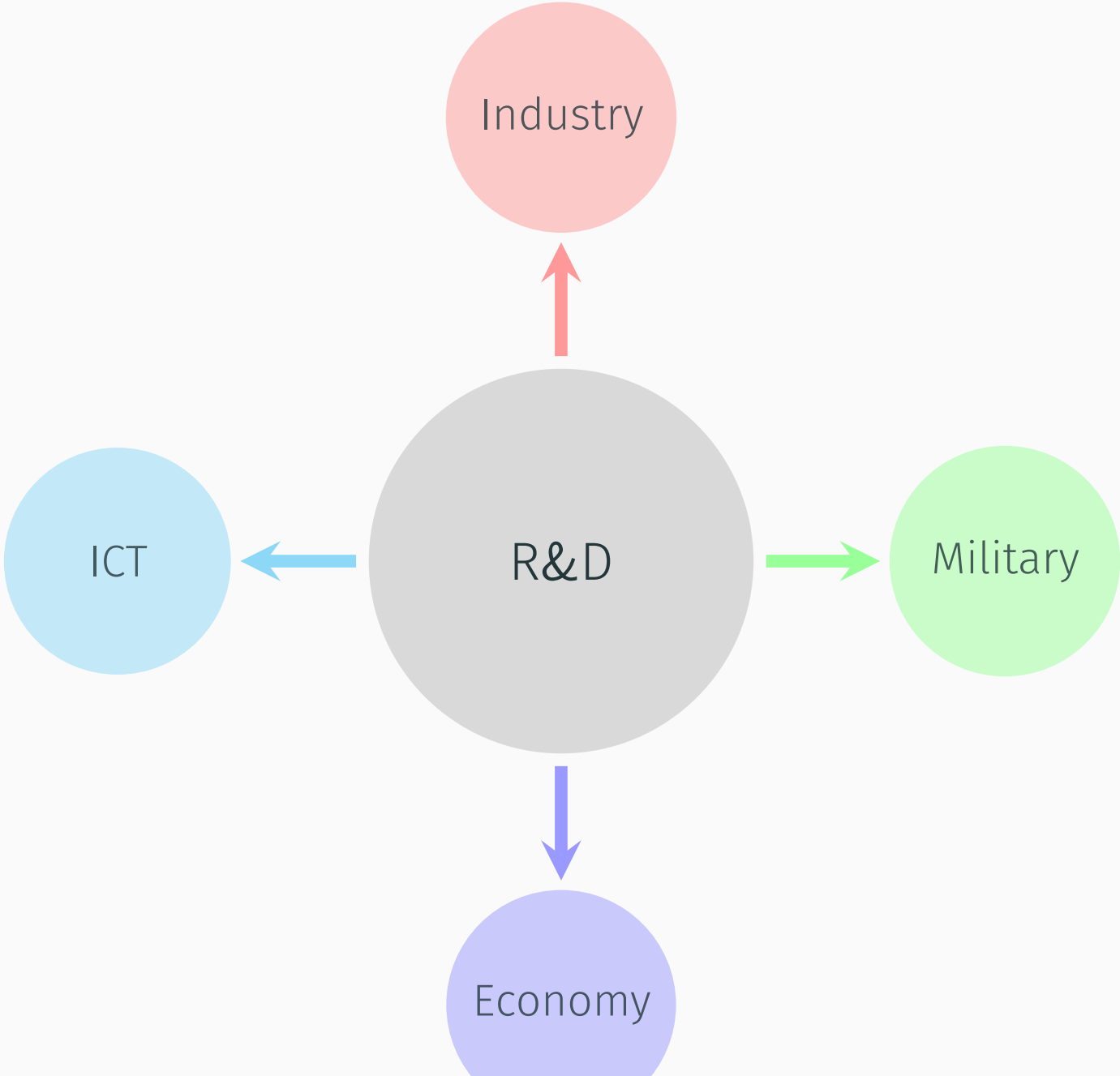
More investment in Research and Development for better Education in the future?

