More investment in R&D for better Education in the (distant) future?

9-th CEAFE/MWET, Rennes, France

June 3 & 4, 2024

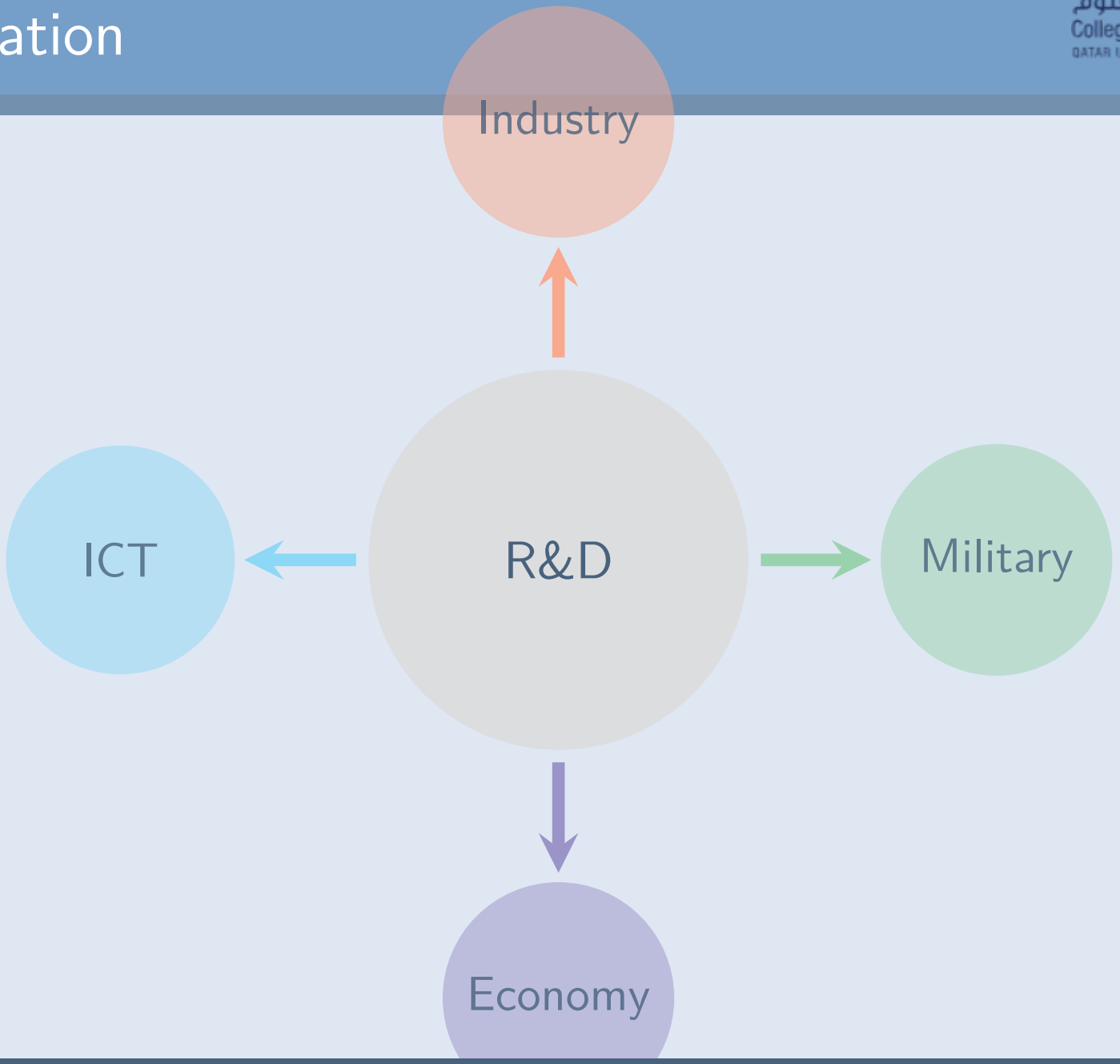
Rim Lahmandi-Ayed
University of Carthage, ESSAI, L.R. MASE (LR21ES21), Tunisia
& CUT, Rennes School of Business, France

Dhafer Malouche
Department of Mathematics, Statistics, and Physics,
College of Arts and Sciences,
Qatar University, Qatar

