BILL & MELINDA GATES foundation

HEALTHY BIRTH, GROWTH & DEVELOPMENT

knowledge integration

Country Segmentation: Tool and Application

Dave King and Sofia Trommlerova

October 2nd, 2017

exaptive

ERASMUS UNIVERSITY ROTTERDAM

Population & Surveillance Data Integration (PSDI)

Sofia Trommlerová
Postdoctoral Researcher
Erasmus University Rotterdam

Dave King
Founder, Ceo
Exaptive, Inc.

Matt Coatney
VP, Services
Exaptive, Inc.

Matthias Rieger
Assistant Professor
Dept. of Economics –
Erasmus University
Rotterdam

Nicholas Kassebaum
Assistant Professor
IHME –
Univ. of Washington

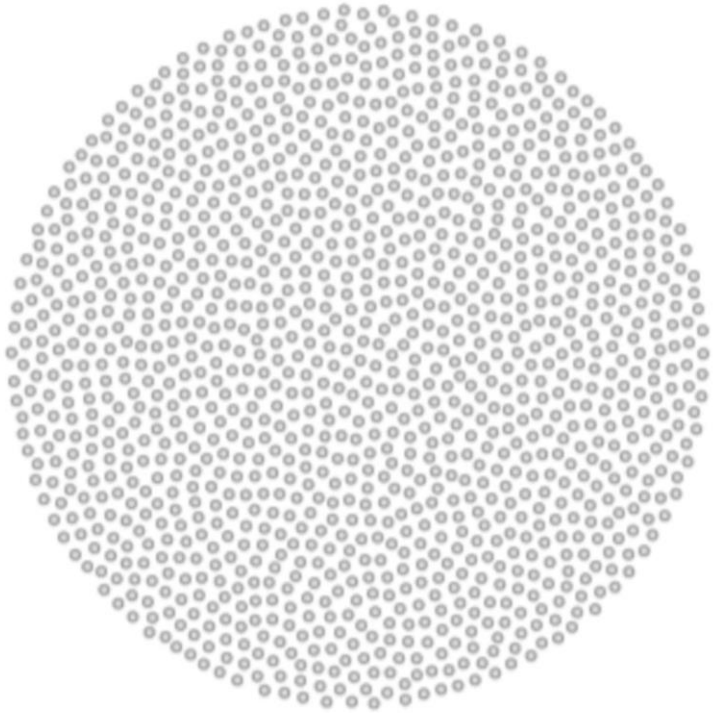
David Rice
Software
Developer
Exaptive, Inc.

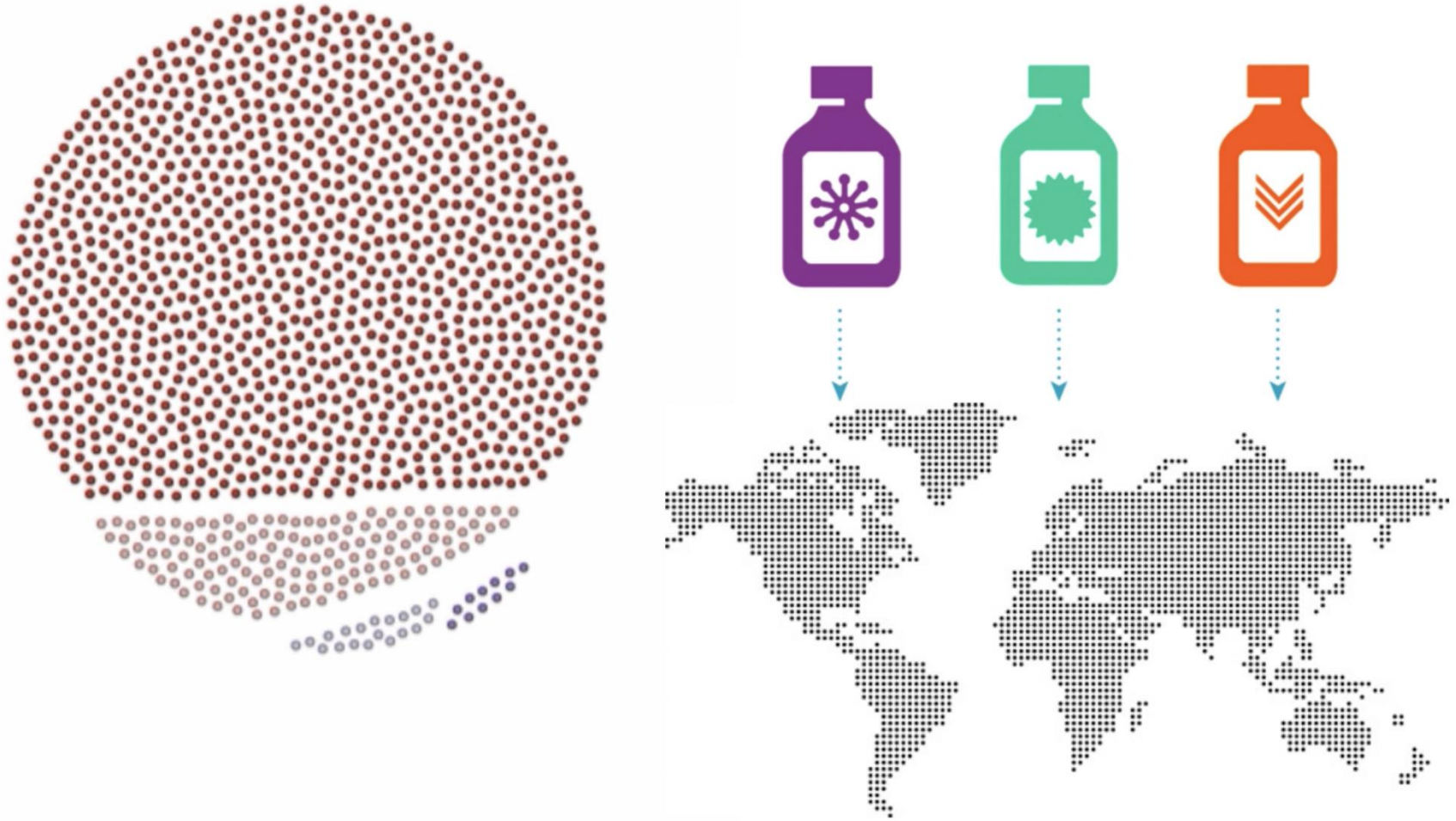
Mark Wissler
Software
Developer &
Architect
Exaptive, Inc.

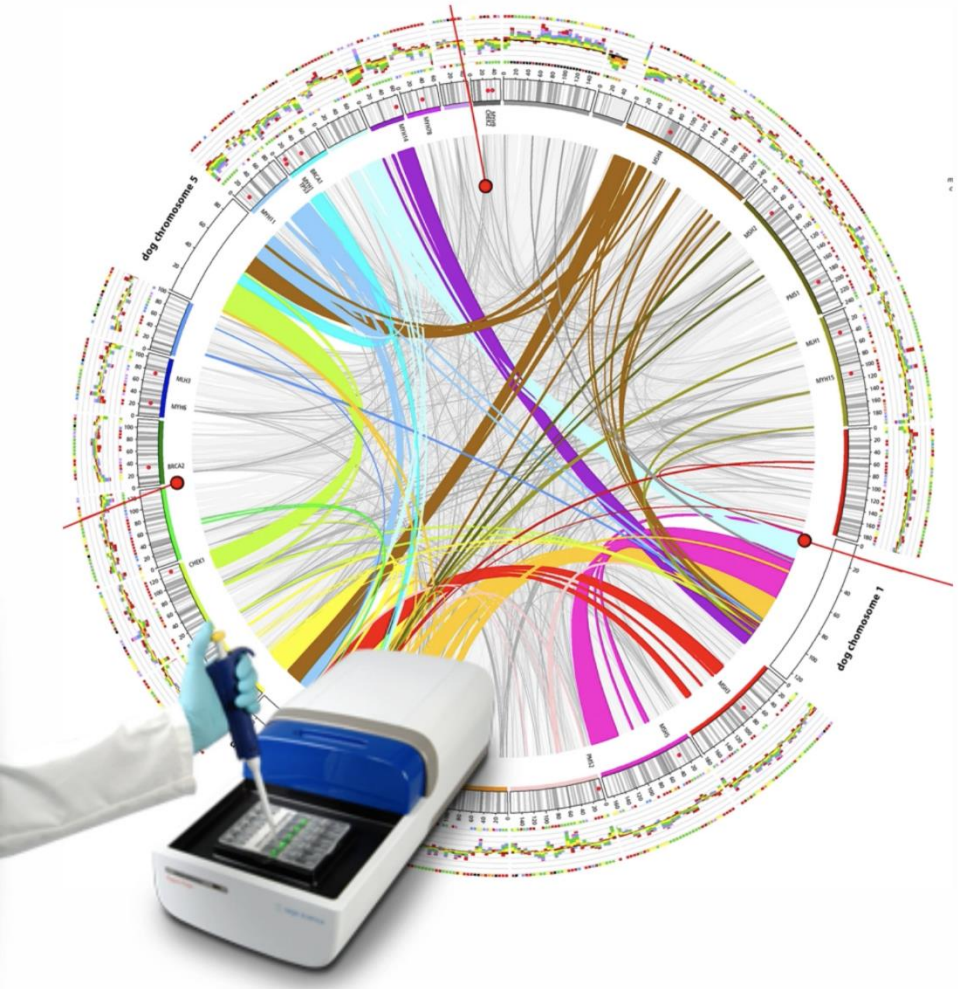
Sofia Trommlerová
Postdoctoral
Researcher
Erasmus University
Rotterdam

