MEASURING
THE AI CONTENT OF PUBLICLY FUNDED R&D PROJECTS

A proof of concept for the OECD Fundstat initiative

Fernando Galindo-Rueda
OECD Directorate for Science, Technology and Innovation

(with I.Yamashita*, A. Murakami, S. Cairns**, (STI/STP)
*: NISTEP **: McGill University

27 April 2021
Intelcomp project launch

OECD
BETTER POLICIES FOR BETTER LIVES

Background and objectives

- Background - AI
  - Radical transformation in the field of AI research over the past two decades
  - 2019 OECD Council Recommendation
  - Tracking government investments into AI-related R&D is of particular importance.
  - No comprehensive method exists by which to track and compare AI-related R&D funding across countries and agencies (nor infrastructure for that type of analysis).

    * AI-related R&D contains not only R&D on AI itself, but also close themes (e.g. AI applications in various fields).
- Bigger picture objectives
  - Pilot exercise to assess the feasibility of constructing a multi-country infrastructure on R&D project funding for analytical purposes (“Fundstat”)
  - Procedures and initial findings from an experimental text-based analysis of project-level R&D funding data – AI as a “case study”
  - Focused on measuring the extent and features of government support for AI-related R&D

Objectives: Building a new OECD data analysis infrastructure on R&D funding

OECD Working Party of National Experts on Science and Technology Indicators

AI policy interest, OECD council recommendation

\+

Build R&D Funding analysis infrastructure

“Fundstat” proof of concept

Operational OECD infrastructure (data, tools, community) serving policy user needs

Approach

- Quantitative case study approach, applying text mining tools to funding databases to identify AI-related R&D

1. Identification of key AI terms
2. Classification of documents as AI related
3. Topic modelling of AI related projects

- Used project-level funding data from 13 databases from eight OECD countries (Australia, Canada, France, Japan, Netherlands, Spain, United Kingdom, United States) and the EU

R&D and innovation funding databases from authorities or agencies in eight countries and the EU used in this analysis

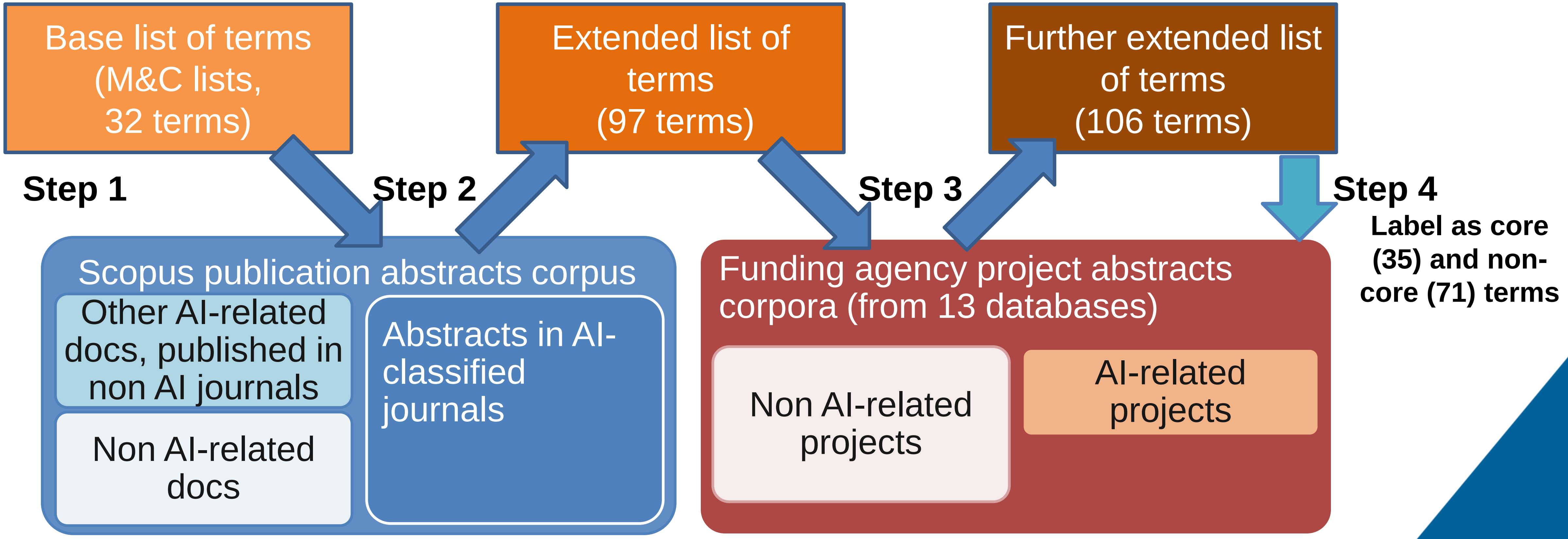
- Australian Research Council (**ARC**).
- Canadian Institutes of Health Research (**CIHR**) and Natural Sciences and Engineering Research Council (**NSERC**).
- The programmes under the Spanish National Plan for Scientific and Technological Research and Innovation (**PlanEst**), covering multiple state-level bodies.
- French National Research Agency (**ANR**).
- UK's Gateway to Research (GtR), which contains data for seven research councils (**GtR_RC**) and Innovate UK (**GtR_Inno**) .
- Japan's Agency for Medical Research and Development (**AMED**) and Database of Grants-in-Aid for Scientific Research (**KAKEN**) .
- Dutch Research Council (**NWO**).
- US' National Institutes of Health (**NIH**) and National Science Foundation (**NSF**).
- European Commission's Funding Programmes covered by the Community Research and Development Information Service (**CORDIS**).