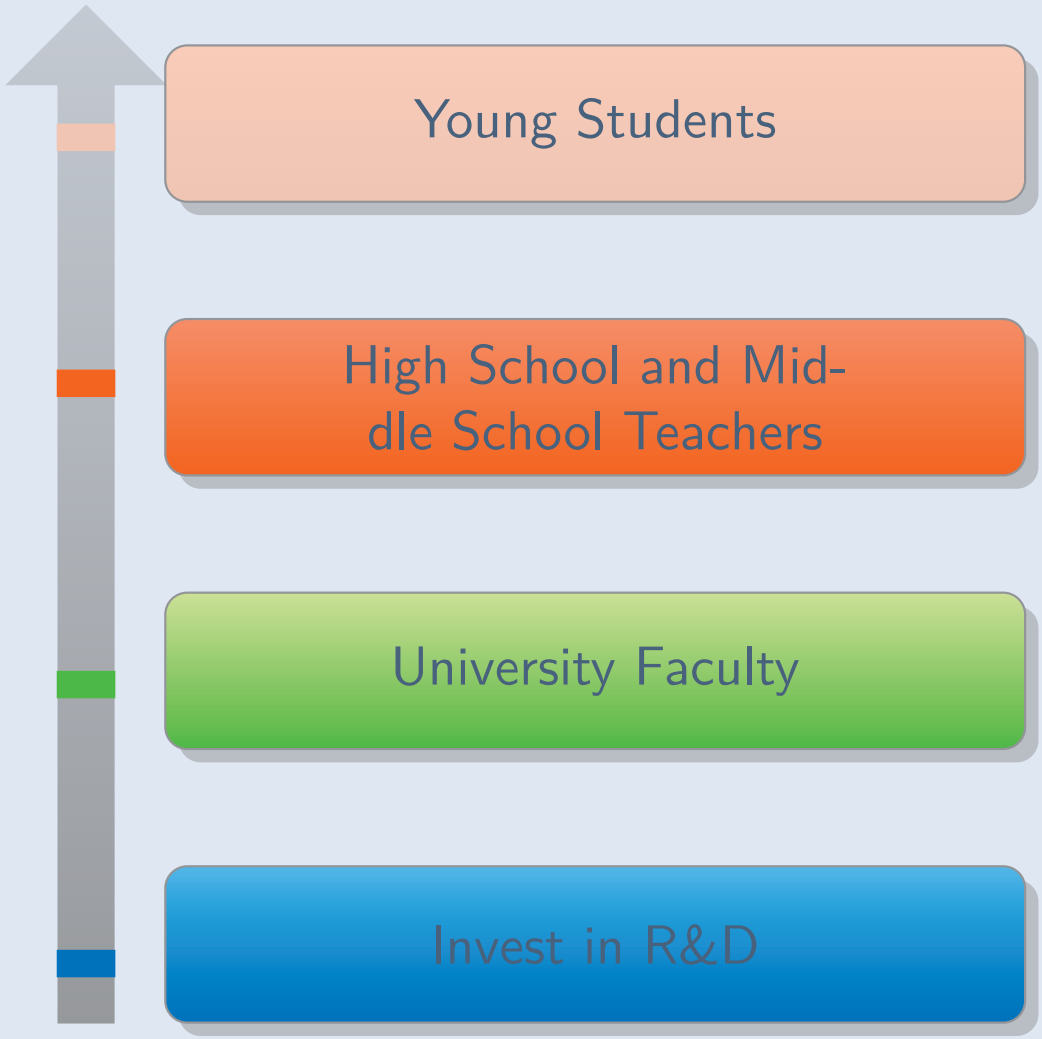
Outline

Introduction

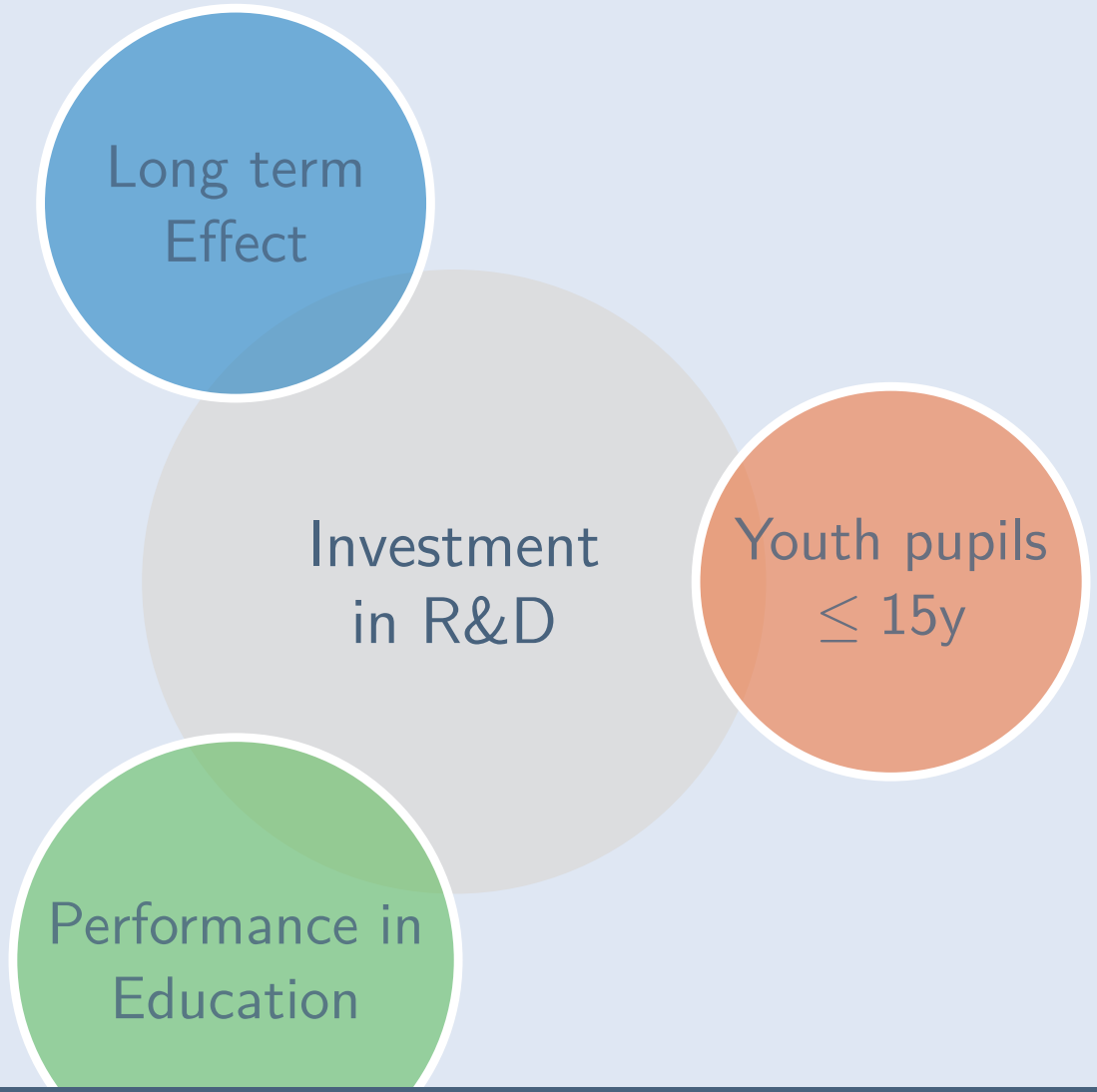
Motivation



Investing in R&D for Early Education



Problem


Long term Effect
Investment in R&D
Youth pupils ≤ 15y
Performance in Education

Statistical proof

RQ.1 What is the relationship between R&D expenditure and educational performance in early education/small classes for reading, mathematics, and science?

RQ.2 How long is the delay between RD expenditure and an observable impact on students' educational performance?

RQ.3 By increasing investment in R&D by 1%, what educational performance improvement can a country expect after this delay, and to what extent do countries differ in this respect?