Rim Lahmandi-Ayed and Dhafer Malouche

CEAFE, Beit El Hikma, June 21st, 2018

ESSAI-MASE-Carthage University

Motivation

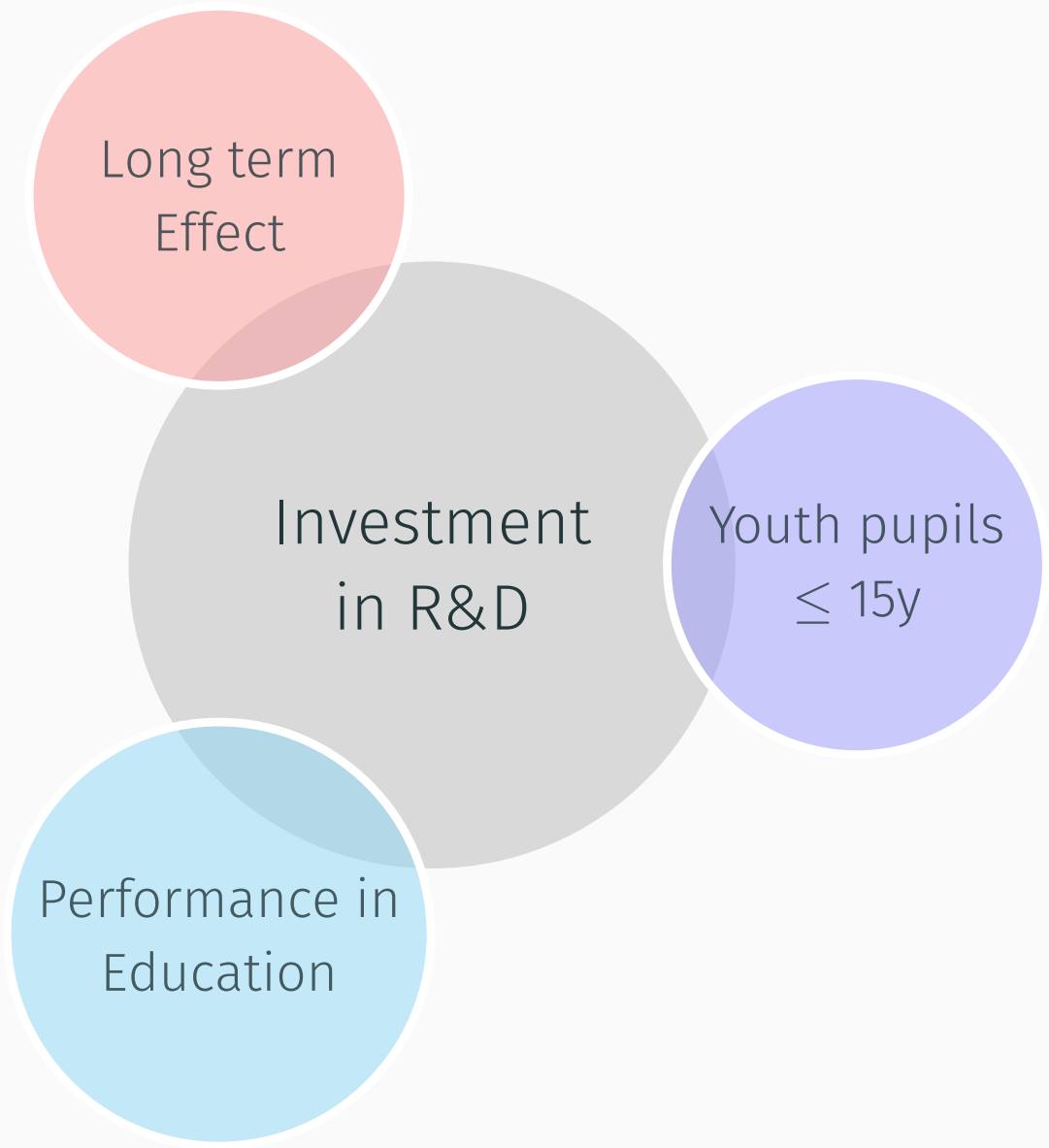


Motivation

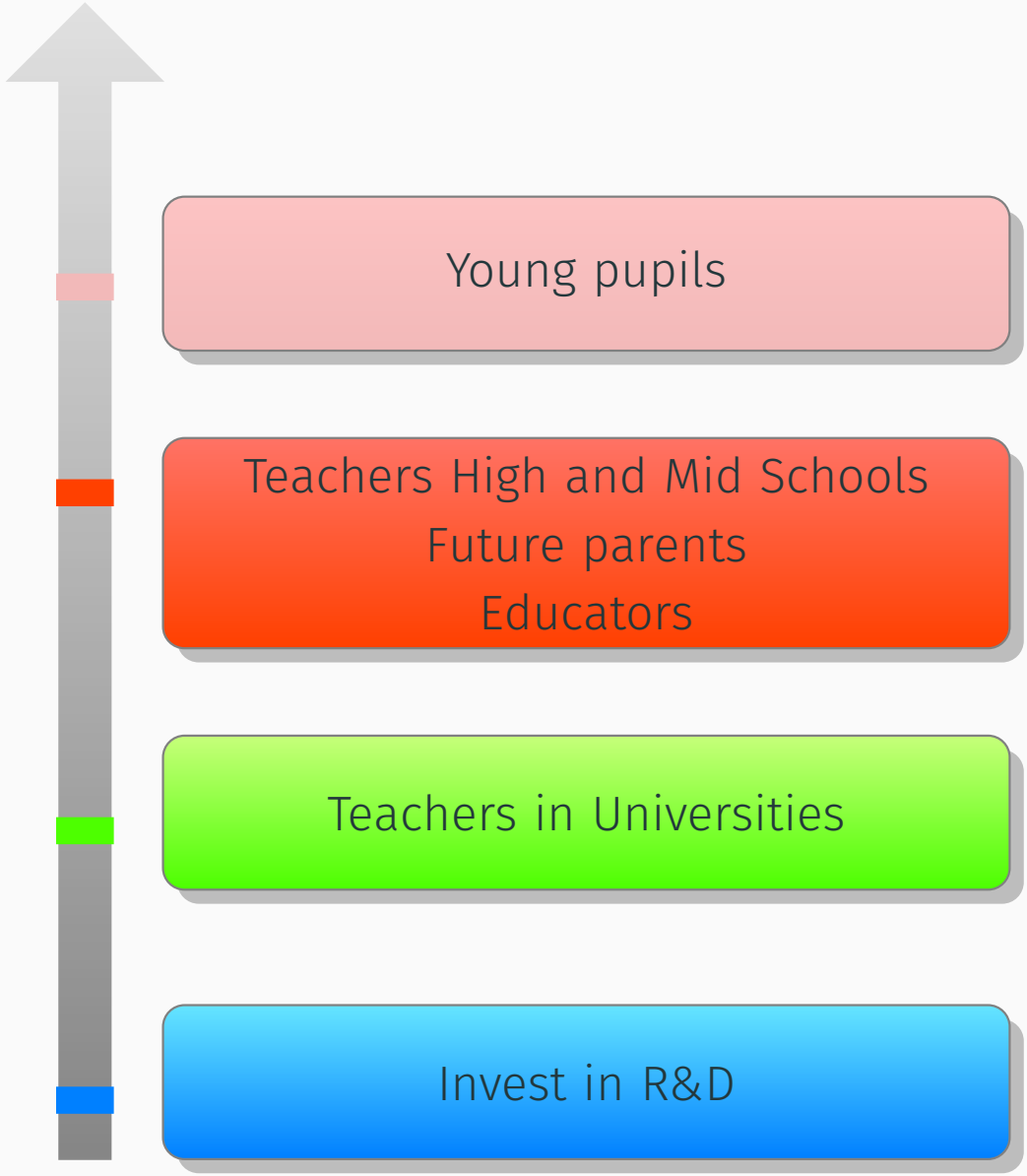
Arnold, 2012[1]: *we know surprisingly little about their long-term effects. This is a pity, because the conventional justifications for state intervention in research depend upon phenomena that are inherently long term in nature.*

[1] E. Arnold, Understanding long-term impacts of rd funding: The EU framework program, Research Evaluation 21 (2012) 332343.

