POTSDAM INSTITUTE FOR CLIMATE IMPACT RESEARCH

PIK

Developing, using, and evaluating the next generation of DESTs to produce AI-Powered LES for Climate & Health

The DESTINY project

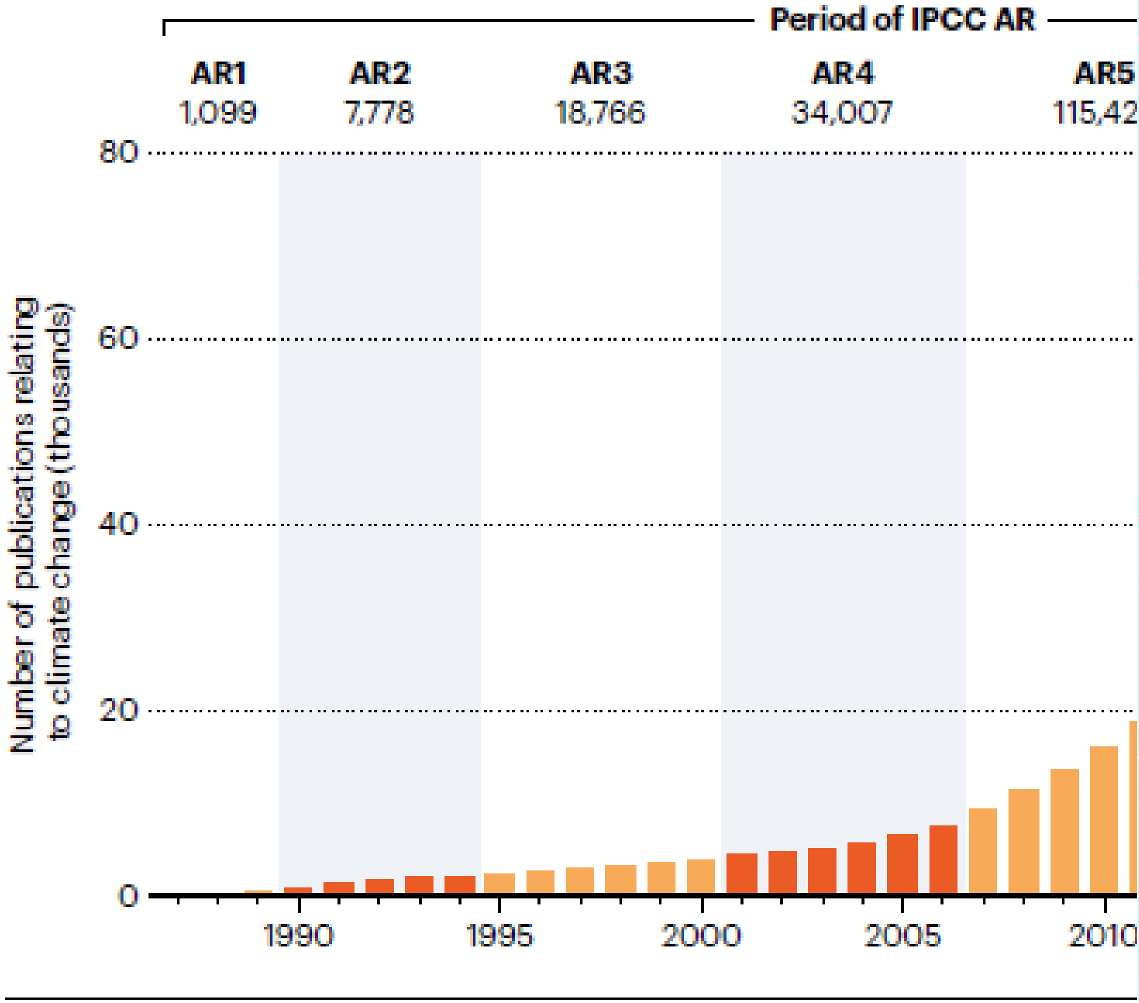
Dr. Maria-Inti Metzendorf

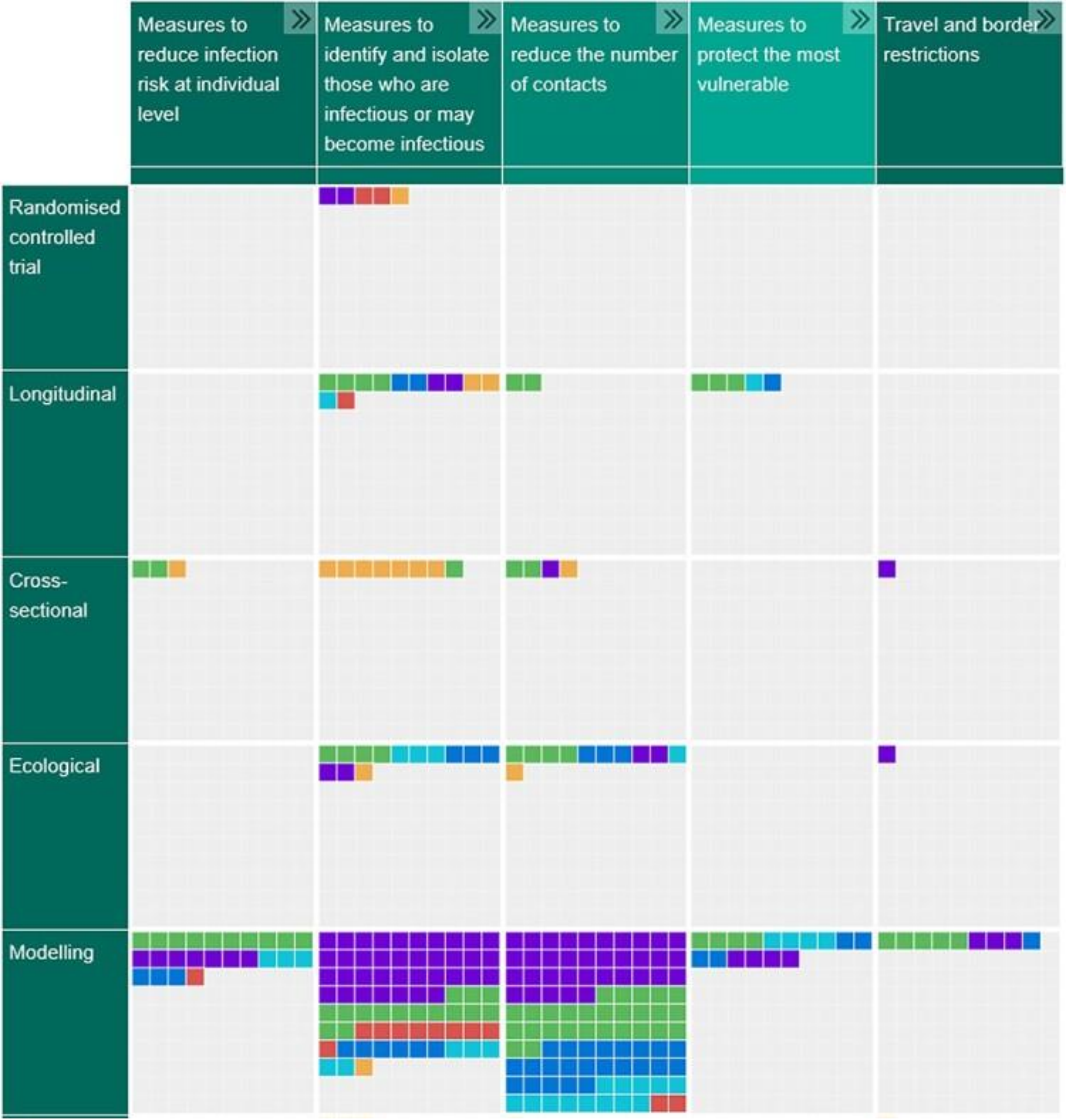
3 March 2026

wellcome

Funded by Wellcome

Leibniz Association


Period of IPCC AR
AR1
1,099
AR2
7,778
AR3
18,766
AR4
34,007
AR5
115,42
80
60
40
20
0
Number of publications relating to climate change (thousands)
1990
1995
2000
2005
2010


Measures to reduce infection risk at individual level
Measures to identify and isolate those who are infectious or may become infectious
Measures to reduce the number of contacts
Measures to protect the most vulnerable
Travel and border restrictions
Randomised controlled trial
Longitudinal
Cross-sectional
Ecological
Modelling

DESTINY

Digital Evidence Synthesis Tool Innovation Yielding Improvements in Climate & Health

DESTINY is co-developing a **new generation of digital evidence synthesis tools (DESTs)**

by showcasing the delivery of rigorous **living evidence** in climate and health

that matters to policymakers & other evidence users.