HBGDki

The segmentation tool






dog chromosome 5
dog chromosome 1

GenotypePhenotype.jpg (696 × | Gates Country Genotype Explo × | Dave

Secure | https://labs.exaptive.city/x/gates/CountryGenotyping

Cluster Mode | Distance Mode

Use a curated dataset

DHS, IPUMS, NIMS, FAO and IYCF

Include stunting-related variables in analysis

Download Dataset

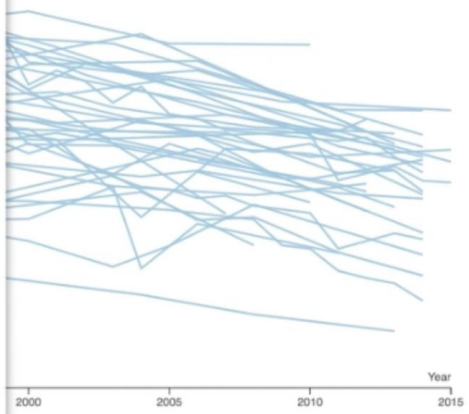
OR

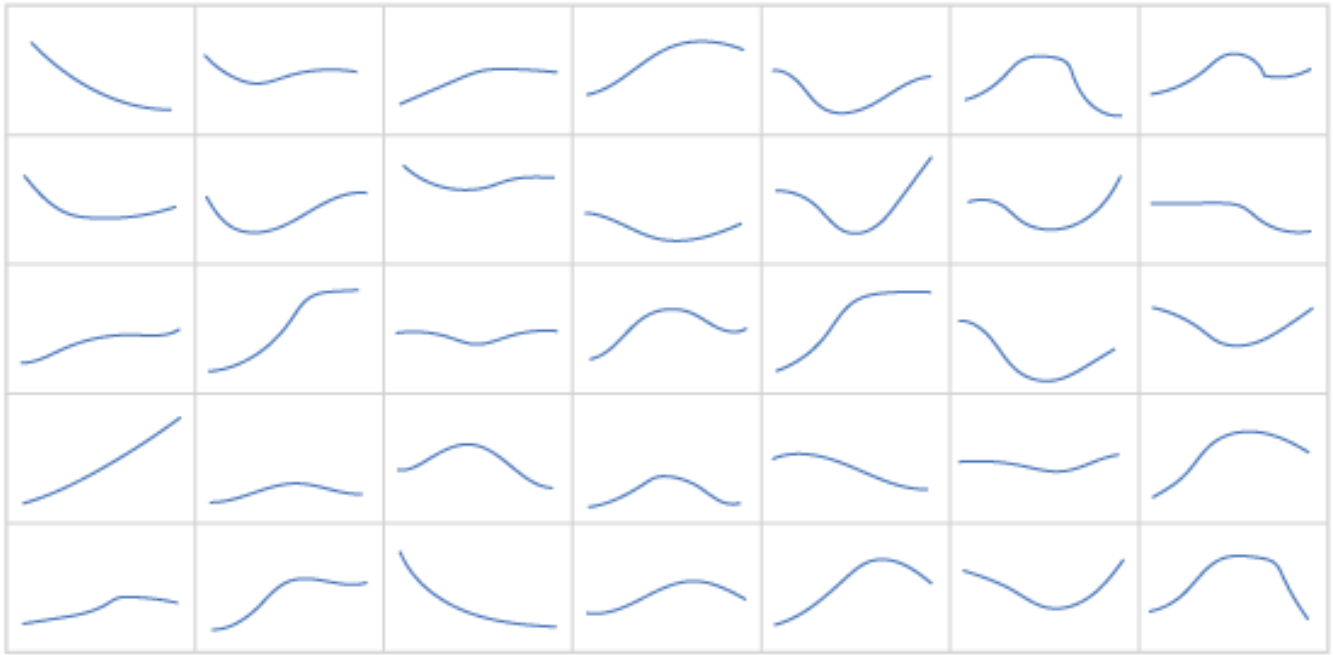
Use your own dataset

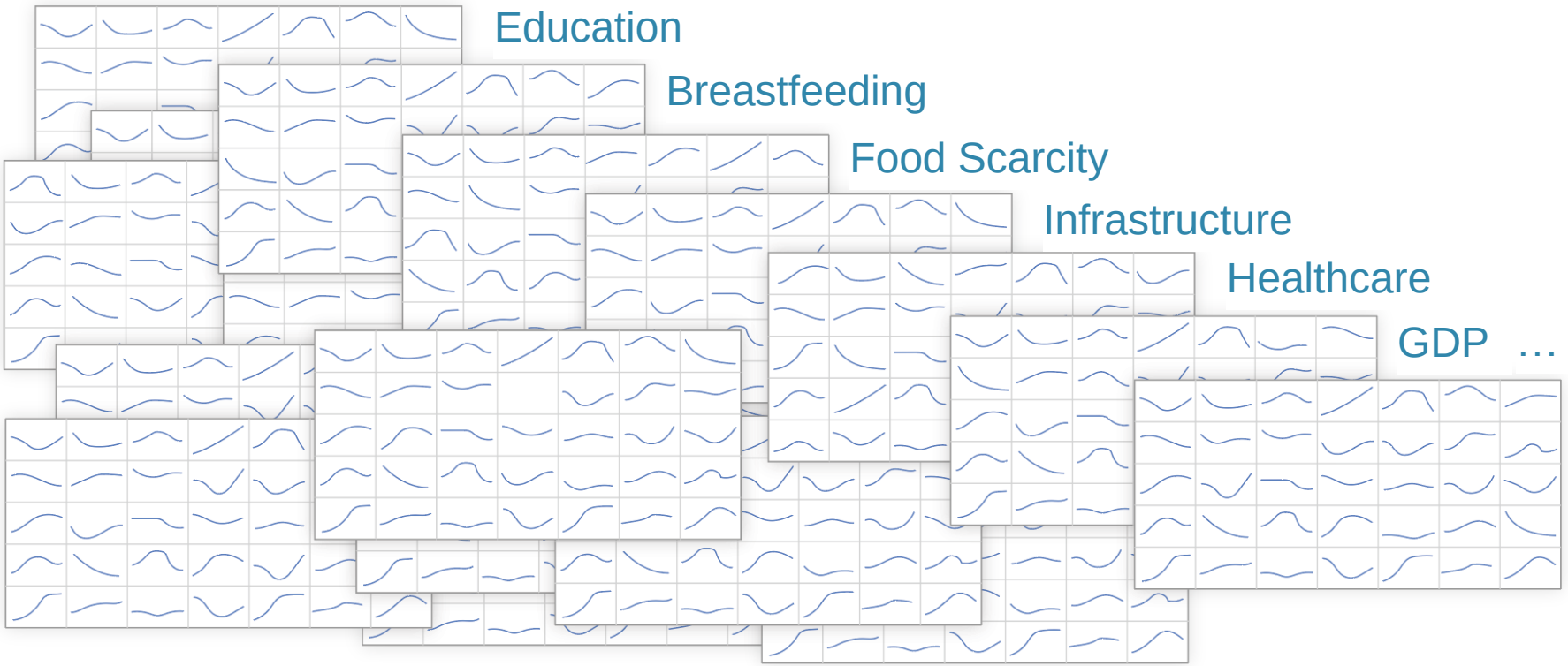
DROP FILE TO UPLOAD (OR CLICK)

