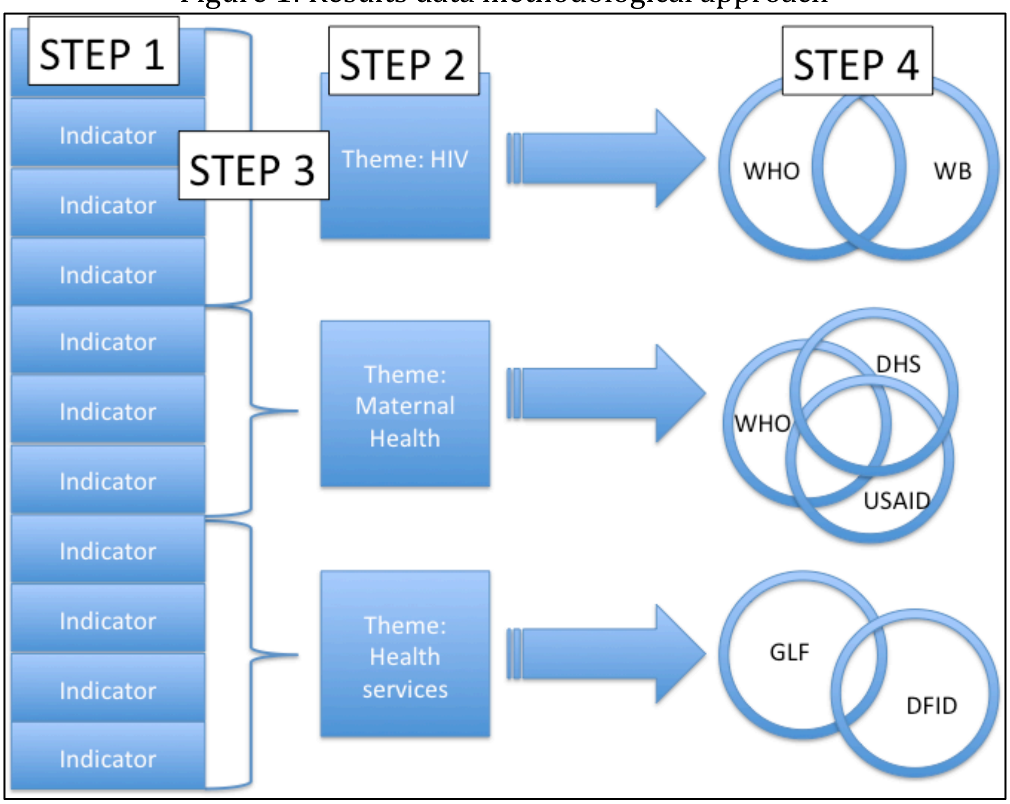
Development Gateway
Results data crosswalk

Summary
Our main goal here is to compare results information across organizations, countries, years, etc. At its core, this is a sample-size-boosting exercise with the added benefits of having results information at a "central" location and being able to compare it across organizations. This should allow for better data use as well as better coordination among development organizations.

Definitions
- **"Development organizations"**—organizations involved in international development, particularly those who fund health initiatives. Examples include the WHO, World Bank, and USAID.
- **"Results information"**—information reported to these organizations as to the results of their initiatives. This data can take on many forms—for example, USAID may receive a simple written report from its analysts in the field as to the success or progress of a specific project.
- **"Comparability"**—the degree to which results information from similar initiatives can be compared. This could be across or within organizations—for example, the World Bank may have HIV prevention projects in both Burundi and Afghanistan while WHO has similar projects in Thailand and Nigeria. Comparability is the degree to which reported information from all of these projects can be grouped together.

Approach
There are four main steps to this exercise as shown in the figure below:

Figure 1: Results data methodological approach



Step 1: Gathering data
To complete the crosswalk activity, the first step is to gather and categorize types of data. In this effort, there are two main types of data we're interested in—indicator data and results data. Often there is

considerable overlap between these data types, the main difference usually being the level of aggregation. Results data is at the project or initiative level and is usually found in project reports or final evaluations. Indicator data is at an aggregate level and is usually found in online databases such as the World Development Indicators.

At Step 1, we are not as concerned about data types, although it is important to gather as much metadata information about indicators as possible. This will help us distinguish levels of aggregation later on in the data compilation and comparison process.

