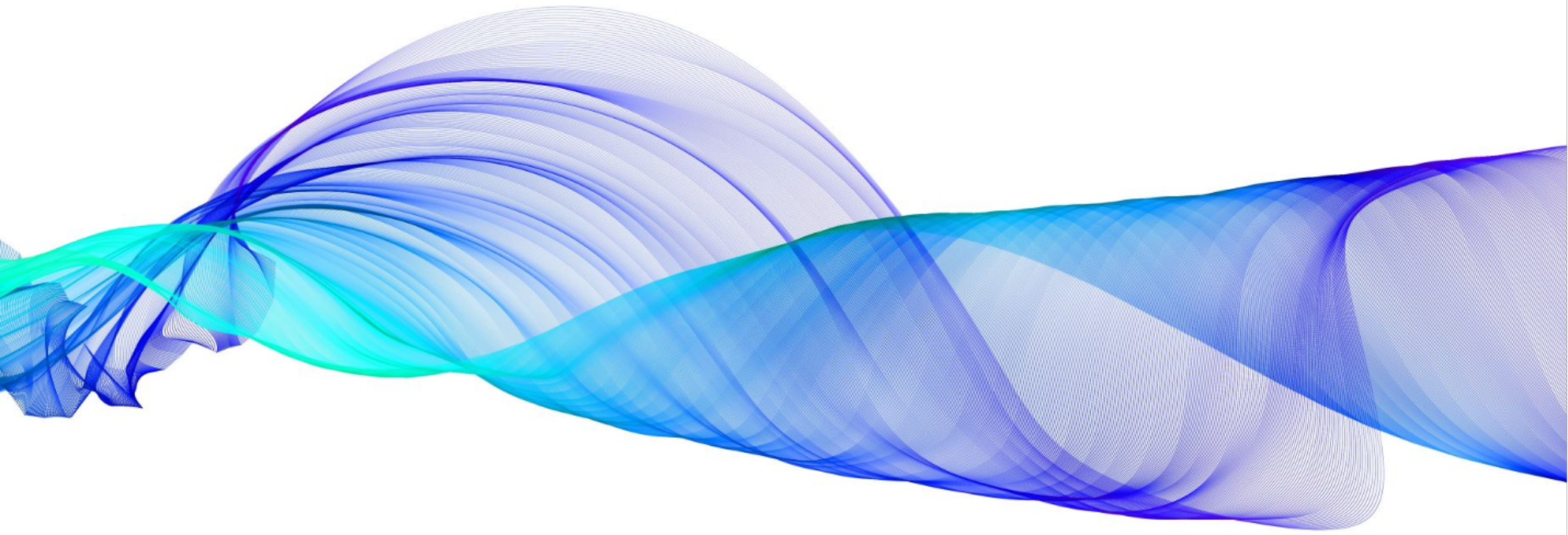


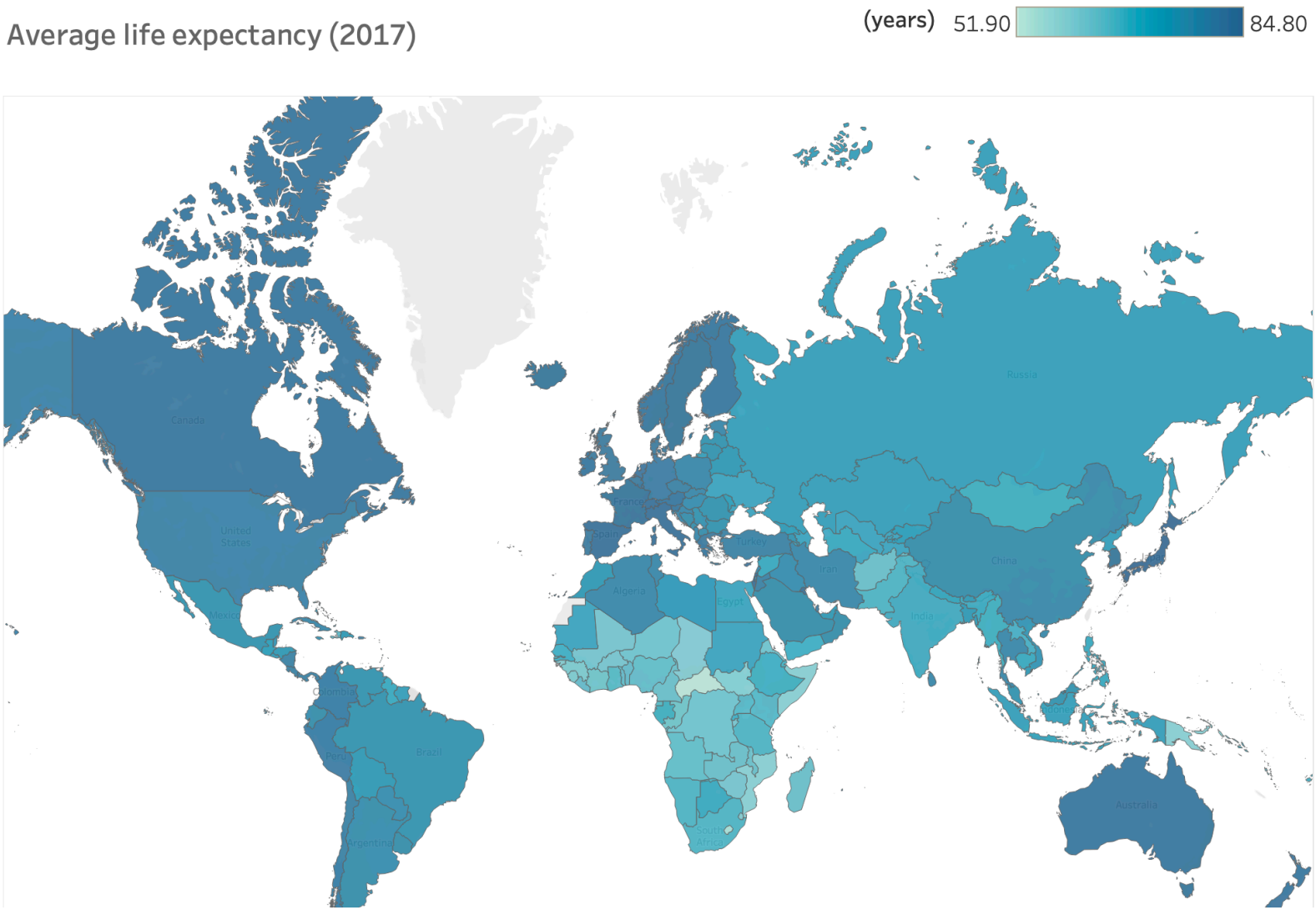
18/12/2020 – DAFT – Final project

Life expectancy across the globe