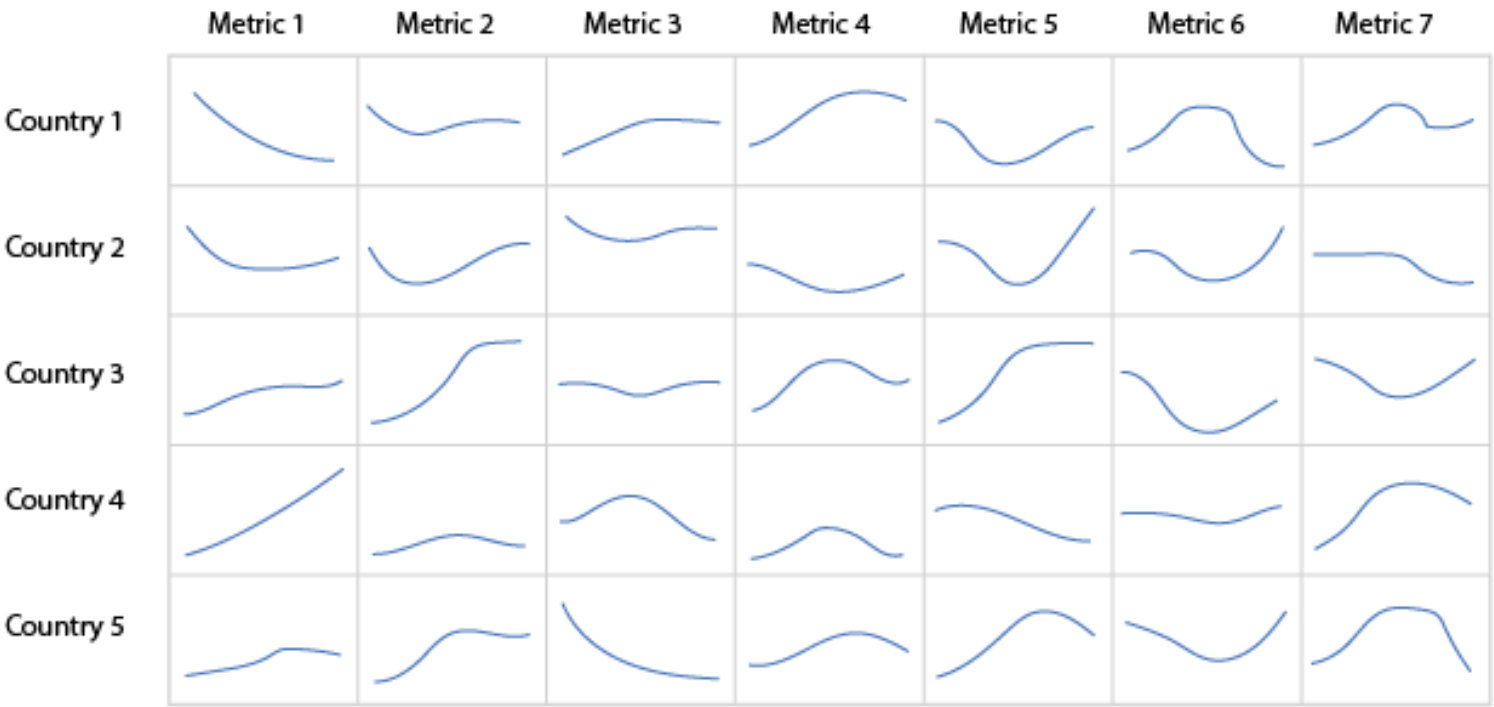
▶ RUN CLUSTERING

Clear Filters

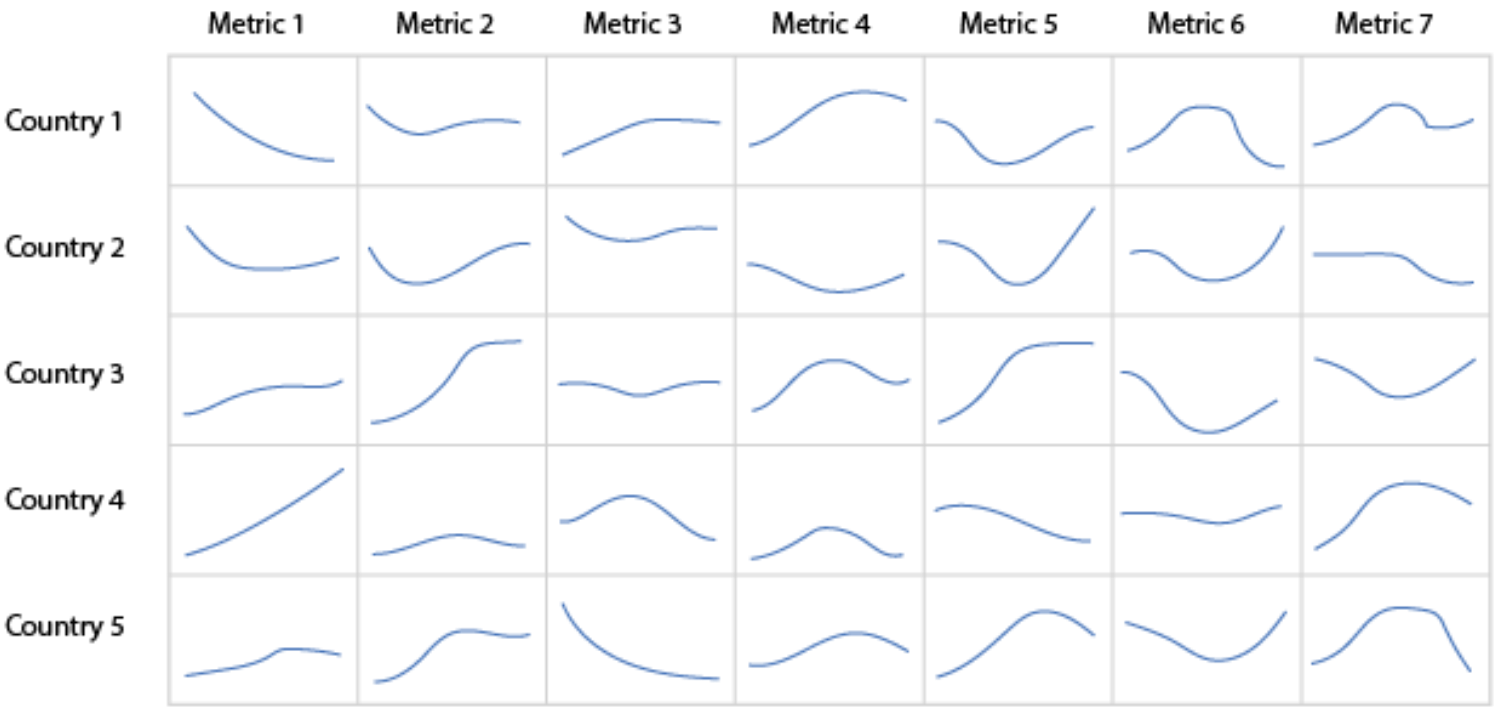


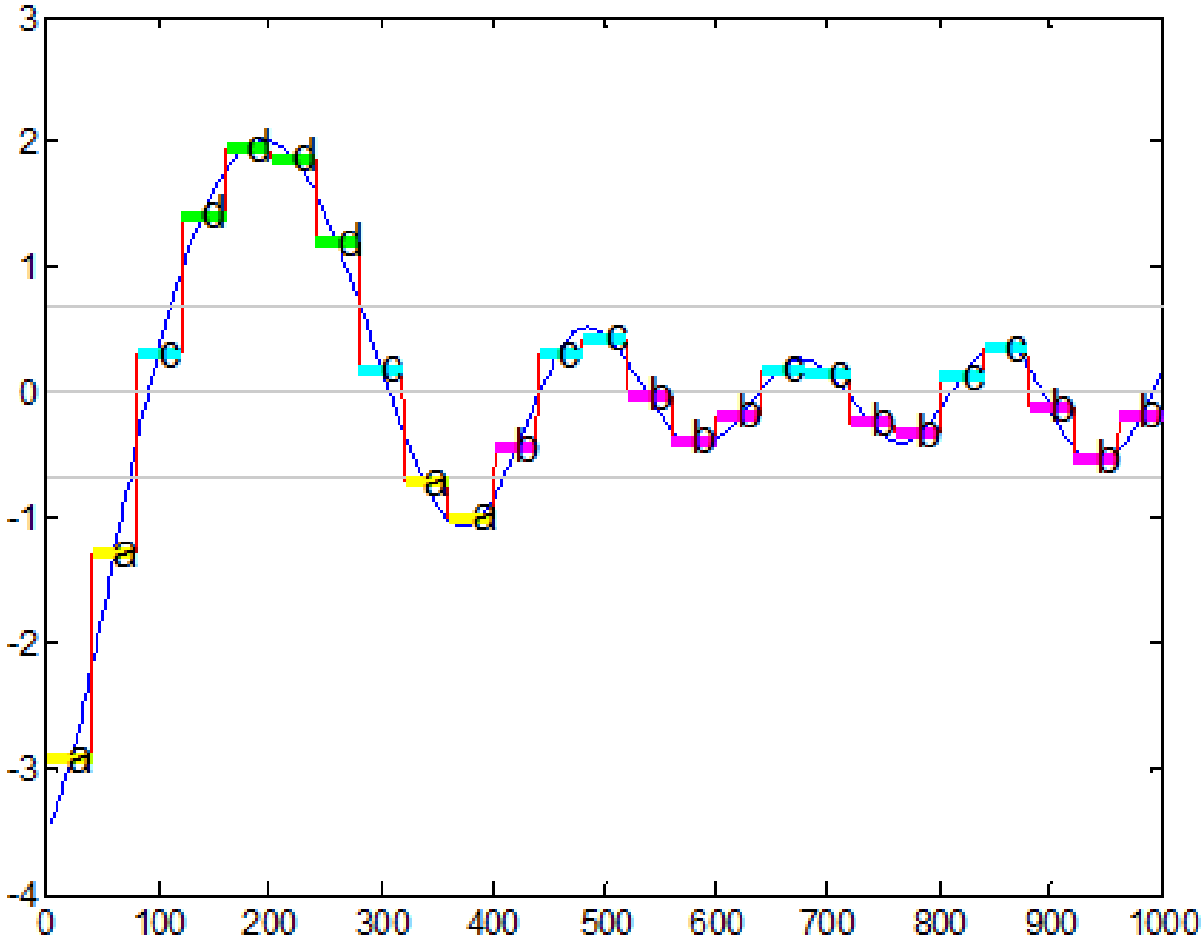


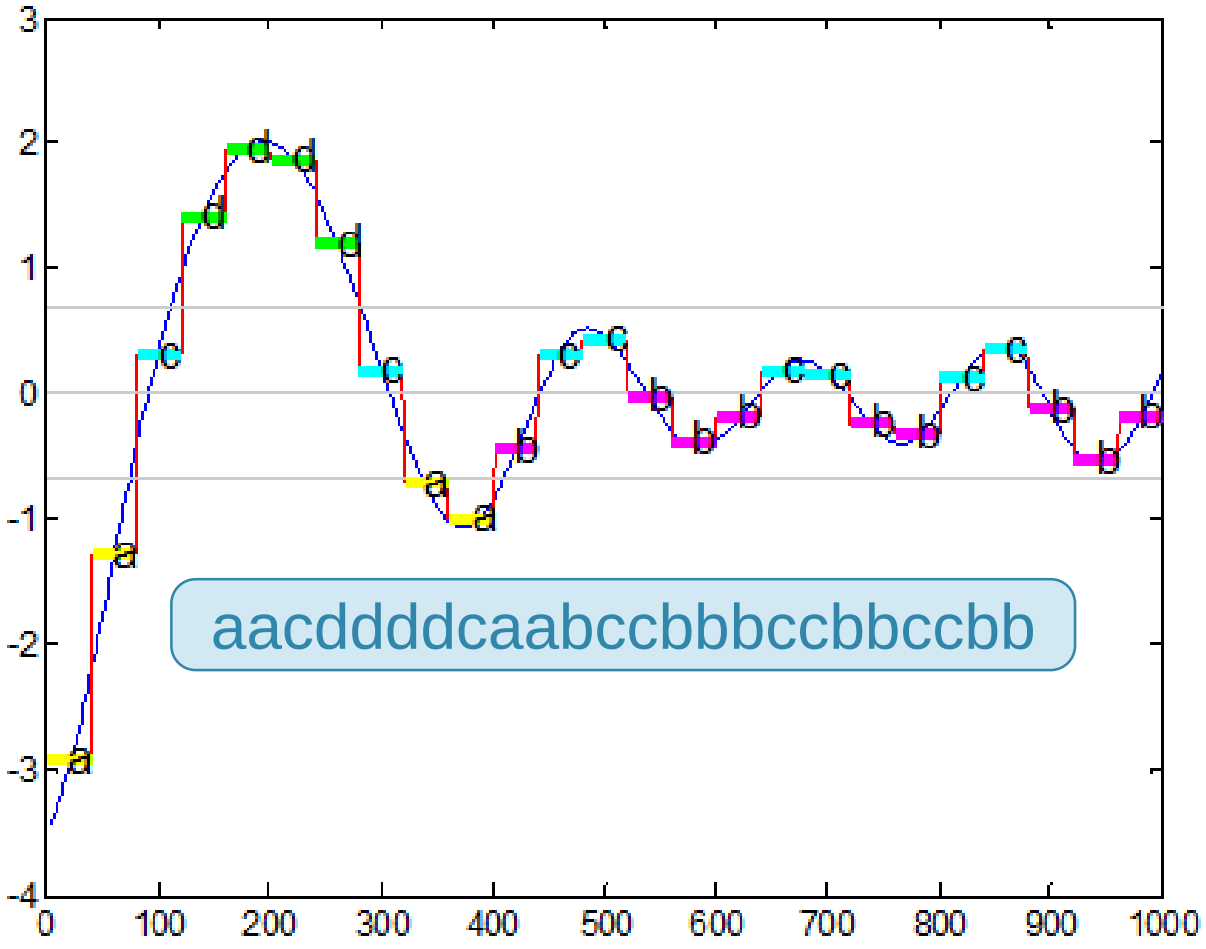

Education
Breastfeeding
Food Scarcity
Infrastructure
Healthcare
GDP
…


Metric 1
Metric 2
Metric 3
Metric 4
Metric 5
Metric 6
Metric 7
Country 1
Country 2
Country 3
Country 4
Country 5