Data

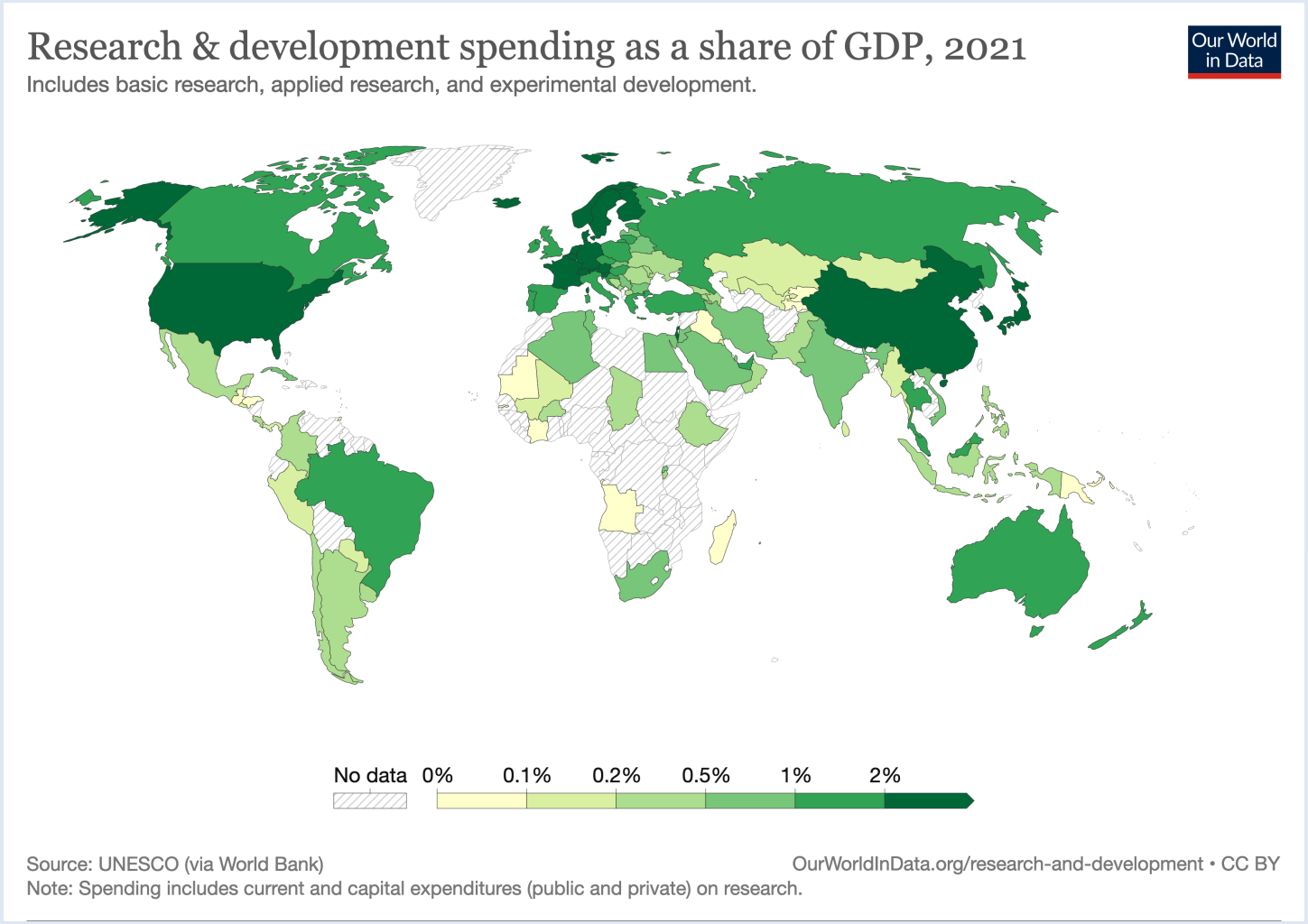
Measuring the expenditure in R&D (**Expend**)

- GB.XPD.RSDV.GD.ZS index (Source WDI)
- It stands for Gross Domestic Expenditure on Research and Development (R&D) as a percentage of Gross Domestic Product (GDP).
- It serves as an indicator of a country's investment in R&D activities.
- Provides insights into the level of emphasis a country places on innovation, scientific research, and technological development.
- Used to assess a country's focus on advancing its knowledge and technology sectors.

R&D Expenditure (% of GDP)

Country	Year	R&D Expenditure (% of GDP)
United States	2020	3.45%
Germany	2020	3.14%
France	2020	2.36%
United Kingdom	2019	1.71%
Brazil	2019	1.21%
Turkey	2020	1.10%
Tunisia	2019	0.75%
Algeria	2015	0.23%

Global map


Research & development spending as a share of GDP, 2021
Includes basic research, applied research, and experimental development.
Our World in Data
No data
0%
0.1%
0.2%
0.5%
1%
2%
Source: UNESCO (via World Bank)
OurWorldInData.org/research-and-development • CC BY
Note: Spending includes current and capital expenditures (public and private) on research.

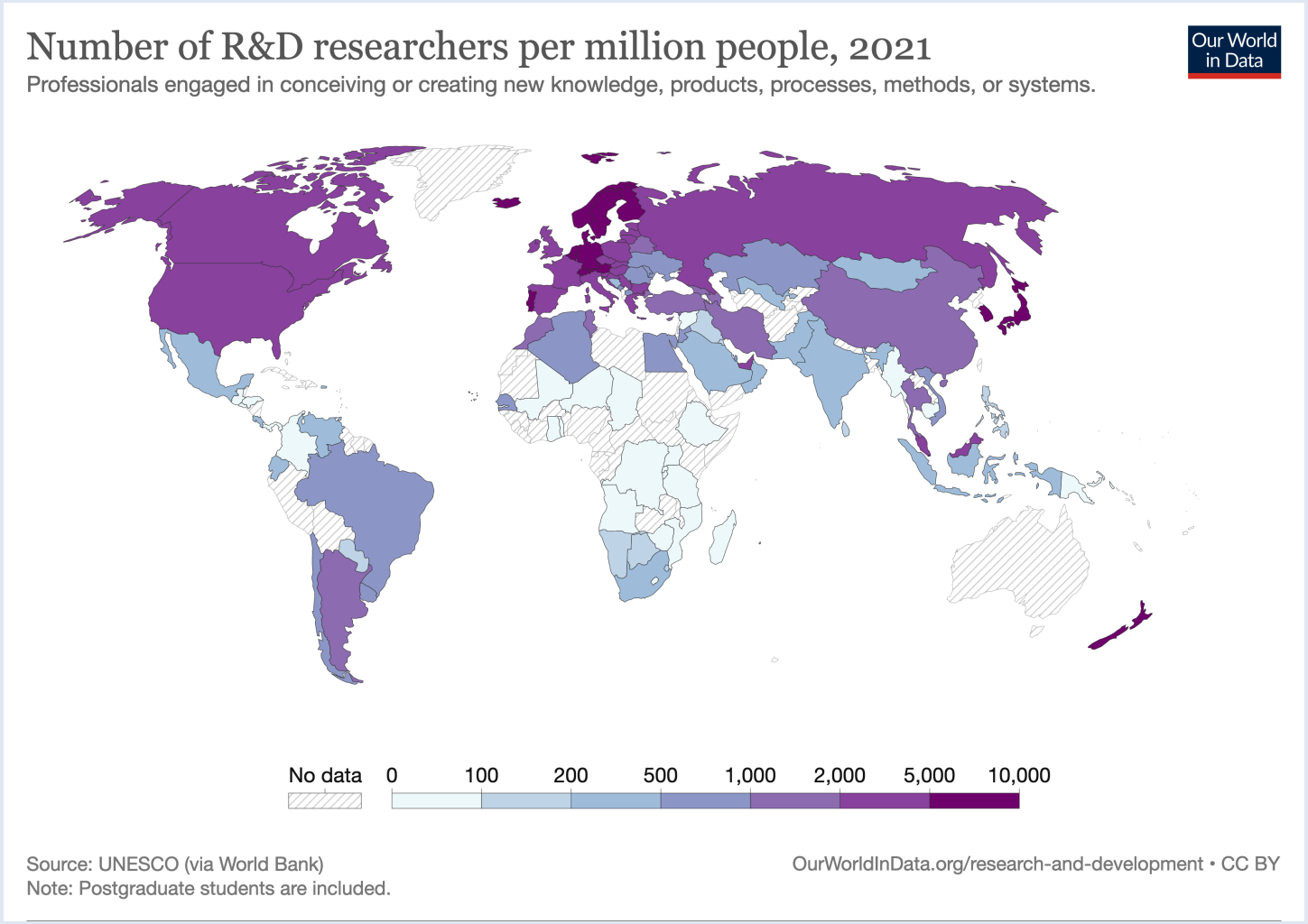
Number of researchers in R&D (**NumbRD**)