Main features of the databases analysed

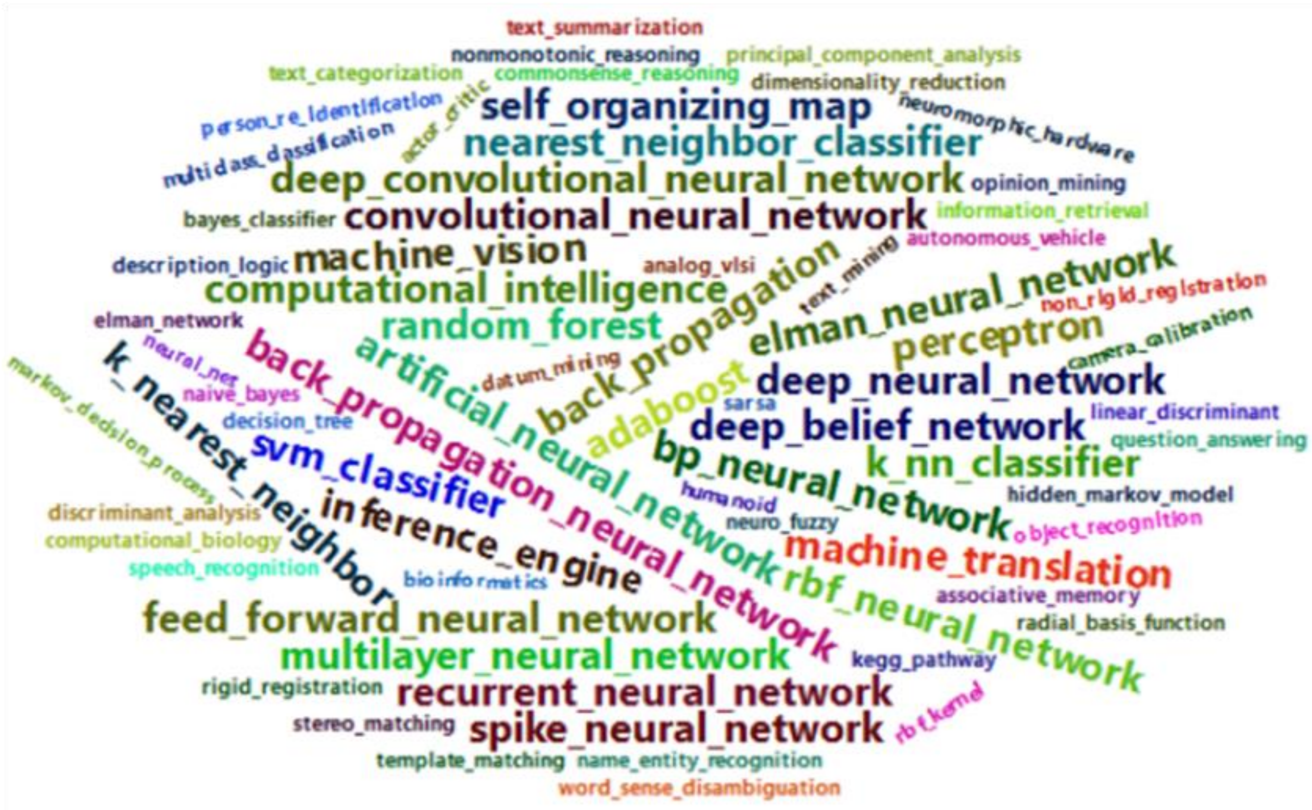
Database	Countries/ Region	Available period	Number of projects	Total amount of funding (USD Million)	Language	Data access	Analysis approach
ARC	Australia	2002-2019	26 677	8 994	English	Open	Pooled OECD
CIHR	Canada	2001-2018	56 778	14 147	English or French	Open	Pooled OECD
NSERC	Canada	2001-2017	175 945	3 402	English or French	Open	Pooled OECD
PlanEst	Spain	2004-2016	67 770	22 256	Spanish	Confidential	Distributed
ANR	France	2005-2019	20 123	6 506	French	Open	Pooled OECD
GtR_Inno	United Kingdom	2008-2019	18 424	14 281	English	Open	Pooled OECD
GtR_RC	United Kingdom	2006-2019	80 736	46 280	English	Open	Pooled OECD
AMED	Japan	2015-2018	4 765	4 213	Japanese	Open	Pooled OECD
KAKEN	Japan	2001-2018	466 709	33 750	Japanese or English	Open	Pooled OECD
NWO	Netherlands	2016-2019	7 177	2 186	English or Dutch	Confidential	Distributed
NIH	United States	2001-2019	1 428 472	497 955	English	Open	Pooled OECD
NSF	United States	2001-2019	224 307	114 883	English	Open	Pooled OECD
CORDIS	European Union	2001-2019	72 061	142 864	English	Open	Distributed

Outline of Key AI term identification

- Step 1: Obtain base list of terms from previous studies
- Step 2: Extend base list of terms by analysing Scopus (only AI-classified journals)
- Step 3: Further extend the list by analysing the 13 funding databases
- Step 4: Label key terms as “core” and “non-core” to tag and classify AI-related projects



Key AI terms expansion



Base key AI terms from two key terms sets (MeSH and Cockburn)

Semi-automatically retrieved additional key AI terms from AI journals and funding databases

Classification rule

- A document is selected as (likely to be) AI-related if

➢ **At least one core key term** found within its title or abstract;