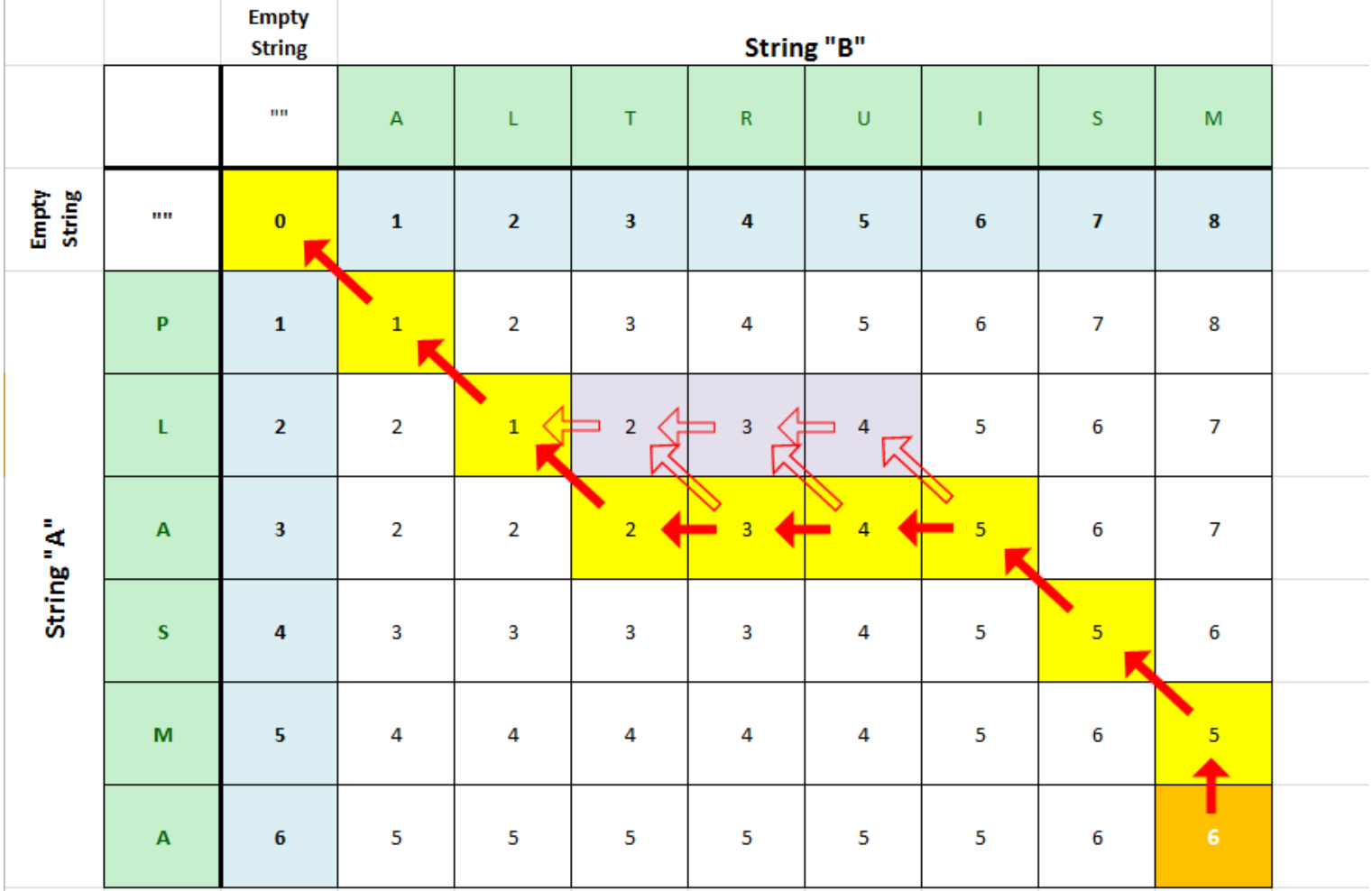
	Metric 1	Metric 2	Metric 3	Metric 4	Metric 5	Metric 6	Metric 7
Country 1	1	3	5	3	4	1	5
Country 2	2	2	3	1	5	1	2
Country 3	2	3	4	2	4	1	2
Country 4	1	2	1	3	3	2	3
Country 5	3	1	4	5	5	4	4

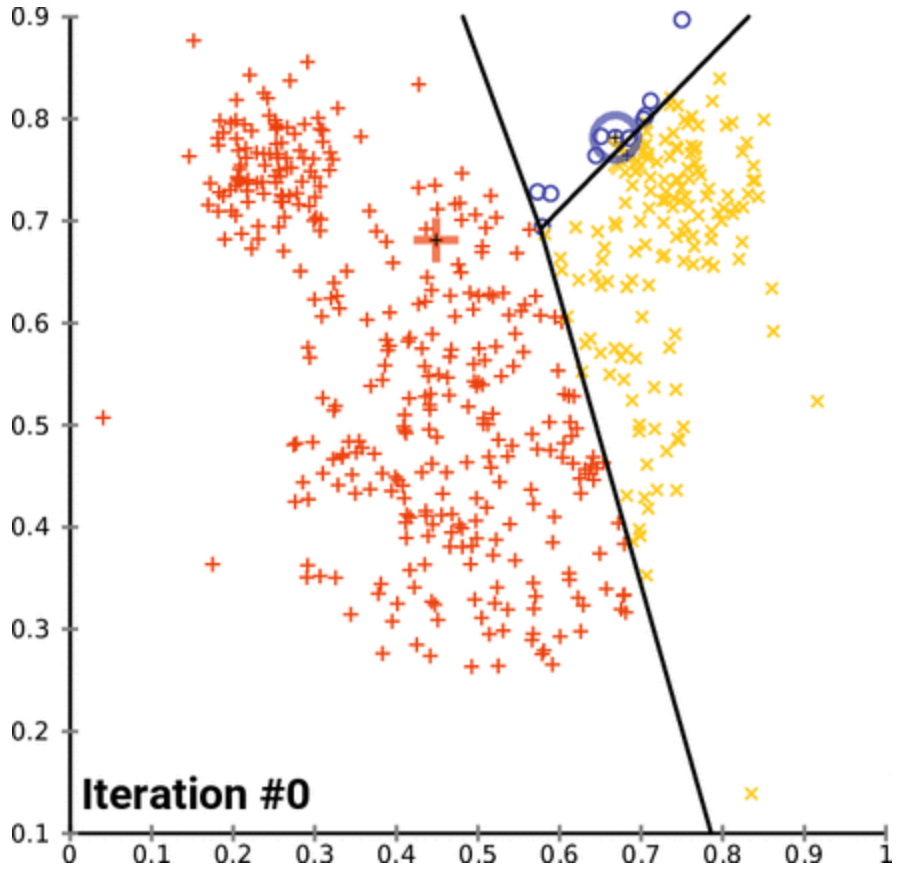

Metric 1
Metric 2
Metric 3
Metric 4
Metric 5
Metric 6
Metric 7
Country 1
Country 2
Country 3
Country 4
Country 5


3
2
1
0
-1
-2
-3
-4
0
100
200
300
400
500
600
700
800
900
1000


3
2
1
0
-1
-2
-3
-4
0
100
200
300
400
500
600
700
800
900
1000
aacddddcaabccbbbccbbccbb

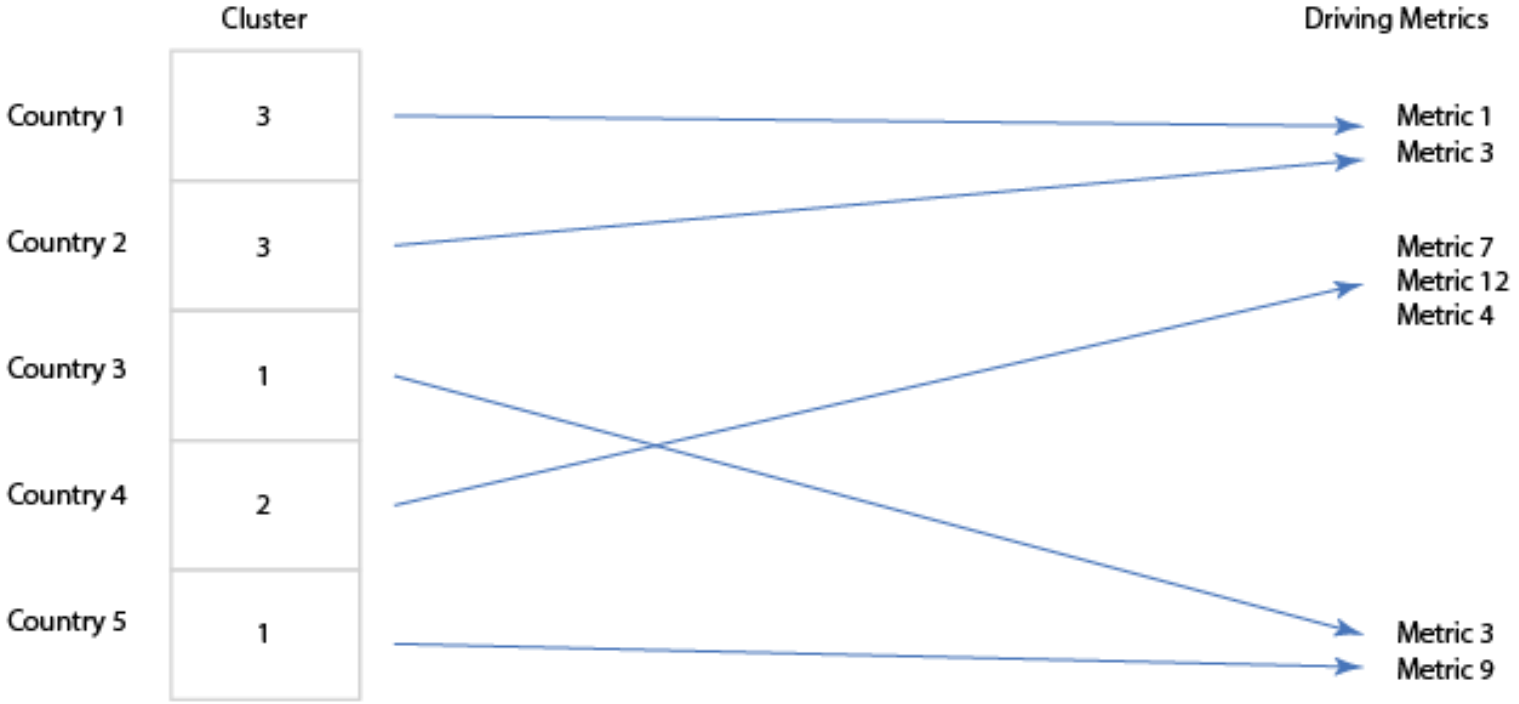
	Metric 1	Metric 2	Metric 3	Metric 4	Metric 5	Metric 6	Metric 7
Country 1	aabcdaffg	ddedghh	jjhhhssd	aabbccdd	aajjkllsss	ccddeefd	dkifdded
Country 2	daffeghg	deghhslif	aaddssd	aaassdg	eeddffss	ccddeeffd	edfkggh
Country 3	aabcdaffg	ddsseefh	eedfjkk	aabcdaffg	aabcdaffg	aabcdaffg	ddeeffgh
Country 4	aabcdaffg	aabbees	bbcccdde	aabcdaffg	aabcdaffg	ggffdkkk	cccddeeh
Country 5	aabcdaffg	jjdffhgl	llefdggs	aabcdaffg	aabcdaffg	aabcdaffg	dddeffg