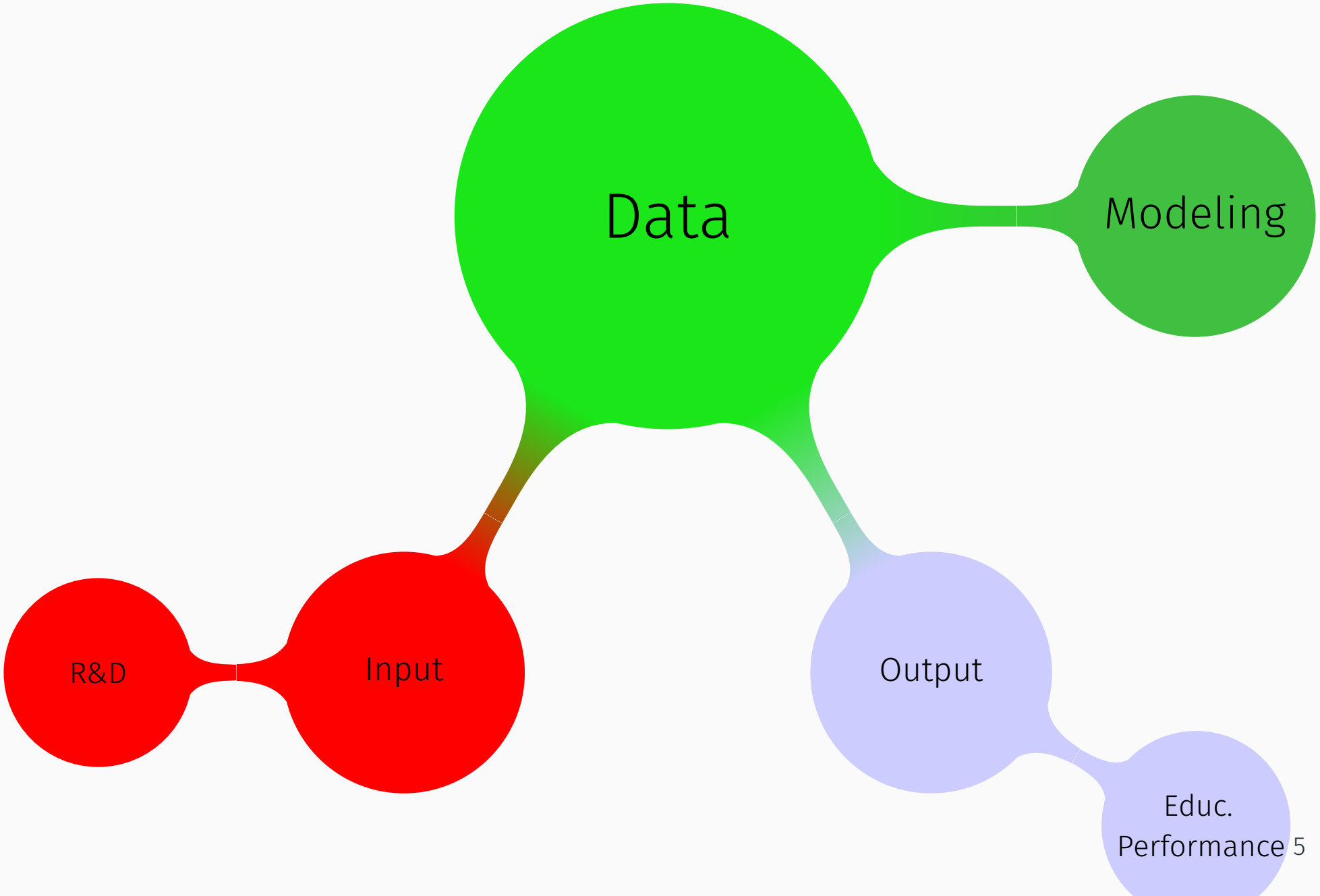
Problem


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

Why we are interested in this problem?



Statistical proof



Input Data: R&D Investment

Word Development Indicators[2]