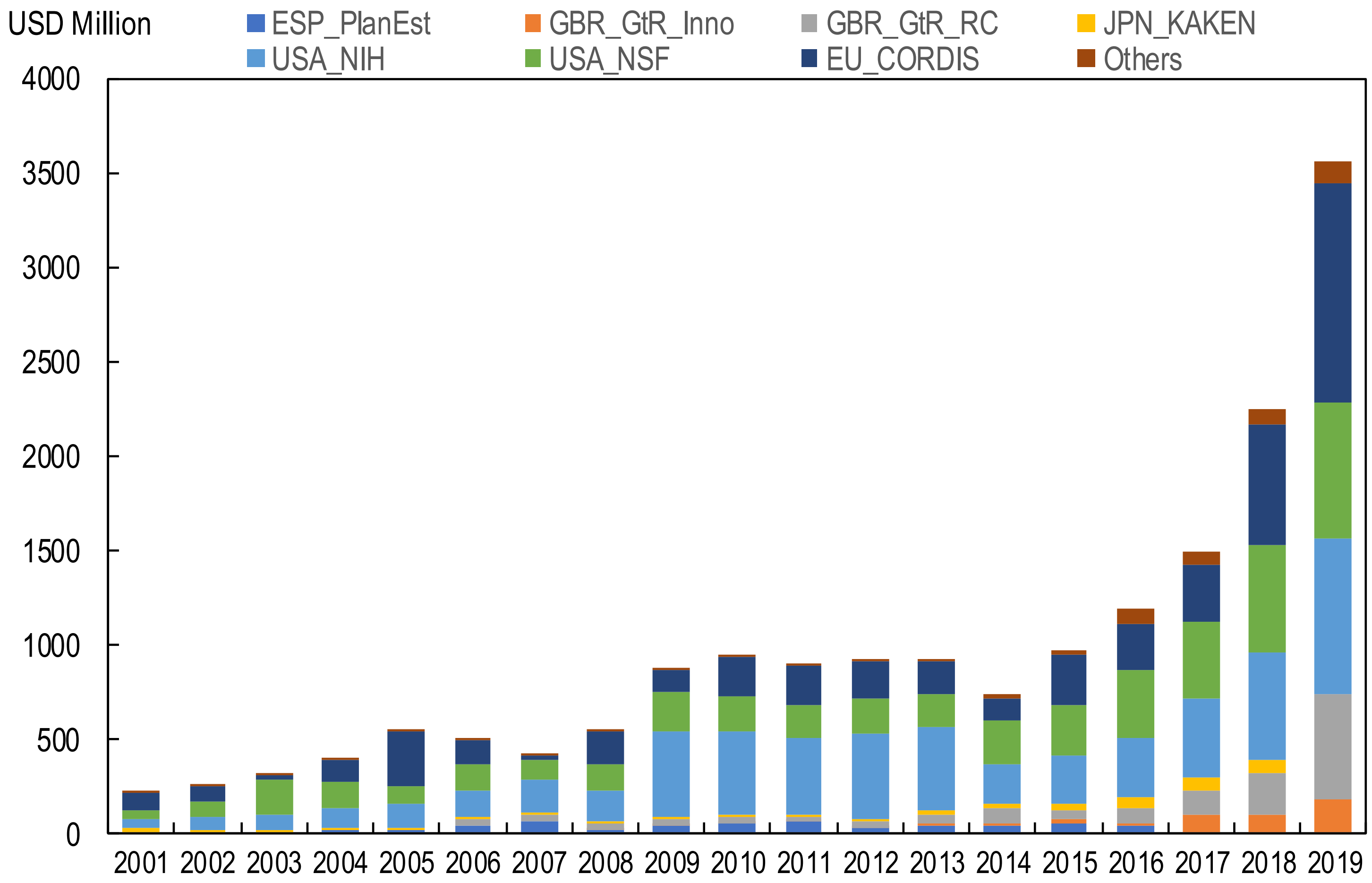
or

➢ **Two or more distinct non-core terms** found.

* An additional special rule is applied for excluding “bioinformatics” and “computational biology” combination, which does not necessarily select documents relevant to AI.

Funding trends in AI-related R&D projects

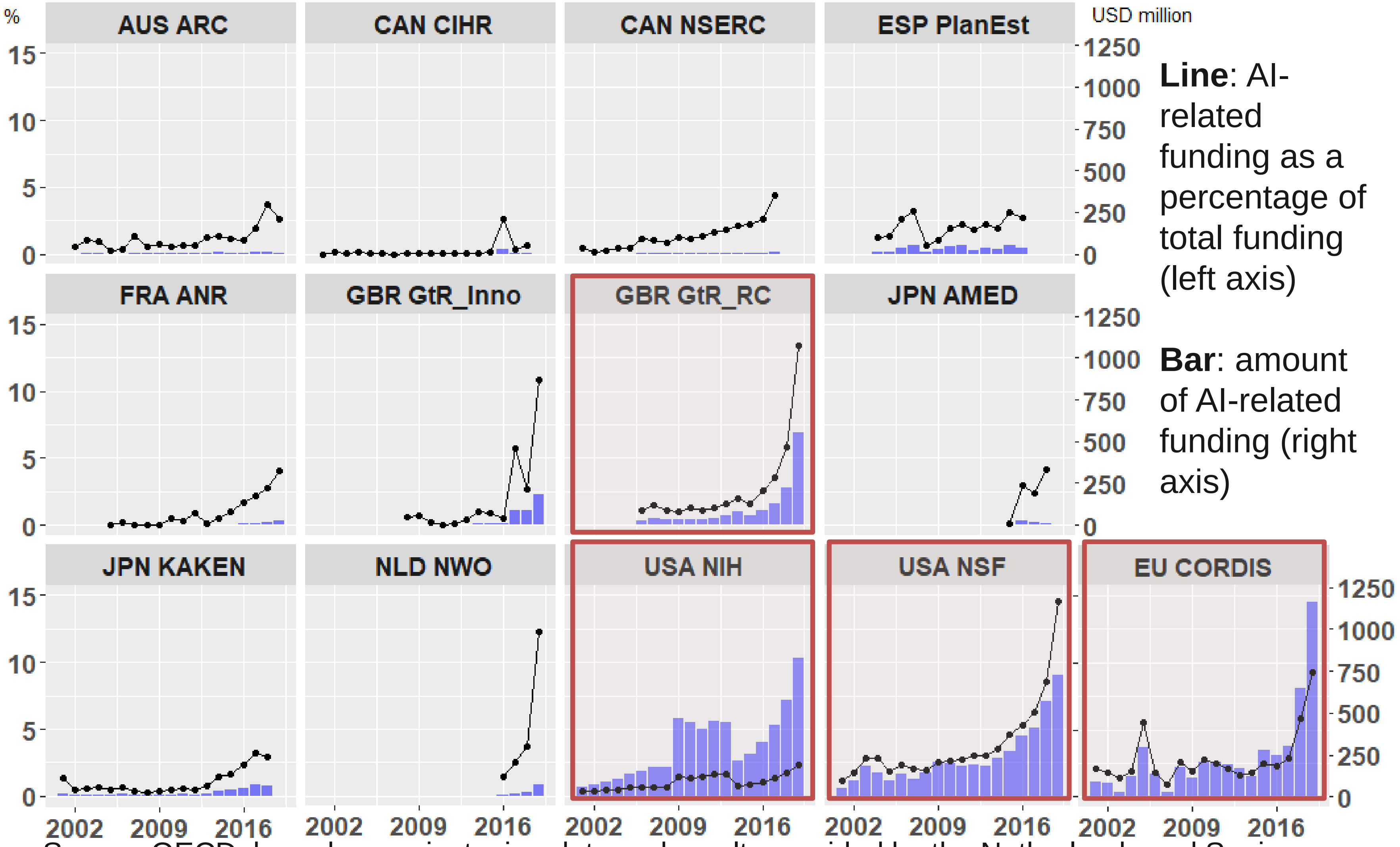
- For agencies with data available over a common period (2008 to 2018)[1], the total volume of AI-related R&D project funding increased from USD **525** to **2,210** million.