Step 1: What we need
[We need all the data…]
[We also need a way to pull data out of PDFs…]

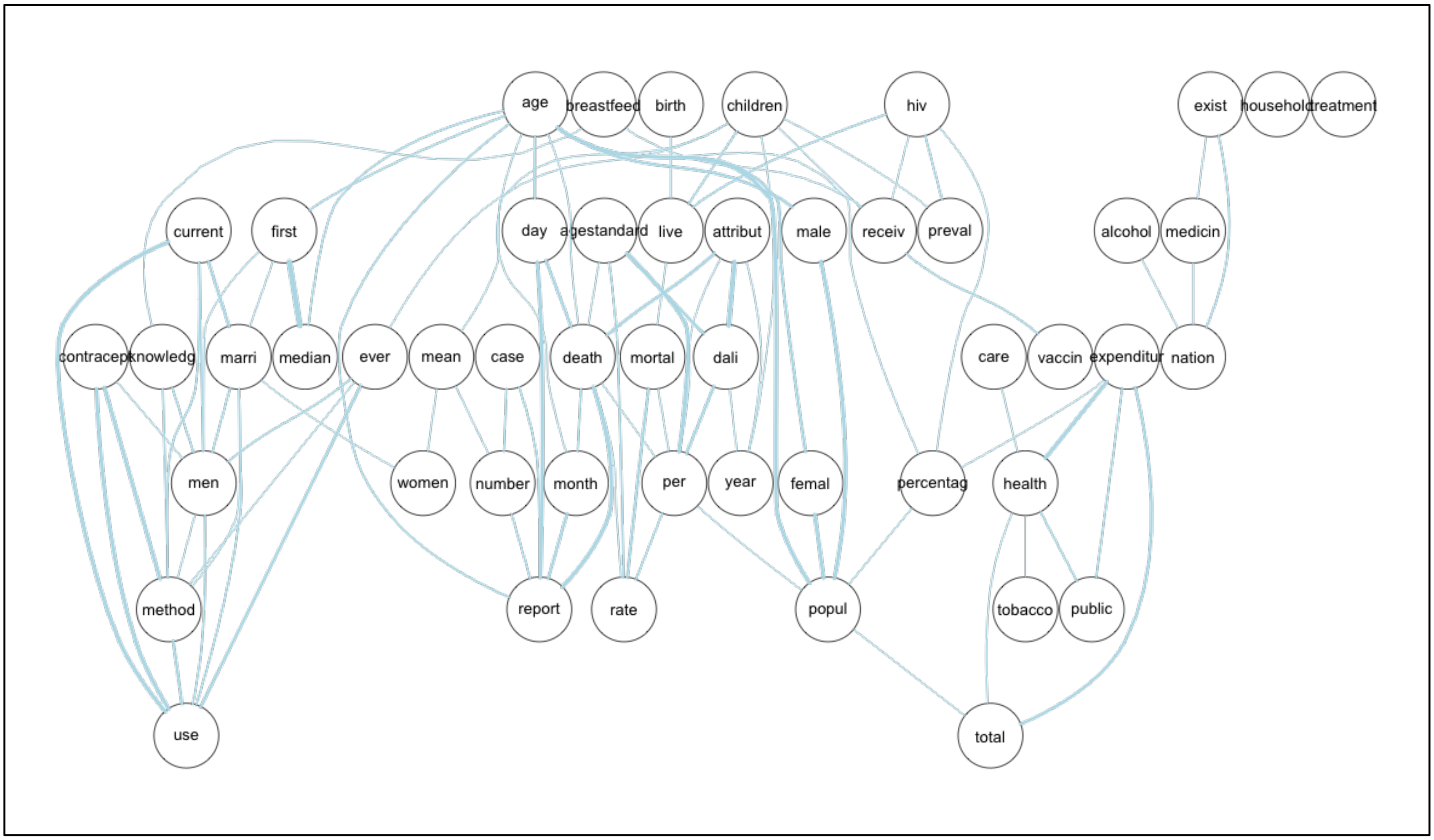
Step 2: Creating themes and classification schema
There are two main ways to create groups for data organization. The first is to create them "by hand" using knowledge of the field or by finding a standardized list. In the health sector, for example, the WHO, in conjunction with several other aid donors, has created a list of the Top 100 health indicators of interest. This list is already organized into themes and subthemes, which can be used for data organization.