- "SP.POP.SCIE.RD.P6" index
- It represents the number of researchers in R&D per M people in a country.
- It measures the number of professionals engaged in the creation of new knowledge, products, processes, methods, or systems.
- Provides insights into a country's focus on research and development.
- Useful for assessing the human capital available for scientific and technological advancement.

Researchers in R&D (per M people)

Country	Year	Researchers per M
Germany	2020	5393.146
France	2020	4926.189
United Kingdom	2019	4683.766
United States	2019	4821.228
Turkey	2020	1775.347
Tunisia	2020	1659.923
Brazil	2015	295.235
Algeria	2016	168.719

Global map


Number of R&D researchers per million people, 2021
Professionals engaged in conceiving or creating new knowledge, products, processes, methods, or systems.
Our World in Data
No data
0
100
200
500
1,000
2,000
5,000
10,000
Source: UNESCO (via World Bank)
Note: Postgraduate students are included.
OurWorldInData.org/research-and-development • CC BY

Input Data: R&D Investment, Word Development Indicators

- **Expend**: Research and Development expenditure (% of GDP)
- **NumbRD**: Number of researchers by a one M person.
- **Period from 1997 to 2014**

Input Data: Expenditure per Researcher

- Total Expenditure in Dollars

$$\textbf{TotExp}(\mathit{US\$}) = \textbf{Expend} \times \textbf{GDP}(\mathit{US\$}) \times 10^{-2}.$$

- Total Number of Researchers

$$\textbf{TotRD} = \textbf{NumbRD} \times \textbf{Pop} \times 10^{-6}.$$

- Expenditure per Researcher

$$\textbf{ExpOneRD}(\mathit{US\$}) = \frac{\textbf{TotExp}(\mathit{US\$})}{\textbf{TotRD}}.$$

R&D Expenditure (US$)

Country	Year	R&D Expenditure in GDP (USD)
United States	2020	$726.62 B
Germany	2020	$122.30 B
France	2020	$62.15 B
United Kingdom	2019	$48.80 B
Brazil	2019	$22.63 B
Turkey	2020	$7.84 B
Tunisia	2019	$313.62 M
Algeria	2016	$126.07 M

Total Number of Researchers

Country	Year	Total Number of Researchers
United States	2020	1,582,953
Germany	2020	448,499
France	2020	332,868
United Kingdom	2019	313,046
Turkey	2020	149,370
Tunisia	2020	20,188
Algeria	2016	5,560

World Development Indicators: Estimated Cost per Researcher

Country	Year	Estimated Cost per Researcher (USD)
United States	2020	$458,913
Germany	2020	$272,549
France	2020	$186,743
United Kingdom	2019	$155,910
Turkey	2020	$52,484
Tunisia	2019	$15,534
Algeria	2015	$22,672

Output Data: PISA 2015

- Program for International Student Assessment (PISA).
- A triennial international survey
  - Evaluate education systems worldwide
  - 15-year-old students.
  - 72 countries are tested in science, mathematics, reading, collaborative problem solving and financial literacy.
- The Organization for Economic Co-operation and Development (OECD)
- Output data: Performance in mathematics, reading, and science obtained in the 2015 study.

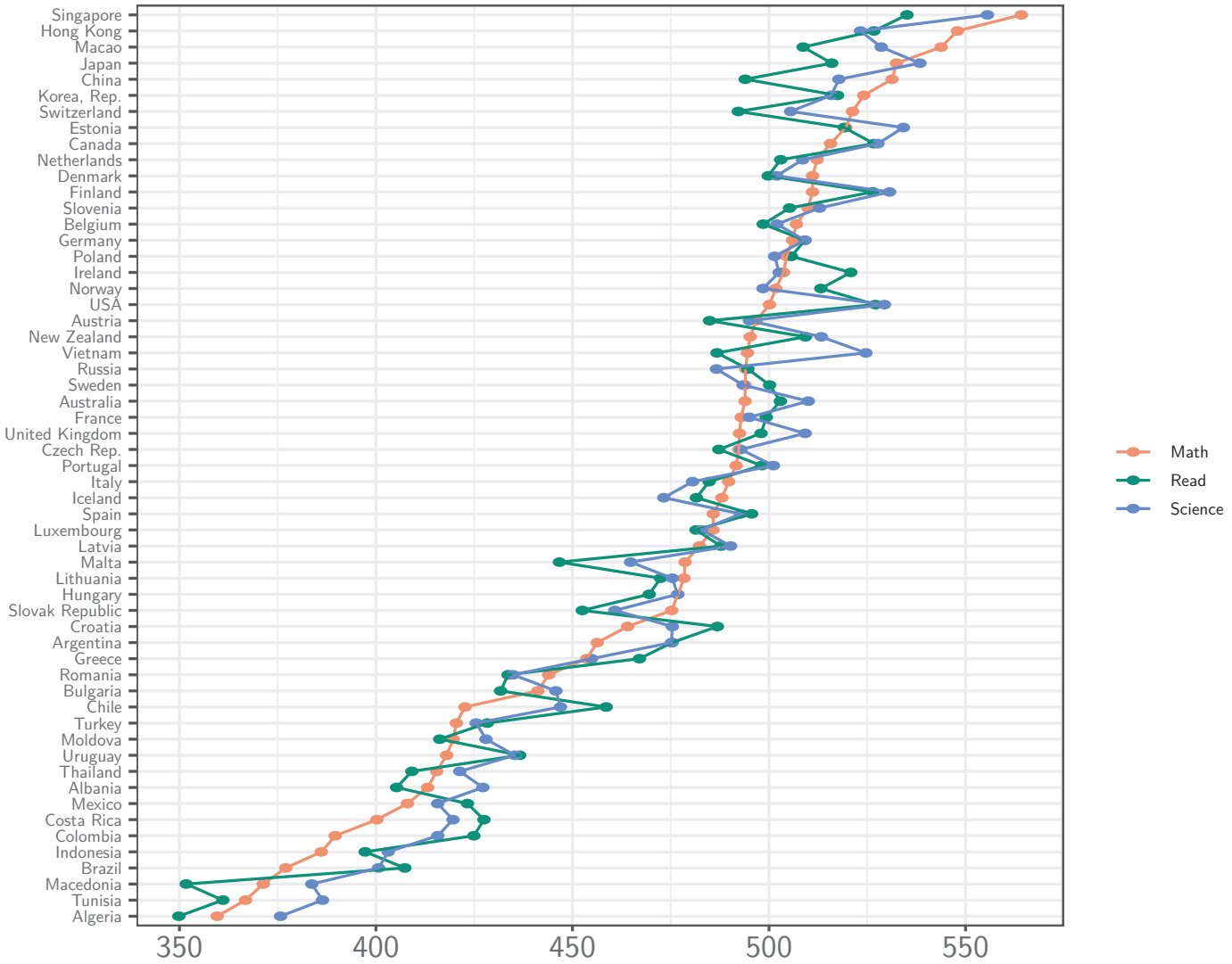
Merging Datasets: Input, Output, and Scope

