GRAND CHALLENGES | OCTOBER 1–4, 2017 | WASHINGTON, DC, USA

From Precision Medicine to Precision Policy: A tool for highly-dimensional country segmentation

Being able to generate, interrogate, and simulate country peer groups based on large numbers of longitudinal metrics offers a new lens by which to view regional difference in intervention efficacy.

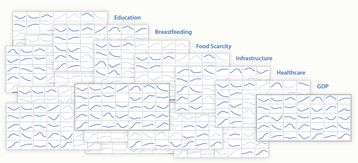
DAVID KING, MATTHEW COATNEY, MARK WISSLER
Exaptive, Inc., Oklahoma City, OK, USA

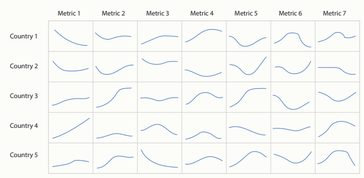
Motivation: Precision medicine has used data-driven patient stratification to improve patient outcomes. Can similar approaches improve the accuracy with which policy recommendations are matched to the places where they will be most effective?

"It is much more important to know what sort of a patient has a disease than what sort of a disease a patient has."
—WILLIAM OSLER

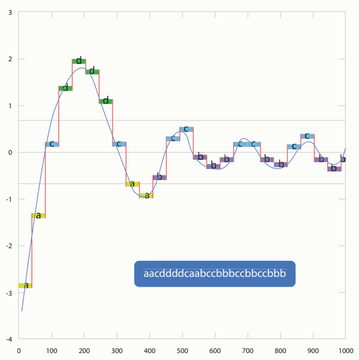
APPROACH

Assemble longitudinal metrics:



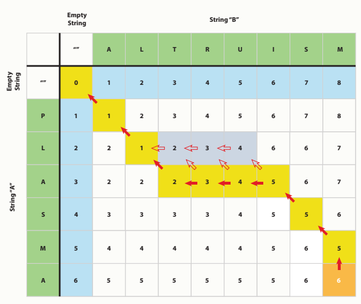


Time-series metrics converted to strings using symbolic aggregate approximation (SAX)



	Metric 1	Metric 2	Metric 3	Metric 4	Metric 5	Metric 6	Metric 7
Country 1	aabcdaffg	ddedghh	jjhhhssd	aabbccdd	aajjkillsss	ccddeefd	dkiifdded
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	jjdfthgl	llefdggs	aabcdaffg	aabcdaffg	aabcdaffg	dddeffg

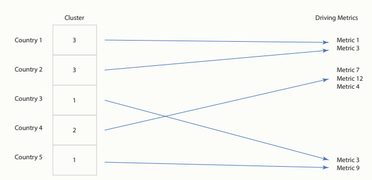
Distances between strings calcualted using Levenshtein distance calculation, to allow for k-means clustering a single metric across all countries



	Metric 1	Metric 2	Metric 3	Metric 4	Metric 5	Metric 6	Metric 7
Country 1	1	ddedghh	jjhhhssd	aabbccdd	aajjkillsss	ccddeefd	dkiifdded
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	jjdfthgl	llefdggs	aabcdaffg	aabcdaffg	aabcdaffg	dddeffg

Repeating for each metric allows for transforming a matrix of longitudinal curves for countries into a numeric matrix, which can then be clustered into a single set of clusters with unique driving metrics:

	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

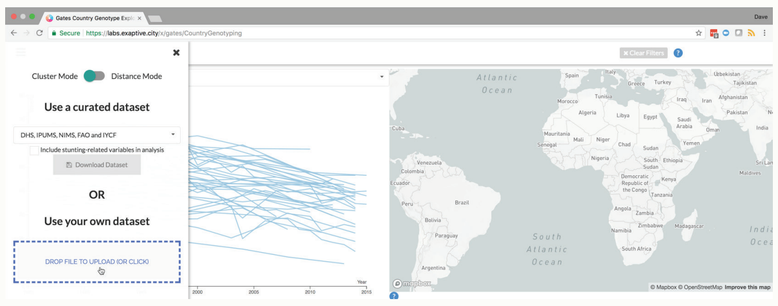


Since the algorithm takes into account the shape of the curves as well as the magnitudes, the same metric may appear as a driver in multiple clusters because of different trajectories.