DESTINY consortium

PIK

FUTURE EVIDENCE FOUNDATION

Cochrane

ACRES THE CENTER FOR RAPID EVIDENCE SYNTHESIS

covidence

ebase Effective Basic Services

EPPI Centre Evidence for Policy & Practice

UCL

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Campbell Collaboration

University of Cape Town • iYunivesithi yaseKapa • Universiteit van Kaapstad •

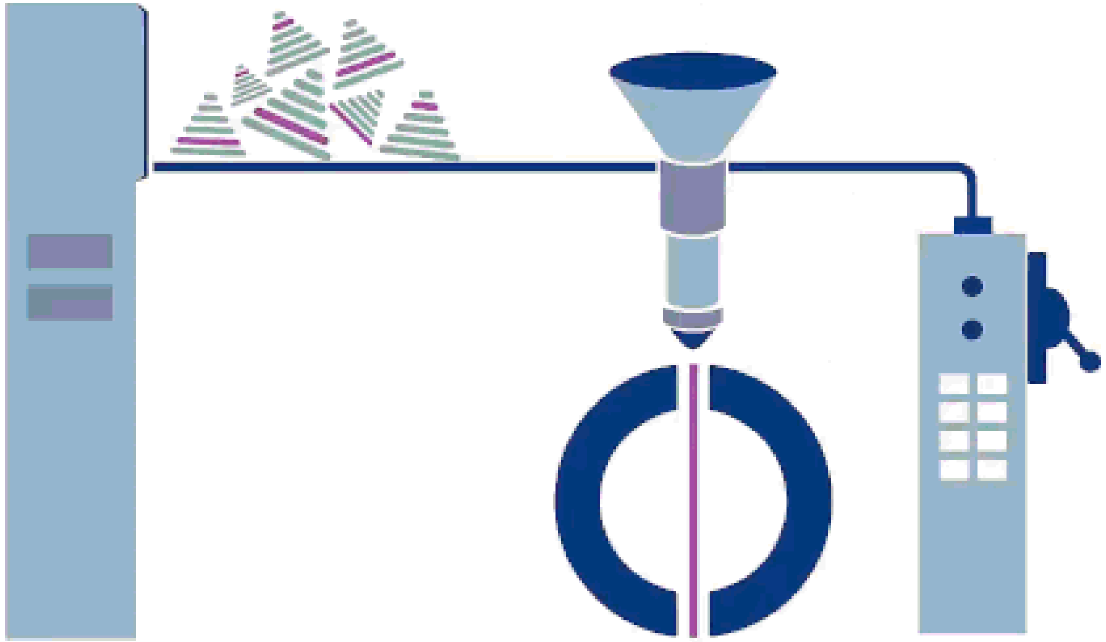
SPES BONA

ASCEND

wellcome Funded by Wellcome

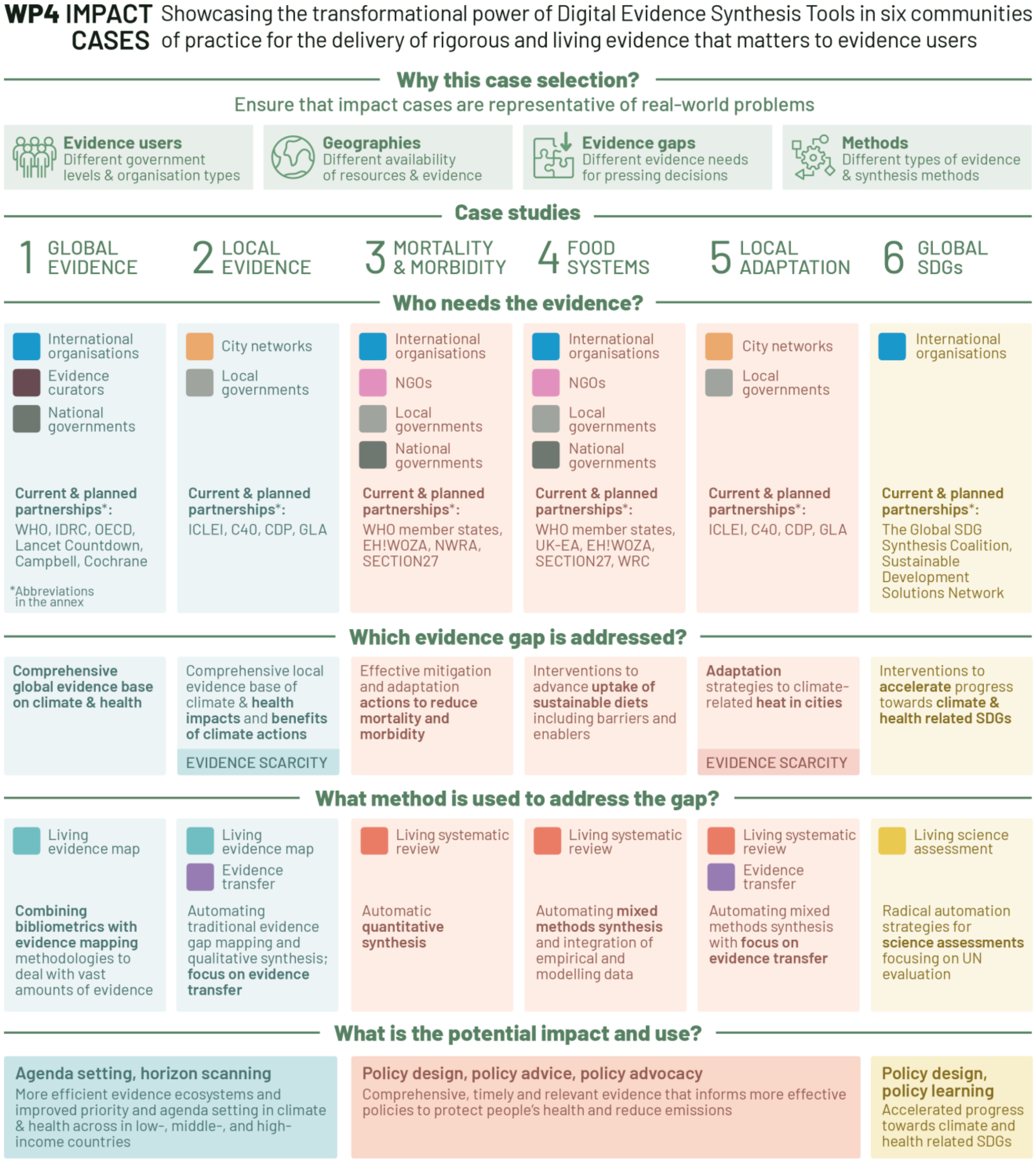
DESTINY

Living evidence for climate & health



- **New DESTs** – explore AI to create faster, cheaper, and more useful evidence synthesis tools (WP2)
- **Responsible use** – ensure safe and responsible DEST applications without compromising standards (WP3)
- **Impact through co-production**
  – work with decision-makers to apply DESTs in key impact cases (WP1, WP4)
- **Mainstreaming DESTs**
  – help users, producers, and funders establish best practices (WP5)

Six DESTINY Impact Cases



WP4 IMPACT CASES Showcasing the transformational power of Digital Evidence Synthesis Tools in six communities of practice for the delivery of rigorous and living evidence that matters to evidence users

Why this case selection?

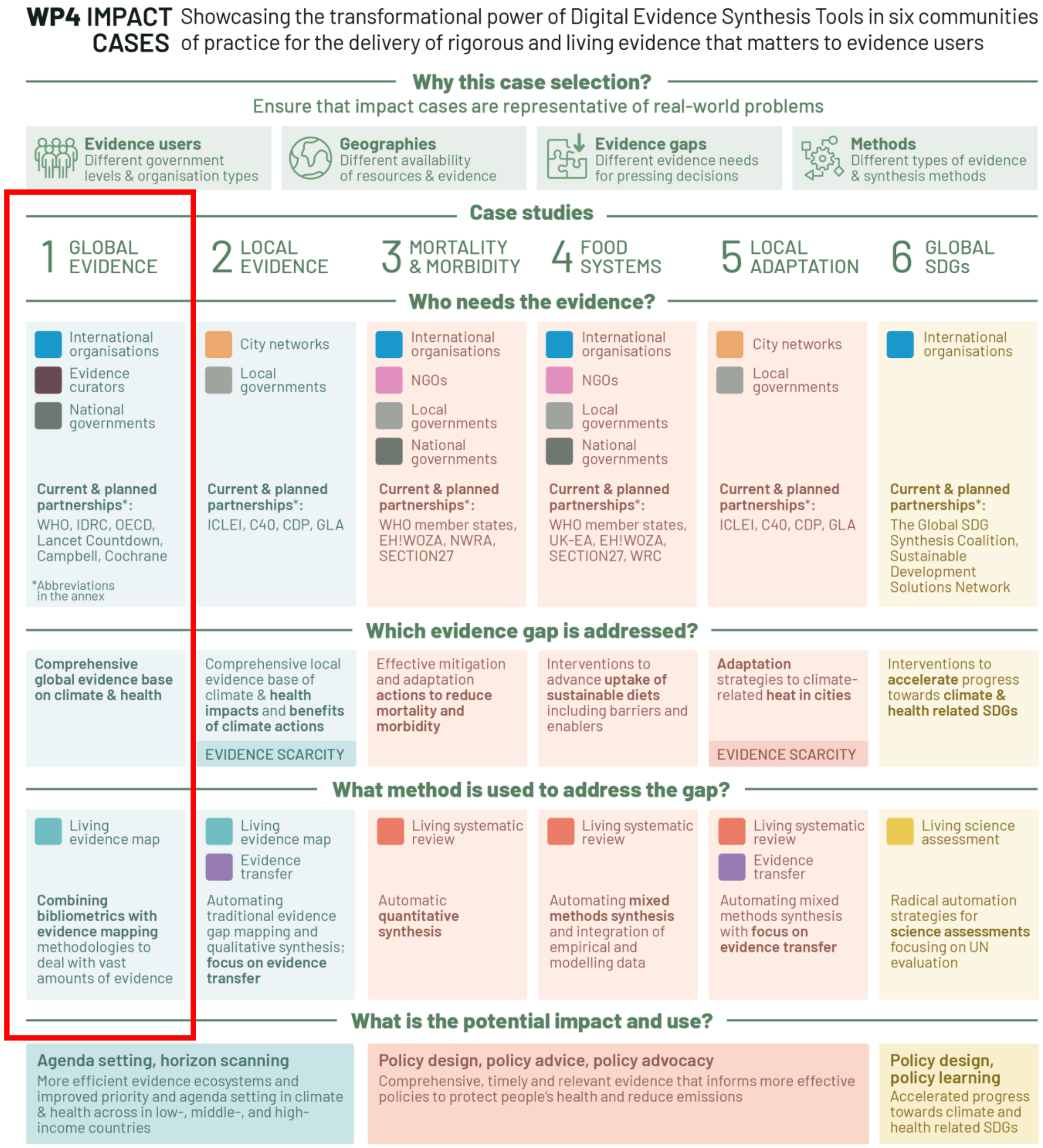
Ensure that impact cases are representative of real-world problems

- **Evidence users** Different government levels & organisation types
- **Geographies** Different availability of resources & evidence
- **Evidence gaps** Different evidence needs for pressing decisions
- **Methods** Different types of evidence & synthesis methods

Case studies

	1 GLOBAL EVIDENCE	2 LOCAL EVIDENCE	3 MORTALITY & MORBIDITY	4 FOOD SYSTEMS	5 LOCAL ADAPTATION	6 GLOBAL SDGs
Who needs the evidence?	■ International organisations ■ Evidence curators ■ National governments	■ City networks ■ Local governments	■ International organisations ■ NGOs ■ Local governments ■ National governments	■ International organisations ■ NGOs ■ Local governments ■ National governments	■ City networks ■ Local governments	■ International organisations
	Current & planned partnerships*: WHO, IDRC, OECD, Lancet Countdown, Campbell, Cochrane *Abbreviations in the annex	**Current & planned partnerships*:** ICLEI, C40, CDP, GLA	**Current & planned partnerships*:** WHO member states, EH!WOZA, NWRA, SECTION27	**Current & planned partnerships*:** WHO member states, UK-EA, EH!WOZA, SECTION27, WRC	**Current & planned partnerships*:** ICLEI, C40, CDP, GLA	**Current & planned partnerships*:** The Global SDG Synthesis Coalition, Sustainable Development Solutions Network
Which evidence gap is addressed?	**Comprehensive global evidence base on climate & health**	Comprehensive local evidence base of climate & **health impacts** and **benefits of climate actions** EVIDENCE SCARCITY	Effective mitigation and adaptation **actions to reduce mortality and morbidity**	Interventions to advance **uptake of sustainable diets** including barriers and enablers	**Adaptation** strategies to climate-related **heat in cities** EVIDENCE SCARCITY	Interventions to **accelerate** progress towards **climate & health related SDGs**
What method is used to address the gap?	■ Living evidence map **Combining bibliometrics with evidence mapping** methodologies to deal with vast amounts of evidence	■ Living evidence map ■ Evidence transfer Automating traditional evidence gap mapping and qualitative synthesis; **focus on evidence transfer**	■ Living systematic review Automatic **quantitative synthesis**	■ Living systematic review Automating **mixed methods synthesis** and integration of empirical and modelling data	■ Living systematic review ■ Evidence transfer Automating mixed methods synthesis with **focus on evidence transfer**	■ Living science assessment Radical automation strategies for **science assessments** focusing on UN evaluation
What is the potential impact and use?	**Agenda setting, horizon scanning** More efficient evidence ecosystems and improved priority and agenda setting in climate & health across in low-, middle-, and high-income countries		**Policy design, policy advice, policy advocacy** Comprehensive, timely and relevant evidence that informs more effective policies to protect people's health and reduce emissions			**Policy design, policy learning** Accelerated progress towards climate and health related SDGs

Living evidence map of the global climate and health literature



WP4 IMPACT CASES Showcasing the transformational power of Digital Evidence Synthesis Tools in six communities of practice for the delivery of rigorous and living evidence that matters to evidence users

Why this case selection?