- **Input Data:** WDI data and RD Indicators, focusing on Expenditure per Researcher from 1997 to 2014. This dataset covers approximately 130 countries.
- **Output Data:** PISA Scores from 2015 in Reading, Math, and Science. This dataset includes 72 countries.
- **Scope:** By merging both datasets, we focus on 57 countries for which data is partially available in both the input and output datasets.

Countries in the data

- **Africa**: Algeria, Tunisia
- **Asia**: Hong Kong, Indonesia, Japan, Korea, Macao, China, Singapore, Thailand, Vietnam
- **Europe**: Albania, Austria, Belgium, Bulgaria, Switzerland, Czech Republic, Germany, Denmark, Spain, Estonia, Finland, France, United Kingdom, Greece, Croatia, Hungary, Ireland, Iceland, Italy, Lithuania, Luxembourg, Latvia, Moldova, North Macedonia, Malta, Netherlands, Norway, Poland, Portugal, Romania, Russian Federation, Slovak Republic, Slovenia, Sweden
- **North America**: Canada, USA , Mexico
- **South America**: Brazil, Chile, Colombia, Costa Rica, Argentina, Uruguay
- **Oceania**: Australia, New Zealand

PISA scores are correlated



The merged data $\mathcal{D}$

$\mathcal{D}$ is an $n-$sample of observations of the random vector

$$[Y, \mathbf{X}] = [Y, (X_{1997}, \ldots, X_{2014})],$$

where

- Y Reading PISA score.
- $X_{1997}, \ldots, X_{2014}$: log(**ExpOneRD**) variables from $t = 1997, \ldots, 2014$.

Gaussian Bayesian Network

Modeling Approach

- Let f be a function and $S \subseteq \{1997, \ldots, 2014\}$ such that:

$$Y = f(X_s,\, s \in S)$$

- The function f will be estimated using **Gaussian Bayesian Networks** (GBN).

- $\max(S)$ represents the lag in the impact of R&D on educational performance.

- Our objectives are:
  - To estimate $\widehat{S}$ as an approximation of S.
  - To determine whether $\widehat{S}$ is empty or not.

What's a Gaussian Bayesian Network

- Bayesian Networks (BN) are Directed Acyclic Graphs (DAG) used to read the relationships between the variables in the random vector $[Y, \mathbf{X}]$.
- BN is a couple $G = (V, E)$ where V is the set of nodes and E is the set of directed edges.
  i. $\forall\, v \in V$, represents one variable from $\{Y\} \cup \{X_t, t = 1997, \ldots, 2014\}$
  ii. $E \subseteq V \times V$ such that if $(v, v') \in E$ then $(v', v) \notin E$
- $\forall\, v \in V$: $\theta(v)$ is the variable in $[Y, \mathbf{X}]$ represented by the node v in the DAG G.
- A Gaussian BN is a BN where $\Theta = (\theta(v),\ v \in V)$ is a Gaussian random vector.

Markov proprety

if f is the density of $[Y, \mathbf{X}]$,

- f satisfies the factorized Markov (FMP) propriety according to G if

$$f(\Theta) = \prod_{v \in V} g(\theta(v) \mid \Theta(pa(v)))$$

  where $pa(v) = \{v' \in V, \text{ such that } (v', v) \in E\}$: parents of v: parents of v.

- If f satisfies the FMP, then f satisfies the pairwise Markov propriety:

$$v \not\sim v' \text{ then } \theta(v) \perp\!\!\!\perp \theta(v') \mid \Theta(pa(v))$$

  where $\Theta(pa(v)) = (\theta(u),\, u \in pa(v))$

Learning Bayesian Network

- A score that measures the goodness of fit of the model to the data:

  Bayesian Information Criteria: $\mathrm{BIC} = \log(n)k - 2\log(\hat{L})$.

  where
  - $\hat{L} =$ is the maximized value of the likelihood function
  - n is the sample size
  - k is the number of parameters

- We usually estimate the BN that corresponds to the minimum of a score (BIC): It's the learning procedure

Learning Bayesian Network

- The learning procedure is an NP-hard problem
- Our DAGs or BN should not contain edges
  - from $X_{t'}$ to X_t when $t' > t$,
  - or edges from Y to any of the X_t when t and t' belong to $\{1997, \ldots, 2014\}$
- The set of possible DAGs has a cardinality equal to

$$18! = 18 \times 17 \times \ldots \times 1 = 6.402374 \times 10^{15},$$

instead of $2^{\binom{18}{2}} = 2^{153}$

Two families of Learning BN

- Constraint-based Algorithms:
  - PC-algorithm
  - Two steps:
    1. Conditional Independence hypothesis testing,
    2. learning directions using the V-structure principal.
  - The final result is a partially directed graph with undirected and directed arrows.
- Score-based algorithms
  - Heuristic optimization techniques in order to search for a minimum score.
  - Hill-Climbing with random restarts (Bouckaert, 1995).
  - Start from an initial BN and add or remove an edge until the score can no longer be improved.
- Hybrid algorithms: a composition between constraint-based and score-based algorithms, Max-Min Hill-Climbing algorithm (MMHC) (Tsamardinos et al., 2006).

Scoring Metrics

- **Log-Likelihood (LL)** Measures the likelihood of the observed data.
- **Akaike Information Criterion (AIC)** Balances likelihood and the number of parameters.
- **Bayesian Information Criterion (BIC)** Similar to AIC, the different penalty for complexity.
- **Bayesian Dirichlet equivalent uniform (BDeu)** Assumes a uniform prior over structures.
- **Minimum Description Length (MDL)** Aims for the model that best compresses the data.