Empty String
String "B"
""
A
L
T
R
U
I
S
M
Empty String
""
0
1
2
3
4
5
6
7
8
String "A"
P
1
1
2
3
4
5
6
7
8
L
2
2
1
2
3
4
5
6
7
A
3
2
2
2
3
4
5
6
7
S
4
3
3
3
3
4
5
5
6
M
5
4
4
4
4
4
5
6
5
A
6
5
5
5
5
5
5
6
6

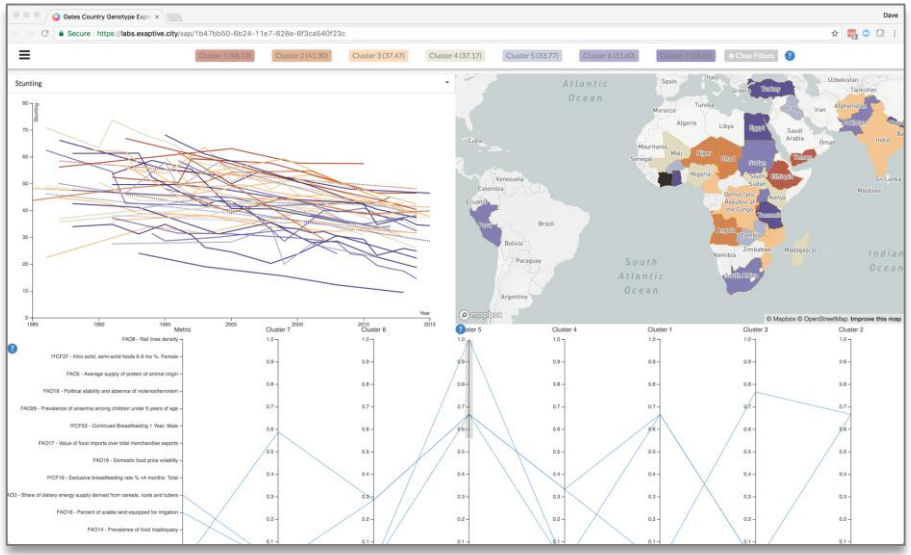

0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Iteration #0

	Metric 1	Metric 2	Metric 3	Metric 4	Metric 5	Metric 6	Metric 7
Country 1	1	ddedghh	jjhhhssd	aabbccdd	aajjkllsss	ccddeefd	dkifdded
Country 2	2	deghhslif	aaddssd	aaassdg	eeddffss	ccddeeffd	edfkggh
Country 3	2	ddsseefh	eedfjkk	aabcdaffg	aabcdaffg	aabcdaffg	ddeeffgh
Country 4	1	aabbees	bbcccdde	aabcdaffg	aabcdaffg	ggffdkkk	cccddeeh
Country 5	3	jjdffhgl	llefdggs	aabcdaffg	aabcdaffg	aabcdaffg	dddeffg

	Metric 1	Metric 2	Metric 3	Metric 4	Metric 5	Metric 6	Metric 7
Country 1	1	3	5	3	4	1	5
Country 2	2	2	3	1	5	1	2
Country 3	2	3	4	2	4	1	2
Country 4	1	2	1	3	3	2	3
Country 5	3	1	4	5	5	4	4


Cluster
Driving Metrics
Country 1
3
Country 2
3
Country 3
1
Country 4
2
Country 5
1
Metric 1
Metric 3
Metric 7
Metric 12
Metric 4
Metric 3
Metric 9

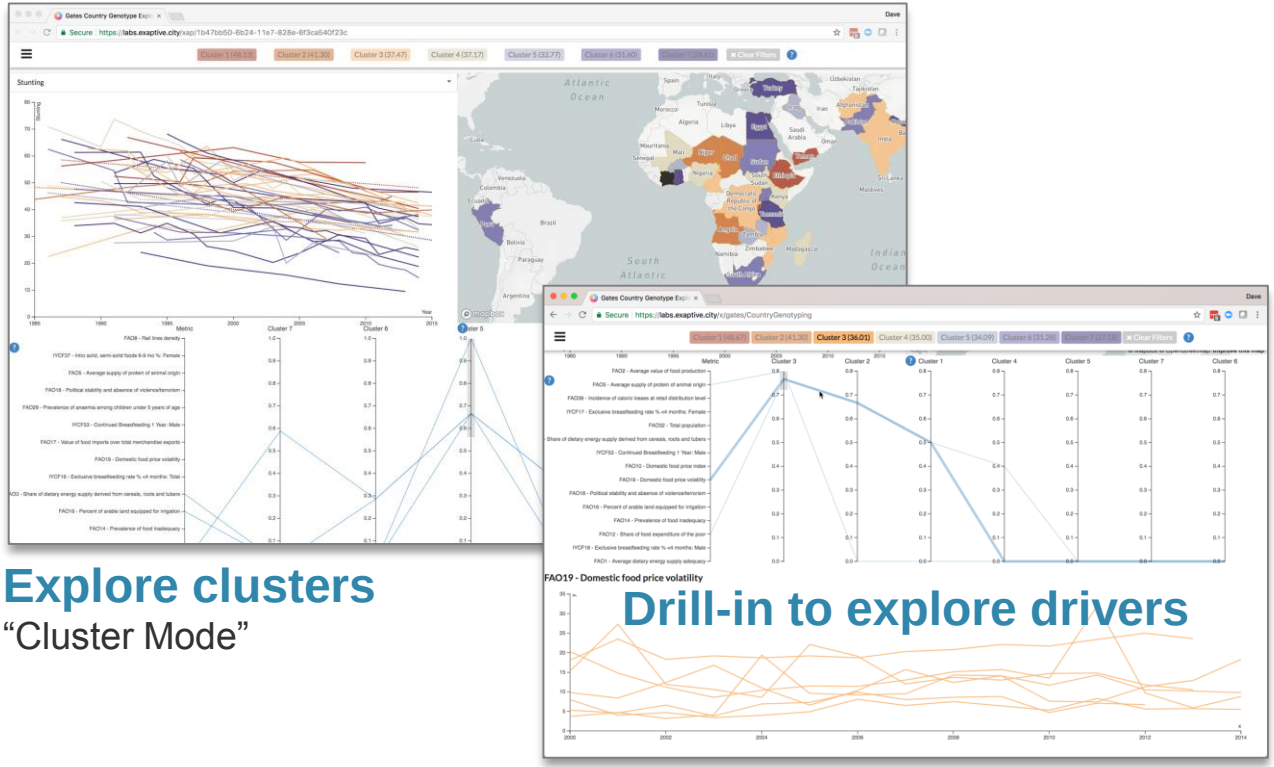
Country Segmentation interactive tool



Explore clusters

"Cluster Mode"

Country Segmentation interactive tool

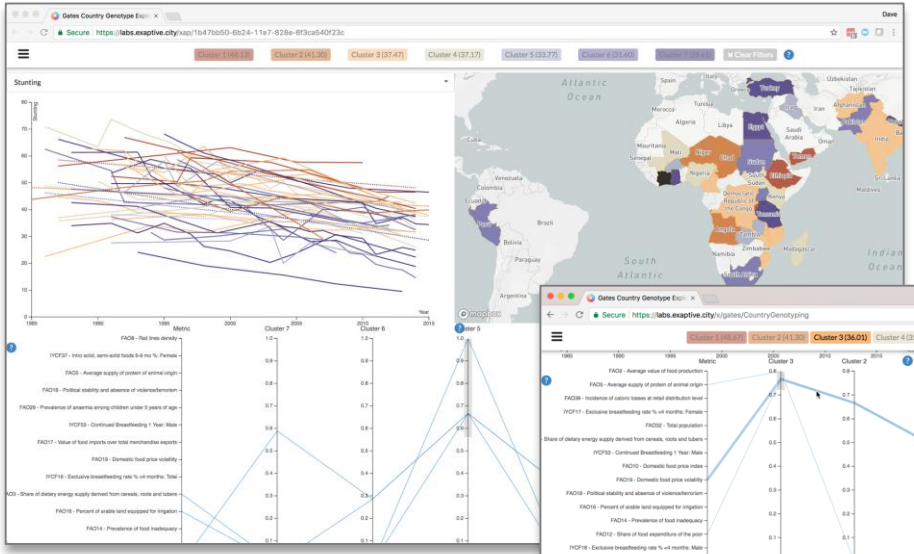


Explore clusters

"Cluster Mode"

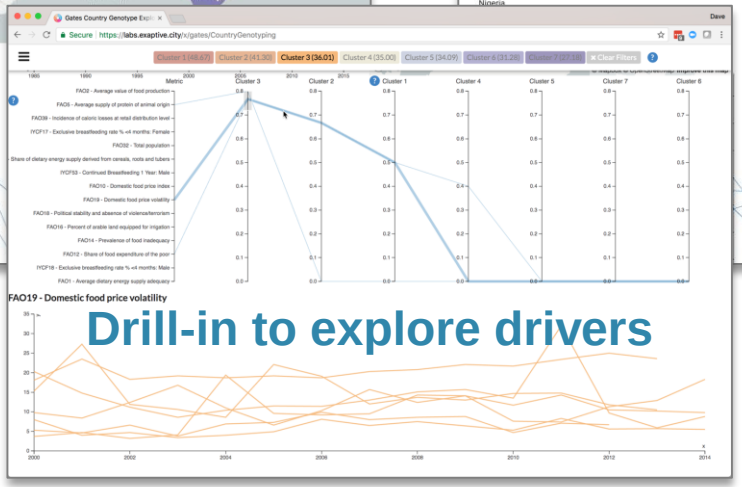
Drill-in to explore drivers

Country Segmentation interactive tool

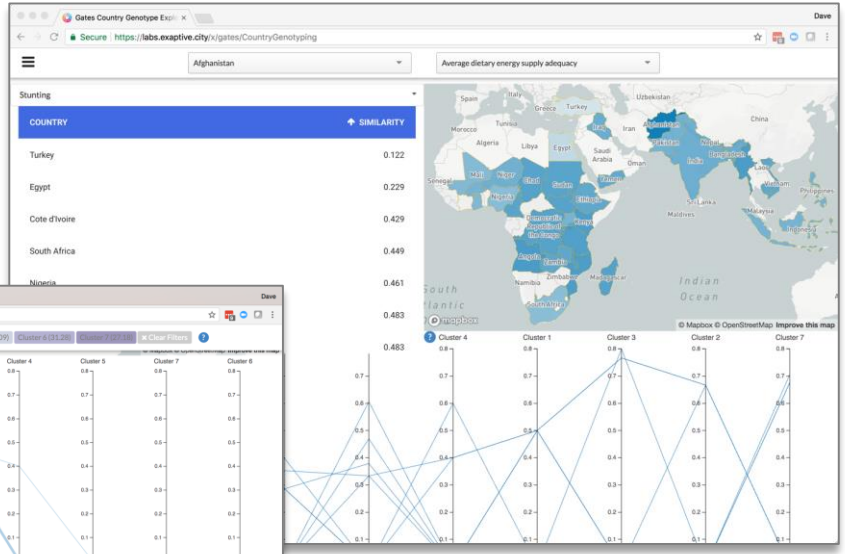


Explore clusters

"Cluster Mode"