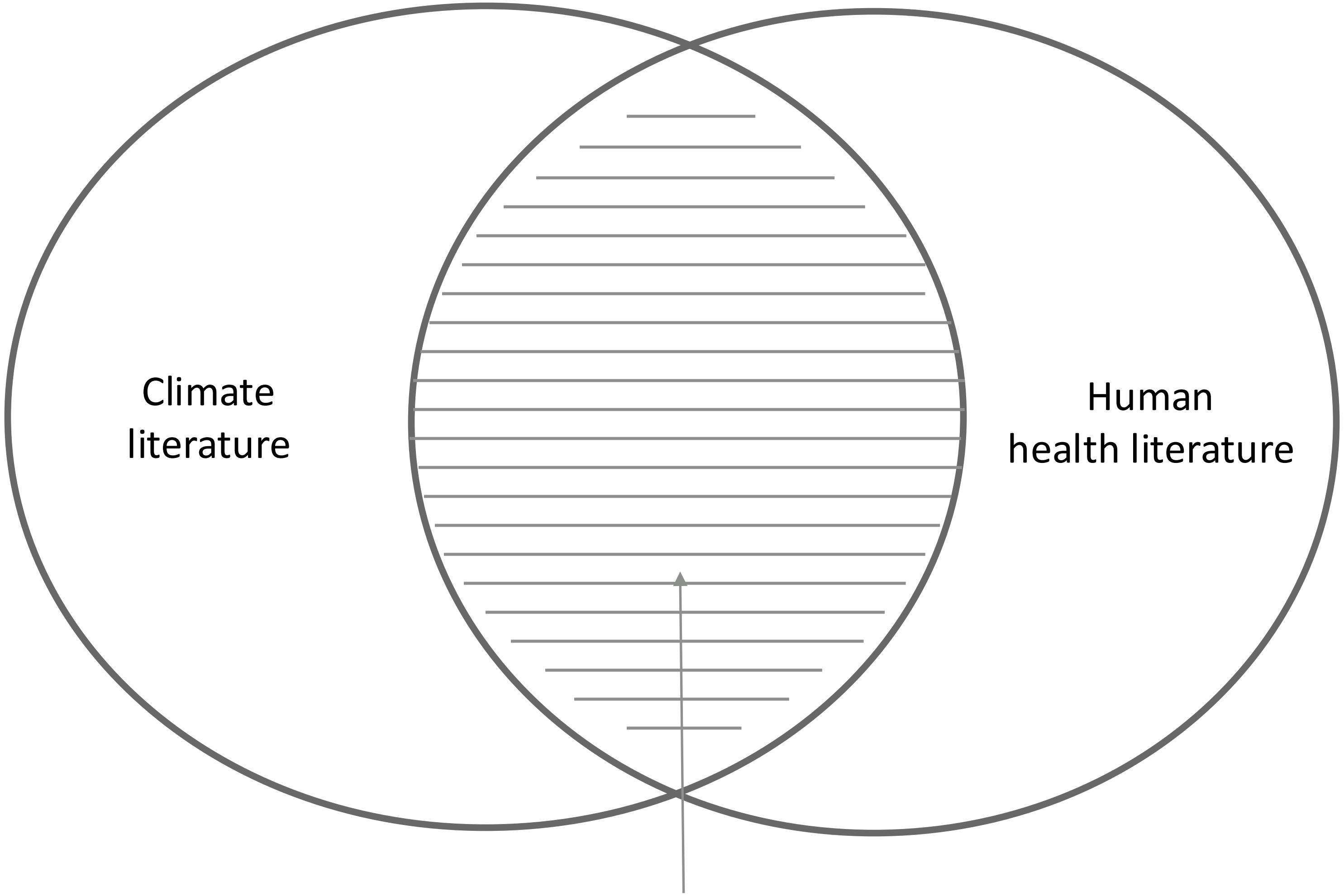
Ensure that impact cases are representative of real-world problems

- **Evidence users** Different government levels & organisation types
- **Geographies** Different availability of resources & evidence
- **Evidence gaps** Different evidence needs for pressing decisions
- **Methods** Different types of evidence & synthesis methods

Case studies

	1 GLOBAL EVIDENCE	2 LOCAL EVIDENCE	3 MORTALITY & MORBIDITY	4 FOOD SYSTEMS	5 LOCAL ADAPTATION	6 GLOBAL SDGs
Who needs the evidence?	International organisations Evidence curators National governments **Current & planned partnerships***: WHO, IDRC, OECD, Lancet Countdown, Campbell, Cochrane *Abbreviations in the annex	City networks Local governments **Current & planned partnerships***: ICLEI, C40, CDP, GLA	International organisations NGOs Local governments National governments **Current & planned partnerships***: WHO member states, EH!WOZA, NWRA, SECTION27	International organisations NGOs Local governments National governments **Current & planned partnerships***: WHO member states, UK-EA, EH!WOZA, SECTION27, WRC	City networks Local governments **Current & planned partnerships***: ICLEI, C40, CDP, GLA	International organisations **Current & planned partnerships***: The Global SDG Synthesis Coalition, Sustainable Development Solutions Network
Which evidence gap is addressed?	**Comprehensive global evidence base on climate & health**	Comprehensive local evidence base of climate & **health impacts** and **benefits of climate actions** EVIDENCE SCARCITY	Effective mitigation and adaptation **actions to reduce mortality and morbidity**	Interventions to advance **uptake of sustainable diets** including barriers and enablers	**Adaptation** strategies to climate-related **heat in cities** EVIDENCE SCARCITY	Interventions to **accelerate** progress towards **climate & health related SDGs**
What method is used to address the gap?	Living evidence map **Combining bibliometrics with evidence mapping** methodologies to deal with vast amounts of evidence	Living evidence map Evidence transfer Automating traditional evidence gap mapping and qualitative synthesis; **focus on evidence transfer**	Living systematic review Automatic **quantitative synthesis**	Living systematic review Automating **mixed methods synthesis** and integration of empirical and modelling data	Living systematic review Evidence transfer Automating mixed methods synthesis with **focus on evidence transfer**	Living science assessment Radical automation strategies for **science assessments** focusing on UN evaluation
What is the potential impact and use?	**Agenda setting, horizon scanning** More efficient evidence ecosystems and improved priority and agenda setting in climate & health across in low-, middle-, and high-income countries		**Policy design, policy advice, policy advocacy** Comprehensive, timely and relevant evidence that informs more effective policies to protect people's health and reduce emissions			**Policy design, policy learning** Accelerated progress towards climate and health related SDGs