1: Excluding Canada's NSERC, Spain's PlanEst, Japan's AMED, and the Netherland's NWO

Estimated AI-related R&D funding within selected agencies, 2001-2019

- USA NIH, USA NSF, and EU CORDIS are the largest AI-related R&D funders, followed by GBR GtR_RC.



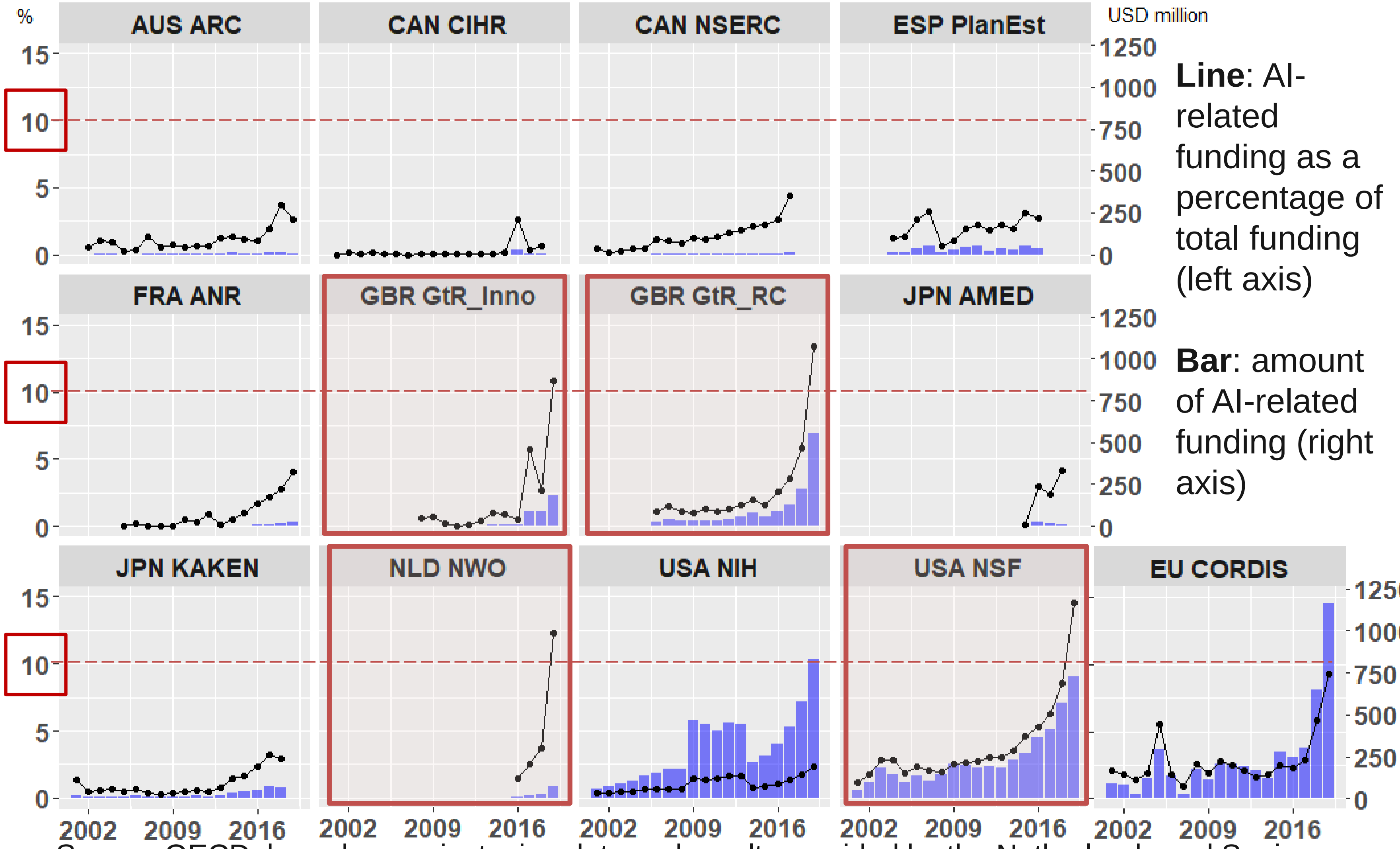
Line: AI-related funding as a percentage of total funding (left axis)

Bar: amount of AI-related funding (right axis)

Source: OECD, based on project microdata and results provided by the Netherlands and Spain.

Estimated AI-related R&D funding within selected agencies, 2001-2019

- GBR GtR_Inno, GBR GtR_RC, NLD NWO, and USA NSF devote the highest proportions of their funding to AI-relatefd R&D (more than 10% of their total funding in recent years).



Source: OECD, based on project microdata and results provided by the Netherlands and Spain.