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

[2] https://datacatalog.worldbank.org/dataset/world-development-indicators

Input Data: Expenditure per Researcher

- Total amount of Expenditure in dollars

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

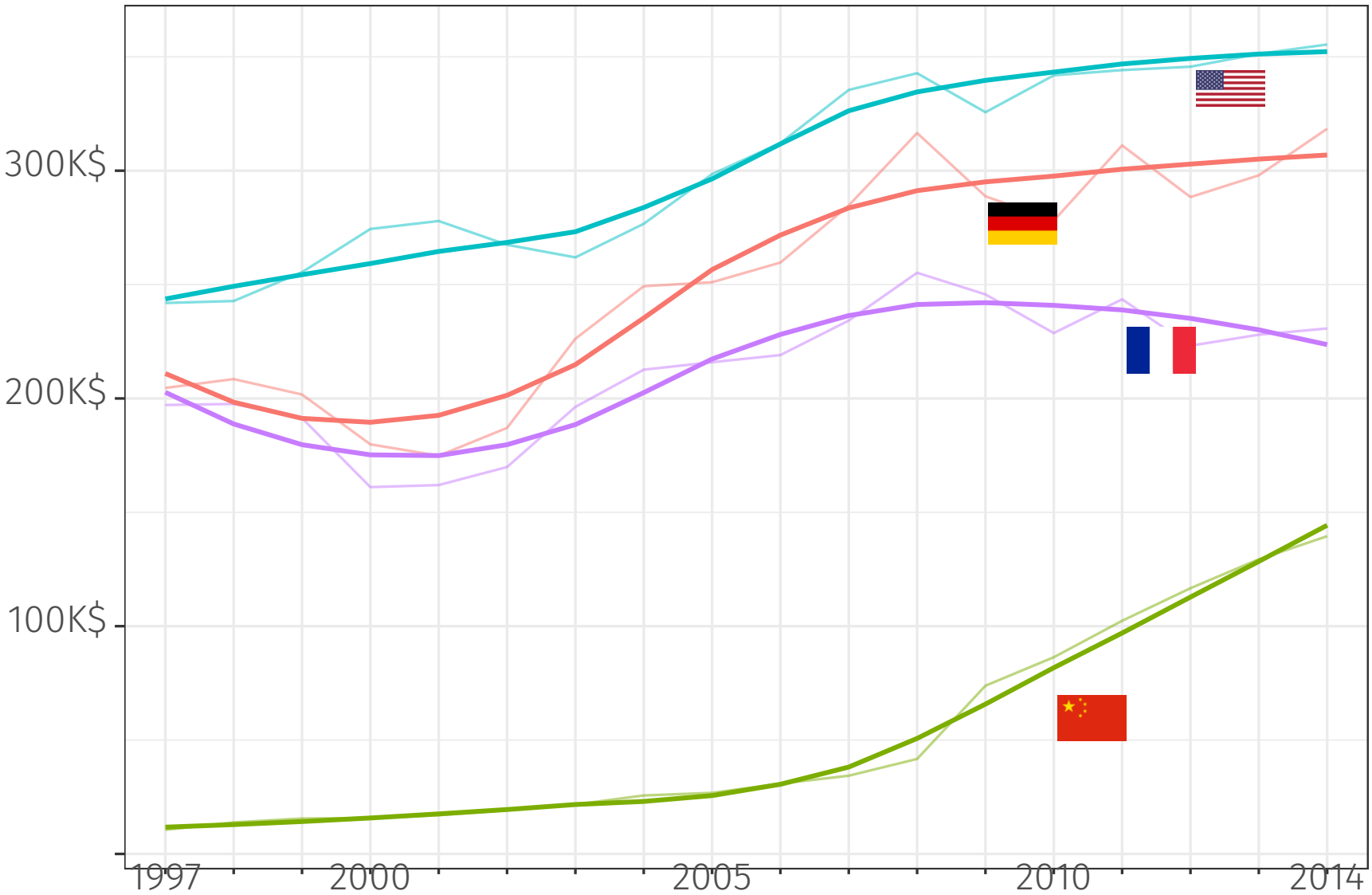
- Total Number of Researchers

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

- Expenditure per Researcher

$$\mathbf{ExpOneRD}(US\$) = \frac{\mathbf{TotExp}(US\$)}{\mathbf{TotRD}},$$

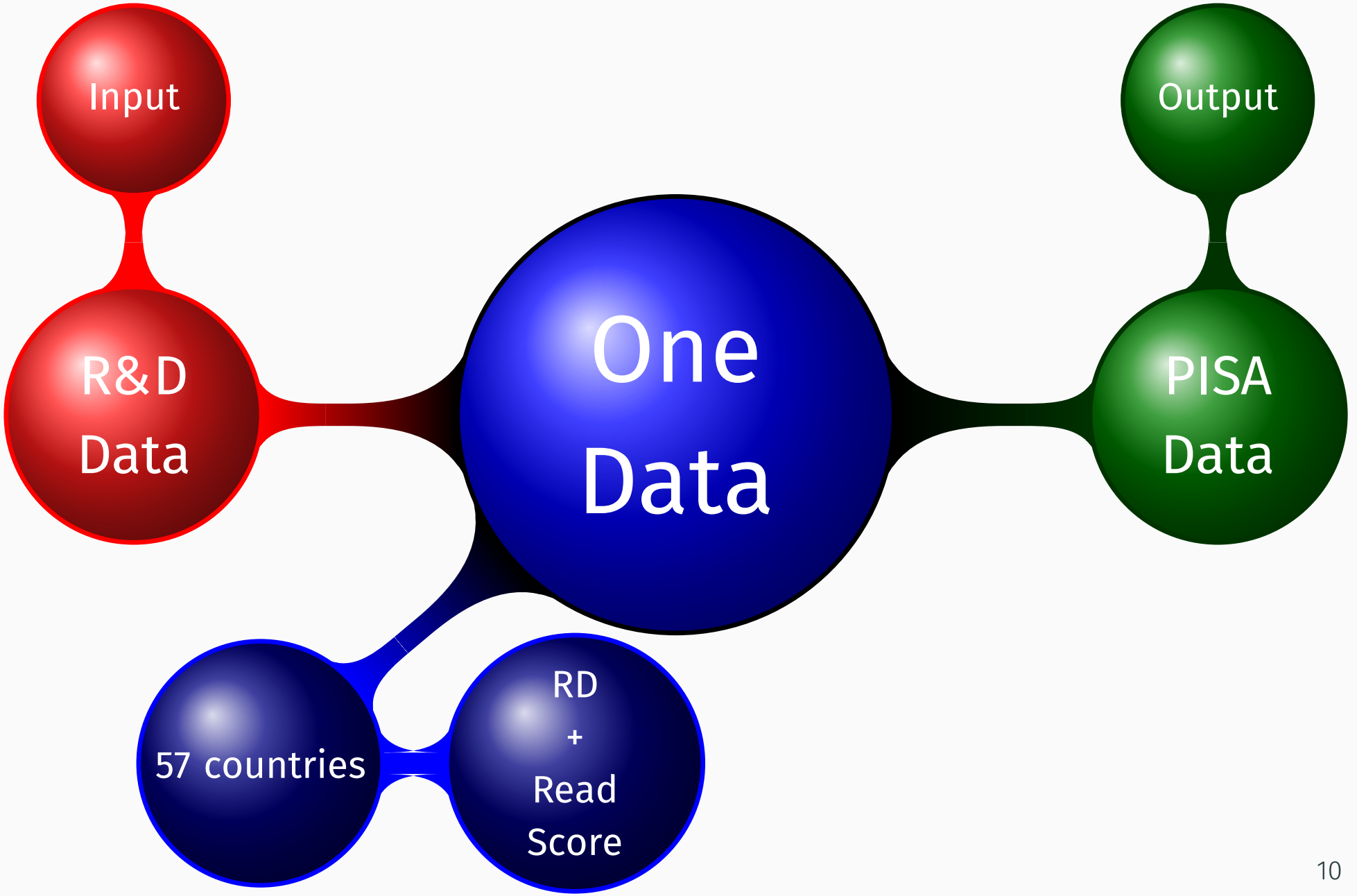
Example



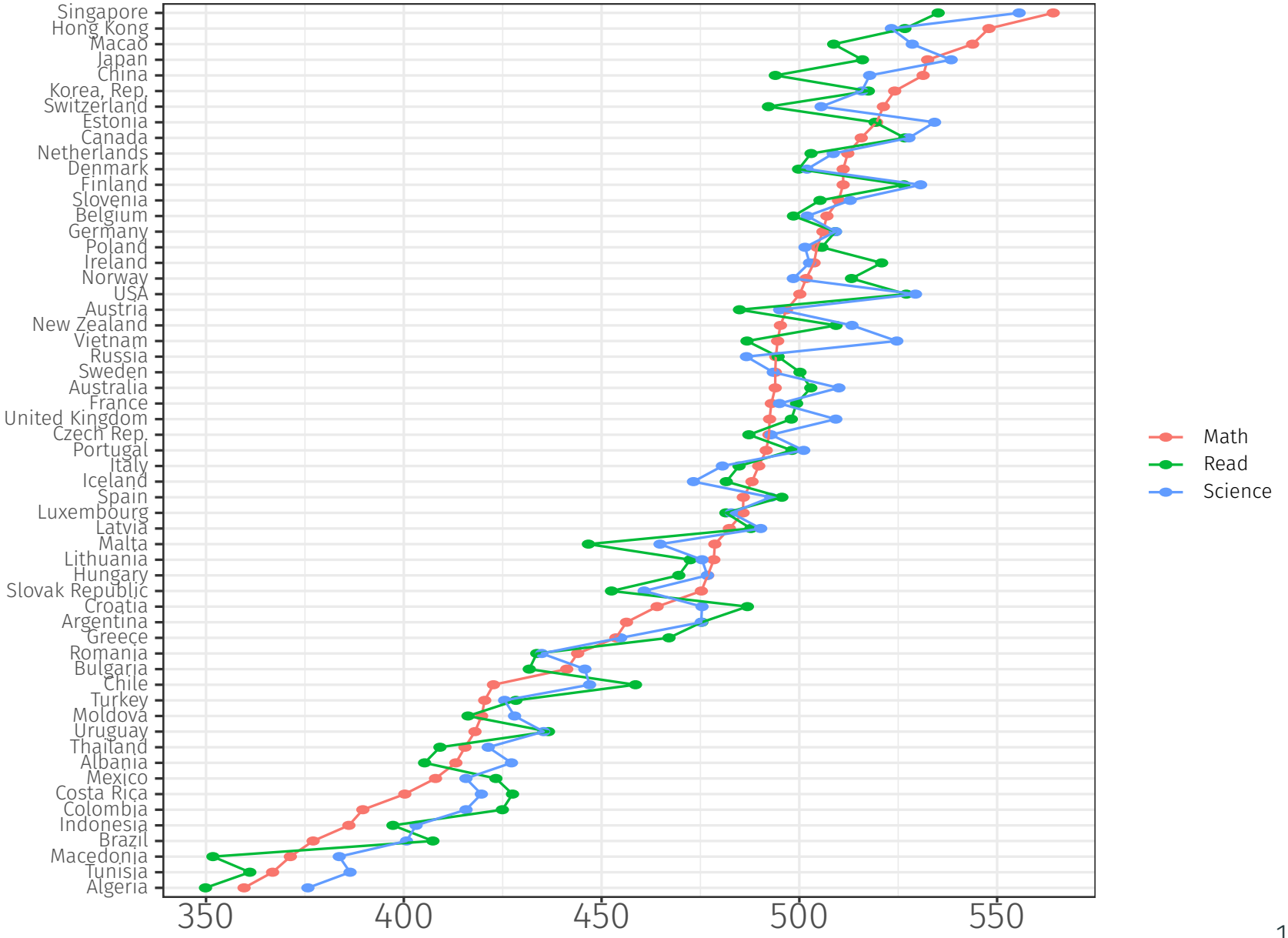
Output Data: PISA 2015

- 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.