Jamil Akili

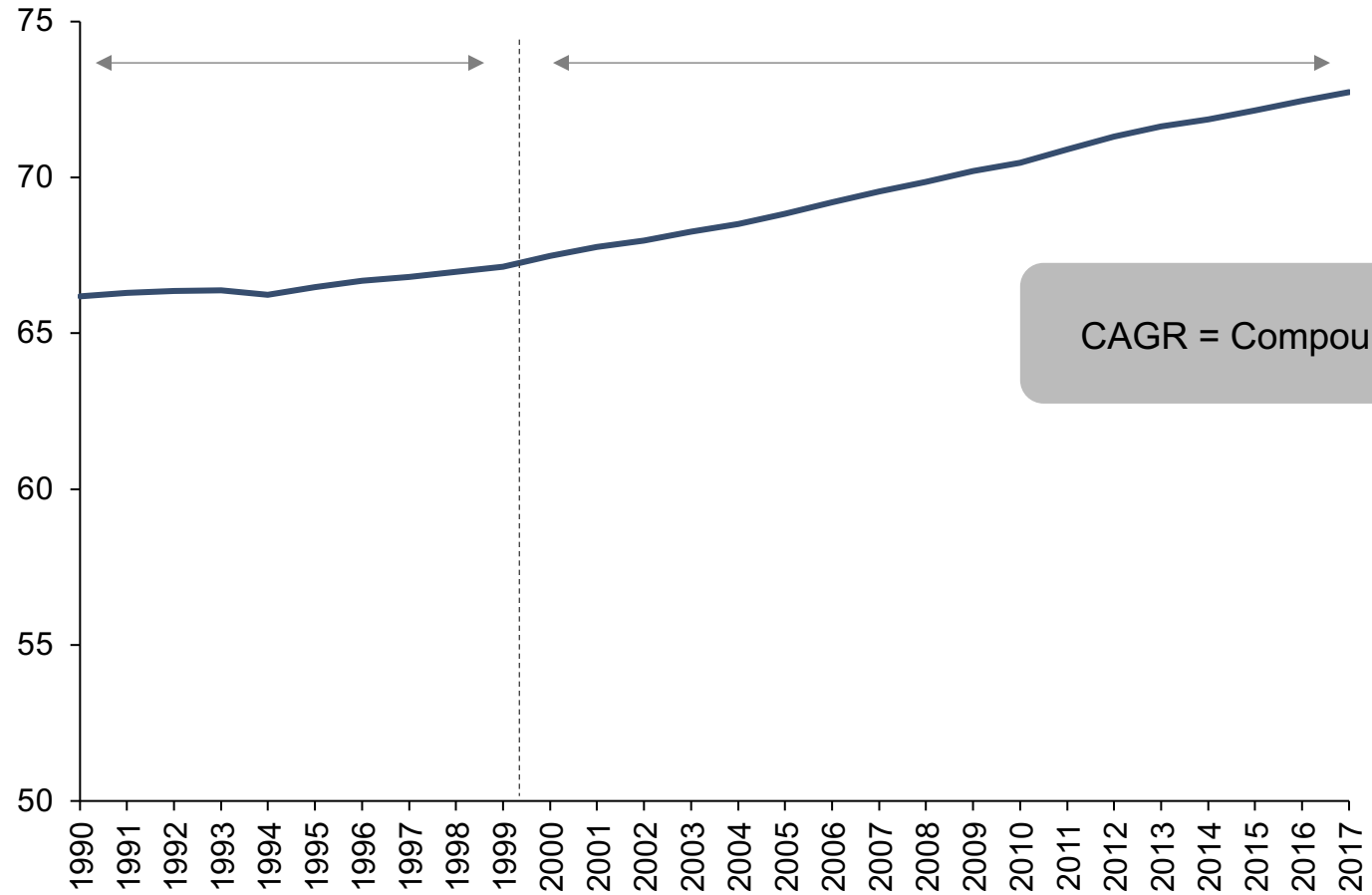
IRON HACK

Global life expectancy values today mostly reflect the gap between developed and developing countries



Global average life expectancy has been growing at c.0.5% p.a. since 2000 after almost stagnating in the 90-ies