The scope in simple terms

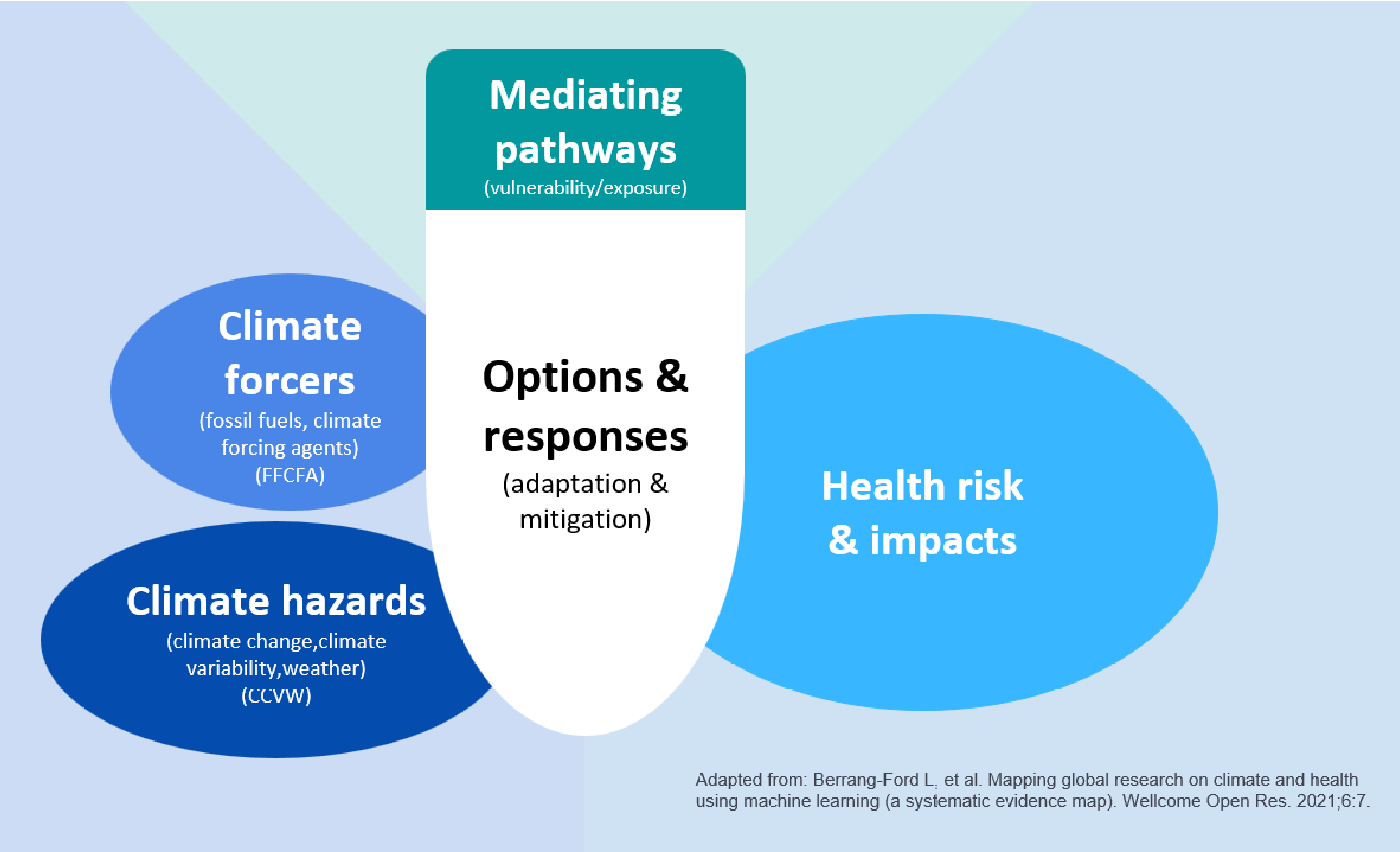


Climate literature

Human health literature

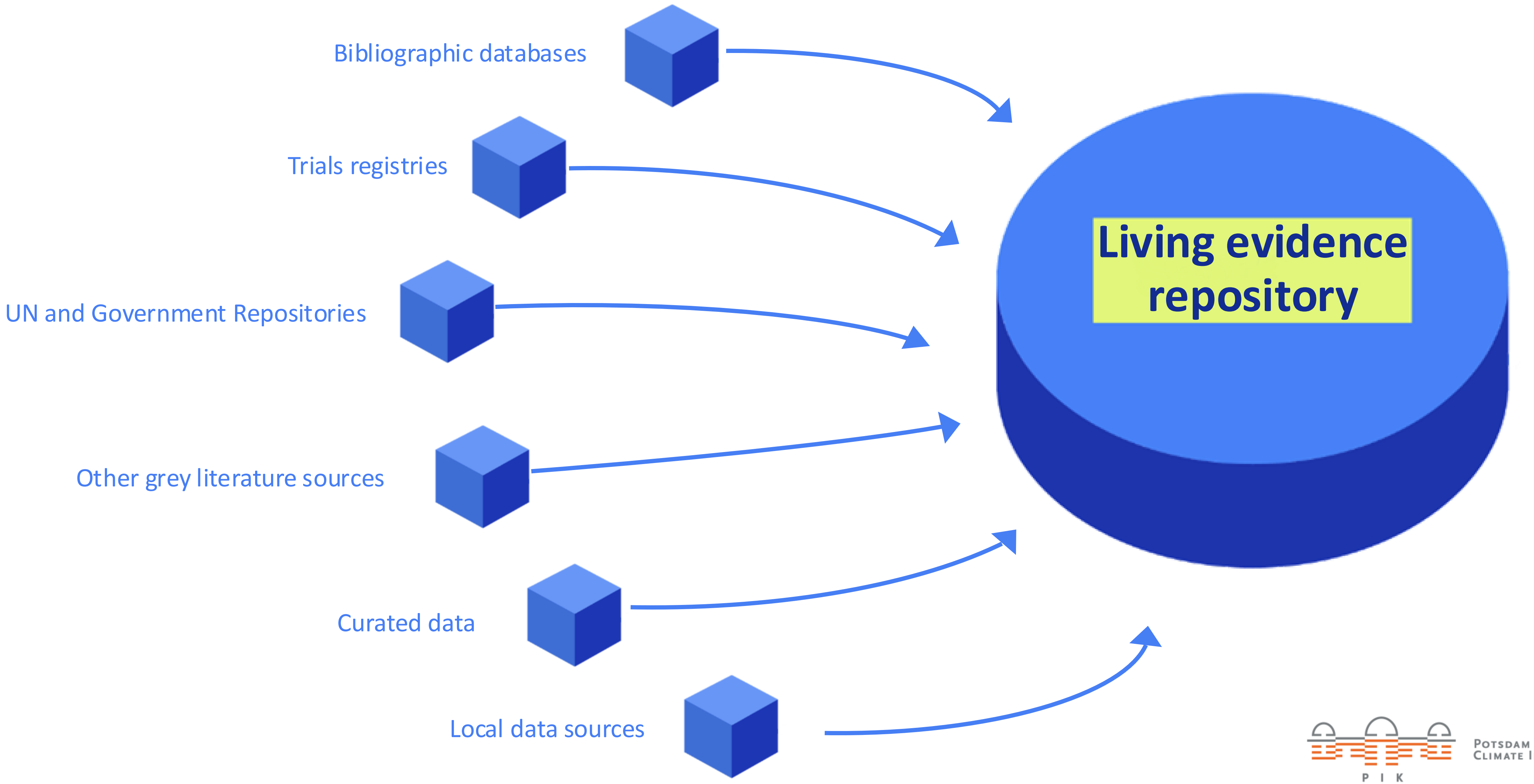
CLIMATE & HEALTH
(~ 500,000 records)

Key concepts



Adapted from: Berrang-Ford L, et al. Mapping global research on climate and health using machine learning (a systematic evidence map). Wellcome Open Res. 2021;6:7.

Evidence sources/types of studies



Bibliographic databases

Trials registries

UN and Government Repositories

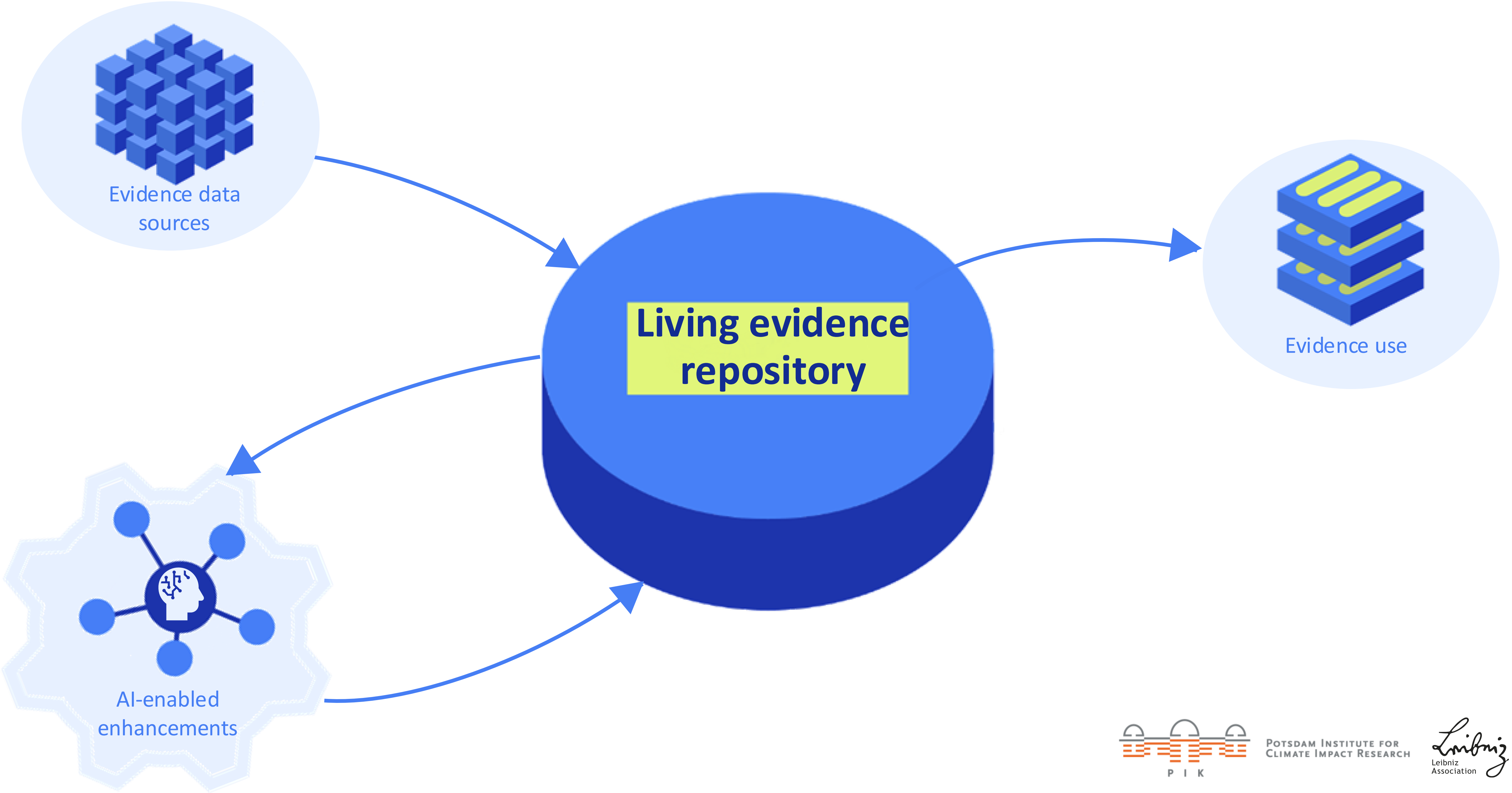
Other grey literature sources

Curated data

Local data sources

Living evidence repository

A living data repository for CC & Health



Potsdam Institute for Climate Impact Research

PIK

Leibniz Association

Taxonomy

Population (sub-categories: vulnerability, gender/sex, age)

Topic (sub-categories: mitigation, adaptation, impact);

Climate factors (sub-categories: climate drivers, extreme weather events, climate-forcing agents, fossil fuels)

Interventions/measures/responses (sub-categories: policy, technology/infrastructure, behavior/culture, ecosystem-based, institutional);

Sectors (sub-categories: agriculture, fishing and forestry; buildings and housing; education; energy and extractives; finance; healthcare; social protection; industry, trade and services; information and communications technologies; public administration; transportation; water, sanitation and waste management);

Governance scale (sub-categories: global, international, national, sub-national, household and individual);

Geographic location of affected area, and derive location characteristics, e.g. income category (sub-categories: low, lower middle, upper middle, high);

Geographic feature (sub-categories: urban, rural, desert, forest, freshwater, grassland, island, mountain, ocean/coastal, polar, rainforest, valley, temperate, tropical, wetland);