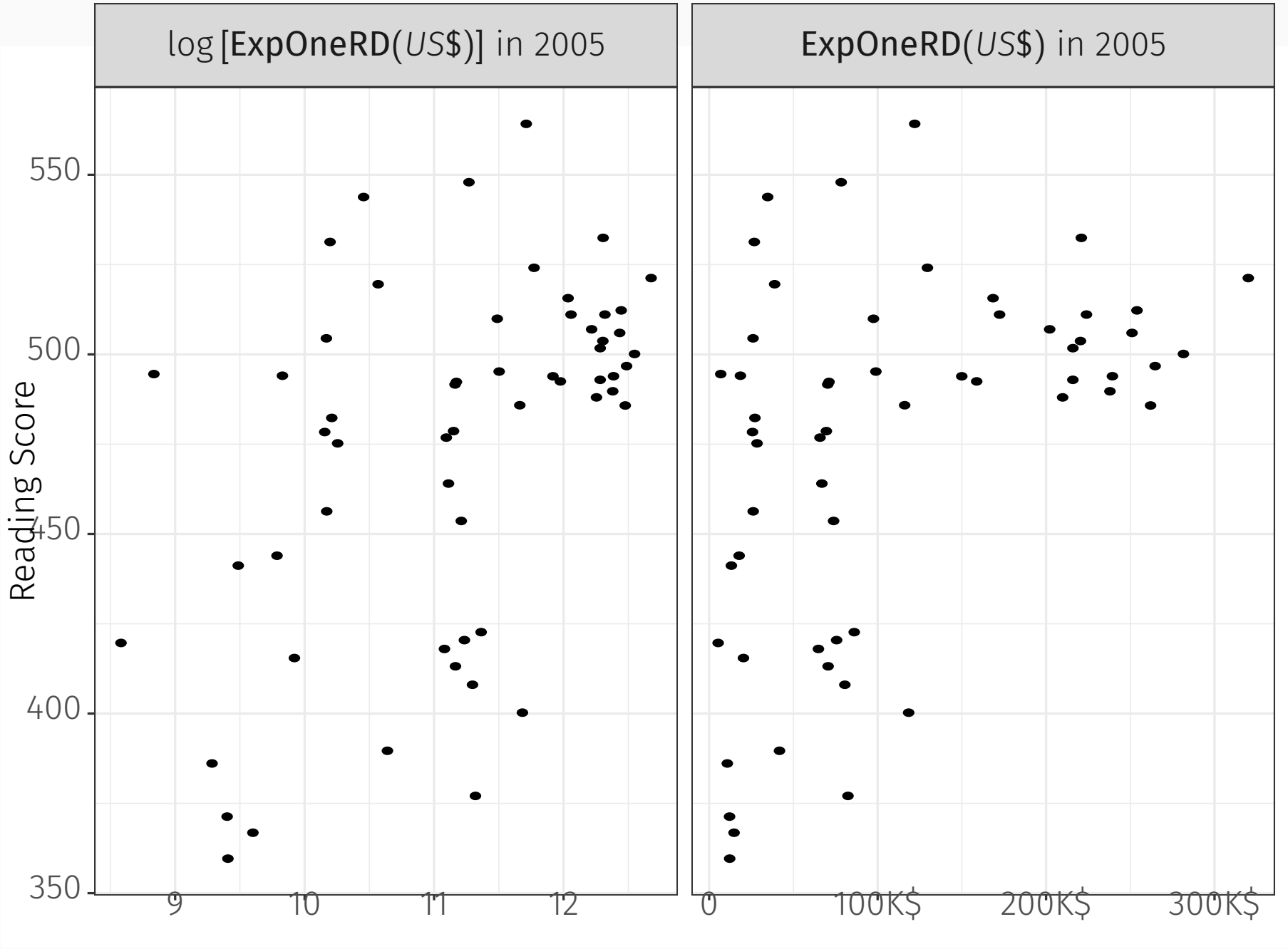
Merging both datasets



PISA scores are correlated



Logarithmic transformation



The final 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 $t = 1997, \ldots, 2014$.

Modeling

- $\exists f$ and $S \subseteq \{1997, \ldots, 2014\}$ such that

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

- f will be estimated using **Gaussian Bayesian Networks** (GBN)
- $\max(S)$ is the lag of the impact of R&D on the Performance of the Education
- We will then:
  - find $\widehat{S}$ an estimation of S
  - be interested if $\widehat{S}$ is $\emptyset$ 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 propriety

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 $p(a) = \{v' \in V, \text{ such that } (v', v) \in E\}$: 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:

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

  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): the learning procedure

Learning Bayesian Network

- The learning procedure is a 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 $2^{\binom{18}{2}} = 1.141798 \times 10^{46}$

Two families of Learning BN

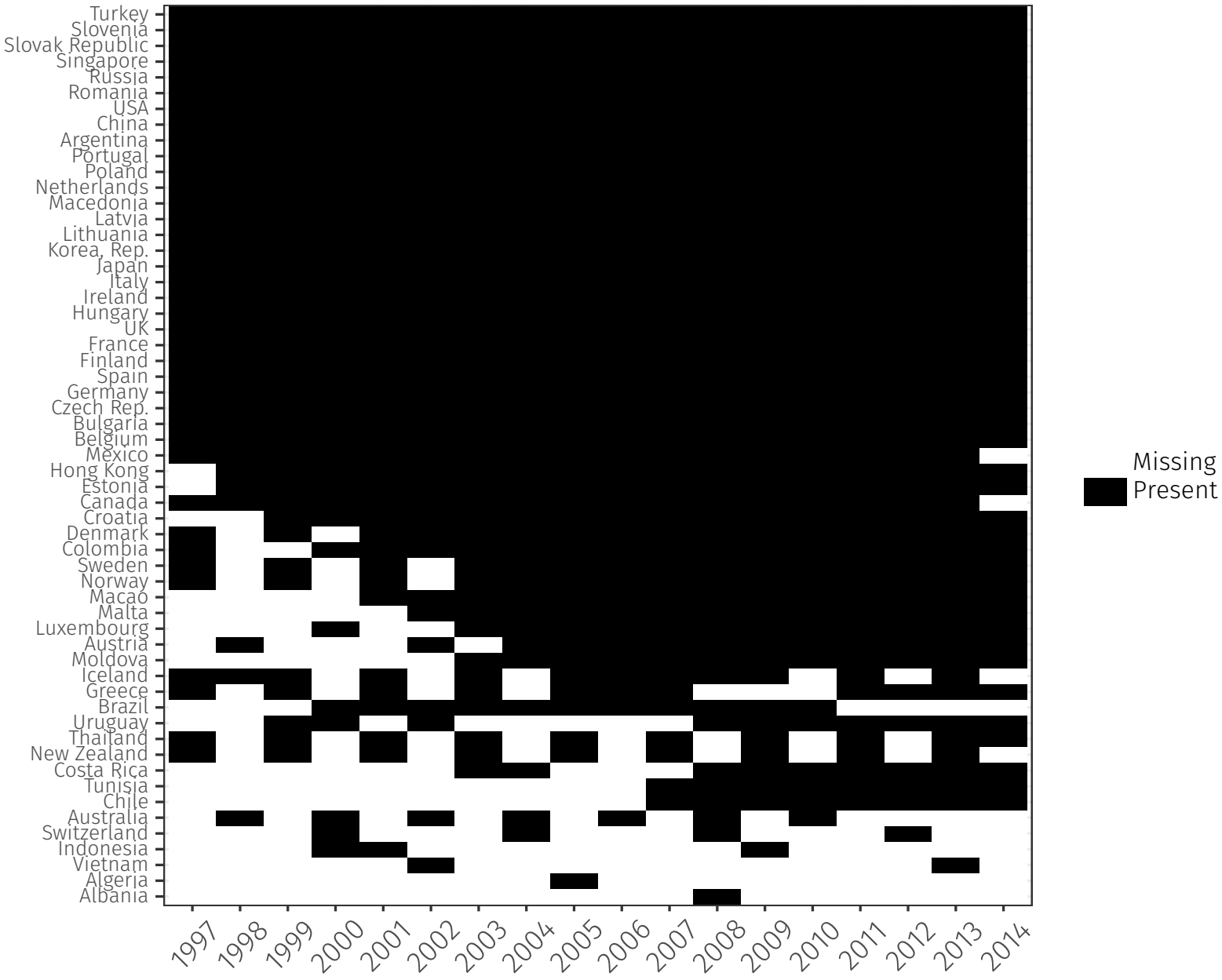
- Constraint-based Algorithms:
  - PC-algorithm
  - `pcalg` package in R
  - Two steps:
    1. Conditional Independence hypothesis testing,
    2. learning directions using the V-structure principal.
  - The final result is a partially directed graph with undirected and directed arrows.
- Score-based algorithms
  - Heuristic optimization techniques in order to search a minimum of a score.
  - Hill-Climbing with random restarts (Bouckaert, 1995).
  - Start from an initial BN and by adding or removing 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).
- `bnlearn` package

Strength of the links

- Bootstrap method (Efron and Tibshirani, 1993)
- 500 of bootstrap replicates and use for every sample the Hill-Climbing algorithm.
- Compute the strength of an edge: the frequency of its presence in each of the 500 estimations.
- Detect the *strongest* link among R&D variables with the Reading PISA score variable: measuring the lag of the impact

No estimation when the data contains missing values!