Table 1: Top 100 health indicators (sample)

Indicator name	**Theme**	**Subtheme**
Life expectancy at birth	Health status	Mortality by age and sex
Adult mortality rate between 15 and 60 years of age	Health status	Mortality by age and sex
Under-five mortality rate	Health status	Mortality by age and sex
Antenatal care coverage	Service coverage	Reproductive, maternal, newborn, child and adolescent
Births attended by skilled health personnel	Service coverage	Reproductive, maternal, newborn, child and adolescent
HIV care coverage	Service coverage	HIV
Antiretroviral therapy (ART) coverage	Service coverage	HIV
HIV viral load suppression	Service coverage	HIV

Although a helpful baseline, many of these indicators are highly specific and may miss related variables that could be informative. Thus, a second option is to run classification algorithms to find patterns in the data. This is advantageous since it can be more flexible and use more data, however this can also be a drawback—more data overemphasizes "common" indicators and often neglects more specialized ones such as disease variants (e.g. cervical vs. prostate cancer) or poor health covariates (e.g. insecticide use). Even using heavier weights or normalizing methods (such as the commonly used TF/IDF method) doesn't give us the specificity we would like. The figure below demonstrates this—it shows the most common terms and associations in a group of almost 4,000 indicators (comprised of WHO, DHS, World Bank, and Global Fund data). This figure shows the top 50 most common words, which were words that appeared at least 70 times in the original list. Even when adjusting this aggregation and these associations using TF/IDF weights, uninformative terms such as "current," "nation," and "attribute" still appear.

Figure 2: Association of terms (WHO, DHS, WB, and GF)



Latent Dirichlet Allocation ([LDA](LDA)), which is basically a different way of searching for patterns in word frequencies, produces similar results. The table below shows the top 20 predicted themes using term frequency weights and LDA analysis.

Table 2: LDA predicted themes

Topic number	Theme	Number of indicators classified
1	HIV	172
2	Report	221
3	Percentage	157
4	Men	171
5	Nation	207
6	Tobacco	158
7	Month	167
8	Birth	207
9	Rate	179
10	Children	194
11	Total	159
12	Age	205
13	Women	174
14	Public	215
15	Per	223
16	Number	139
17	Use	313
18	Period	178
19	Health	135

20	Household	204

You'll notice that LDA is pretty good at producing groups of similar size, but this isn't actually what we would prefer. Again, although HIV or maternity-related variables may comprise large groups, we'll want to look beyond basic patterns since our end-goal is crosswalking indicators instead of finding overall patterns.

More metadata can help—i.e. indicator descriptions—but still tends to overemphasize frequencies. For example, a more detailed list of just over 400 indicators (using data from WHO, UNICEF, and others) shows more intelligible clustering, but topic analysis through LDA shows that this set of indicators is heavily weighted towards HIV and other sexually-oriented variables. The figure below shows the most prominent relationships in this list and the table shows LDA-predicted topics. Notice that "HIV" is predicted twice and that along with the "sex" category comprise about 40% of the list.

Figure 3: Association of terms (WHO, UNICEF, and others)