Average global life expectancy
years lived

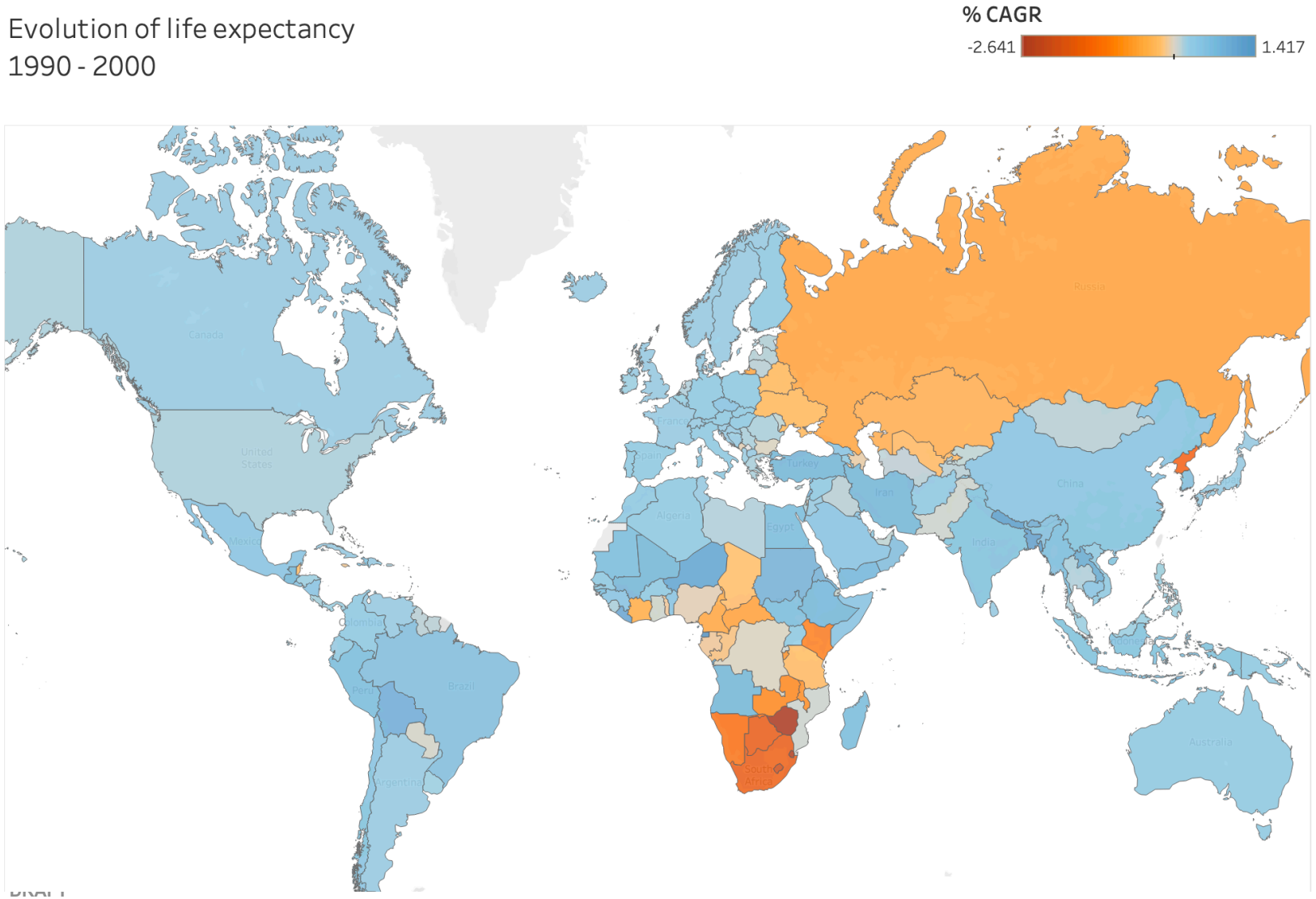


CAGR%	(1990-2000)	(00-17)
	0.19	**0.44**

CAGR = Compound Annual Growth Rate

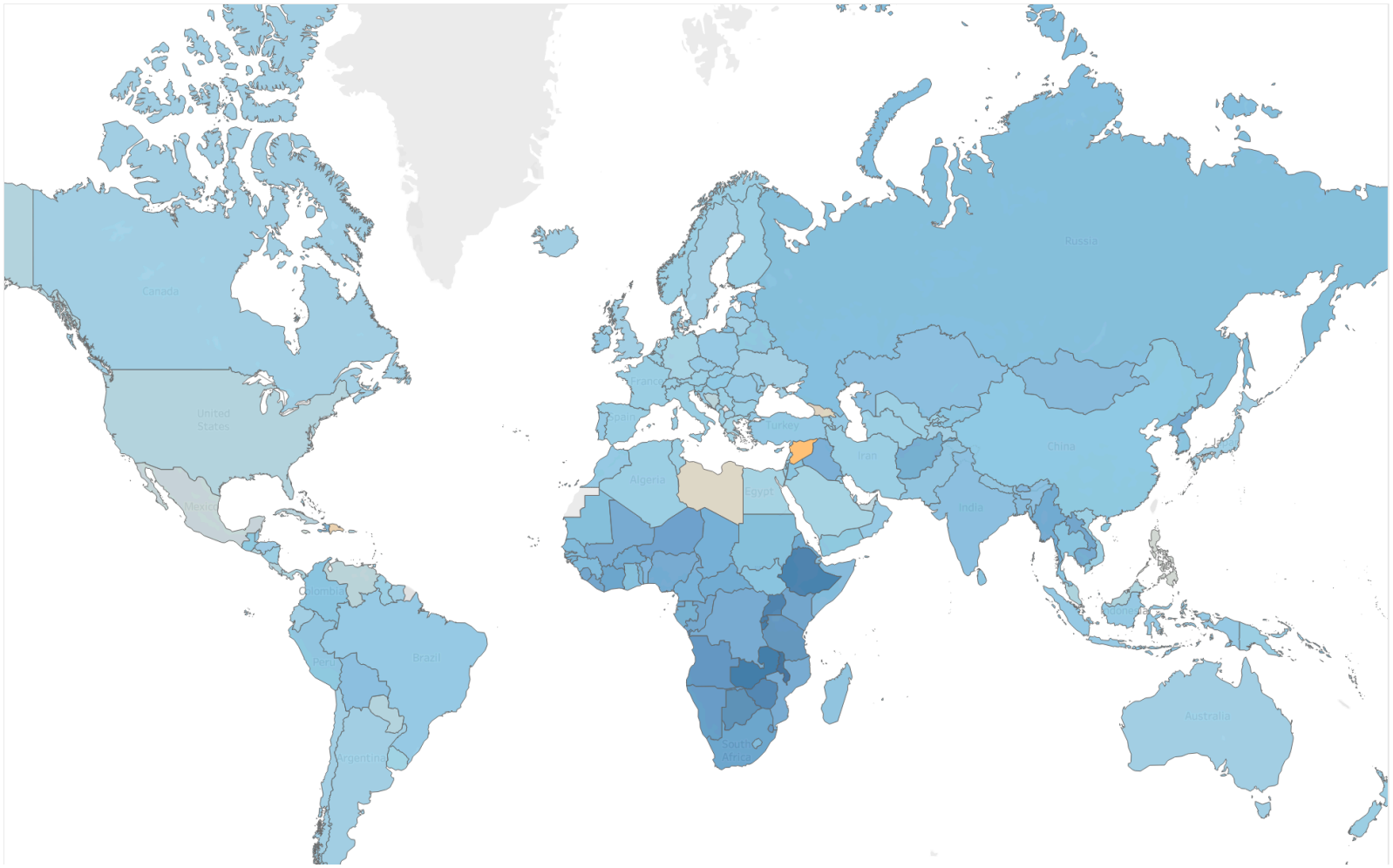
During the 1990s, average life expectancy was still decreasing in many countries...

Evolution of life expectancy
1990 - 2000



… but the trend has been postive across the globe since 2000

Evolution of life expectancy
2000 - 2017



Project goal

What are the main factors influencing life expectancy?

Is it possible to use machine learning to predict the impact of public policies on life expectancy?

Initial approach: work with the kaggle database

- Silver medal on Kaggle, referenced on several websites focused on machine learning datasets
- c.3000 rows and 22 columns of data
  - Each row is a combination of a country and a year (2000 – 2015)

Country	Year	Status	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepatitis B	Measles	BMI	under-five deaths	Polio	Total expenditure	Diphtheria	HIV/AIDS	
