil**-induced **liver damage** have been observed. The first case is of an acute type of damage, proven by rechallenge; the second presents a clinical and histologic picture resembling **chronic active hepatitis**, with spontaneous remission.
Entity	D011441, Chemical, “Propylthiouracil”, 0-16
Entity	D011441, Chemical, “propylthiouracil”, 54-70
Entity	D056486, Disease, “hepatic damage”, 25-39
Entity	D056486, Disease, “liver damage”, 79-91
Entity	D006521, Disease, “chronic active hepatitis”, 246-270
Relation	D011441-D056486
Relation	D011441-D006521

Information Extraction from Chemical Patents

2-Phenyl-2H-imidazo[1,5-a]pyridinium tetrafluoroborate (1)

The general synthesis starts with the slow addition of excess concentrated **hydrochloric acid** to **aniline** (4.66 g, 4.6 mL, 50.0 mmol) dissolved in a small amount of **methylene chloride** under rigorous stirring. A solid immediately formed, which was collected, washed with diethyl ether and dried at 40 °C at <10 mbar for two hours. Then the **hydrochloride salt** was dissolved in 100 mL **ethanol**, and 37 wt% aqueous **formaldehyde** solution (2.25 g, 2.1 mL, 75.0 mmol) as well as **2-pyridinecarboxyaldehyde** (5.36 g, 4.8 mL, 50.0 mmol) were added.

2-(4-Methoxyphenyl)-2H-imidazo[1,5-a]pyridinium chloride monohydrate (3)

The synthesis followed *the general procedure as given for **1*** but without salt metathesis to *the corresponding tetrafluoroborate salt*. **4-Methoxyaniline** (6.16 g, 50.0 mmol) was used as *amine*.

Product 1: 2-Phenyl-2H-imidazo[1,5-a]pyridinium tetrafluoroborate
Stage 1:
Reactant 1: hydrochloric acid
Reactant 2: aniline
Solvent 3: methylene chloride
Product: hydrochloride salt[1]
Stage 2: collected, washed with diethyl ether
Stage 3:
Reactant 4: hydrochloride salt[1]
Solvent 5: ethanol
Solvent 6: aqueous formaldehyde solution
Reactant 7: 2-pyridinecarboxyaldehyde
Product 3: 2-(4-Methoxyphenyl)-2H-imidazo[1,5-a]pyridinium chloride monohydrate

- Pull out key entities and events
- Identify roles of entities
- Resolve references and analogous reactions
- Structure chemical information
  - ➤ Search
  - ➤ Compare
  - ➤ Synthesise
  - ➤ Connect
  - ➤ Discover
  - ➤ Characterise

ChEMU
Cheminformatics Elsevier
Melbourne Universities

https://chemu.eng.unimelb.edu.au/

A Chemical Reaction Snippet

10.0 g (35.0 mmol) of **2-tert-butyl 4-ethyl 5-amino-3-methylthiophene-2,4-dicarboxylate** (Example 1A) were dissolved in 500 ml of **dichloromethane** and 11.4 g (70.1 mmol) of **N,N'-carbonyldiimidazole** (CDI) and 19.6 ml (140 mmol) of **triethylamine** were added

ID	Type	Text span
T1	Starting _material	2-tert-butyl 4-ethyl 5-amino-3-methylthiophene-2, 4-dicarboxylate
T2	Solvent	dichloromethane
T3	Starting _material	N,N'-carbonyldiimidazole
T4	Reagent	triethylamine
T5	Trigger	dissolved
T6	Trigger	added

ID	Event type	Event trigger	Argument _1	Argument _2	Argument _3
E1	Reaction _step	T5	Theme:T1	Theme:T2	
E2	Reaction _step	T6	Theme:E1	Theme:T3	Theme:T4

Task 1 – NER – in Red
Task 2 – Event extraction – in Purple

Quick aside on anaphora:

Rich anaphora phenomena in procedural texts

1. To the solution of Compound (4) (0.815 g, 1.30 mmol) in THF (4.9 ml) in a flask were added acetic acid (9.8 ml) and water (4.9 ml).
2. The mixture was stirred for 3 hrs at 50°C and then cooled to 0°C.
3. 2N-sodium hydroxide aqueous solution was added to the mixture until the pH of the mixture became 9.
4. The mixture was extracted with ethyl acetate for 3 times.
5. The combined organic layer was washed with water and saturated aqueous sodium chloride.
6. The organic layer was dried over anhydrous magnesium sulfate and evaporated.

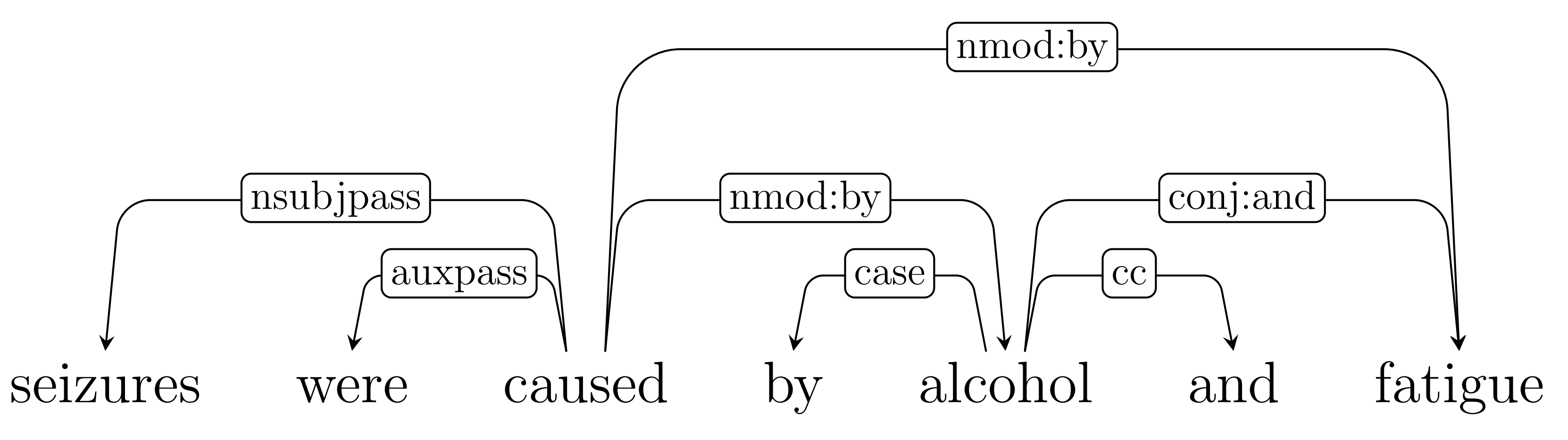
An example from chemical patents

1. Preheat the oven to 400F.
2. Lightly grease a baking sheet.
3. Place the biscuits on the prepared baking sheet and use the palm of your hand to flatten the dough to 1/4 inch in thickness.
4. Divide the sauce evenly among the biscuits, top with a pinch of the oregano, then layer the mozzarella, pepperoni (if using), and Parmesan cheese.
5. Make sure the cheese is covering and bake until the biscuits are golden, about 15 minutes.
6. Allow the biscuits to cool slightly and serve warm.

An example from recipes

slide figures courtesy Biaoyan Fang

Machine learning of entity relations with Approximate Subgraph Matching



Shortest path between *seizures* and *fatigue* is through *caused*

$\phi_{dist}=(0.9)^2$ $\phi_{forward}=0.9$ $\phi_{backward}=0.9$ $\phi_{nsubjpass}=0.9$ $\phi_{nmod:by}=0.9$