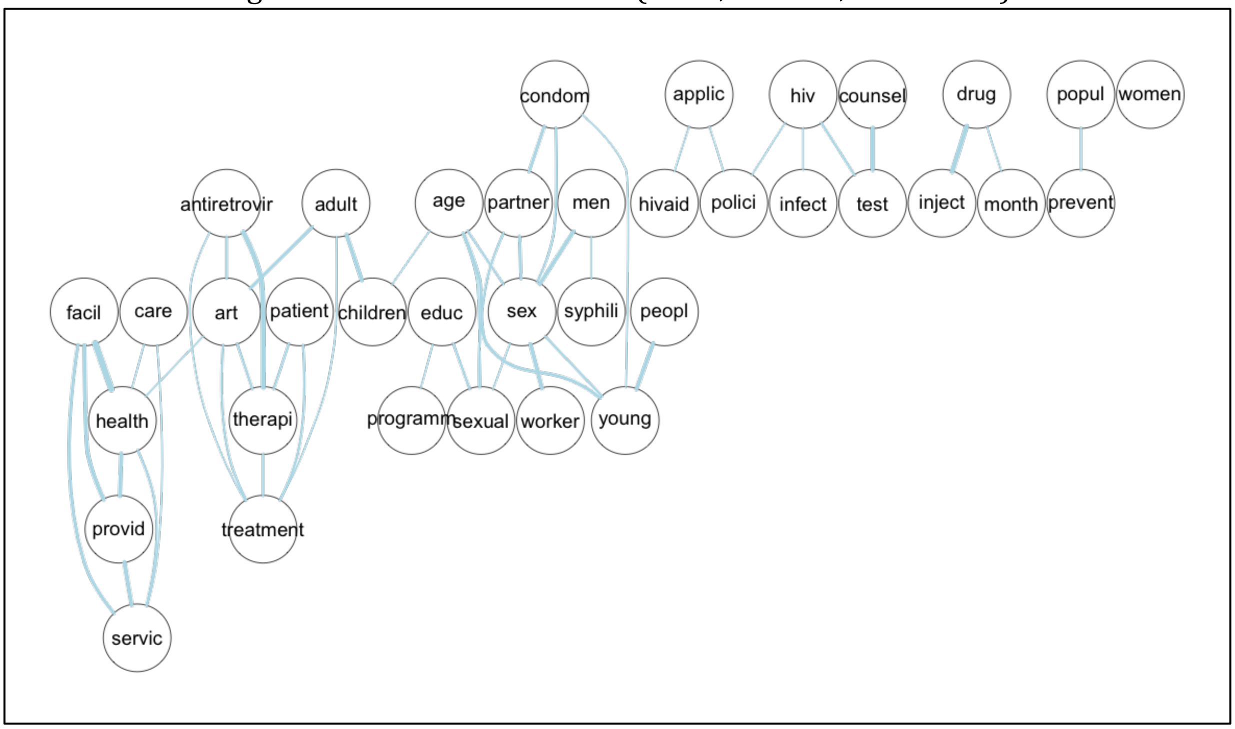


Table 3: LDA predicted themes (small set)

Topic number	Theme	Number of indicators classified
1	Sex	65
2	Number	36
3	HIV	50
4	Service	56
5	Children	36
6	HIV	32
7	Therapist	38
8	Nutrition	16
9	Food	12
10	Young	31

In addition, words such as "number" and "service" are still listed as common themes even though these wouldn't make sense as topic categories.

With these considerations, it is likely most beneficial to use a combined approach. We can use the Top 100 as a baseline and verify its accuracy and usefulness using word frequency analysis. Additionally, we can update word frequency results using the Top 100 and expertise from people in field of interest. By combining the strengths of each categorization type, we can ensure the formation of detailed groups. This will be important for the next phase of our analysis—classifying new indicators based on our categorical list.

Step 2: What we need [We need a solid list of categories and key terms…]

Step 3: Data classification

Once we have a working list of key themes and indicators, we can start classifying new indicator data. For this we will be using various Machine Learning methods such as Support Vector Machines (SVM) and neural networks. We won't go incredibly in-depth as to the statistical specifications of these different algorithms in this report; suffice to say they are methods that "learn" from one set of classifications and then apply these learned rules to new data. This is why a solid list of key indicators is important—the better and more detailed our initial list, the better our algorithms will be at predicting the categories of new indicator lists.

For example, consider the following rough-draft analysis. We use two lists for training the data—an unedited list of the Top 100 indicators and the same list with more metadata. Once we have trained the data, we then classify a list of almost 2,000 WHO indicators using seven different algorithms (and TF/ IDF weights). The following tables show a sample of WHO indicators and how they were classified using the different Top 100 lists and algorithms.

Table 4: WHO classified using Top 100 (indicator names only)