Strength of the Links

- Bootstrap Method (Efron and Tibshirani, 1993)
- Use 500 bootstrap replicates, applying the Hill-Climbing algorithm to each sample.
- Compute the strength of an edge as the frequency of its occurrence across the 500 estimations.
- Identify the *strongest* link between R&D variables and the Reading PISA score variable, measuring the lag of the impact.

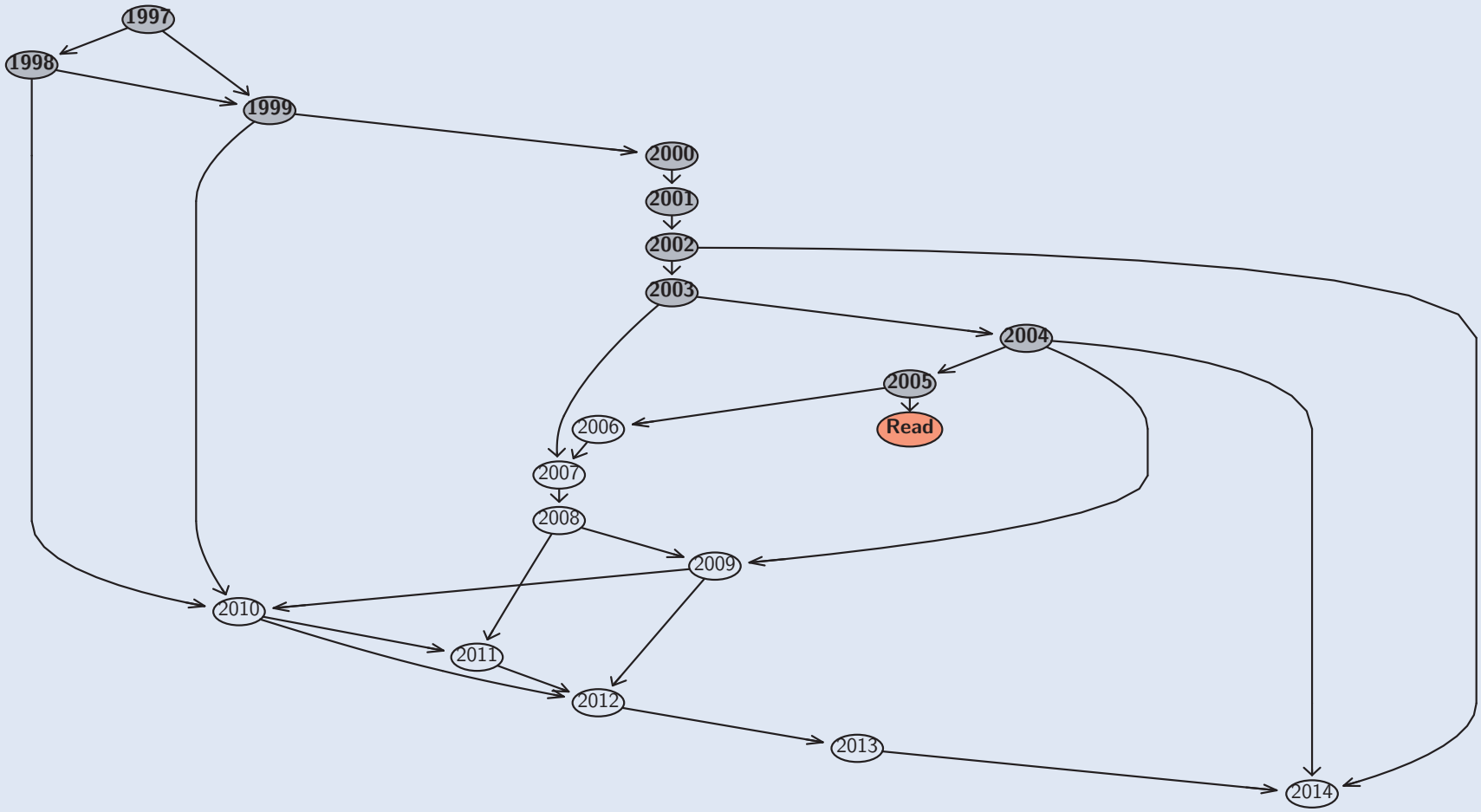
Missing values imputation procedure is required

Methodology

1. Impute missing values using a Bayesian Network procedure.
2. Initialize a Bayesian Network using a Hybrid algorithm: Max-Min Hill Climbing (MMHC) with BIC score and CI-Independence testing as criteria.
3. Refine the Bayesian Network using Hill Climbing (HC) with random starts, employing BIC score as the evaluation metric.
4. Perform 500 replicates of the estimations from step 2 to assess the strength of the probabilistic relationships expressed by the network arcs.
5. Apply model averaging to construct a network that includes only statistically significant arcs.

Results

Estimated Bayesian Network



Estimated regression models

	Dependent variable:							
	'1998'	'1999'	'2000'	'2001'	'2003'	'2004'	'2005'	**Read**
'1997'	0.956***	−0.406***						
	(0.015)	(0.073)						
'1998'		1.445***						
		(0.075)						
'1999'			0.971***					
			(0.014)					
'2000'				0.987***				
				(0.014)				
'2002'					0.980***			
					(0.017)			
'2003'						0.952***		
						(0.013)		
'2004'							0.915***	
							(0.033)	
'2005'								**25.139*** **
								(4.620)
Const.	0.481***	−0.448***	0.275*	0.125	0.372**	0.672***	0.984***	**192.574*** **
	(0.159)	(0.096)	(0.148)	(0.149)	(0.184)	(0.147)	(0.366)	**(51.807)**
R^2	0.987	0.996	0.989	0.990	0.984	0.989	0.934	**0.350**
Adj. R^2	0.987	0.996	0.989	0.989	0.984	0.989	0.933	**0.338**
df	55	54	55	55	55	55	55	**55**
F Stat.	4,343.5***	7,412.7***	5,135.3***	5,226.0***	3,377.6***	5,126.5***	782.2***	**29.6*** **
Note:								*p<0.1; **p<0.05; ***p<0.01

Contribution & Efficiency of the investment in R&D

- Estimated Regression Model

$$\widehat{\text{Read}} = 192.574 + 25.139 \times \log(\textbf{ExpOneRd}(2005))$$

- **Contribution** of the investment in R&D in the explanation of the Performance of the Education System of a country w