Topic modelling for AI-related documents

- Goals
  - To examine what topics frequently appear in the AI-related documents.
  - To infer what types of research are supported by the funding organisations (e.g. what technologies are often studied and for what purposes)

- Steps:
  - Apply a topic modelling algorithm to find prominent topics in a collection of documents
  - Associate each document to a topic by probability measures produced by the algorithm

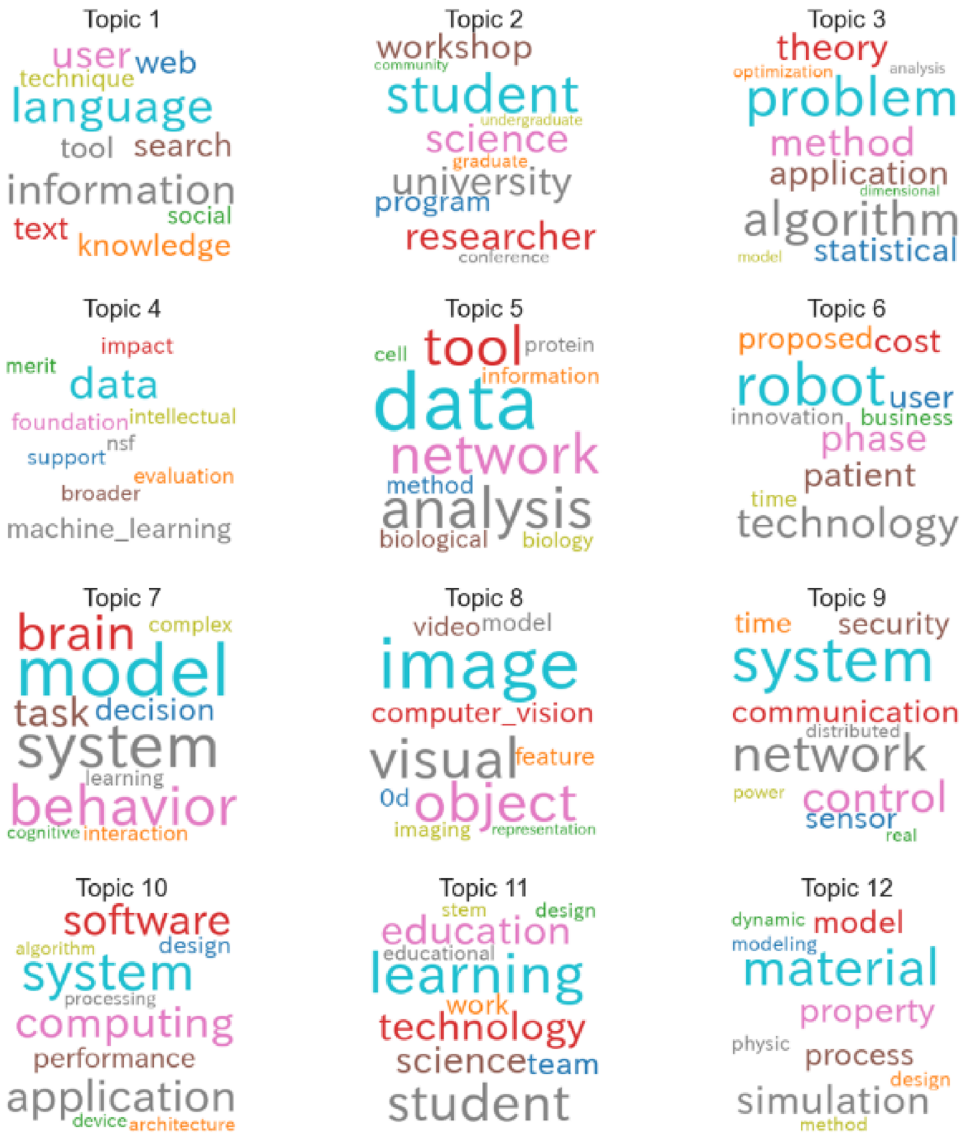
Topic modelling for AI-related documents

- For each funding database, 9 or 12 topics were generated.

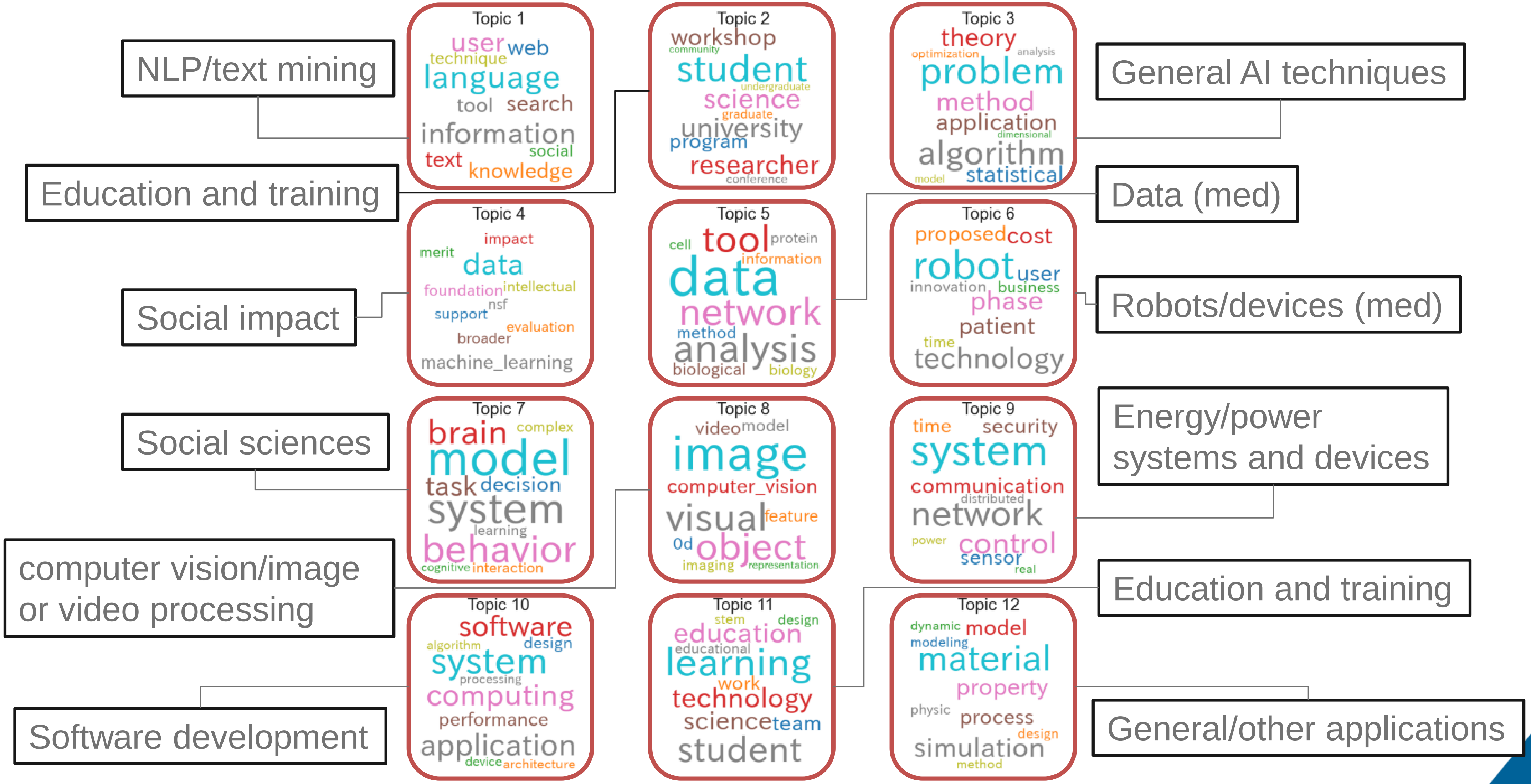
To compare the different funding databases