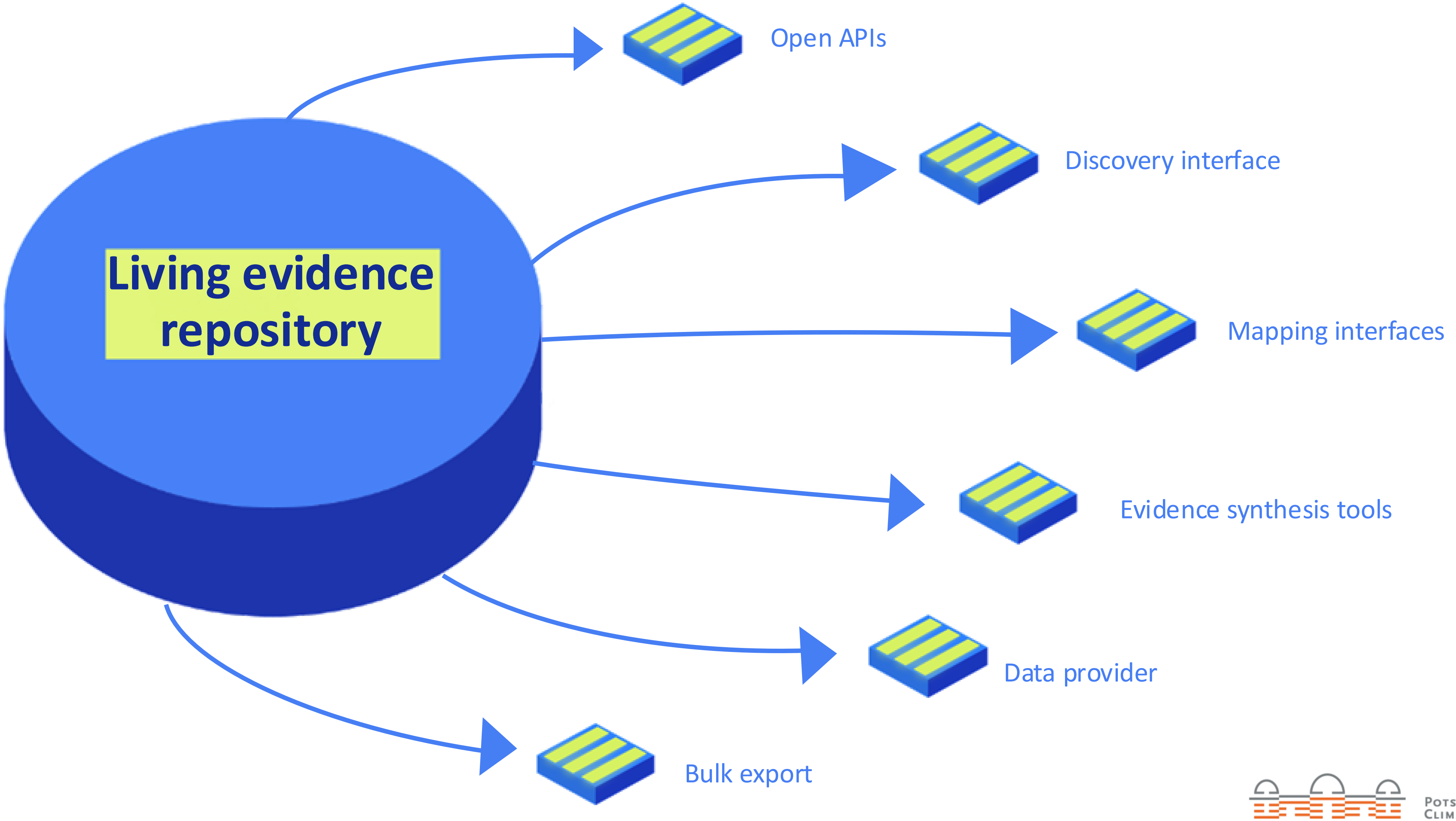
Exposure (sub-categories: heat stress, air quality, food supply and safety, water quality and quantity, extreme weather events, vector distribution and ecology, social factors (forced displacement, violence, conflict));

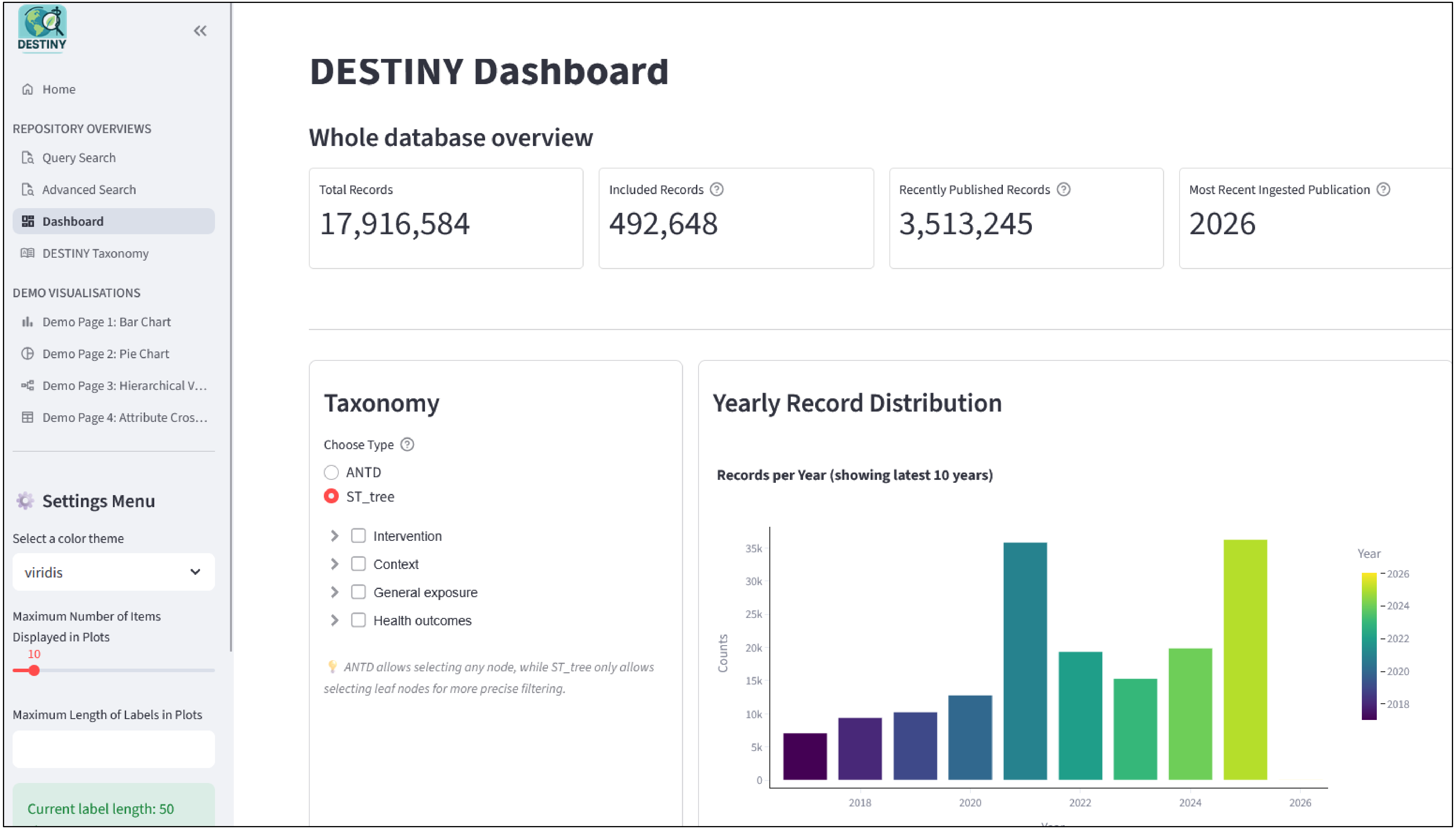
Actors (sub-categories: international organizations, government, private sector, civil society, households and individuals);

Methods (sub-categories: primary research, evidence synthesis);

Outcomes (sub-categories: death/mortality, outpatient visits, hospitalization, and the first three levels of the ICD-11).

Evidence products / outputs





DESTINY

Home

REPOSITORY OVERVIEWS

Query Search

Advanced Search

Dashboard

DESTINY Taxonomy

DEMO VISUALISATIONS

Demo Page 1: Bar Chart

Demo Page 2: Pie Chart

Demo Page 3: Hierarchical V...

Demo Page 4: Attribute Cros...

Settings Menu

Select a color theme

viridis

Maximum Number of Items Displayed in Plots

10

Maximum Length of Labels in Plots

Current label length: 50

DESTINY Dashboard

Whole database overview

Total Records	Included Records ⓘ	Recently Published Records ⓘ	Most Recent Ingested Publication ⓘ
17,916,584	492,648	3,513,245	2026

Taxonomy

Choose Type ⓘ

○ ANTD

● ST_tree

› □ Intervention

› □ Context

› □ General exposure

› □ Health outcomes

ANTD allows selecting any node, while ST_tree only allows selecting leaf nodes for more precise filtering.

Yearly Record Distribution

Records per Year (showing latest 10 years)

Thank you and over to Max…

maria-inti.metzendorf@pik-potsdam.de

Tools and Evaluation

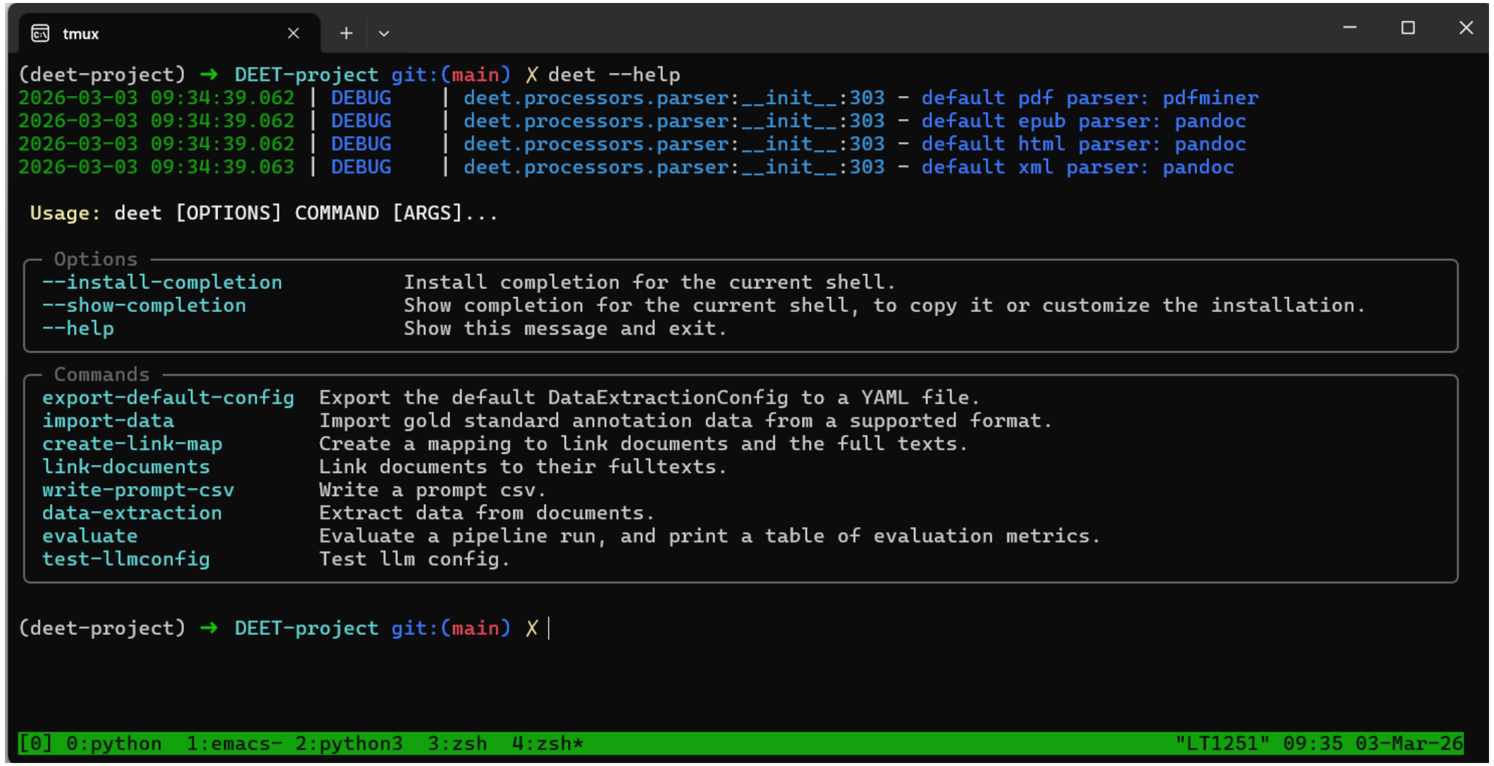
Enabling systematic reviewers and evidence users to conduct and access trustworthy and responsible AI-assisted evidence synthesis