Missing values in the data



Bayesian Network Iterative Imputation Algorithm

- $\mathcal{X}$ is the R&D data
- n and p are respectively the number of rows and columns in $\mathcal{X}$.
- $\mathcal{S}_T = \{1, \ldots, n\} \times \{1, \ldots, p\}$
- $\mathcal{S}_{NA}$ be the set of indexes of the missing observations in $\mathcal{X}$,

$$\mathcal{S}_{NA} \subset \mathcal{S}_T$$

- $\mathcal{X}_0$ is the complete version of $\mathcal{X}$ using KNN algorithm.

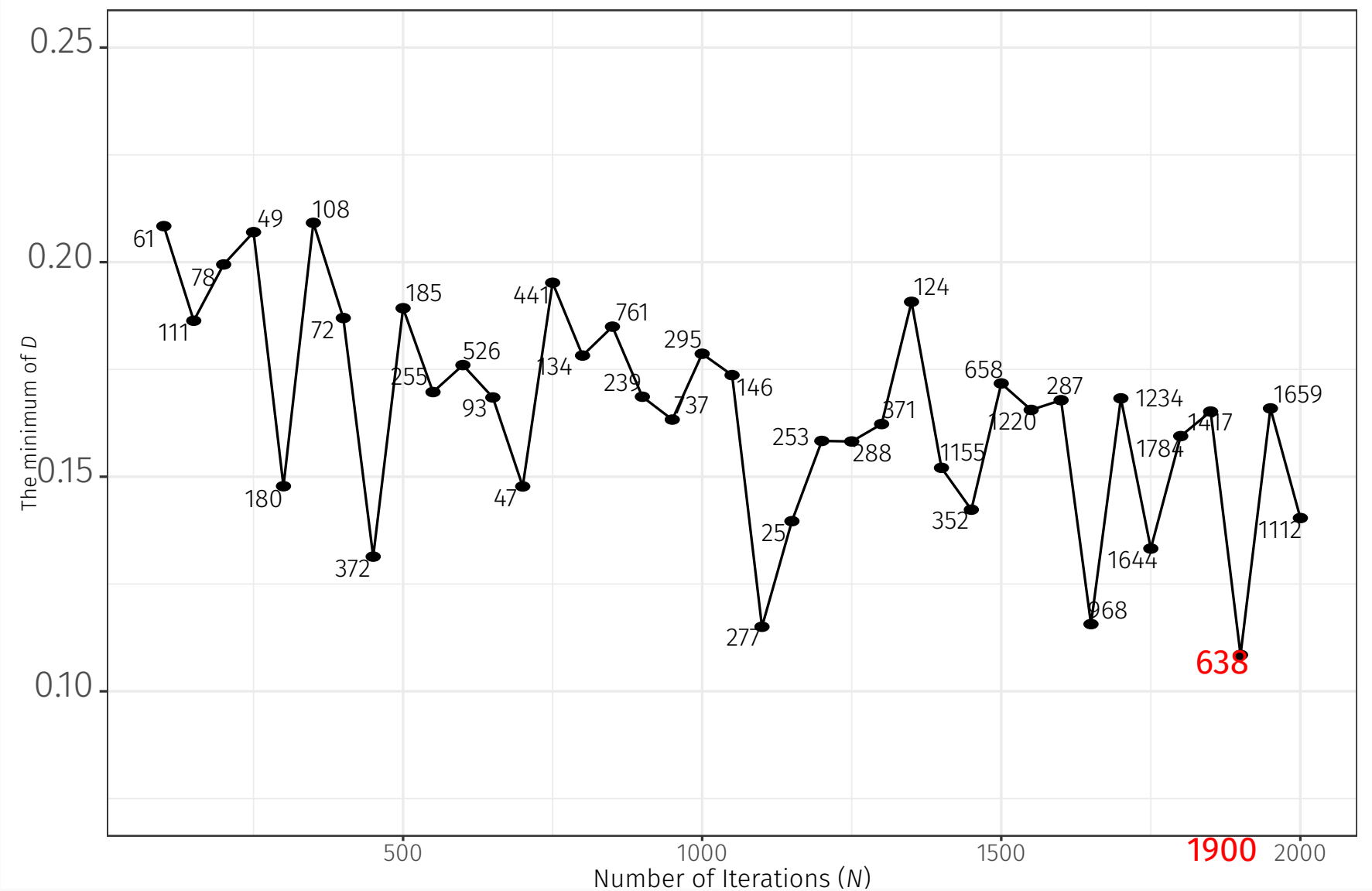
Bayesian Network Iterative Imputation Algorithm (BNII)

let $i = 0$, $\mathcal{X}_0$, $d_0 = 10$, $N = 1000$
while $i \leq N$ **do**
    Sample $\mathcal{S}_r \subset \mathcal{S}_T \setminus \mathcal{S}_{NA}$, $|\mathcal{S}_r| = 50$
    $\mathcal{X}' = \mathcal{X}$ where $\mathcal{X}'[\mathcal{S}_r \cup \mathcal{S}_{NA}] = NA$.
    Estimate and fit the $\widehat{G}$ using HC algorithm from $\mathcal{X}_i$.
    Impute $\mathcal{X}'$ using $\widehat{G}$ and obtain a new complete version of $\mathcal{X}$.
    Let's denote it by $\mathcal{X}_c$.
    Compute