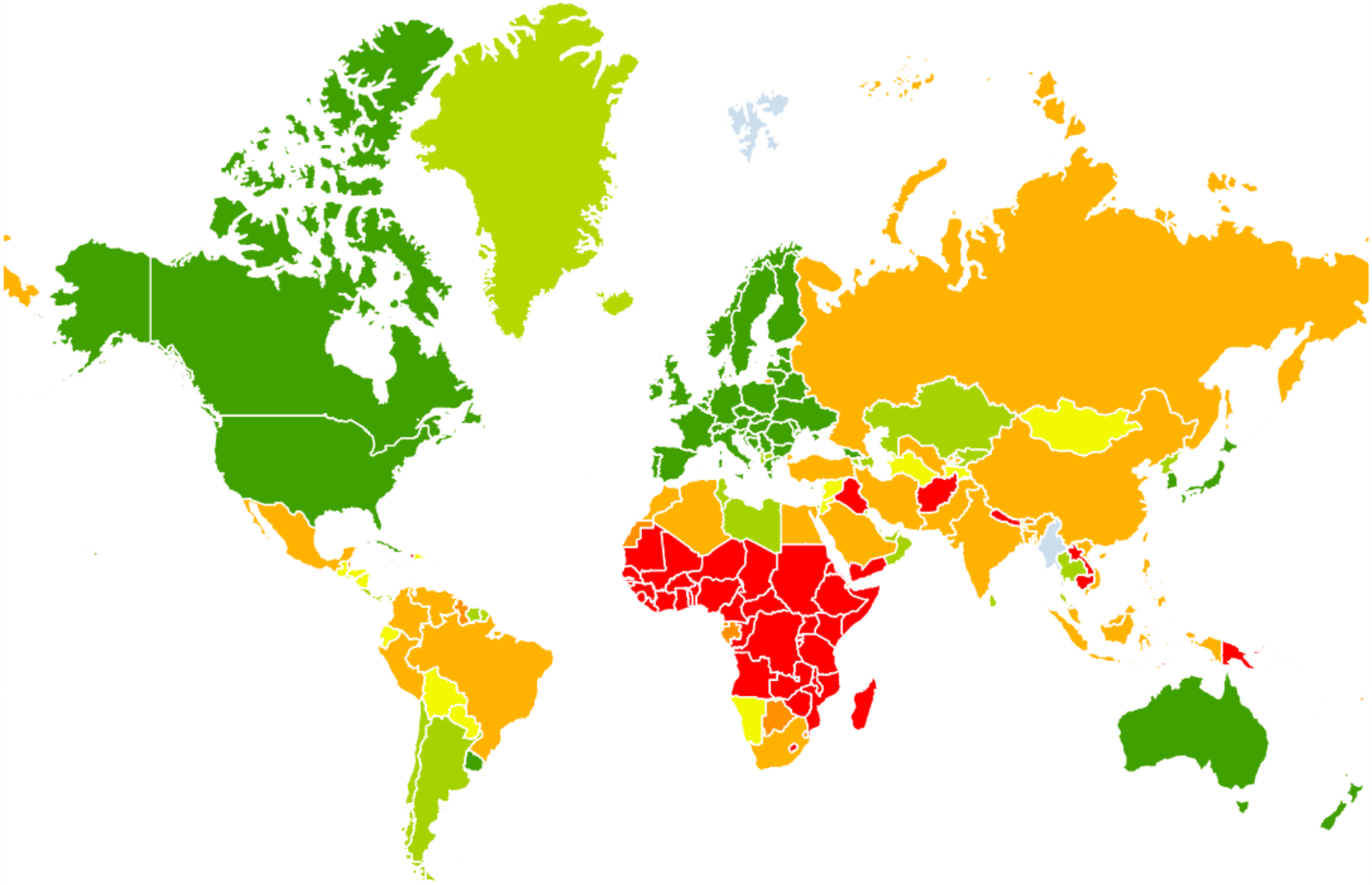


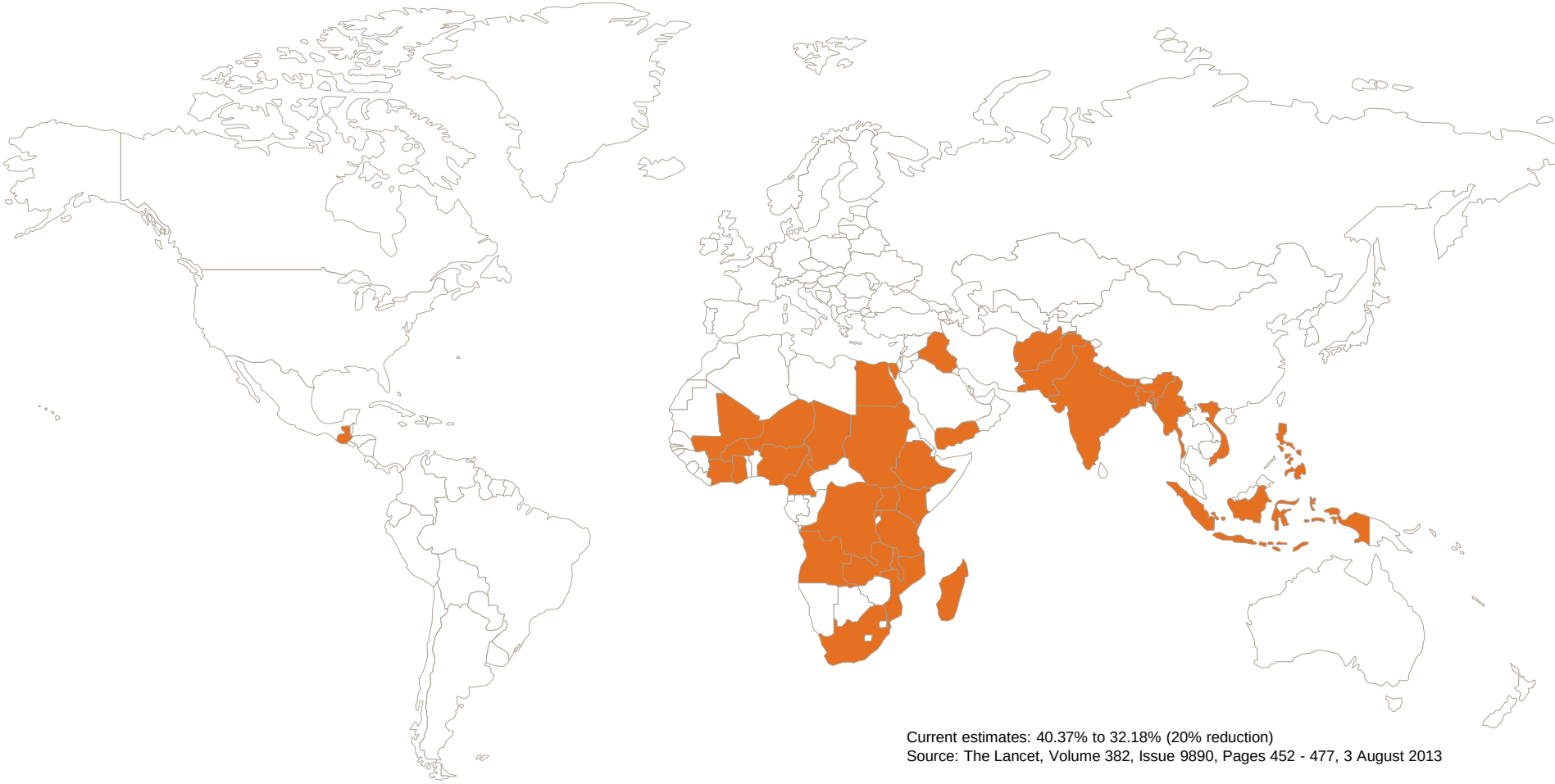
Drill-in to explore drivers



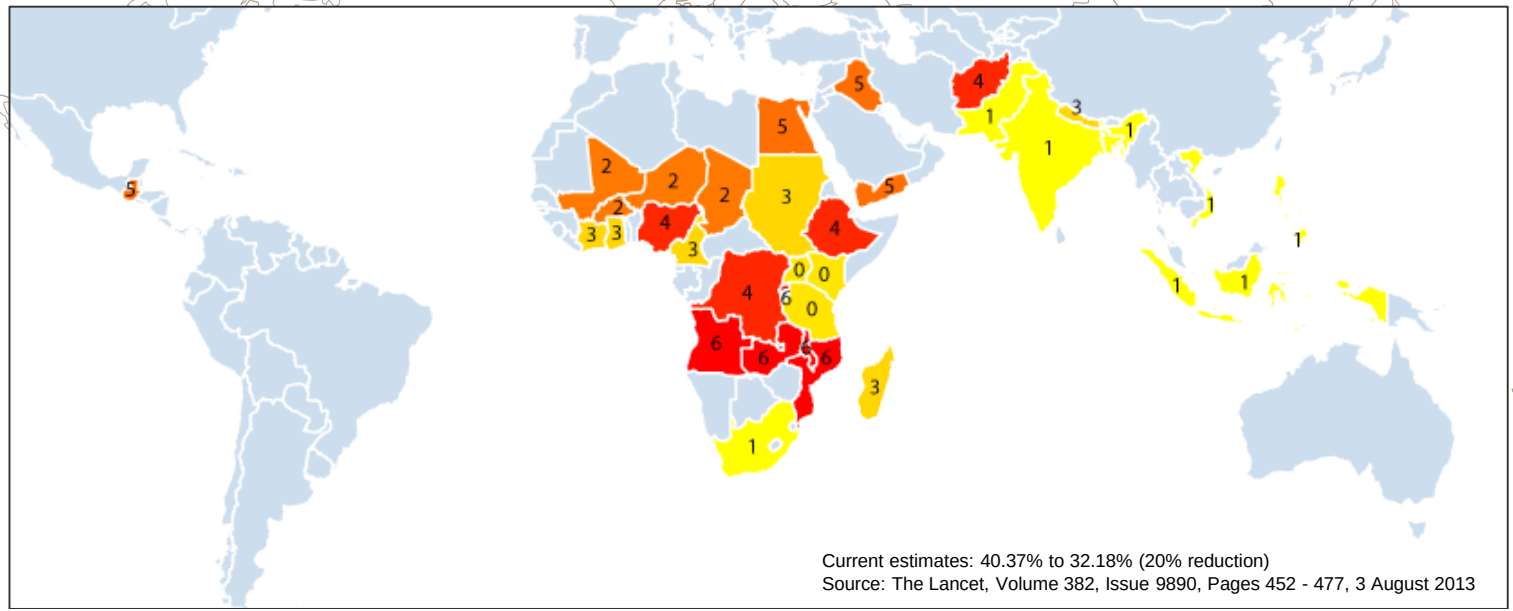
Explore similarity

"Distance Mode"