DEET

DEET

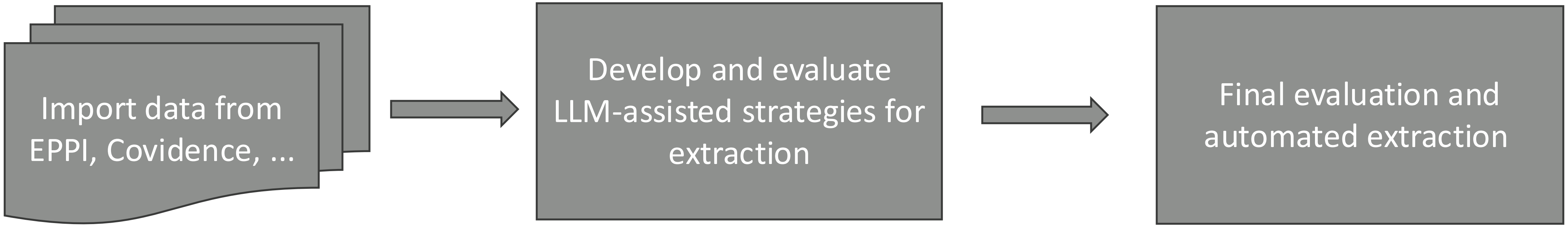
Data Extraction Evaluation Toolkit

- There are **many** platforms which allow you to use LLMs to screen, categorise, or extract data from studies
- There are **none** that are **open source**, that give you complete **control of models and prompts**, and that encourage **good evaluation practice**



```
(deet-project) ➜ DEET-project git:(main) ✗ deet --help
2026-03-03 09:34:39.062 | DEBUG    | deet.processors.parser:__init__:303 - default pdf parser: pdfminer
2026-03-03 09:34:39.062 | DEBUG    | deet.processors.parser:__init__:303 - default epub parser: pandoc
2026-03-03 09:34:39.062 | DEBUG    | deet.processors.parser:__init__:303 - default html parser: pandoc
2026-03-03 09:34:39.063 | DEBUG    | deet.processors.parser:__init__:303 - default xml parser: pandoc

 Usage: deet [OPTIONS] COMMAND [ARGS]...

╭─ Options ────────────────────────────────────────────────────────────────────────────────────────────╮
│ --install-completion          Install completion for the current shell.                              │
│ --show-completion             Show completion for the current shell, to copy it or customize the installation. │
│ --help                        Show this message and exit.                                            │
╰──────────────────────────────────────────────────────────────────────────────────────────────────────╯
╭─ Commands ───────────────────────────────────────────────────────────────────────────────────────────╮
│ export-default-config  Export the default DataExtractionConfig to a YAML file.                      │
│ import-data            Import gold standard annotation data from a supported format.                │
│ create-link-map        Create a mapping to link documents and the full texts.                       │
│ link-documents         Link documents to their fulltexts.                                           │
│ write-prompt-csv       Write a prompt csv.                                                          │
│ data-extraction        Extract data from documents.                                                 │
│ evaluate               Evaluate a pipeline run, and print a table of evaluation metrics.            │
│ test-llmconfig         Test llm config.                                                             │
╰──────────────────────────────────────────────────────────────────────────────────────────────────────╯

(deet-project) ➜ DEET-project git:(main) ✗ |



[0] 0:python  1:emacs- 2:python3  3:zsh  4:zsh*                                  "LT1251" 09:35 03-Mar-26
```

DEET Workflow

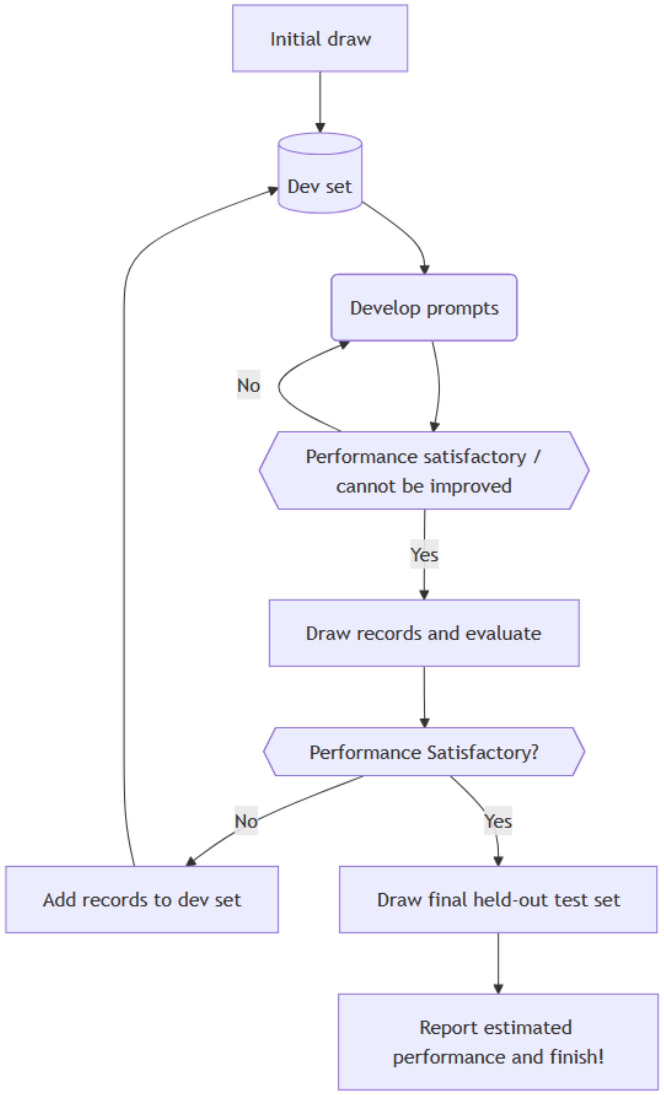


Import data from EPPI, Covidence, ...

Develop and evaluate LLM-assisted strategies for extraction

Final evaluation and automated extraction

Chunked evaluation workflow



Initial draw
Dev set
Develop prompts
No
Performance satisfactory / cannot be improved
Yes
Draw records and evaluate
Performance Satisfactory?
No
Yes
Add records to dev set
Draw final held-out test set
Report estimated performance and finish!

- We formalise an evaluation protocol that
  - strictly separates development/training from testing
  - gives a realistic estimation of available work savings

Source: https://destiny-evidence.github.io/evaluation-book/index-1/#chunked-evaluation-data, adjusted from Homiar, A., et al. (2025). Development and evaluation of prompts for a large language model to screen titles and abstracts in a living systematic review. *BMJ Mental Health, 28*(1), e301762. 10.1136/bmjment-2025-301762

Future of DEET

- Initial release in coming weeks
  - **Can be used via CLI, or as a python package**
- **Self-hosted UI**
- **DESTINY-hosted UI**

Open operational questions for evaluation

- How many documents do I need in my development sets?
- How do I deal with the fact that "gold-standard" annotations are not solid gold?
- How can I communicate the implications of any evaluation I do on the answer to my review's research question?

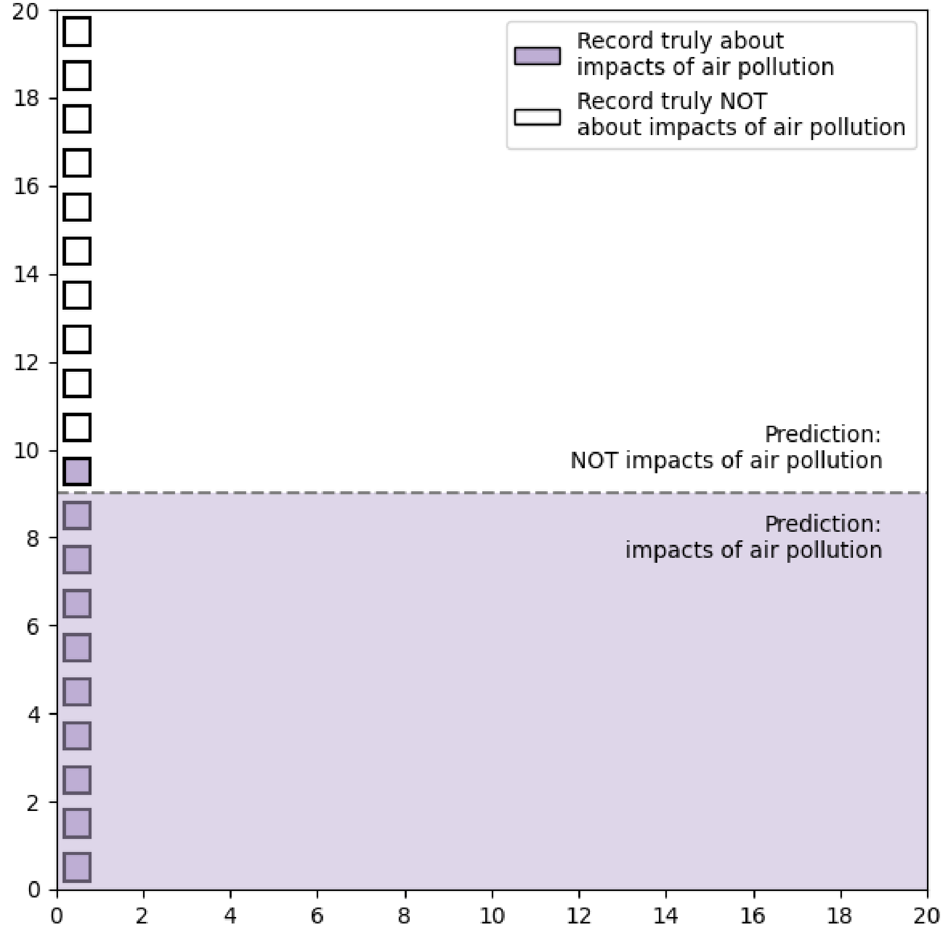
A wider research agenda on the evaluation for AI-assisted systematic reviews