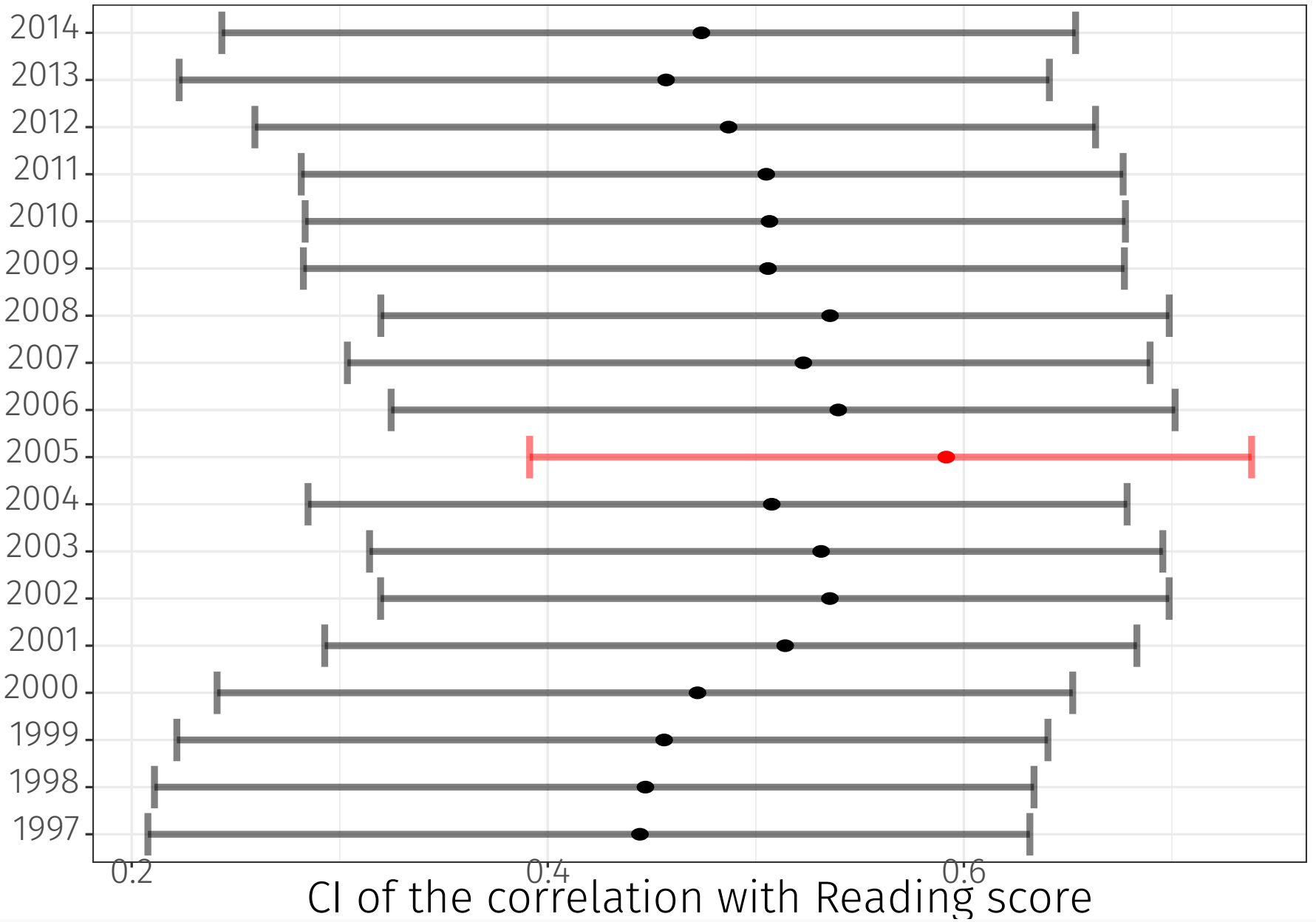
$$D = \sum_{s \in \mathcal{S}_r} (\mathcal{X}(s) - \mathcal{X}_c(s))^2$$

    if $d_i \geq D$ **then**
        $i \leftarrow i + 1$, $d_i \leftarrow D$ and $\mathcal{X}_i = \mathcal{X}_c$.
    else
        $i \leftarrow N + 1$
return $\mathcal{X}_i$ a complete version of the data