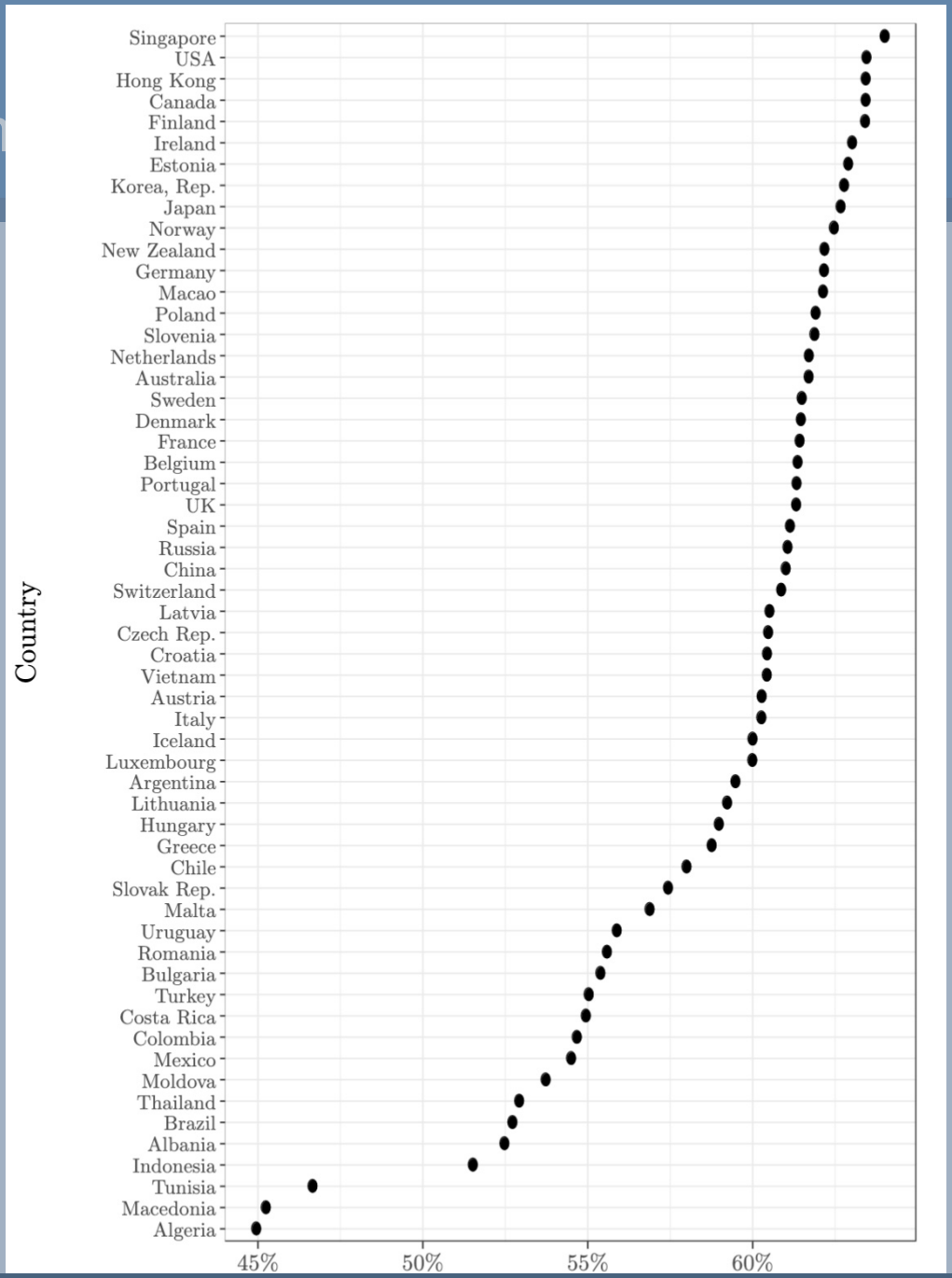
$$\text{Contribution of R\&D}(w) = \frac{\text{Read}(w) - 192.574}{\text{Read}(w)}$$

- **Efficiency** of the investment in R&D in the explanation of the Performance of the Education System of a country w

$$\text{Efficiency of R\&D}(w) = \frac{\text{Read}(w) - \widehat{\text{Read}}(w)}{\text{Read}(w)}$$