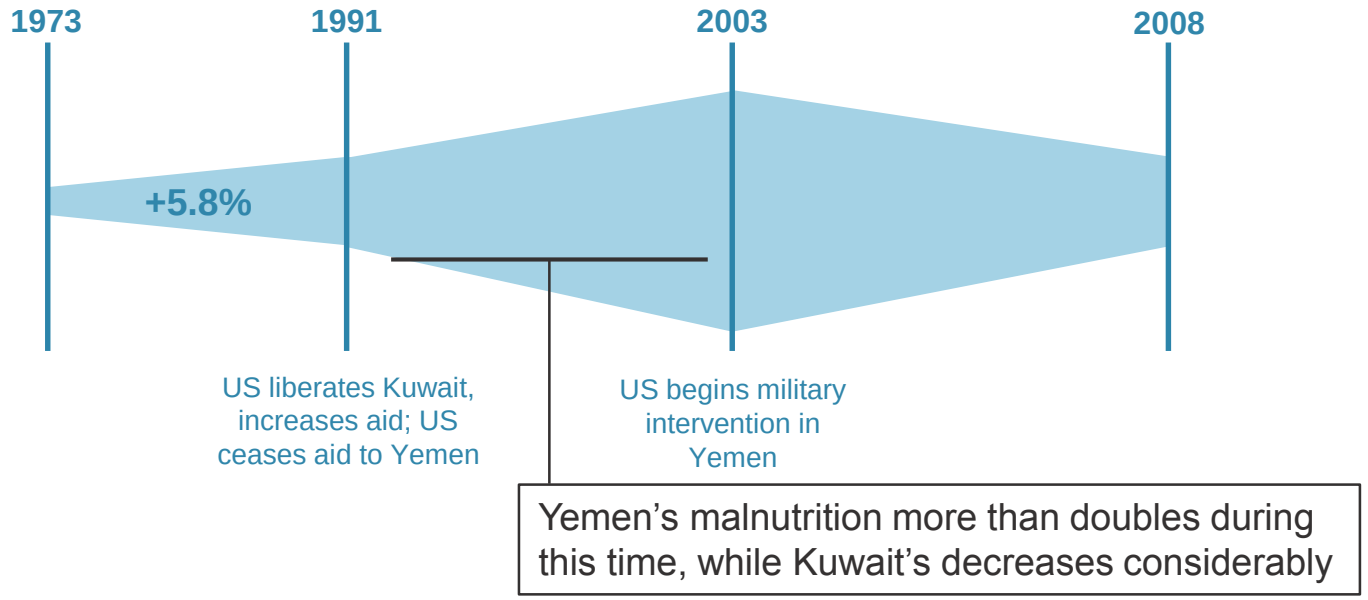




Current estimates: 40.37% to 32.18% (20% reduction)
Source: The Lancet, Volume 382, Issue 9890, Pages 452 - 477, 3 August 2013


Current estimates: 40.37% to 32.18% (20% reduction)
Source: The Lancet, Volume 382, Issue 9890, Pages 452 - 477, 3 August 2013

Looking at shifts in segmentation distance over time: Kuwait vs. Yemen



Kuwait in Iraq cluster
Yemen in Afghanistan cluster

Kuwait joins cluster with India and Egypt

Yemen joins cluster with Ghana and Cote d'Ivoire

Application to child stunting

Positive/Negative exemplar analysis: Overview

Objective:

Identify which variables might be driving the large improvements or deterioration in child stunting in exceptional “exemplar” countries.

Steps:

1. Identify positive and negative exemplars in terms of child stunting:
   - *Positive exemplars:* Countries that will be able to reach the World Health Assembly (WHA) target of reducing child stunting rate by 40% between 2012 and 2025
   - *Negative exemplars:* Countries where child stunting is deteriorating
2. Run segmentation algorithm using a data set of potential determinants of child stunting
3. Overlay the exemplars and the segmentation clusters

Idea behind this approach:

Variables that drive the segmentation clusters with positive and negative exemplars might be the ones that actually drove the exceptional changes in the overperforming/underperforming countries.

Positive/Negative exemplar analysis: Data

1. Identification of exemplars:
UNICEF-WHO-WB stunting rates

- Coverage:
  - 39 focus countries years 1973-2014
  - ➔ 2-23 observations per country, median 6
- Sample:
  - 1973-2014 – all data available
  - 2000-2014 – recent developments

2. Creation of segmentation clusters:
FAO and WHO-UNICEF data

- FAO:
  - 43 variables
  - ➔ food intake, dietary energy, food production and prices
  - ➔ socio-economic conditions, infrastructure, and demography
- WHO-UNICEF Infant & Young Child Feeding data:
  - 60 variables
  - ➔ breastfeeding, dietary diversity, meal frequency

Positive/Negative exemplar analysis: Identification method

Estimation method:
Country fixed effects model with linear country-specific time trends
Dependent variable (stunting) is transformed into logs ➔ coefficients capture annual change in percent

Regression model:
Let c denote country in survey year t; N will denote sample size. We estimate equation:

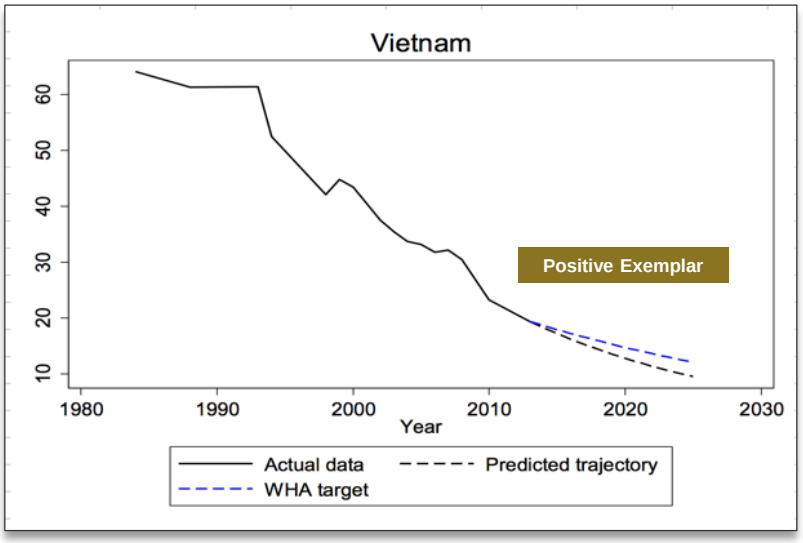
$$\text{Log}\,(\,\text{stunting\_rate}_{ct}\,) = t_c * \beta_c + \alpha_c + \varepsilon_{ct}$$

- Log (stunting_rate ct) is the N ×1 vector of logarithmic transformations of stunting rates
- α_c is country fixed effect
- β_c is annual change in stunting rate in country c in percent

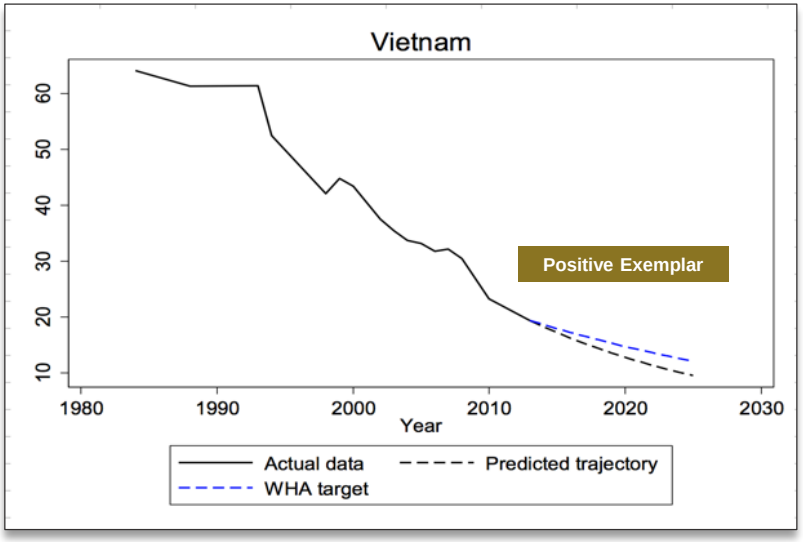
Categorization:

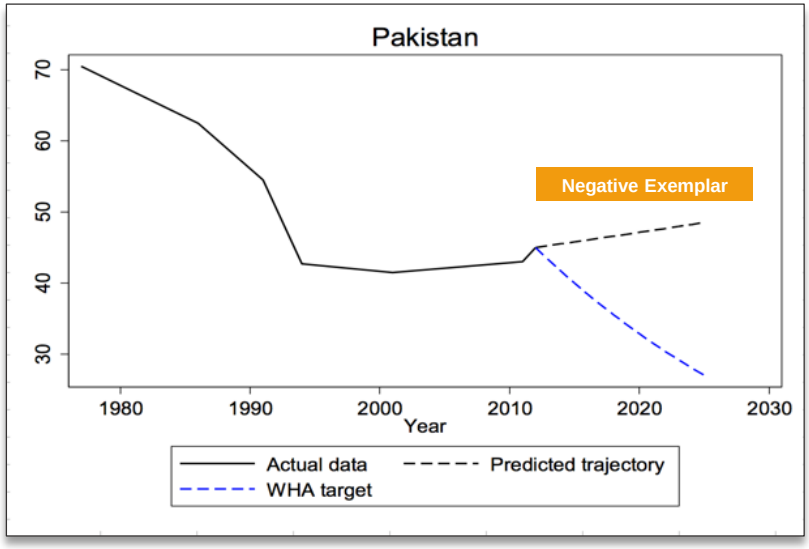
- positive exemplar: $\beta_c < -3.85$
  ➔ 3.85% is the annual rate necessary to achieve WHA target of reducing stunting rate by 40% in 2012-2025
- negative exemplar: $\beta_c > 0$