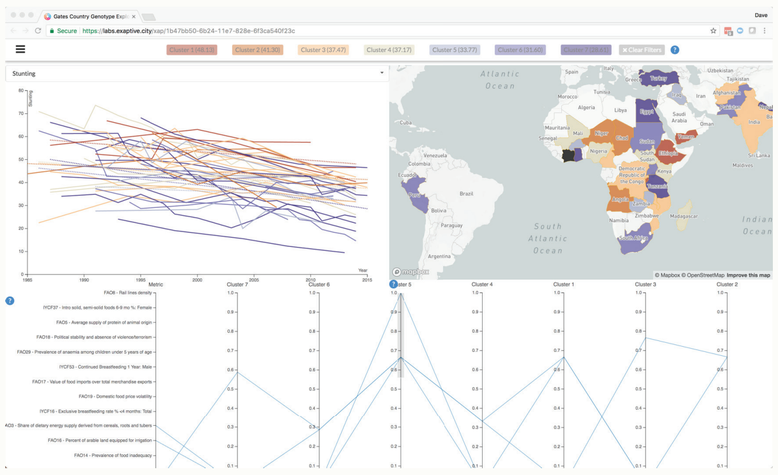
The complexity of interrogating the drivers behind the clusters required the generation of an interactive tool for visualization of the country clusters and capability to drill-in to the underlying metrics.

INTERACTIVE TOOL

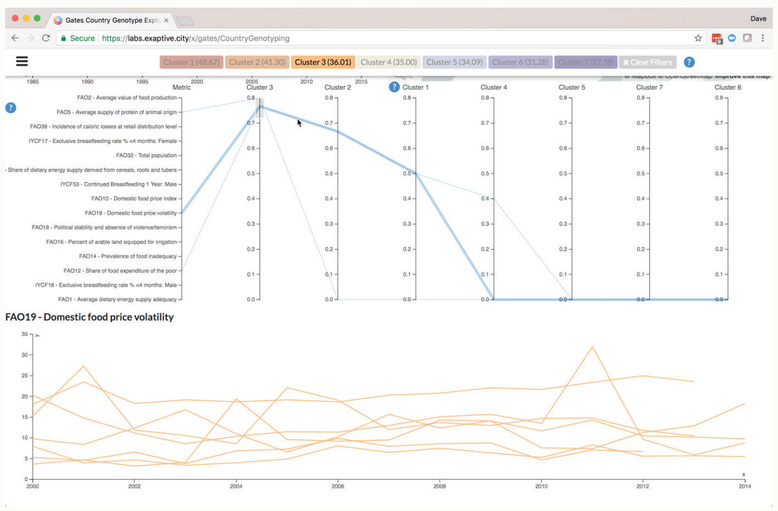
Tool supports use of pre-assembled curated datasets, or the ability to upload an arbitrary dataset of country curves.



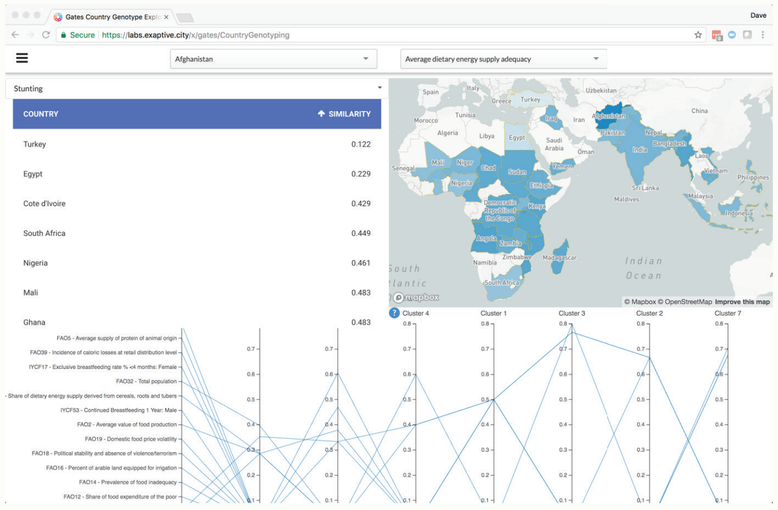
Real-time processing can cluster hundreds of metrics on all the countries in the world in under a minute:



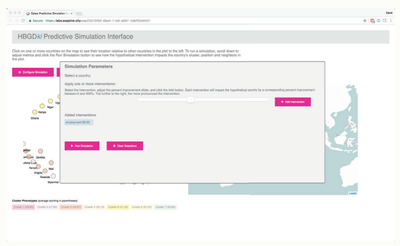
An interactive parallel-coordinates visualization allows for exploring the relative contribution of each metric to each cluster, and for drilling in to individual metric curves:



To address the issue of clustering leading to "hard-line" boundaries that can lead to misinterpretation of similarity of cluster members, the tool also supports a "smooth distance" mode for viewing the underlying distance scores for each country across metrics:

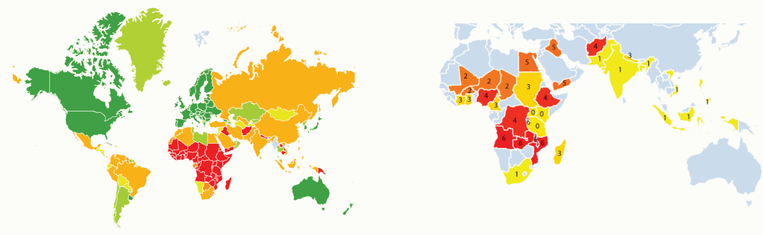


Simulation capability allows for testing how hypothetical intervention scenarios could shift a country into a different peer group:

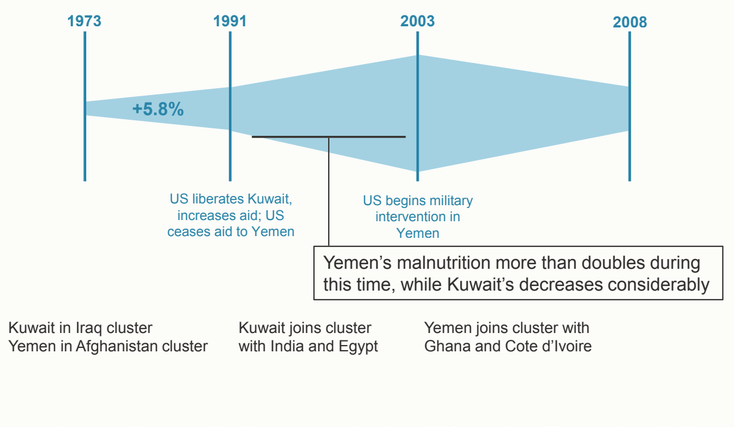


USE-CASES

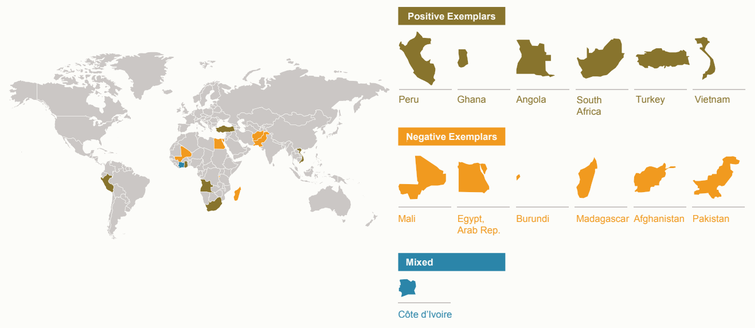
Clustering of all countries in the world based on GapMinder dataset provides a sanity check on the algorithm by revealing nuanced 1st, 2nd, 3rd world type cohorts despite the algorithm assigning relatively low weight to GDP metric. When applied to the 39 countries bearing the brunt of the world's stunting burden, clusters emerged that were independent of geography:



Using the algorithm to find countries whose relative position had changed the most over time illuminated Kuwait and Yemen as having periods of increased divergence, and then convergence that aligned with known periods of policy intervention:



The tool was also applied to an analysis of "Stunting Exemplars" that was performed by Sofia Trommlerová and Matthias Rieger of Erasmus University Rotterdam using UNICEF-WHO-WB global stunting data, classifying countries as positive or negative based on whether a linear regression predicted they would decrease stunting as per the WHO's targets for 2025, or increase it instead:



In parallel, countries were clustered using the segmentation tool, based on FAO and IYFC data. The FAO data included 43 variables related to food intake, dietary energy, food production and prices, socio-economic conditions, infrastructure, and demography. The IYCF data included 60 variables related to breastfeeding, dietary diversity, and meal frequency. The resulting clusters, when superimposed over the exemplar categorization show postive exemplars and negative exemplars coming from distinct segments:

(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
(1) Linear Regression on Log(Stunting)	+	Peru South Africa	Turkey	Vietnam	Angola			Ghana
	-					Burundi Madagascar Pakistan	Mali	Afghanistan Egypt
	+/-							Ivory Coast

BILL & MELINDA GATES foundation
HBGDki
exaptive