Contribution of the investment in R&D

Contribution

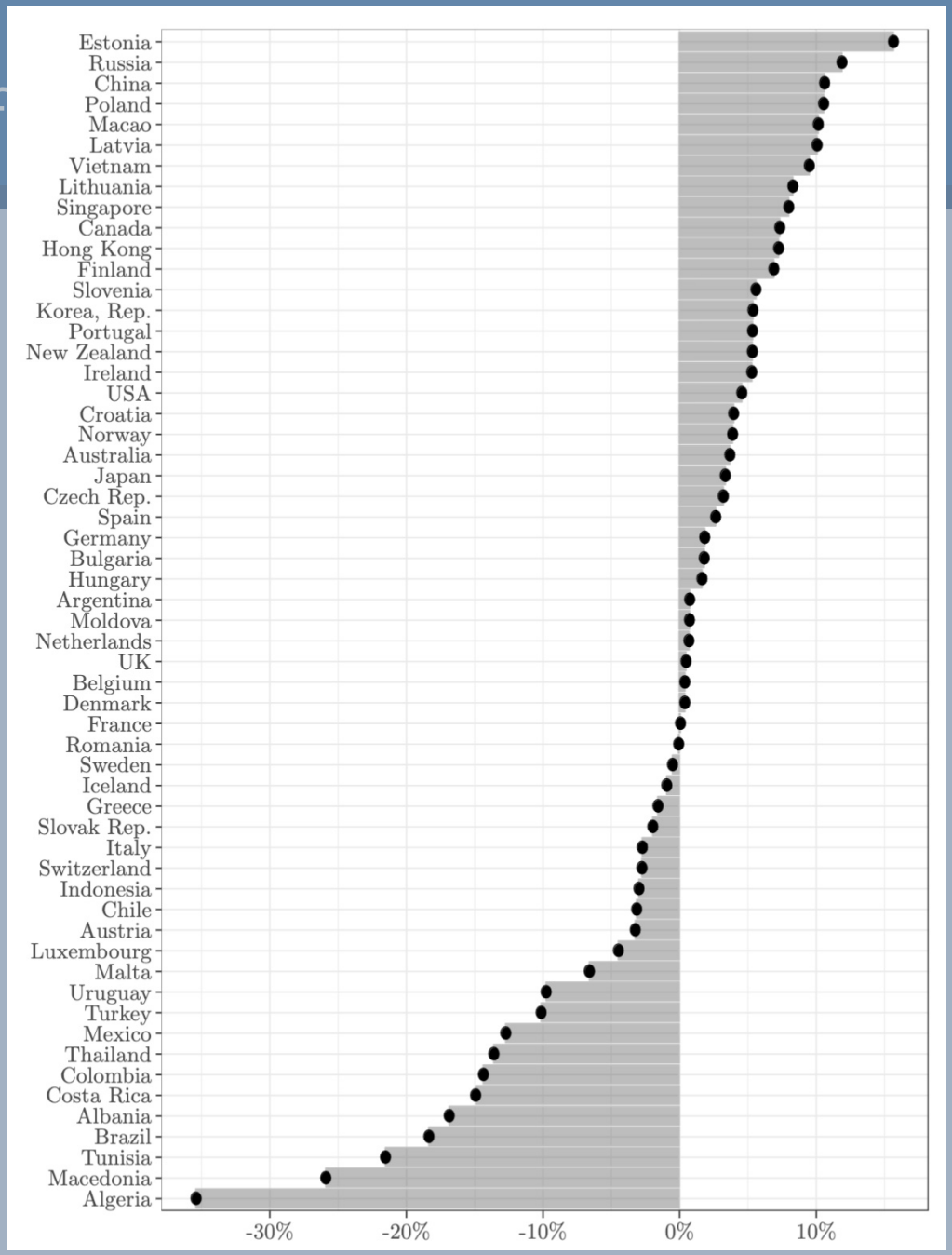


Contribution of the investment in R&D

- Algeria (45%), Tunisia (46%), Brazil (52.5%), Turkey (55%)
- China (61%), Japan (62.5%0), South Korea (62.5%), Singapore (64.5%)
- USA (64%), France (61%), UK (61%)

Efficiency of the investment in R&D

Efficiency of

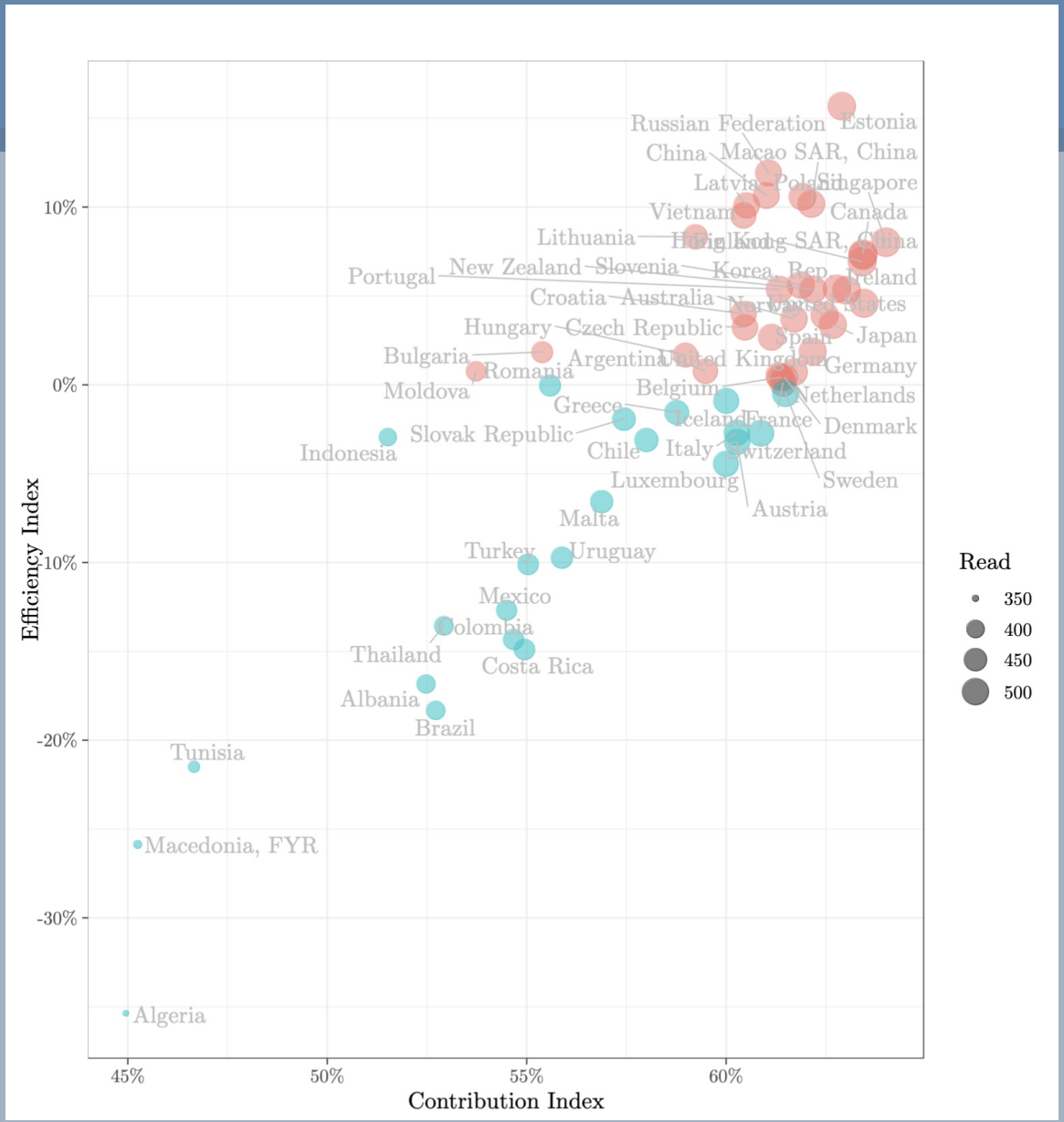


Efficiency of the investment in R&D

- Algeria (-35%), Tunisia (-22%), Brazil (-18%), Turkey (-10%)
- China (+10%), Japan (+3%0), South Korea (5%), Singapore (7.5%)
- USA (5%), France (0%), UK (0%)

Efficiency vs Contribution

Efficiency


Russian Federation
Estonia
China
Macao SAR, China
Latvia
Poland
Singapore
Vietnam
Canada
Lithuania
Hong Kong SAR, China
Portugal
New Zealand
Slovenia
Korea, Rep.
Ireland
Croatia
Australia
Norway
United States
Hungary
Czech Republic
Spain
Japan
Bulgaria
Argentina
United Kingdom
Germany
Romania
Moldova
Belgium
Netherlands
Greece
Iceland
France
Denmark
Slovak Republic
Indonesia
Chile
Italy
Switzerland
Luxembourg
Sweden
Austria
Malta
Turkey
Uruguay
Mexico
Colombia
Thailand
Costa Rica
Albania
Brazil
Tunisia
Macedonia, FYR
Algeria
10%
0%
-10%
-20%
-30%
45%
50%
55%
60%
Efficiency Index
Contribution Index
Read
350
400
450
500

Thank you