Positive/Negative exemplar analysis: Visualization

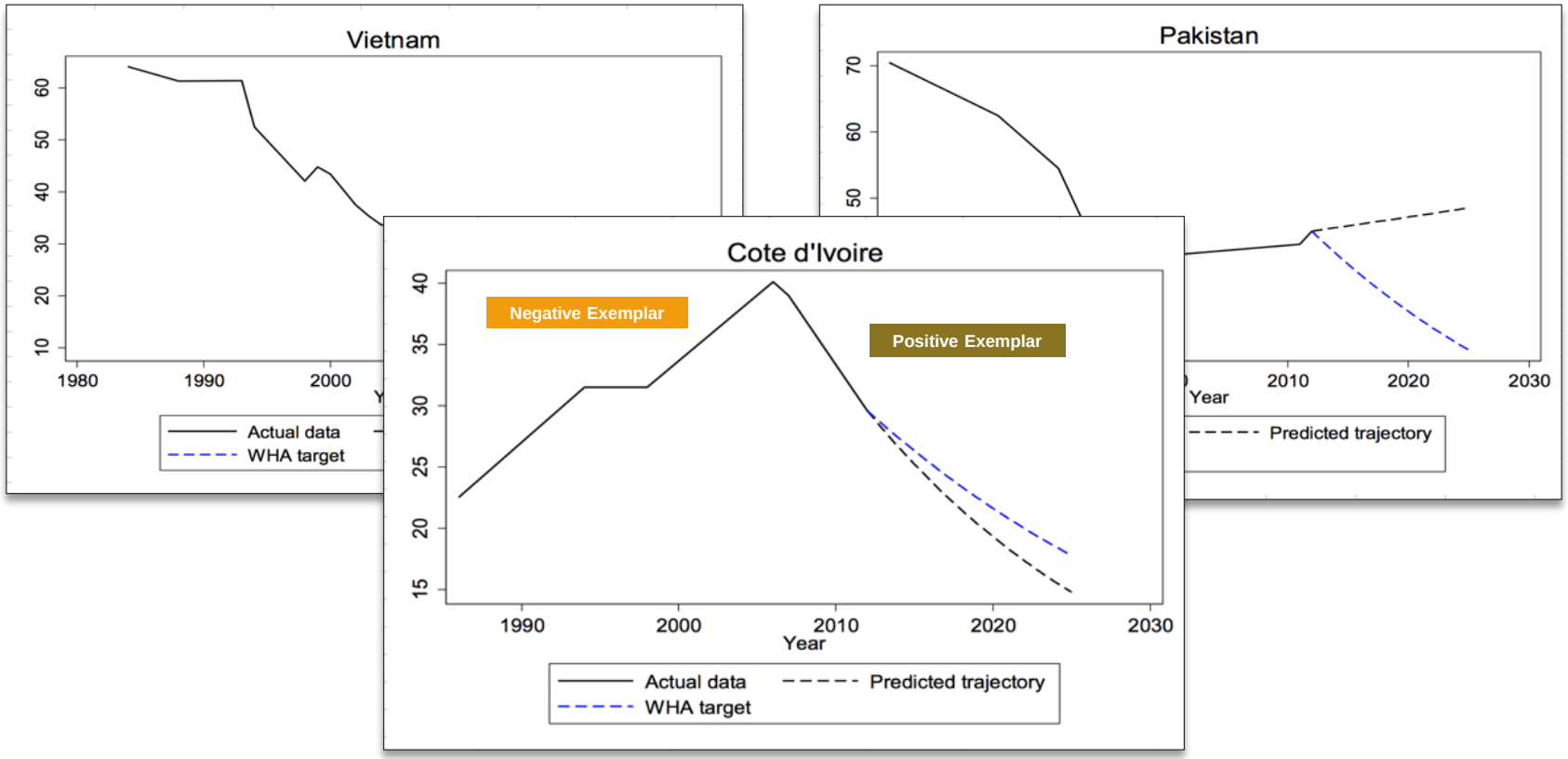


Positive/Negative exemplar analysis: Visualization

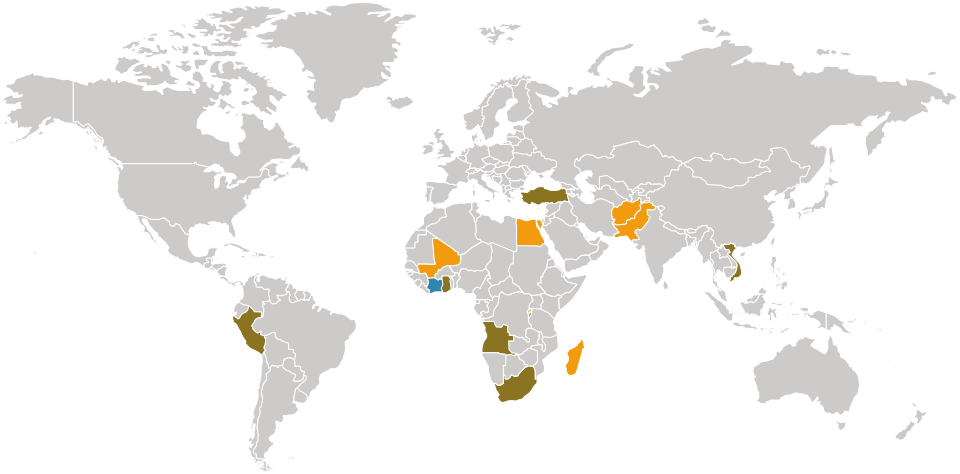




Positive/Negative exemplar analysis: Visualization



Exemplars among the 39 countries that bear the greatest burden of stunting:



Positive Exemplars

Peru | Ghana | Angola | South Africa | Turkey | Vietnam

Negative Exemplars

Mali | Egypt, Arab Rep. | Burundi | Madagascar | Afghanistan | Pakistan

Mixed

Côte d'Ivoire

		(2) SAX Fingerprint Segmentation on Time-Series Trajectories of Non-Stunting Metrics						
		Segmentation cluster 1 FAO1 - Average dietary energy supply adequacy FAO12 - Share of food expenditure of the poor FAO16 - Percent of arable land equipped for irrigation FAO10 - Domestic food price index	**Segmentation cluster 2** FAO12 - Share of food expenditure of the poor FAO16 - Percent of arable land equipped for irrigation FAO19 - Domestic food price volatility	**Segmentation cluster 3** FAO3 - Share of dietary energy supply derived from cereals, roots and tubers FAO12 - Share of food expenditure of the poor FAO10 - Domestic food price index	**Segmentation cluster 4** IYCF2 - Ever Breastfed %: Female IYCF52 - Continued Breastfeeding 1 Year: Female IYCF53 - Continued Breastfeeding 1 Year: Male	**Segmentation cluster 5** FAO5 - Average supply of protein of animal origin FAO12 - Share of food expenditure of the poor FAO19 - Domestic food price volatility	**Segmentation cluster 6** FAO1 - Average dietary energy supply adequacy FAO4 - Average protein supply FAO3 - Share of dietary energy supply derived from cereals, roots and tubers	**Segmentation cluster 7** FAO18 - Political stability and absence of violence/terrorism FAO16 - Percent of arable land equipped for irrigation FAO10 - Domestic food price index
(1) Linear Regression on Log(Stunting)	+							
	-							
	+/-							

(1) Linear Regression on Log(Stunting)	(2) SAX Fingerprint Segmentation on Time-Series Trajectories of Non-Stunting Metrics						
	Segmentation clusters with Positive Exemplars				**Segmentation clusters with Negative Exemplars**		**Mixed**
	Segmentation cluster 7 FAO18 - Political stability and absence of violence/terrorism FAO16 - Percent of arable land equipped for irrigation FAO10 - Domestic food price index	**Segmentation cluster 6** FAO1 - Average dietary energy supply adequacy FAO4 - Average protein supply FAO3 - Share of dietary energy supply derived from cereals, roots and tubers	**Segmentation cluster 4** IYCF2 - Ever Breastfed %: Female IYCF52 - Continued Breastfeeding 1 Year: Female IYCF53 - Continued Breastfeeding 1 Year: Male	**Segmentation cluster 2** FAO12 - Share of food expenditure of the poor FAO16 - Percent of arable land equipped for irrigation FAO19 - Domestic food price volatility	**Segmentation cluster 1** FAO1 - Average dietary energy supply adequacy FAO12 - Share of food expenditure of the poor FAO16 - Percent of arable land equipped for irrigation FAO10- Domestic food price index	**Segmentation cluster 3** FAO3 - Share of dietary energy supply derived from cereals, roots and tubers FAO12 - Share of food expenditure of the poor FAO10 - Domestic food price index	**Segmentation cluster 5** FAO5 - Average supply of protein of animal origin FAO12 - Share of food expenditure of the poor FAO19 - Domestic food price volatility
+	Peru South Africa 	Turkey 	Vietnam 	Angola 			Ghana 
-					Burundi Madagascar Pakistan 	Mali 	Afghanistan Egypt 
+/-							Ivory Coast 

HBGDki

Next steps

Next steps: Test the tool

So far:

- Identify segmentation clusters
- Overlay these clusters with classification of countries based on a different data source

Next steps:

1. **Asses the explanatory power of the segmentation tool**
   - Apply the segmentation algorithm to non-stunting variables measured at country level
   - Estimate individual-level models (i.e. child level):
     - Do segmentation clusters explain extent/probability of child stunting at individual level?
     - Does including the segmentation clusters into the model increase its explanatory power?
     - Are associations between individual stunting and its determinants similar within the segmentation cluster?

2. **Relevance of (different) data sources**
   - Run segmentation algorithm and individual-level models on the same data source

3. **Sensitivity of segmentation clusters to variables included in the algorithm**
   - Exclude input variables one at a time
   - Identify variables that lead to the largest differences in segmentation clusters