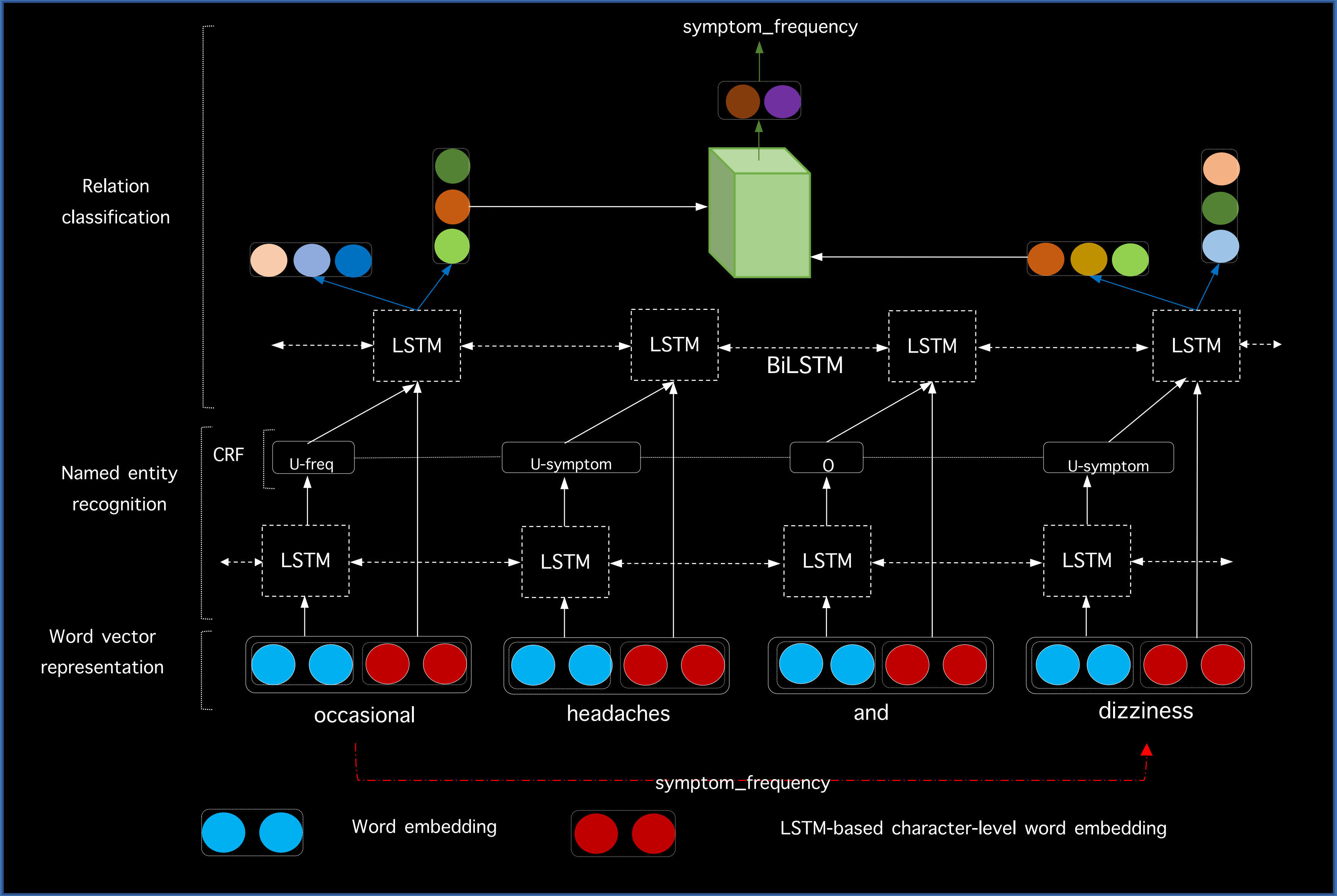
Consider shortest path between vertices x and y in Graph G

Approximate subgraph matching:

Feature map ϕ concatenation of features for similarity of graph along structural, directional, edge dimensions.

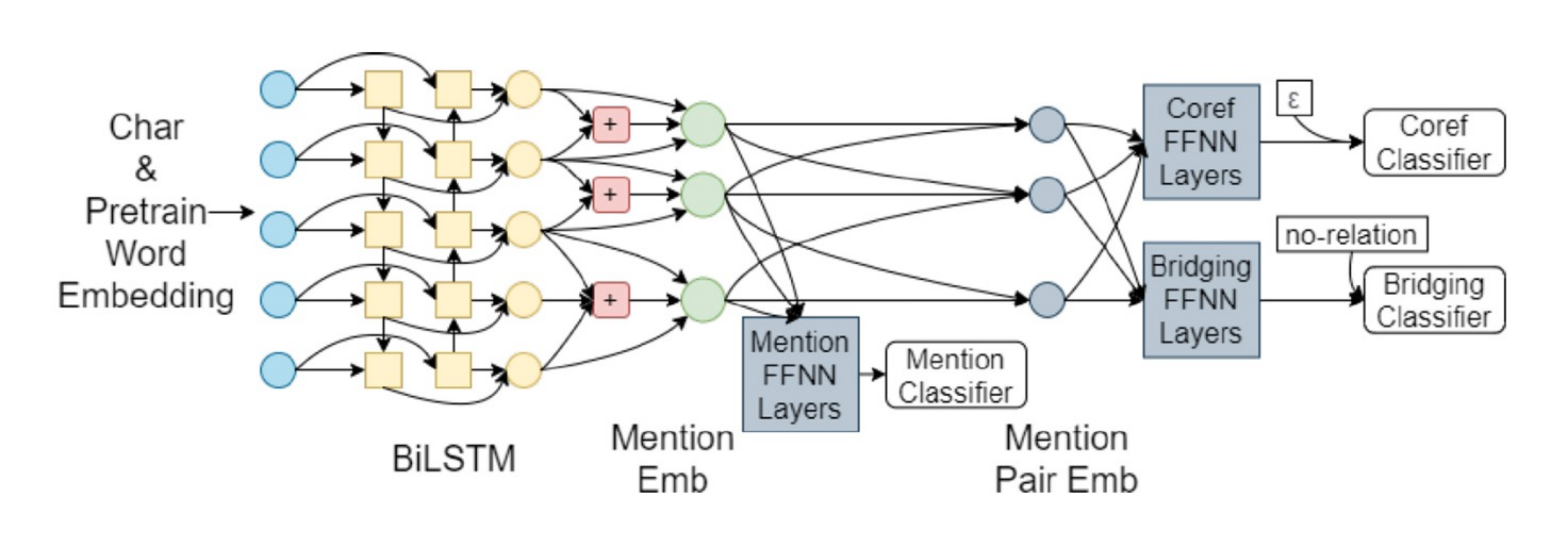
Liu H, Hunter L, Keselj V, Verspoor K. (2013) Approximate Subgraph Matching-based Literature Mining for Biomedical Events and Relations. PLoS ONE 8(4):e60954. DOI: 10.1371/journal.pone.0060954

Methods for Information Extraction



Nguyen and Verspoor. "End-to-end neural relation extraction using deep biaffine attention." *ECIR* 2019.

Methods for Information Extraction



Pipeline

Entity detection using pre-trained word embeddings → Relation classification

slide figures courtesy Biaoyan Fang

Organising knowledge enables semantic search

keywords	→	concepts
AND / OR co-occurrence	→	relations

literature

→

literature
augmented with concepts and relationships

Find papers on

[bariatric surgery]
[type 2 diabetes]
[remission]

[Flurbiprofen]
[metabolized-by]
[CYP2C9]

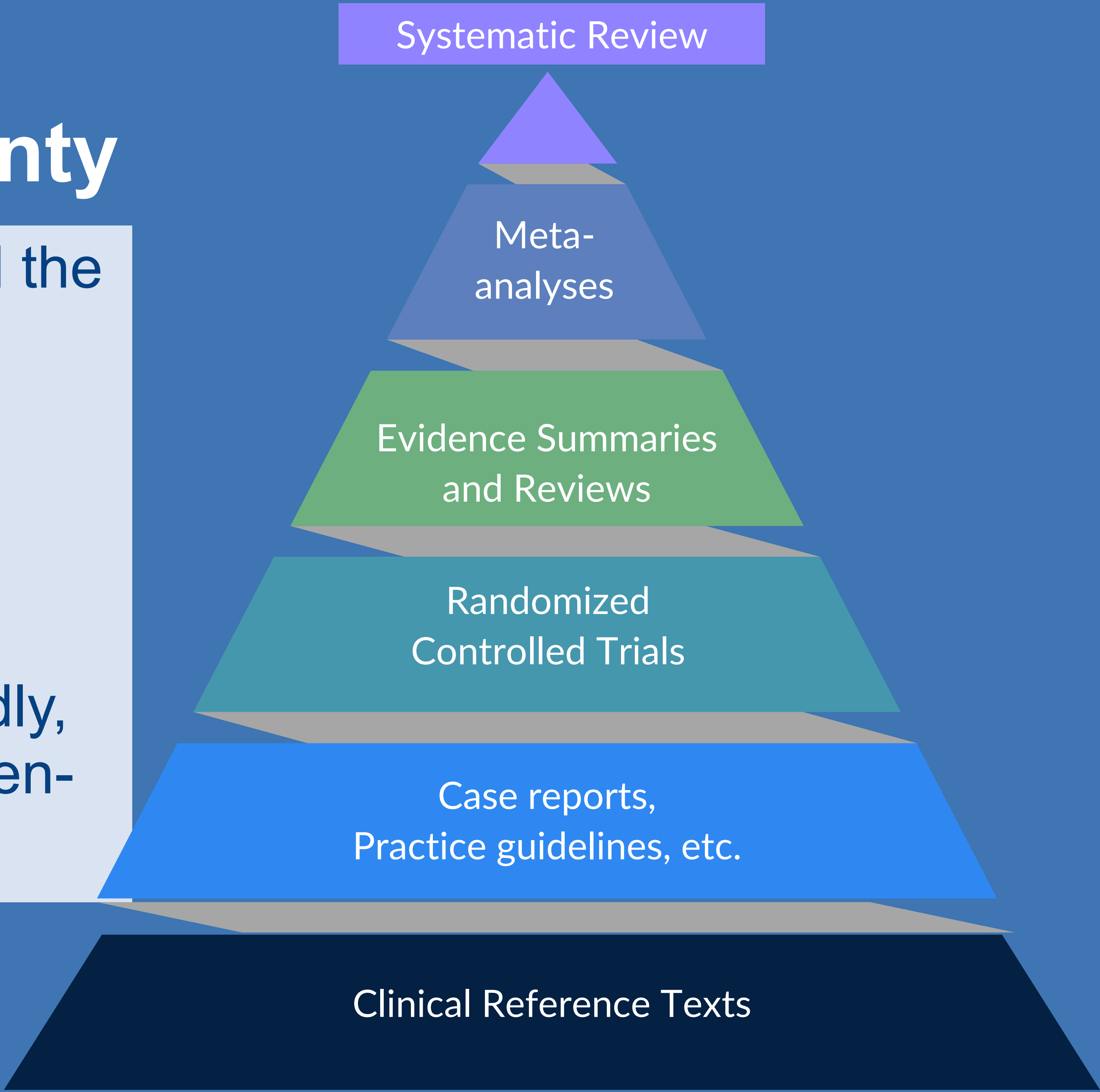
Evidence Exploration

Evidence
– in times of uncertainty

Research is (typically) slow, and the need for rapid accumulation of information is sometimes great

– such as during the COVID-19 pandemic.