- **Dealing with unreliable human annotations**
- Can we correct disagreements between human-annotated and AI-annotated data without biasing our evaluations of the AI? Can we *not* correct disagreements without biasing our evaluations? How should we approach this?
- Under what conditions do more human annotators improve the fidelity of gold-standard data? Does accuracy always approach 1 with more annotators?
- Can we infer the number of annotators we would need to achieve a given level of fidelity to ground truth data from measures of inter-rater reliability?
- Does a ground-truth exist? How do we deal with this epistemologically? What would superhuman performance mean?
- **Managing uncertainty**
- Are binomial confidence intervals for recall and precisions good baselines for estimating confidence intervals around performance metrics?
- Can confidence intervals be narrowed?
    - By using Bayesian statistics and priors based on performance on similar tasks?
    - By estimating jointly across predicted categories / tasks?
    - Automation and the results of systematic reviews
- How do errors and uncertainties around errors compound across tasks, when automation is used for multiple stages of a systematic review?
- How do errors affect the results of systematic reviews?
- How can we incorporate uncertainty around the accuracy of specific tasks (screening, data extraction, criticial appraisal) into the overall uncertainty in our results?
- Optimal stopping criteria / stopping criteria for living reviews / LLMs for screening / alternatives to prioritised screening
- **Optimising the distribution of labour between humans and machines**
- How can we quantify the costs and benefits of conducting evidence synthesis tasks by hand and using different AI approaches?
- How can we manage trade-offs and allocate resources efficiently?

https://destiny-evidence.github.io/evaluation-book/research-questions/

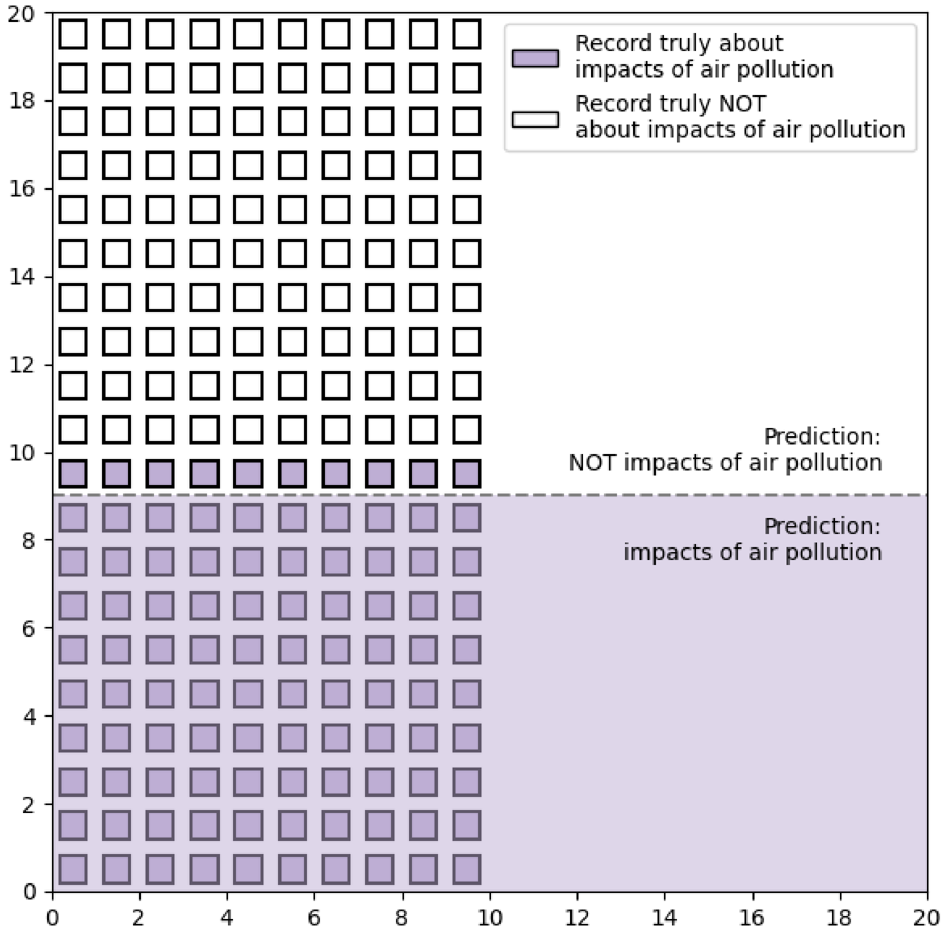
Uncertainty-aware evaluation

Uncertainty in evaluation metrics is often unaccounted for



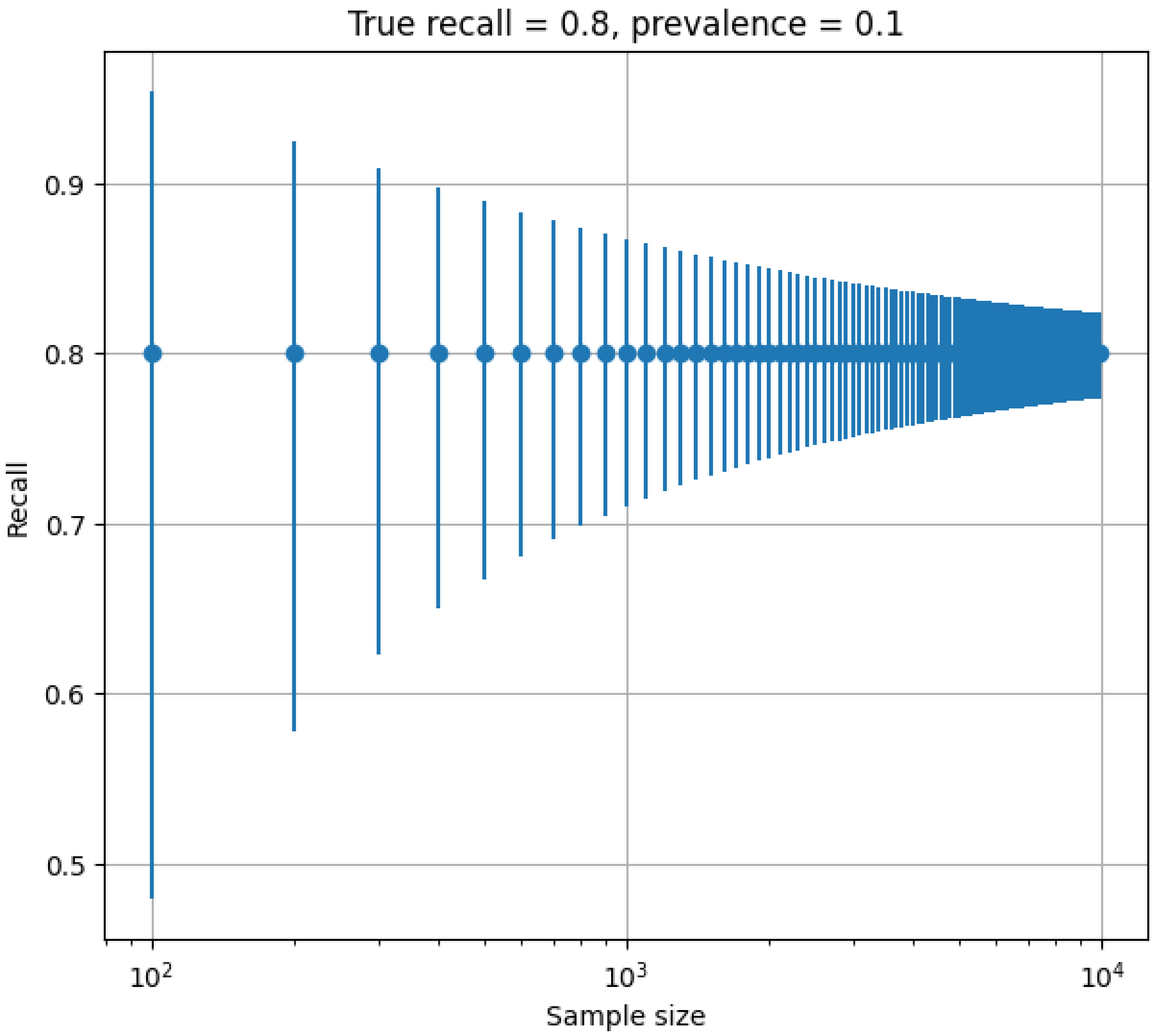
- **Given this evaluation data, we estimate our recall to be 90%**

Uncertainty in evaluation metrics is often unaccounted for



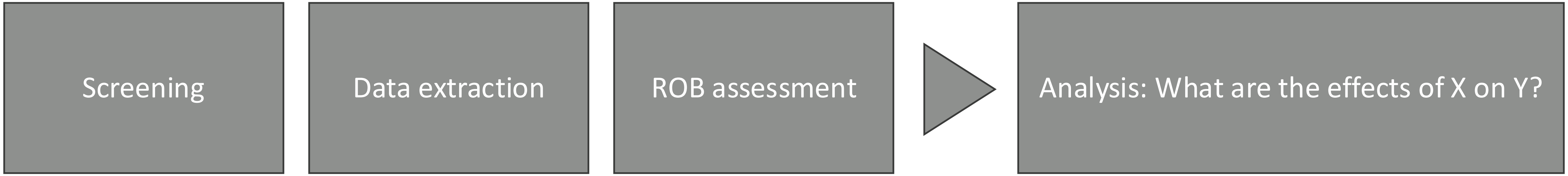
- **Given this evaluation data, we estimate our recall to be 90%**
- **With more evaluation data, we still estimate 90%, but our confidence interval should shrink**

Binomial confidence intervals for evaluation metrics



- We can use binomial confidence intervals (as in diagnostic test accuracy) to estimate proportions
- These are large!
- Can they be narrowed with
  - Bayesian statistics
  - Bootstrapping
  - Joint estimation across classes

When we use AI across the systematic review process, how do errors and uncertainties compound?



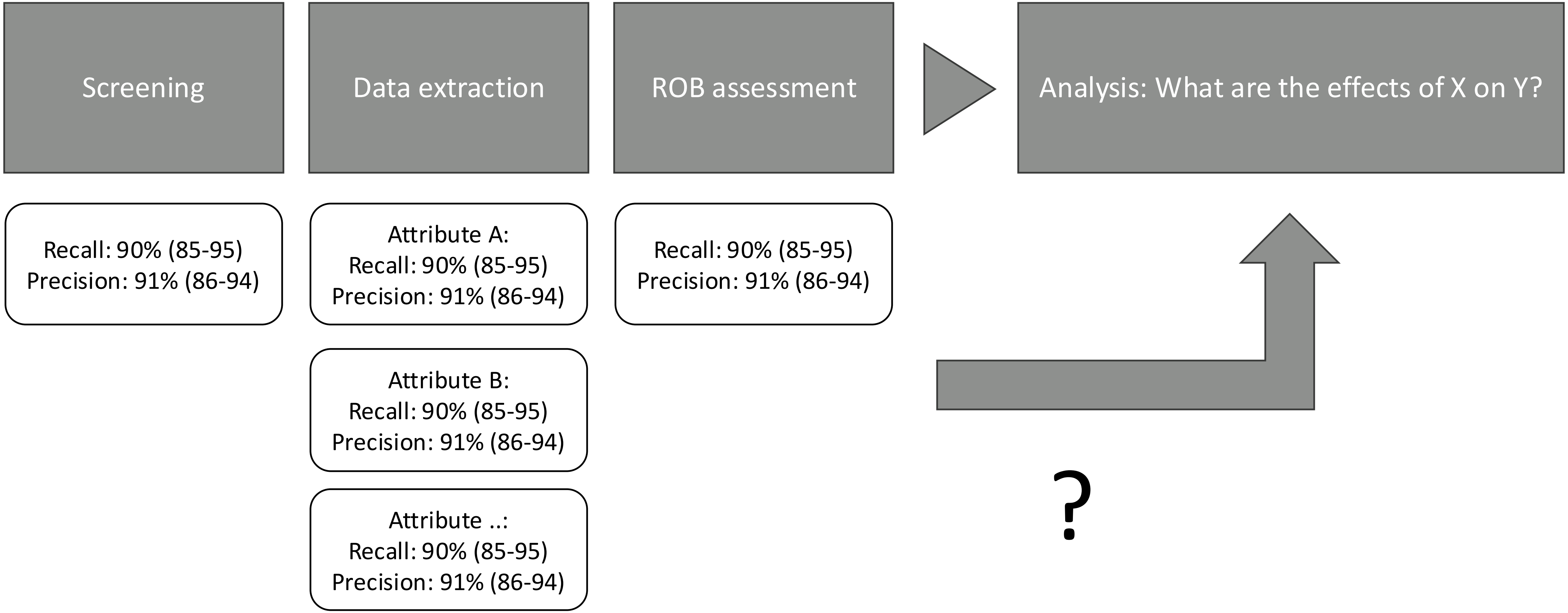
Screening

Data extraction

ROB assessment

Analysis: What are the effects of X on Y?