- Manual labelling of topic subjects was undertaken based on the examination and interpretation of the terms present in each word cloud.

< e.g. USA NSF, 2001-2019 >



Manual labelling of topics



< e.g. USA NSF, 2001-2019 >

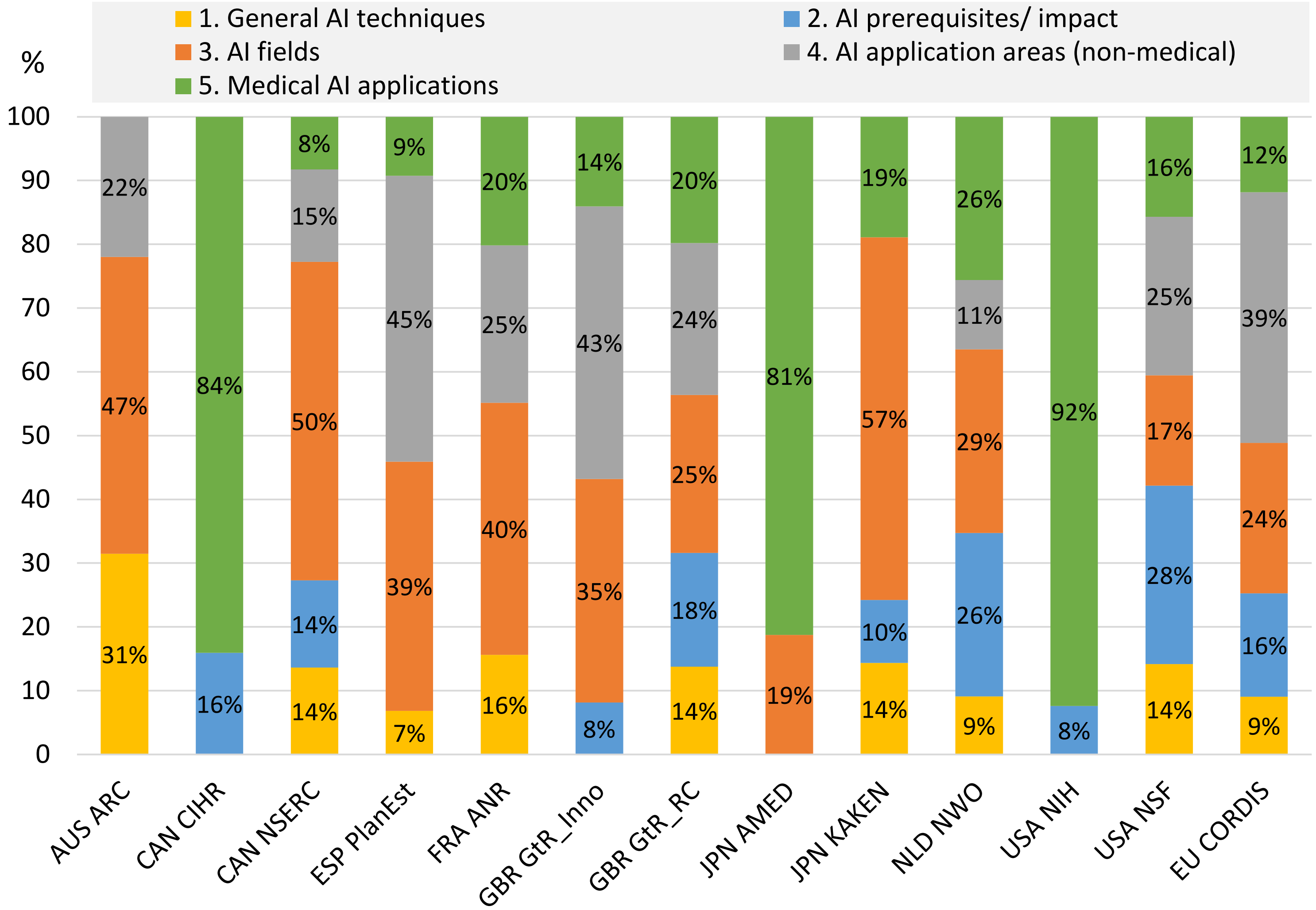
Classification of agency-specific topics into common themes and topics

Five common themes and 21 common topics were identified.