We needed to get answers rapidly, and our questions were very open-ended.

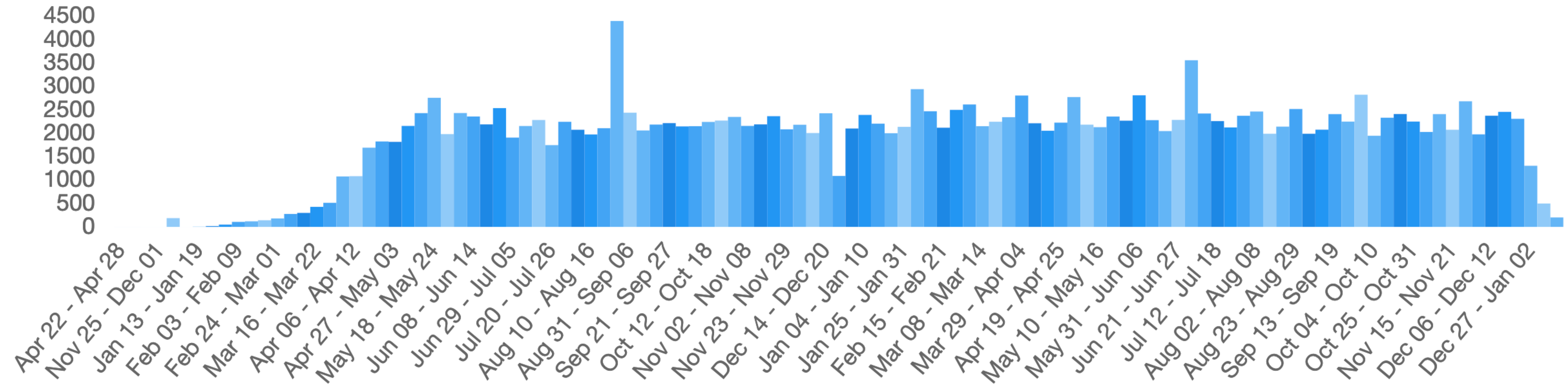


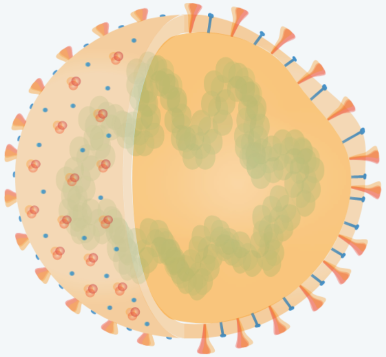
LitCovid

Ex: Remdesivir

NIH NLM

WEEKLY PUBLICATIONS



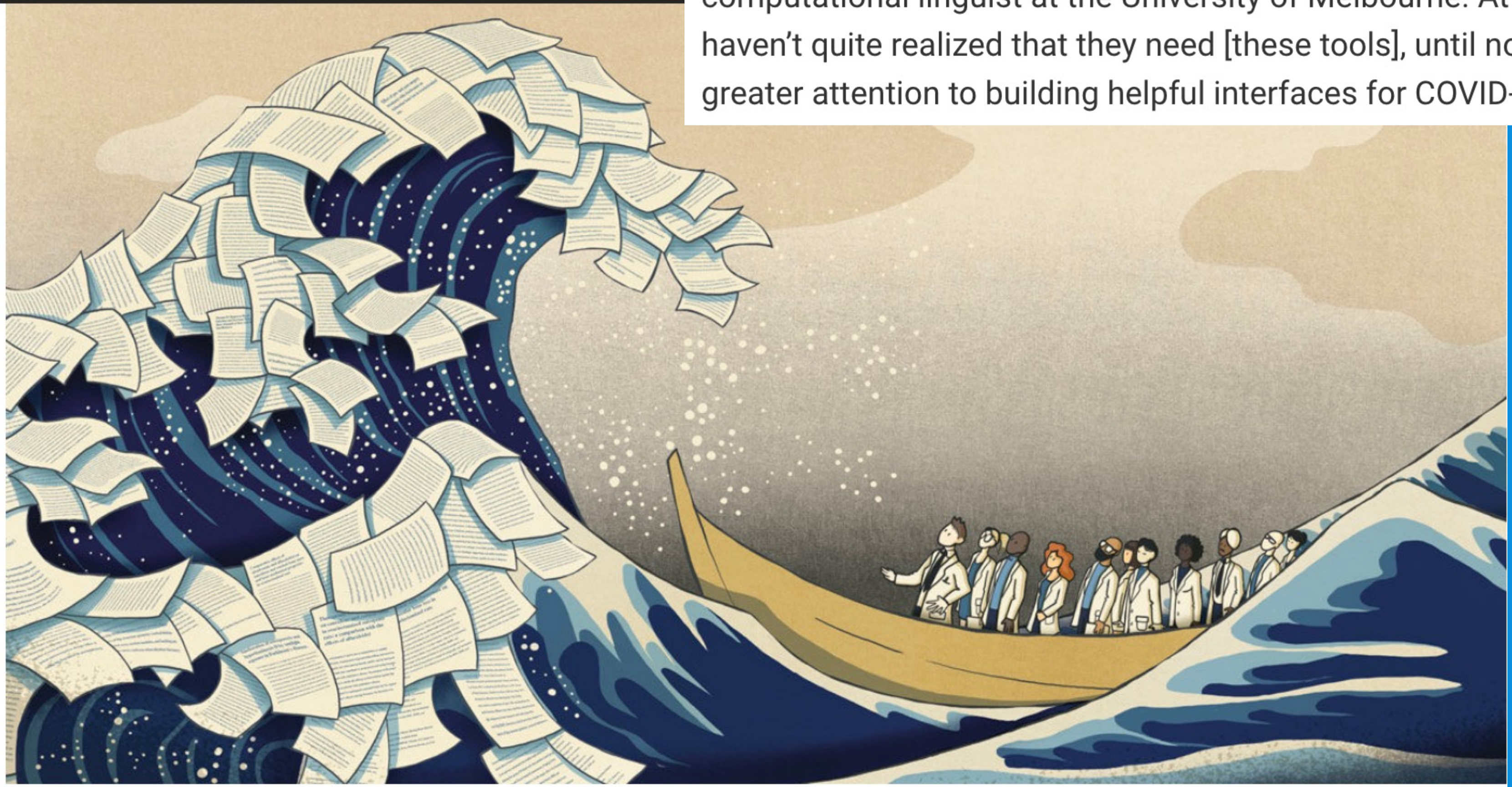


LitCovid is a curated literature hub for tracking up-to-date scientific information about the 2019 novel Coronavirus. It is the most comprehensive resource on the subject, providing a central access to 209561 (and growing) relevant articles in PubMed. The articles are updated daily and are further categorized by different research topics (e.g. transmission) and geographic locations.

Cite | FAQ | Download | Long Covid

Science Contents News

Another challenge is making the tools more user friendly. Although data scientists have spent more than 20 years building tools to mine other topics in scientific literature, they have lagged in fine-tuning ways to help users explore the content of research articles, says Karin Verspoor, a computational linguist at the University of Melbourne. At the same time, “People on the user side haven’t quite realized that they need [these tools], until now,” she says. And that could promote greater attention to building helpful interfaces for COVID-19 and, eventually, other research topics.



SARA GIRONI CARNEVALE

Scientists are drowning in COVID-19 papers. Can new tools keep them afloat?

By **Jeffrey Brainard** | May. 13, 2020 , 12:15 PM

Standard search tools aren’t good enough …

Science discovers text mining!

WHO: “Find new insights”

What has been published about medical care?

What do we know about vaccines and therapeutics?

What do we know about COVID-19 risk factors?

What do we know about non-pharmaceutical interventions?

What do we know about diagnostics and surveillance?