Afghanistan	2015	Developing	65.0	263.0	62	0.01	71.279624	65.0	1154	19.1	83	6.0	8.16	65.0	0.1	5
Afghanistan	2014	Developing	59.9	271.0	64	0.01	73.523582	62.0	492	18.6	86	58.0	8.18	62.0	0.1	6
Afghanistan	2013	Developing	59.9	268.0	66	0.01	73.219243	64.0	430	18.1	89	62.0	8.13	64.0	0.1	6
Afghanistan	2012	Developing	59.5	272.0	69	0.01	78.184215	67.0	2787	17.6	93	67.0	8.52	67.0	0.1	6
Afghanistan	2011	Developing	59.2	275.0	71	0.01	7.097109	68.0	3013	17.2	97	68.0	7.87	68.0	0.1	

  - Columns are comprehensive

```
['country', 'life_expectancy', 'adult_mortality', 'infant_deaths',
 'alcohol', 'percentage_expenditure', 'hepatitis_b', 'measles', 'bmi',
 'under-five_deaths', 'polio', 'total_expenditure', 'diphtheria',
 'hiv/aids', 'gdp', 'population', 'thinness__1-19_years',
 'thinness_5-9_years', 'income_composition_of_resources', 'schooling'],
```

However, this dataset turned out to be unusable

- c.20% null values per columns on average, but also many inconsistencies:

Country	Year		percentage expenditure
Afghanistan	2009	.01	56.76221682
Afghanistan	2010	.01	79.67936736
Afghanistan	2011	.01	7.097108703
Afghanistan	2012	.01	78.1842153
Afghanistan	2013	.01	73.21924272
Afghanistan	2014	.01	73.52358168
Afghanistan	2015	.01	71.27962362
Albania	2006	.31	3.3021542
Albania	2010	.28	41.82275719
Albania	2009	.79	348.0559517
Albania	2003	.29	14.71928882
Albania	2004	.54	221.8428
Albania	2005	.16	26.99312143
Albania	2002	.73	104.5169157

Country	Year	HIV/AIDS	GDP	I
Algeria	2000	0.1	1757.17797	
Algeria	2010	0.1	4463.39468	
Algeria	2011	0.1	5432.2523	
Algeria	2012	0.1	5564.82566	
Algeria	2013	0.1	5471.86677	
Algeria	2015	0.1	4132.76292	
Algeria	2014	0.1	547.8517	
Algeria	2003	0.1	294.33556	
Algeria	2002	0.1	1774.33673	
Algeria	2001	0.1	1732.85798	
Algeria	2006	0.1	3464.6179	
Algeria	2007	0.1	3935.18334	
Algeria	2008	0.1	495.254866	
Algeria	2009	0.1	3868.83123	
Algeria	2005	0.1	31.122378	
Algeria	2004	0.1	2598.9823	

Country	Year	Population
Argentina	2015	43417765
Armenia	2000	369588
Armenia	2006	29585
Armenia	2001	35655
Armenia	2009	2888584
Armenia	2005	2981259
Armenia	2003	31786
Armenia	2002	333897
Armenia	2010	2877311
Armenia	2007	293356
Armenia	2004	3612
Armenia	2011	2875581
Armenia	2013	289359
Armenia	2012	2881922
Armenia	2015	291695
Armenia	2008	29822
Armenia	2014	29622
Australia	2000	19153

- Requires code to detect errors
  - function returning a unique row per country, measuring "*(max – min)/mean*" over the available timeframe

The Kaggle database is unusable

Alternative: building a dataset from scratch and running a multiple linear regression on it

- Constructing dataset from different sources, with a focus on 1990-2017
- Merged datasets used include:
  - Life expectancy, total population, GDP per capita
  - Access to basic sanitation, access to water, doctors per 1 000, child mortality
  - HIV, measles, TB

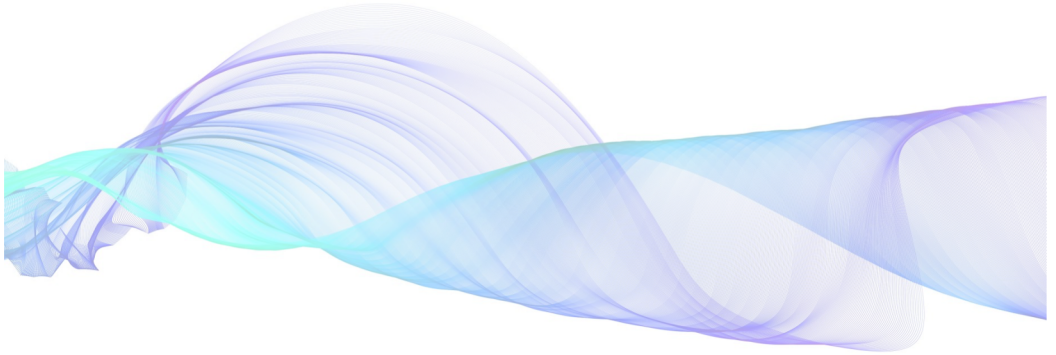
- New dataset (5460 rows × 12 columns)
- Good accuracy for regression models

	model	R2	R2A	MSE	RMSE	MAE
0	linreg	0.90	0.87	8.11	2.85	2.07
1	lasso	0.90	0.87	8.29	2.88	1.99
2	ridge	0.91	0.88	7.69	2.77	2.00
3	elastic	0.90	0.87	8.30	2.88	1.99
4	rfr	0.77	0.70	18.57	4.31	2.41

Key takeaways

- Always pay close attention to data source and collection method (time series)
- Relatively limited influence on life expectancy of indicators related economic development (i.e. *GDP per capita*, *doctors available per 1000 pop*)
- Much higher impact for simpler factors (e.g. *access to basic sanitation*, *access to clean water)* : **UN programs are efficient**
- Explore the impact of other factors like *schooling level*, *poverty/inequality index*, or *public funding of healthcare* → **next step**
- Finalize database by assigning data to missing values (depending on the number of values already available for a given year & country) → **next step**

Thank you!