Common themes	**1. General AI techniques**	**2. AI prerequisites and impact**	**3. AI fields**	**4. AI application areas (non-medical)**	**5. Medical AI applications**
Common topics	1.1 General AI techniques	2.1 Education and training	3.1 Computer vision/image or video processing	4.1 Business	5.1 Treatment and patients (med)
		2.2 Social impact	3.2 NLP/text mining	4.2 Decision support	5.2 Research (med)
		2.3 Cost/production/monitoring	3.3 Big data/data analysis	4.3 Network/service systems	5.3 Diagnosis or imaging (med)
		2.4 Software development	3.4 Robots	4.4 Energy/power systems and devices	5.4 Data (med)
				4.5 Smart technology	5.5 Robots/devices (med)
				4.6 Social sciences	
				4.7 General/other applications	

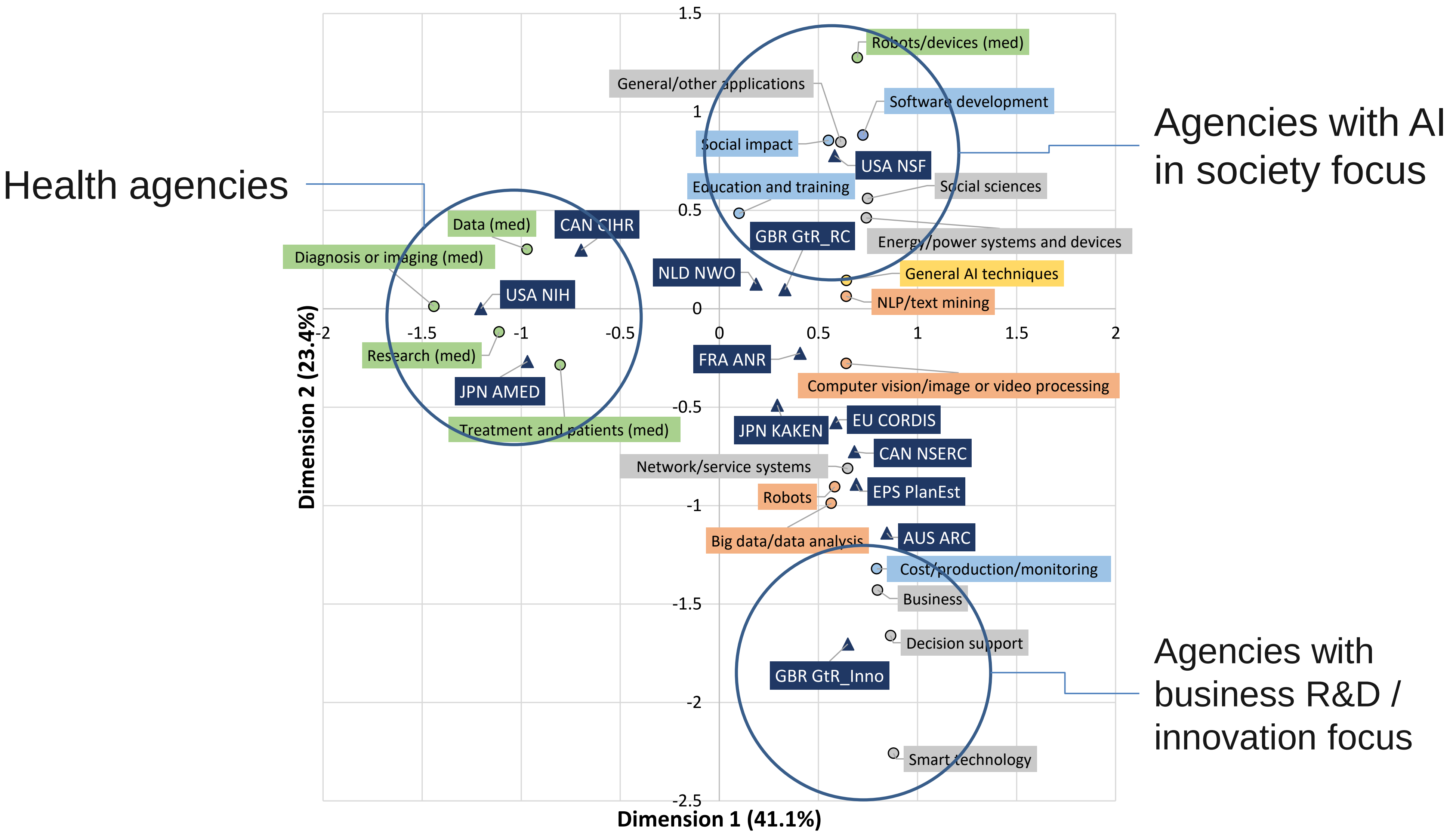
Distribution of documents by common theme within selected agencies

- CAN CIHR, JPN AMED, and USA NIH have a large share of documents that fall under the “medical AI applications” theme.
- AUS ARC, CAN NSERC, FRA ANR, and JPN KAKEN have relatively high shares that fall under the “AI fields” theme.
- More than 40% of all documents in ESP PlanEst and GBR GtR_Inno fall under the “AI application areas (non-medical)” theme.