Indicator	**Actual group**	**SVM**	**Entropy**	**Boosting**	**Bagging**	**Forests**	**Trees**	**Neural Net**
Life expectancy	Mortality by age and sex	Morbidity	Mortality by age and sex	Mortality by age and sex	Nutrition	Noncommunicable diseases	Noncommunicable diseases	Mortality by cause
Infant mortality rate	Mortality by age and sex	Mortality by cause	Mortality by age and sex	Mortality by age and sex	Mortality by age and sex	Mortality by cause	Mortality by age and sex	Quality and safety of care
Under-five mortality rate	Mortality by age and sex	Mortality by cause	Mortality by age and sex	Mortality by age and sex	Mortality by age and sex	Mortality by cause	Mortality by age and sex	Quality and safety of care
Adult mortality rate	Mortality by age and sex	Mortality by cause	Mortality by age and sex	Environmental risk factors	Mortality by age and sex	Mortality by cause	Mortality by age and sex	Quality and safety of care
Neonatal mortality rate	Mortality by age and sex	Mortality by cause	Mortality by age and sex	Mortality by age and sex	Mortality by cause	Mortality by cause	Mortality by age and sex	Quality and safety of care
TB mortality rate	Mortality by cause	Mortality by cause	Mortality by cause	Mortality by cause	Mortality by cause	Mortality by cause	Mortality by age and sex	Quality and safety of care
HIV mortality rate	Mortality by cause	Mortality by cause	Mortality by cause	Mortality by cause	Mortality by cause	Mortality by cause	Mortality by age and sex	Quality and safety of care
Malaria mortality rate	Mortality by cause	Mortality by age and sex	Mortality by cause	Mortality by cause	Mortality by cause	Mortality by cause	Mortality by age and sex	Quality and safety of care
Suicide rates	Mortality by cause	Mortality by cause	Mortality by cause	Mortality by cause	Morbidity	Morbidity	Morbidity	Mortality by cause
Adolescent fertility rate	Fertility	Noncommunicable diseases	Fertility	Fertility	Noncommunicable diseases	Noncommunicable diseases	Morbidity	Mortality by cause
Total fertility rate	Fertility	Noncommunicable diseases	Fertility	Fertility	Fertility	Fertility	Morbidity	HIV
HIV prevalence	Morbidity	Morbidity	Morbidity	HIV	HIV	Morbidity	Morbidity	Reproductive, maternal, newborn, child and adolescent
Malaria incidence	Morbidity	Morbidity	Morbidity	Environmental risk factors	Morbidity	Morbidity	Noncommunicable diseases	Reproductive, maternal, newborn, child

								and adolescent
Ambient air pollution in urban areas	Environmental risk factors	Reproductive, maternal, newborn, child and adolescent	Environmental risk factors	Environmental risk factors	Environmental risk factors	Environmental risk factors	Noncommunicable diseases	Noncommunicable diseases
Total correct	**—**	**5 (36%)**	**14 (100%)**	**11 (79%)**	**9 (64%)**	**7 (50%)**	**5 (36%)**	**1 (7%)**

Side-note: Fertility rates classified as "noncommunicable diseases" and/ or "morbidity"? Do the machines force us to ask tough ethical questions?

About 53% accurate overall (52/98)

Table 5: WHO classified using Top 100 (indicator names and metadata)

Indicator	**Actual group**	**SVM**	**Entropy**	**Boosting**	**Bagging**	**Forests**	**Trees**	**Neural Net**
Life expectancy	Mortality by age and sex	Nutrition	Mortality by age and sex	Morbidity	Immunization	Mortality by age and sex	Health financing	Noncommunicable diseases
Infant mortality rate	Mortality by age and sex	Quality and safety of care	Mortality by age and sex	Mortality by cause	Mortality by cause	Mortality by age and sex	Mortality by age and sex	Morbidity
Under-five mortality rate	Mortality by age and sex	Nutrition	Mortality by age and sex	Mortality by cause	Mortality by cause	Mortality by age and sex	Mortality by age and sex	Morbidity
Adult mortality rate	Mortality by age and sex	Reproductive, maternal, newborn, child and adolescent	Mortality by cause	Mortality by cause	Mortality by cause	Mortality by cause	Mortality by age and sex	Morbidity
Neonatal mortality rate	Mortality by age and sex	Nutrition	Mortality by age and sex	Mortality by cause	Mortality by cause	Mortality by cause	Mortality by age and sex	Morbidity
TB mortality rate	Mortality by cause	Reproductive, maternal, newborn, child and adolescent	Mortality by cause	Mortality by cause	Mortality by cause	Mortality by cause	Mortality by age and sex	Mortality by cause
HIV mortality rate	Mortality by cause	Morbidity	Mortality by cause	Mortality by cause	Mortality by cause	Mortality by cause	Mortality by age and sex	Mortality by age and sex
Malaria mortality rate	Mortality by cause	Reproductive, maternal, newborn, child and adolescent	Mortality by cause	Mortality by cause	Mortality by cause	Mortality by cause	Mortality by age and sex	Morbidity
Suicide rates	Mortality by cause	Morbidity	Mortality by cause	Mortality by cause	Mortality by cause	Mortality by cause	Mortality by age and sex	Morbidity
Adolescent fertility rate	Fertility	Reproductive, maternal, newborn, child and adolescent	Fertility	Fertility	Noncommunicable diseases	Fertility	Mortality by age and sex	Morbidity
Total fertility rate	Fertility	Reproductive, maternal, newborn, child and adolescent	Fertility	Fertility	Access	Mortality by cause	Mortality by age and sex	Morbidity
HIV prevalence	Morbidity	Noncommunicable diseases	Morbidity	Morbidity	HIV	Morbidity	Health financing	Mortality by cause
Malaria incidence	Morbidity	Reproductive, maternal, newborn, child and adolescent	Morbidity	Morbidity	Access	Morbidity	Health financing	Morbidity
Ambient air pollution in urban areas	Environmental risk factors	Noncommunicable diseases	Environmental risk factors	Environmental risk factors	Environmental risk factors	Environmental risk factors	Health financing	HIV
Total correct	**—**	**0 (0%) (-5)**	**13 (93%) (-1)**	**9 (64%) (-2)**	**5 (36%) (-4)**	**11 (79%) (+4)**	**4 (29%) (-1)**	**1 (7%) (0)**