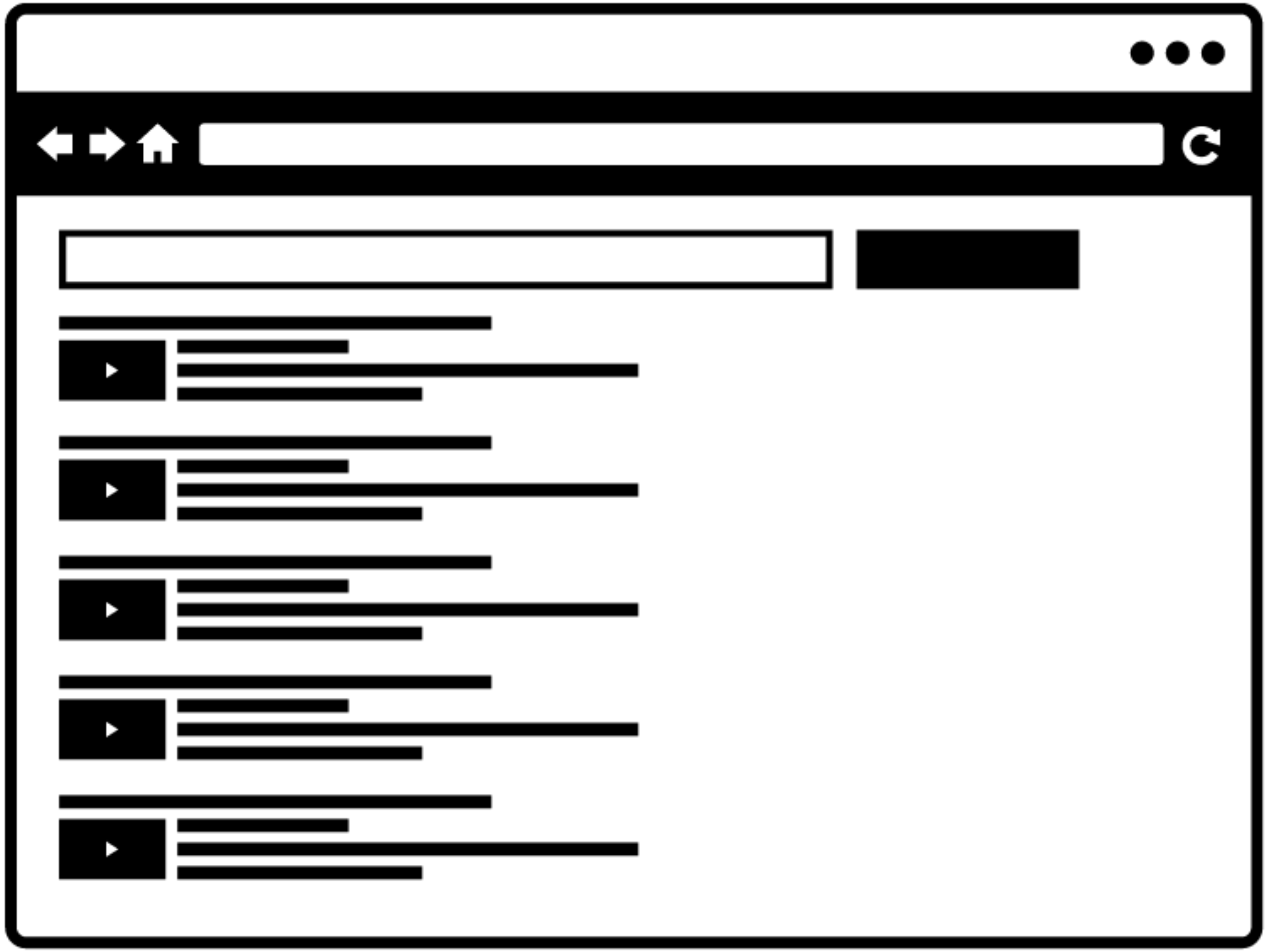
Supporting access to information: Search

targeted queries

well-defined information need

keyword snippet previews

intelligent ranking



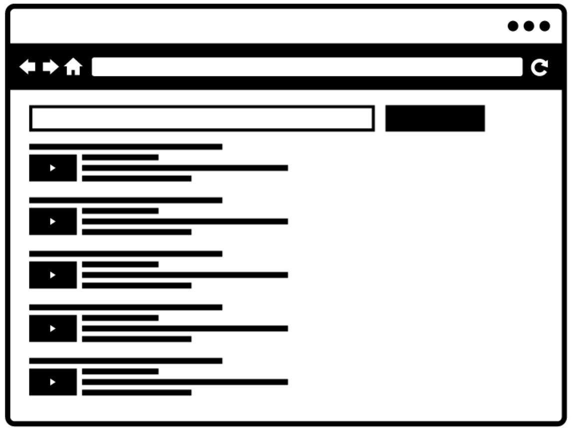
user formulates query

user scans results

limited depth in search

results opened individually

What’s Beyond Search?



document retrieval

information analysis
and synthesis

Transforming documents into information (automatically) requires **AI/NLP**

- categories
- concepts
- relationships
- summaries
- synthesis

Introducing COVID-SEE

http://covid-see.com

search

exploration

concepts

relations

COVID-SEE
Scientific Evidence Explorer

collection-level overview

visual summaries

thematic generalisation

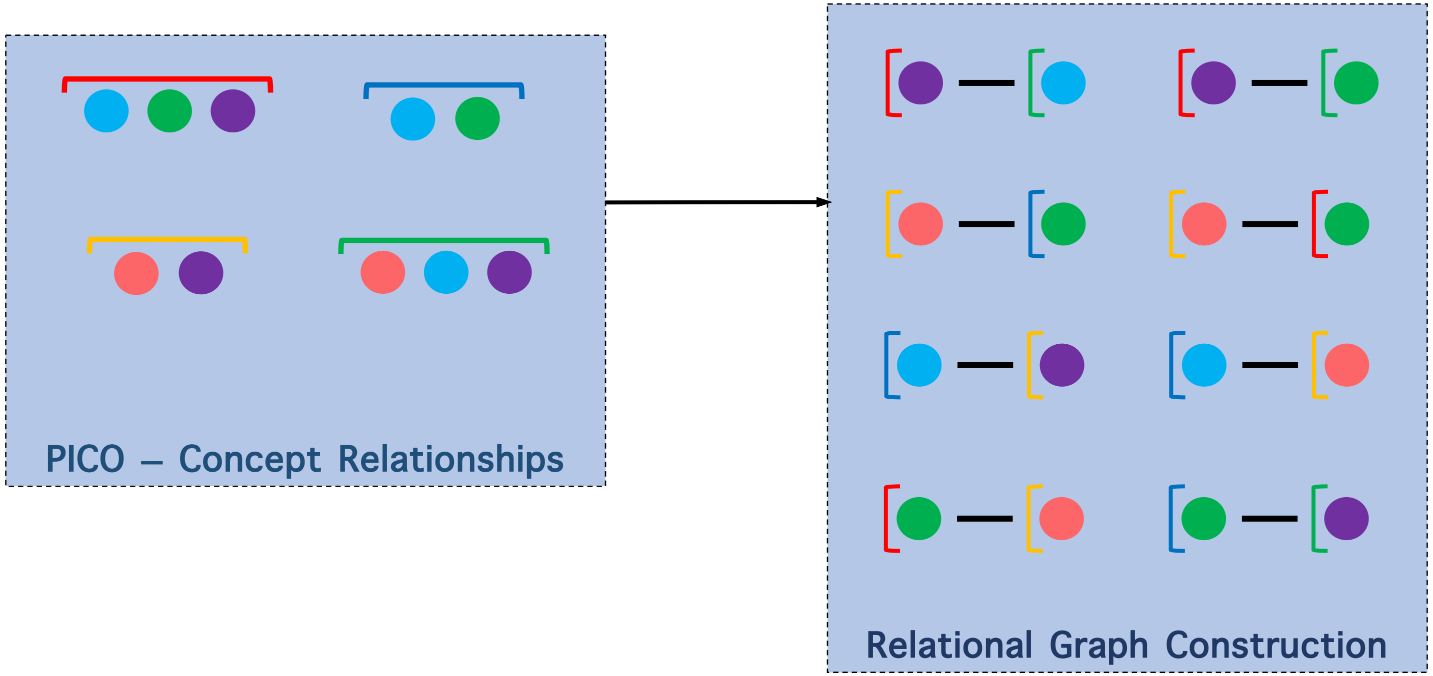
user "briefcase"

Verspoor et al (2021) DOI: https://doi.org/10.1007/978-3-030-72240-1_65

co-occurrence of
PICO concepts:

child$_{POPULATION}$

quarantine$_{INTERVENTION}$



Population concept – Intervention concept

respiratory tract infection – vaccines

child – quarantine

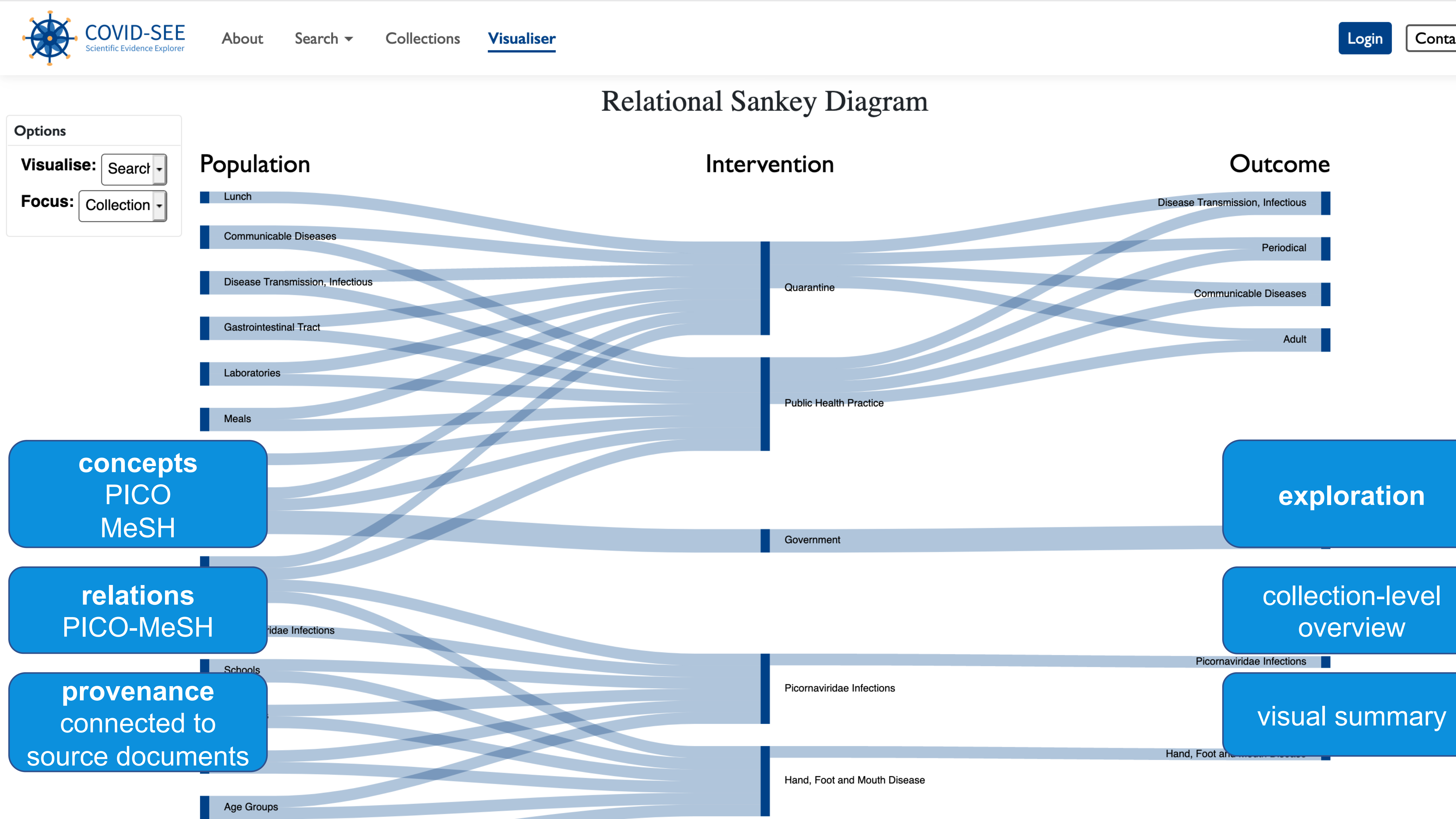
students – health education

Intervention concept – Outcome concept

vaccines – recovery

antiviral agents – cd8+ T-lymphocytes

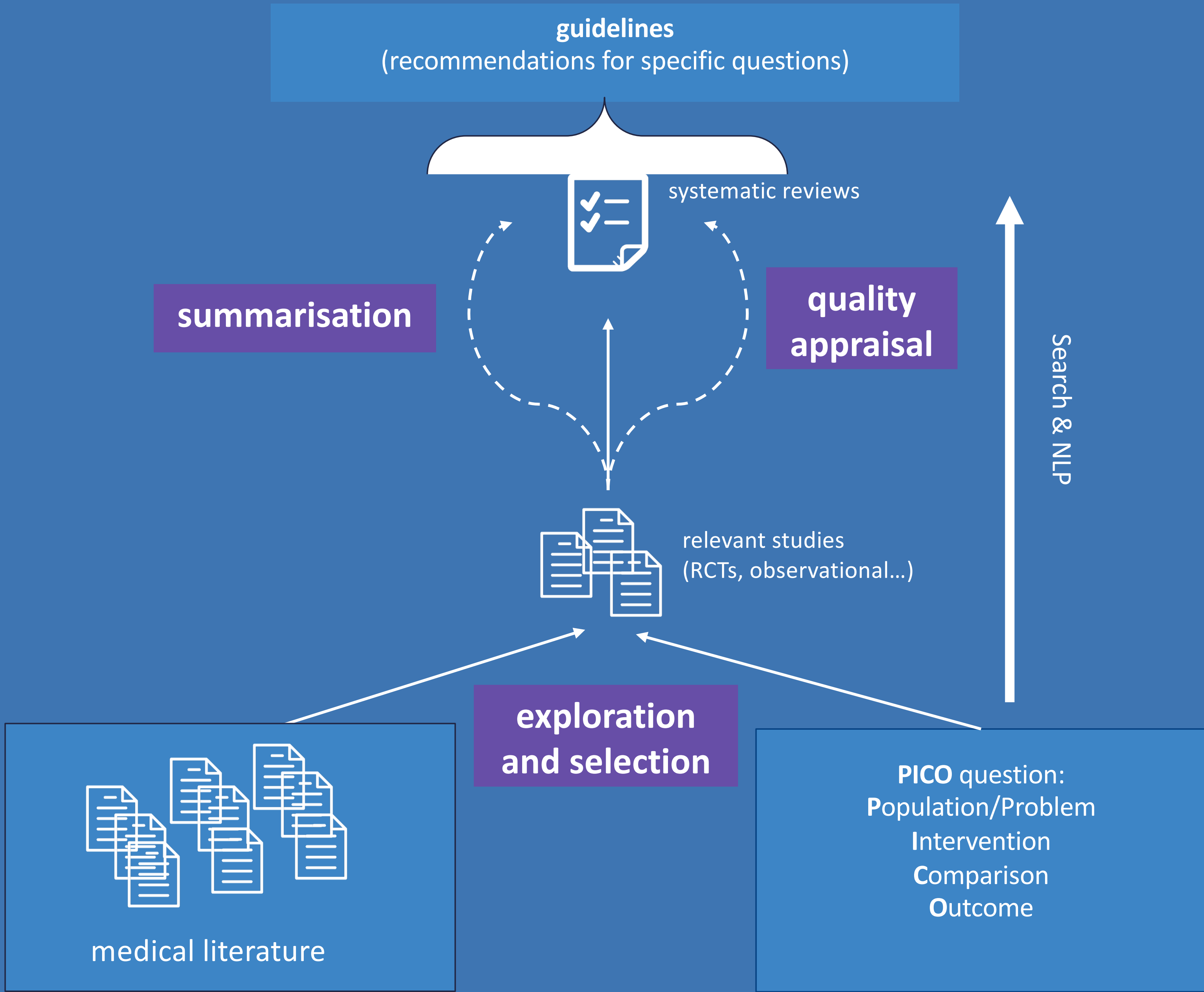
health education – attitudes

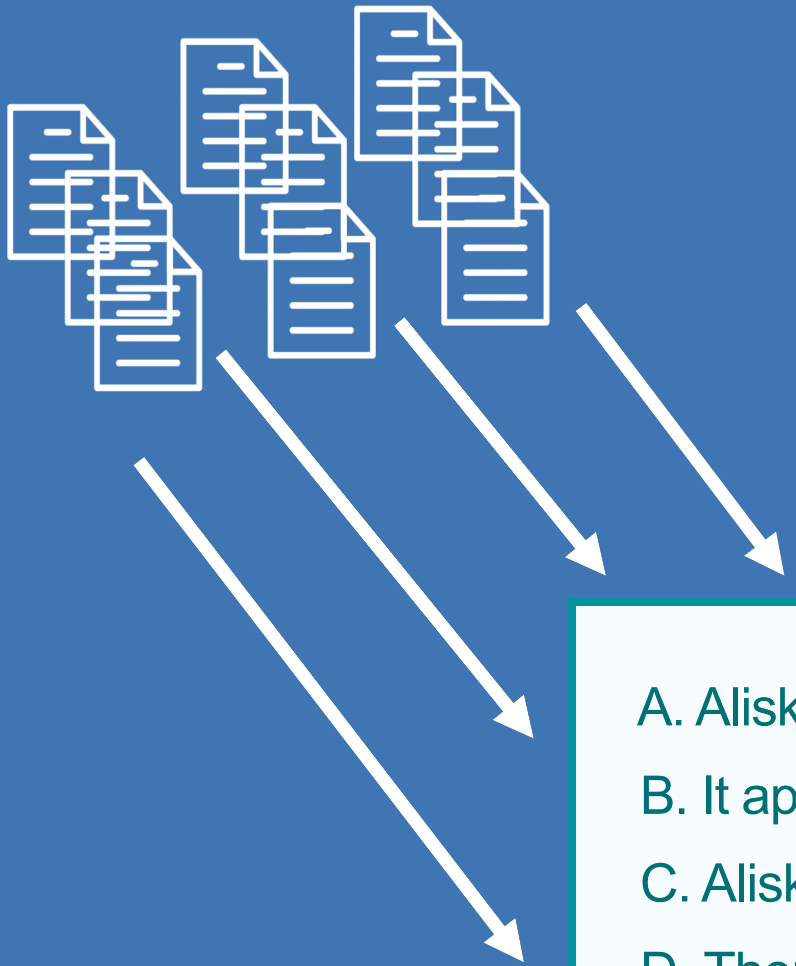

COVID-SEE
Scientific Evidence Explorer
About
Search
Collections
Visualiser
Login
Conta
Relational Sankey Diagram
Options
Visualise:
Search
Focus:
Collection
Population
Intervention
Outcome
Lunch
Communicable Diseases
Disease Transmission, Infectious
Gastrointestinal Tract
Laboratories
Meals
idae Infections
Schools
Age Groups
Quarantine
Public Health Practice
Government
Picornaviridae Infections
Hand, Foot and Mouth Disease
Disease Transmission, Infectious
Periodical
Communicable Diseases
Adult
Picornaviridae Infections
concepts
PICO
MeSH
relations
PICO-MeSH
provenance
connected to
source documents
exploration
collection-level
overview
visual summary

Evidence synthesis

https://www.theguardian.com/world/2022/jan/03/australias-covid-positive-test-rate-why-is-it-rising-and-what-does-it-really-mean

Systematic reviewing





Summarisation requires synthesis

[Aliskiren]
[blood pressure]
[lower]

A. Aliskiren lowers blood pressure
B. It appears that Aliskiren may lower blood pressure
C. Aliskiren does not lower blood pressure
D. There is not enough evidence to confirm if Aliskiren lowers blood pressure
F. Thus it is important to establish if Aliskiren lowers blood pressure

Building blocks of a claim

[Aliskiren, blood pressure]	PICO = **clinical question**
[Aliskiren, blood pressure] + [lower]	PICO + direction = **proposition**
[Aliskiren, blood pressure] + [lower] + [may]	PICO + direction + modality = **claim**

No modality, no claim!