When we use AI across the systematic review process, how do errors and uncertainties compound?



Screening

Data extraction

ROB assessment

Analysis: What are the effects of X on Y?

Recall: 90% (85-95)
Precision: 91% (86-94)

Attribute A:
Recall: 90% (85-95)
Precision: 91% (86-94)

Attribute B:
Recall: 90% (85-95)
Precision: 91% (86-94)

Attribute ..:
Recall: 90% (85-95)
Precision: 91% (86-94)

Recall: 90% (85-95)
Precision: 91% (86-94)

?

Meaningful communication to evidence users

Ultimately we do not care if we are 95% confident we achieve 95% recall in screening, 97% recall on average across attributes in data extraction, …

We want to know how potential errors introduced through automation (OR through human error) affect our findings.

How can we propagate these errors into the confidence interval of our findings?

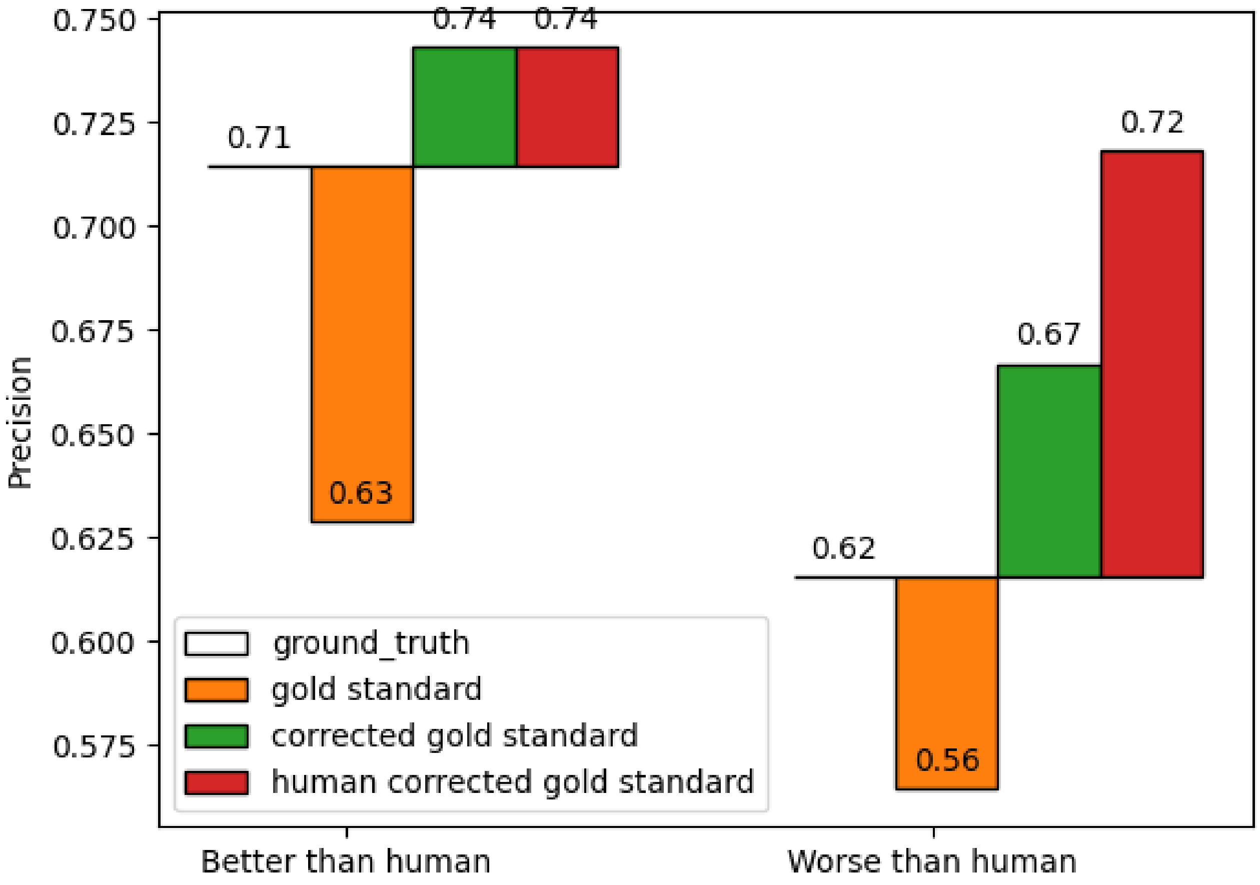
Dealing with imperfect human annotations

What happens if we "correct" disagreements between human annotations and AI predictions?

When human and AI annotations disagree, the evaluation metrics for our AI pipeline go down.

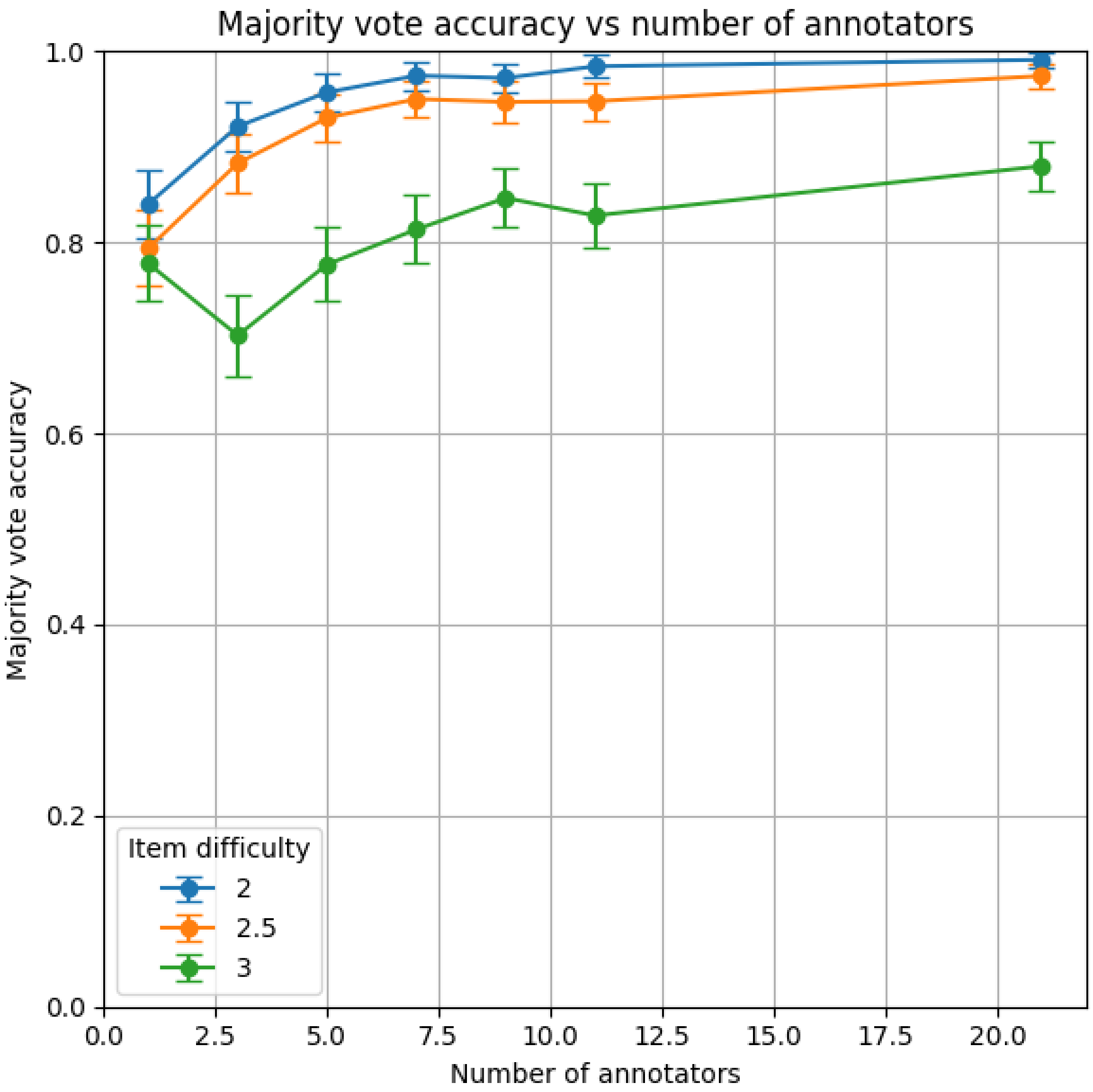
But are these true disagreements? Or were they coding errors?

What happens if we "correct" disagreements between human annotations and AI predictions?



- **Not correcting errors means we underestimate AI performance**
- **Correcting them means we over-estimate performance (especially if AI is working less well than average human)**
    - **This is because we do not catch those cases where the AI and human erred**

How many human annotations do we need?



- **Two human annotations, with disagreements resolved by a 3rd is a convenient heuristic**
- **AI gives us the chance to shift how we allocate human labour**
- **Under non-correlation of errors, accuracy approaches 1 as we add annotators (Condorcet's jury theorem)**
- **But can we estimate the utility of an additional annotator, given observed inter-rater reliability**

Potsdam Institute for Climate Impact Research

Leibniz Association

An optimal distribution of labour between humans and AI

An optimal distribution of labour between humans and AI

If we make progress on the questions above, operational questions around organising work and the distribution of work between humans and AI start to become answerable

We can move towards an optimal distribution of labour between humans and AI that is likely to improve, rather than degrade, the quality and integrity of systematic reviews

Thank you

max.callaghan@pik-potsdam.de

PIK
Potsdam Institute for
Climate Impact Research

Leibniz
Association