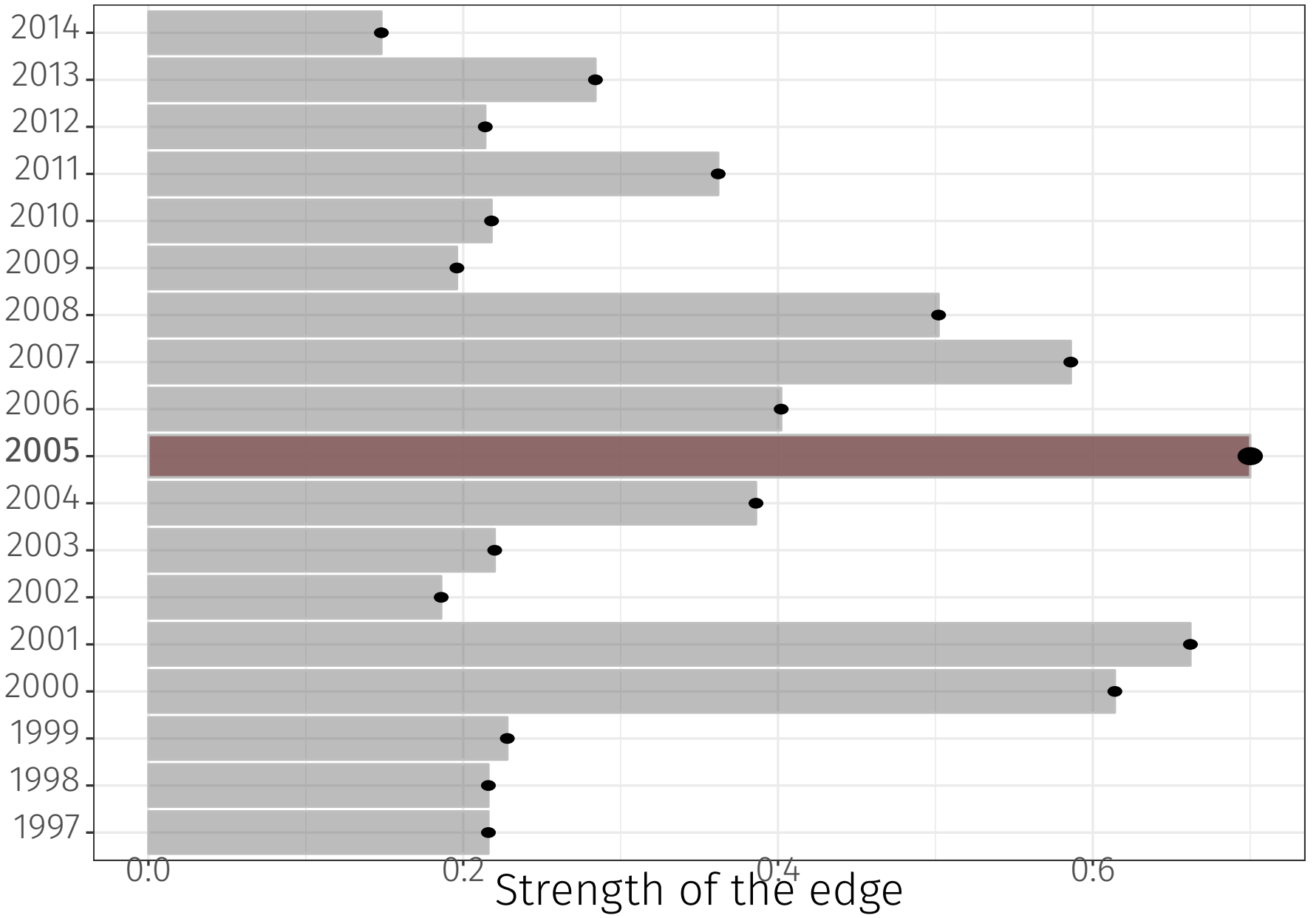
BNII running results



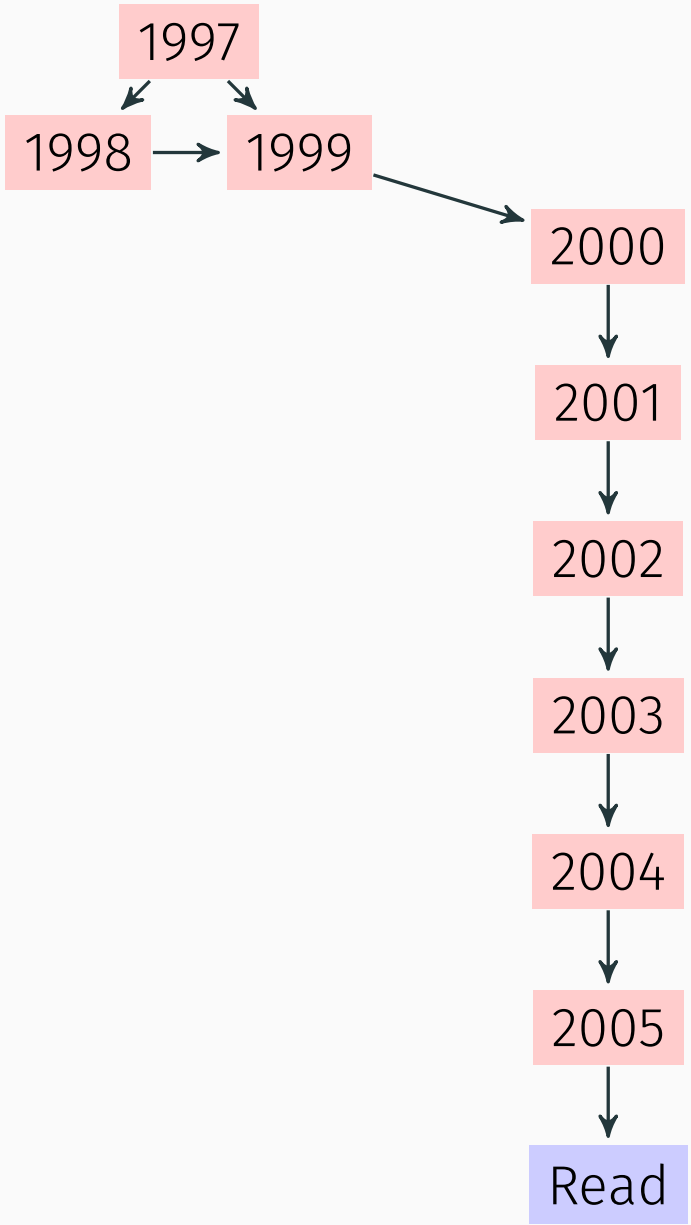
Correlation of R&D. investment with Education Performance



Strength of the edges



Estimated Bayesian Network



Estimated regression models

	Dependent variable:							
	'1998'	'1999'	'2000'	'2001'	'2003'	'2004'	'2005'	Read
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
'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)
Obs.	57	57	57	57	57	57	57	57
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
RSE	0.138	0.077	0.128	0.126	0.153	0.119	0.281	37.542
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***
df	1; 55	2; 54	1; 55	1; 55	1; 55	1; 55	1; 55	1; 55

Note: * p < 0.1; ** p < 0.05; *** p < 0.01

Contribution & Efficiency of the investment in R&D

- Estimated Regression Model

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

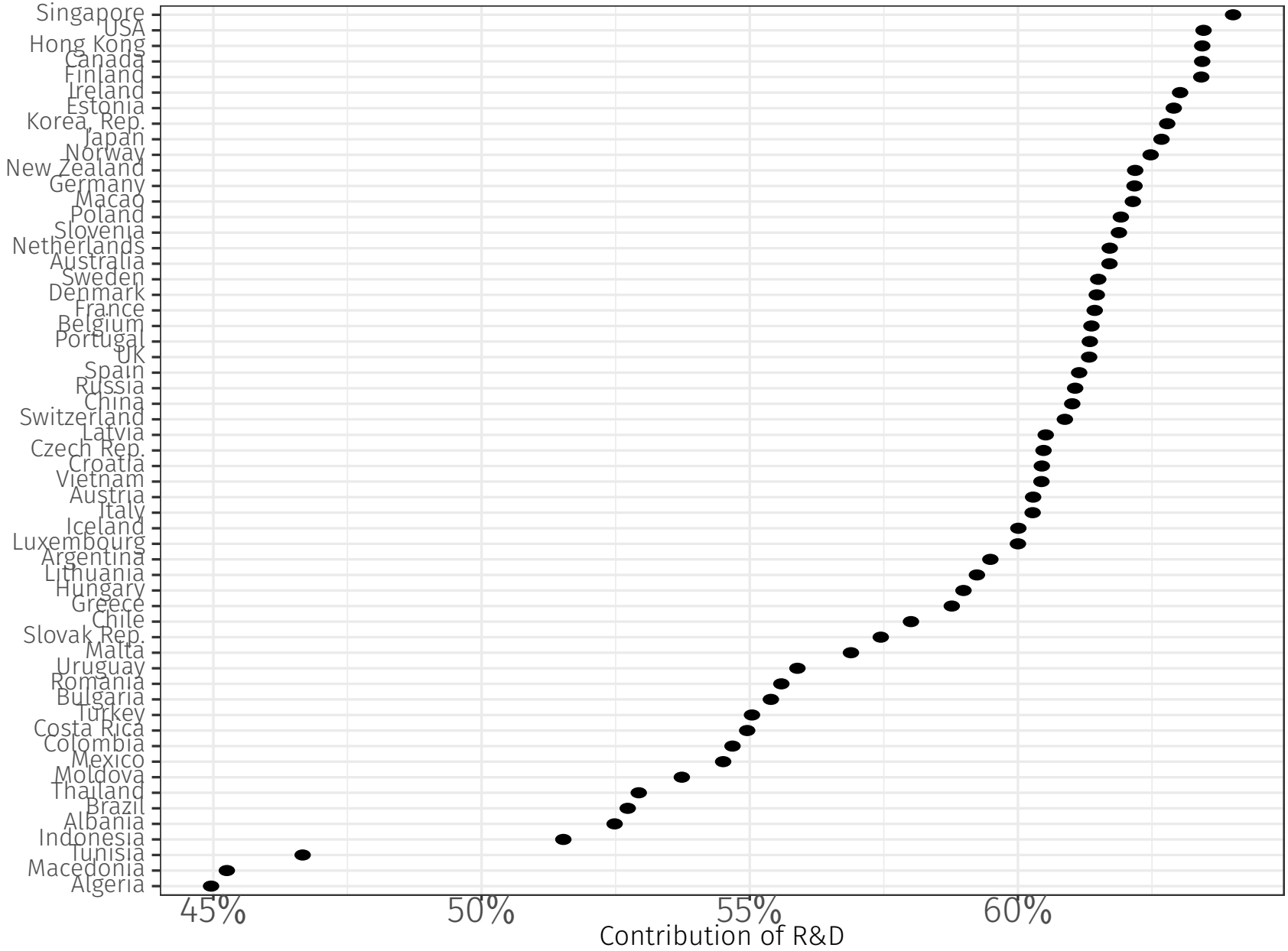
- **Contribution** of the investment in R&D in the explanation of the Performance of the Education System of a country w

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

- **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



Efficiency of the investment in R&D