~~F. Thus it is important to establish if Aliskiren lowers blood pressure~~

slide courtesy Yulia Otmakhova

PICO-level aggregation

Do **robotic companions** help elderly patients?

This pilot study, which compared the benefits of a **robotic cat** and a plush toy cat as interventions for elderly persons with dementia….

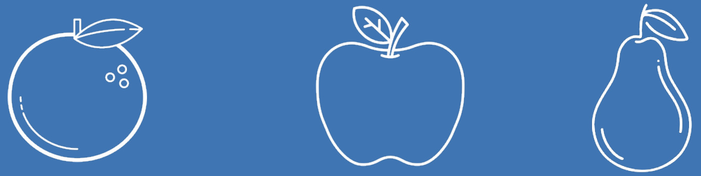
Findings on usability and user experience illustrate that the **robot** has considerable potential to be accepted to support daily living at home.

Socially assistive robot (SAR) technology could assume new roles in health and social care to meet this higher demand.

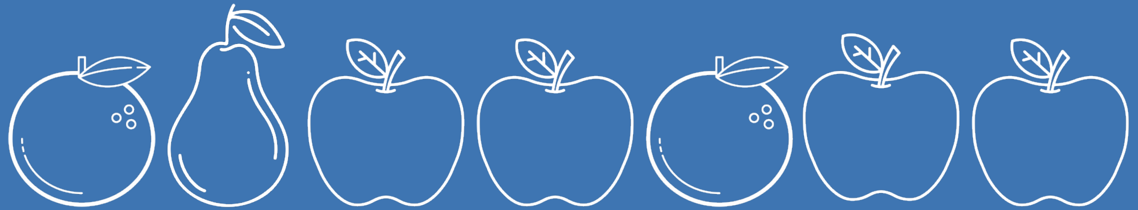
… impact of such low-cost **robotic pets** based on perceptions and experiences of its use with older adults…

The easiest (but still hard).

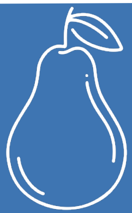
Aggregating direction



(e.g. increases, decreases, no effect)

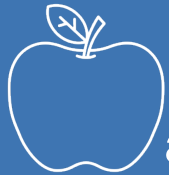
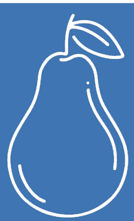
Direction from input documents: 

Bad strategies:

Listing:  and  and 

Majority: 

Good strategies:

Contrasting: Some say that  while others  and some 

Synthesis:  (or something else, it depends)

slide courtesy Yulia Otmakhova

Aggregating modality

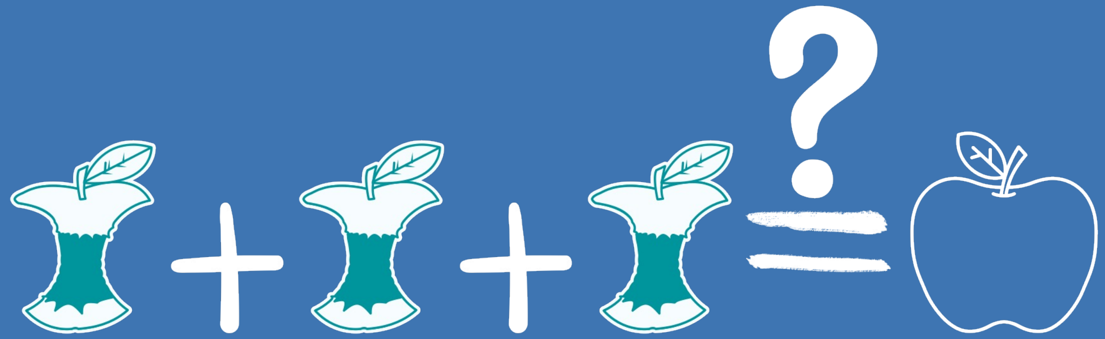
When do we say there is **no evidence**?

Too much conflict?

Weak evidence?

No evidence at all?

How do we aggregate weak to moderate to strong?



slide courtesy Yulia Otmakhova

Quality assessment in evidence synthesis

Population

"Post-menopausal women"

Intervention

What confidence should we have in the estimate (based on flaws in trial design, conduct, analysis or reporting)?

effect: **decreases**

recommend

do not recommend

Impossible to decide without assessing the evidence quality

Inspired by Guyatt et al. (2008). GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *Bmj*, *336*(7650), 924-926.
Suster … Verspoor. (2023) Automating Quality Assessment of Medical Evidence in Systematic Reviews. Journal of Medical Internet Research 2023. DOI: 10.2196/35568

Constructing systematic reviews and quality assessment

Systematic review

Healthcare Question
about diagnosis, screening, prevention, and therapy

Population, Intervention and Comparison

Study 1 | Study 2 | Study 3 | Study 4 | Study 5

Outcome 1 | Outcome 2 | Outcome 3

Generate an estimate of effect for each outcome

Rate the quality of evidence *for each outcome*, across studies

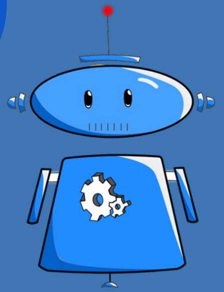
If necessary, reduce the rating for **REASONS** such as risk of bias, imprecision, inconsistency of results, indirectness of evidence, and publication bias

Final rating of evidence quality for each outcome:
HIGH, MODERATE, LOW or VERY LOW

Guideline development

Our goal:

Assume we're given a piece of evidence from a systematic review, predict its quality



Dataset + Tasks + Models with heterogeneous inputs (structured and non-structured)

Grading of Recommendations Assessment, Development and Evaluation (GRADE) framework

Randomised controlled trials: HIGH

Downgrade for:

- Risk of bias
- Inconsistency
- Indirectness
- Imprecision
- Publication bias

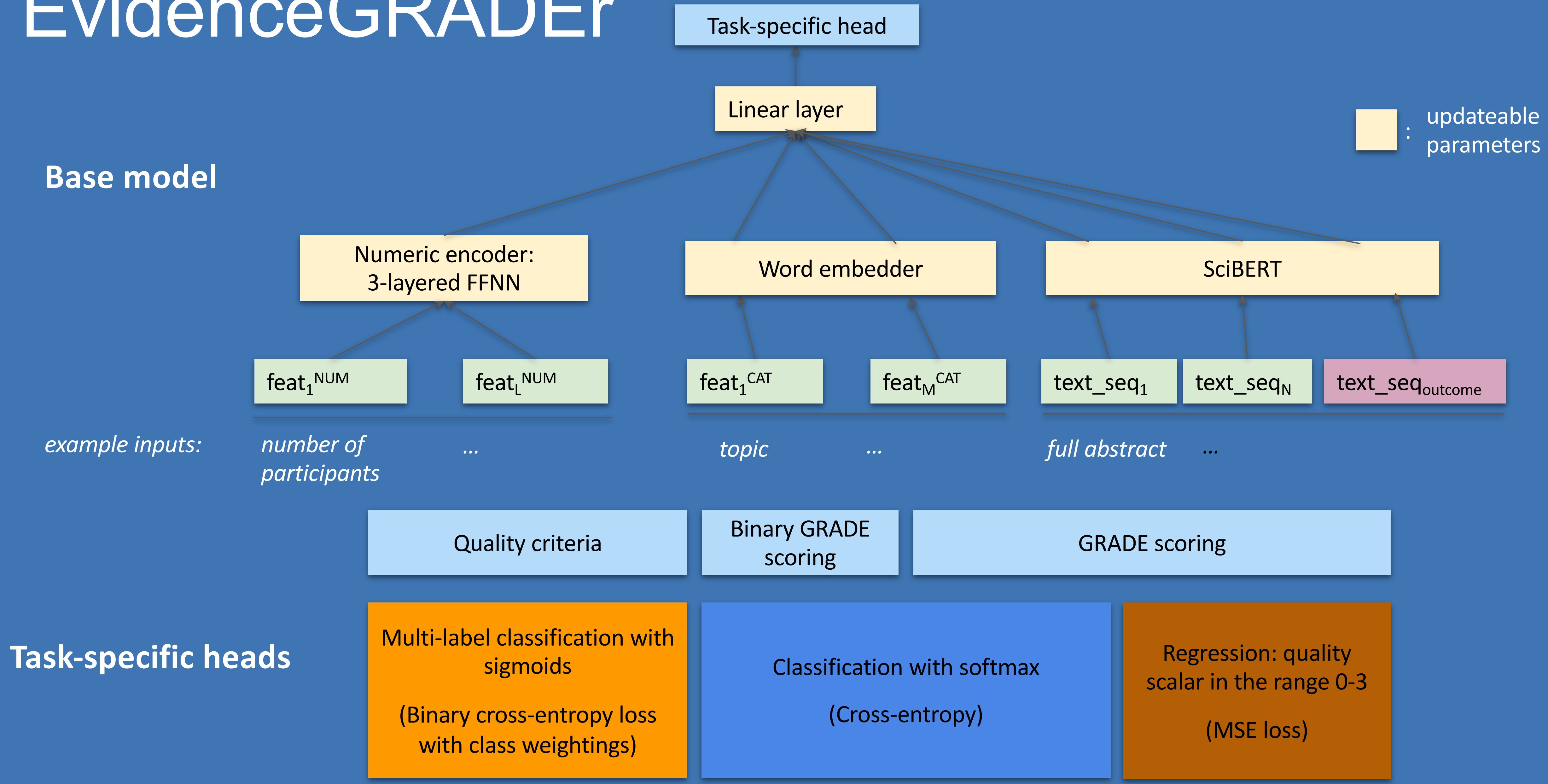
Final grade:

High
Moderate
Low
Very low

Consider other factors affecting recommendation

Make recommendation

EvidenceGRADEr

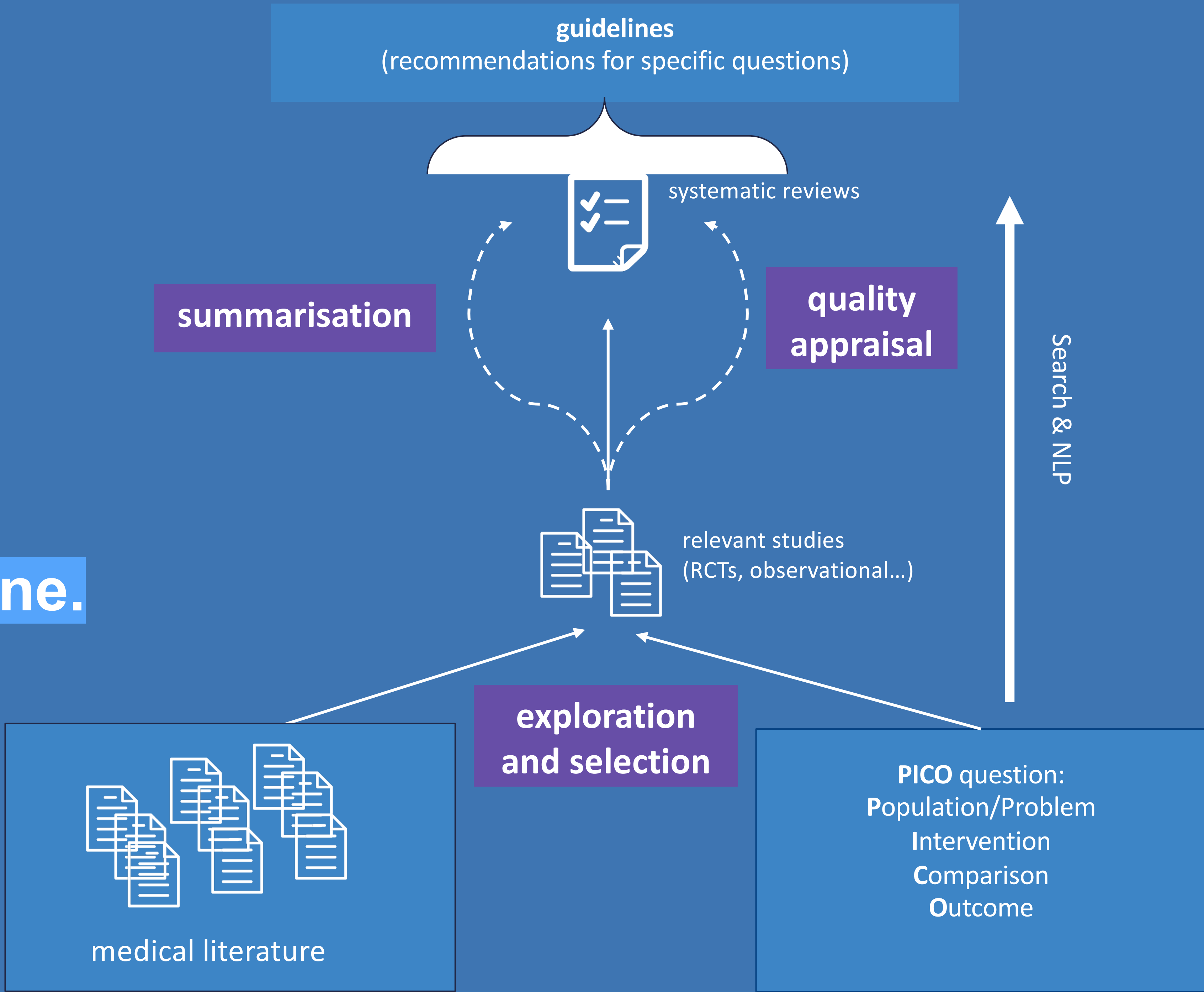


Suster S, Baldwin T, Lau, JH, Jimeno Yepes A, Martinez Iraola D, Otmakhova Y, Verspoor K*. (2023) Automating Quality Assessment of Medical Evidence in Systematic Reviews. Journal of Medical Internet Research 2023.

Systematic Review Automation

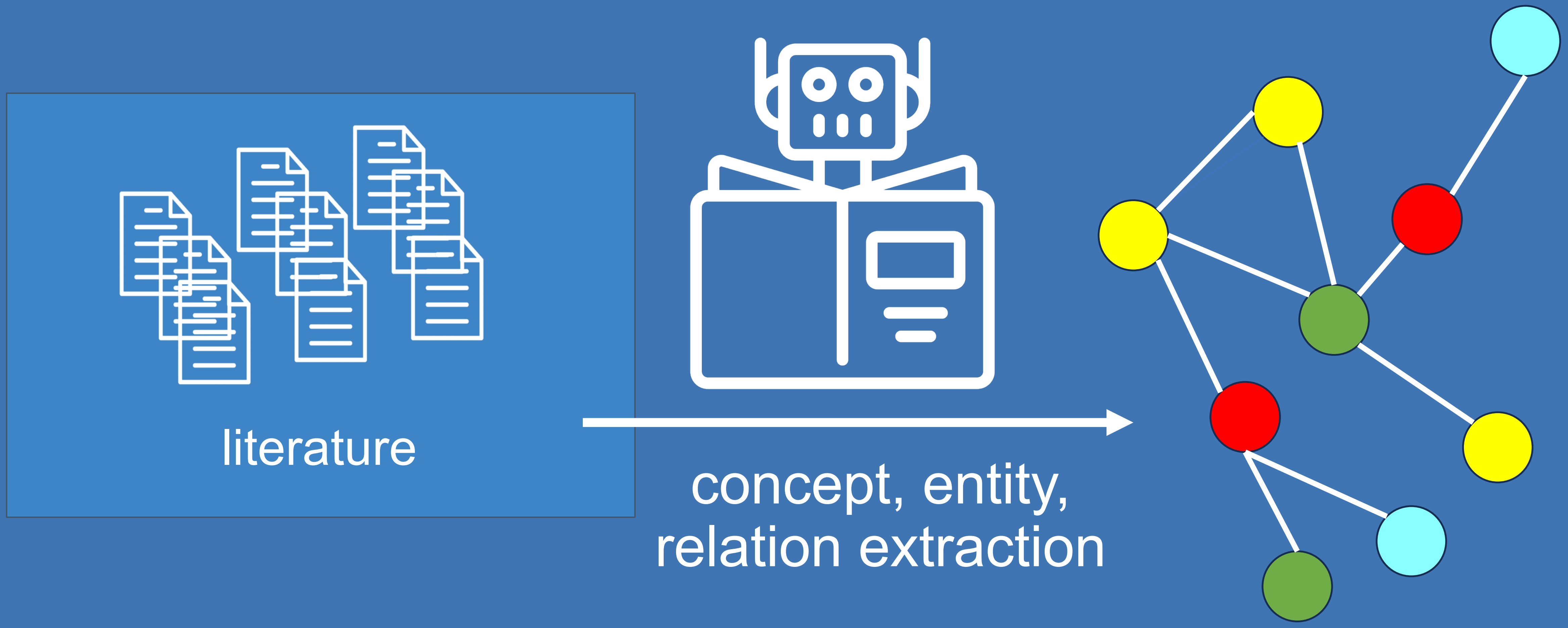
Making progress.

More work to be done.