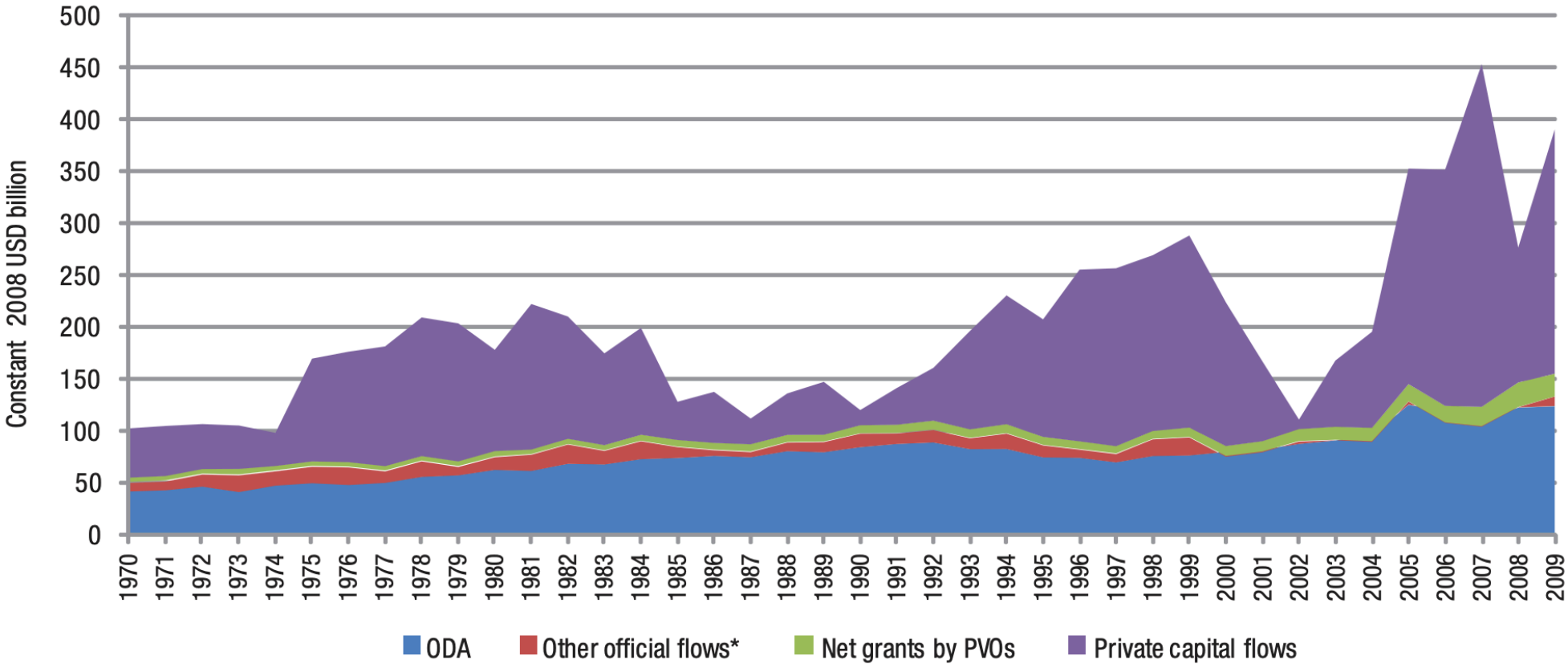
References:

- Kaggle : https://www.kaggle.com/kumarajarshi/life-expectancy-who
- Gapminder : https://www.gapminder.org/data/
- OECD : https://data.oecd.org/healthstat/life-expectancy-at-birth.htm
- WHO : https://www.who.int/data/gho/gho-search

IRON HACK

Annex

Figure 1. DAC members' total net resource flows to developing countries, 1970-2009



* Net OOF flows were negative in 2000-01, 2004 and 2006-07.