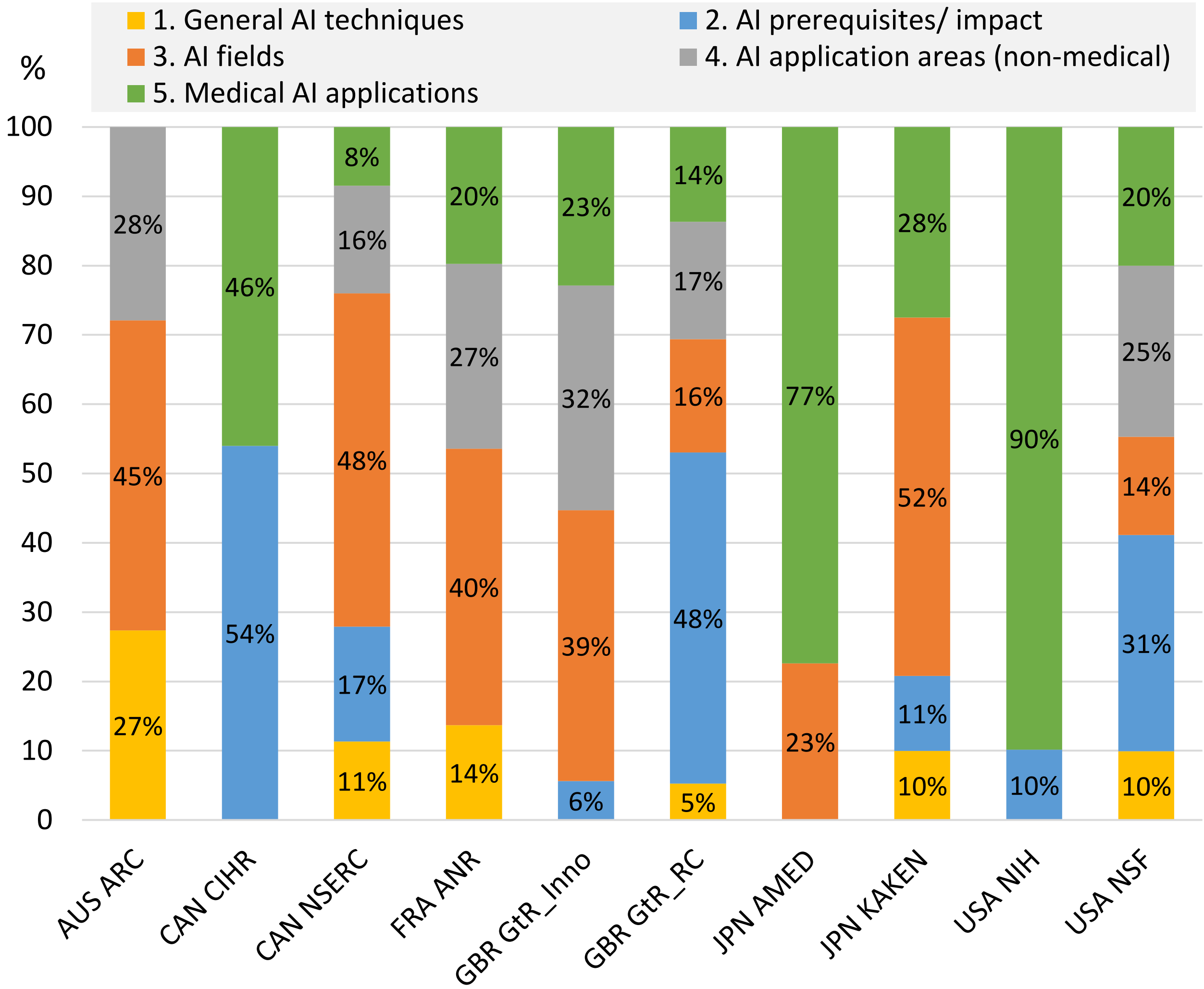
Source: OECD, based on project microdata and results provided by the Netherlands and Spain.

Correspondence analysis VIZ on agencies' AI projects distribution across common topics



Distribution of funding amounts by common theme for 10 agencies

- Both CAN CIHR and GBR GtR_RC dedicated a higher percentage of *funding* to the "AI prereqs and impact" theme than they did research *documents.*
  - ➢ Due to a few projects having received a large amount of funding.
- No other major discrepancies between funding and count percentages.



Source: OECD. The 10 agencies' data was pooled by the OECD.

Conclusions and next steps

- Potential for using project level data to carry out in-depth, internationally coordinated analysis of R&D funding.
- Insights on AI-related R&D funding trends and topics of AI-related R&D, relevant for OECD Council recommendation.
- Next
  - Policy questions on directionality and content
  - From proof of concept to analytical infrastructure
    - The OECD contribution and synergies with different developments, e.g. Intelcomp.

OECD/NESTI: Establishment of Expert Group on the Management and Analysis of R&D and Innovation Administrative Data (MARIAD)

Aim: Analyse content and directionality of R&D funding

Data

- Administrative data (esp. funding)
- Analysis ready/raw
- Qualitative (text) and quantitative
- Open/ confidential

Tools

- Standards
- Text mining methods
- Distributed/pooled

Expert community

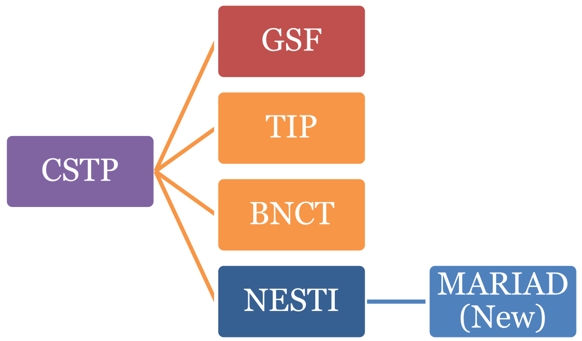
- Access
- Analysis expertise
- Responsibility for admin data
- → Gaps in NESTI

Important role for data/ analysis teams in funding agencies

Proof of concept completed: Analysis of funding of AI related R&D: **DSTI/STP/NESTI(2019)1/REV1** – forthcoming STI WP

MANDATE APPROVED
DSTI/STP/NESTI(2020)5/REV1
CALL FOR NOMINATIONS FORTHCOMING

Position of new group as new level 3 body



Fundstat SDG Funding measurement project

- Goal: identify project abstracts in funding agency databases that are related to one (or many) of the Sustainable Development Goals
- Lit review => Key terms approach not suitable. Machine learning. Training data required.

https://oe.cd/issa2021en

Thank you for your attention

Fernando.galindo-rueda@oecd.org

@galinnovation