Evidence discovery

Analysing knowledge graphs

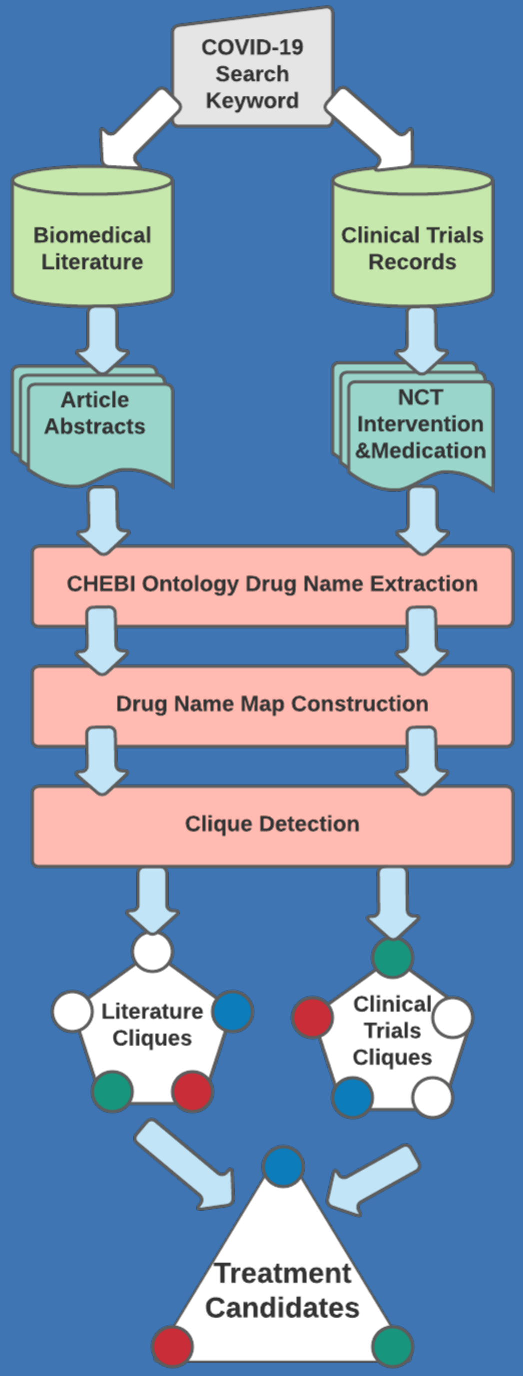


Hypothesis generation from literature

- Information Extraction from Literature + Clinical Trials
- Network construction
  - co-occurrences
  - filtered using Association Analysis
- Network analysis
  - clique detection

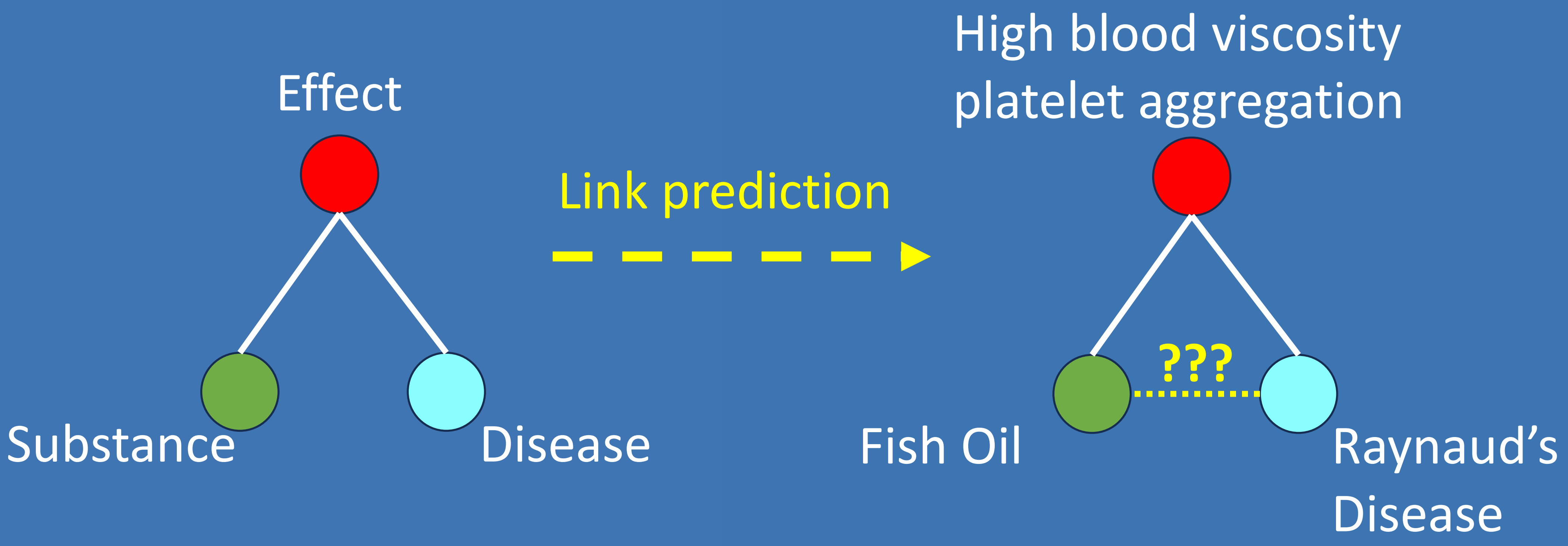
Table 5. Two-drug combinations of COVID-19 treatment candidates identified for further investigation.

Drug 1	Drug 2	Combineability
Estrogen (ChEBI:50114)	Estradiol (ChEBI:23965)	No
Hydroxyethylidene(ChEBI:5801)	Azithromycin (ChEBI:2955)	Possible
Lopinavir (ChEBI:31781)	Ritonavir (ChEBI:45409)	Yes
Ruxolitinib(ChEBI:66919)	Colchicine (ChEBI:23359)	Possible
Hydroxychloroquine (ChEBI:5801)	Favipiravir ChEBI:134722	Possible
Hydroxychloroquine (ChEBI:5801)	Chloroquine ChEBI:3638	No
Azithromycin (ChEBI:2955)	Ivermectin ChEBI:6078	Possible
Hydroxychloroquine (ChEBI:5801)	Lopinavir(ChEBI:31781)	Probably not
Hydroxychloroquine (ChEBI:5801)	Doxycycline(ChEBI:50845)	Possible
Daclatasvir (ChEBI:82977)	Sofosbuvir(ChEBI:85083)	Yes



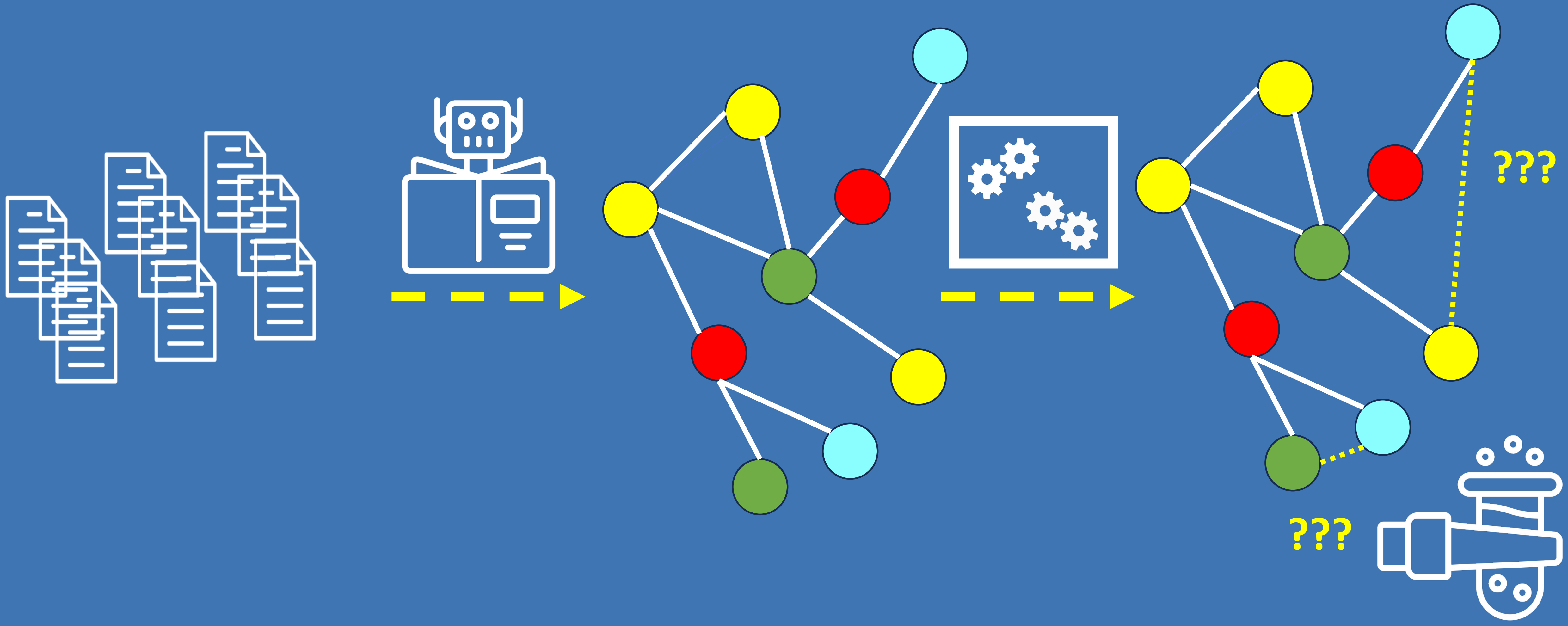
Hamed et al (2022) DOI: https://doi.org/10.3390/pharmaceutics14030567

Literature-based Discovery



→ Hypothesis generation

Literature-based discovery at scale



graph with thousands of nodes,
representing 20 years of research

→ lots of new hypotheses

Conclusions

We need AI to enable learning from the scientific literature, to support evidence detection, exploration, synthesis, and discovery.

AI helps us to find, infer, and utilise knowledge to support ever-improving scientific understanding.

Thank you!

@karinv

Special thanks to:

RMIT University

Vlada Rozova

Estrid He

U. Melbourne

Simon Suster

Yulia Otmakhova

Gracie Pu

Jey Han Lau

Zenan Zhai (Oracle)

Biaoyan Fang (CSIRO)

Trevor Cohn (Google)

Jinghui Liu (CSIRO)

Timothy Baldwin (MBZUAI)

Elsevier

Saber Akhondi

Christian Druckenbrodt

Camilo Thorne