Side-note: Air pollution classified as "HIV"? Do the machines know something we don't?

About 44% accurate overall (43/98)

Now let's use a full list of Top 100 metadata as well as a full list of WHO/ UNICEF metadata. Since Support Vector Machines, Trees, and Neural Networks are consistently bad given our sample, we'll limit this final classification example to Entropy, Boosting, Bagging, and Random Forests.

Table 6: WHO/ UNICEF classified using Top 100 (indicator names and metadata for both sets)

Indicator	Actual group	Entropy	Boosting	Bagging	Forests
Proportion of all deaths attributable to HIV/AIDS	Mortality by cause	Health information	Mortality by cause	Mortality by cause	Noncommunicable diseases
Percentage of health facilities with the capacity to deliver appropriate care to people living with HIV and AIDS	HIV	HIV	Health security	Mortality by cause	Mortality by cause
HIV prevalence among pregnant women	Morbidity	Reproductive, maternal, newborn, child and adolescent	Reproductive, maternal, newborn, child and adolescent	Noncommunicable diseases	Noncommunicable diseases
Migrants: HIV prevalence	Morbidity	Morbidity	Mental health	Morbidity	Morbidity
Prisoners: HIV Prevalence	Morbidity	Morbidity	Mental health	Morbidity	Morbidity
AIDS-related mortality	Mortality by cause	Mortality by cause	Mortality by cause	Mortality by cause	Mortality by cause
Total correct	—	**4 (67%)**	**2 (33%)**	**4 (67%)**	**3 (50%)**

Results are similar, although this is a much smaller sample. Additionally, all three of these tables show results using TF/ IDF weights while there may be good arguments for using simple TF weights when we include a lot of metadata. One final table …

Table 7: WHO/ UNICEF classified using Top 100 (names, metadata, and TF weights)

Indicator	Actual group	Entropy	Boosting	Bagging	Forests
Proportion of all deaths attributable to HIV/AIDS	Mortality by cause	Health information	Mortality by cause	Mortality by cause	Mortality by cause
Percentage of health facilities with the capacity to deliver appropriate care to people living with HIV and AIDS	HIV	HIV	Access	Mortality by cause	HIV
HIV prevalence among pregnant women	Morbidity	Morbidity	Morbidity	Morbidity	Morbidity

Migrants: HIV prevalence	Morbidity	Morbidity	Mental health	Quality and safety of care	Morbidity
Prisoners: HIV Prevalence	Morbidity	Morbidity	Mental health	Quality and safety of care	Morbidity
AIDS-related mortality	Mortality by cause	Mortality by cause	Mortality by cause	Mortality by cause	Mortality by cause
Total correct	**—**	**5 (83%) (+1)**	**3 (50%) (-1)**	**3 (50%) (-1)**	**6 (100%) (+3)**

Step 4: Data crosswalking
The unicorn of data …

A. Tagging and classifying
   a. Unsupervised (i.e. LDA) classification for cross-organizational comparisons

Our original approach was to calculate word frequencies and use these as topical keywords. However, this approach is biased towards frequent words that are not pertinent to health topics—e.g. percentage, number, year, total, rate, etc. This is also the case with other unsupervised methods such as LDA—methods that use term frequency will overemphasize unhelpful terms. For example, the string "Meningitis - number of reported cases" will likely be classified with the word "number" instead of "Meningitis," which is the more informative keyword.

There are a few ways to deal with this:

1. Run an "anti-copier" string loop—this basically looks for unique words in strings and makes this the "string topic" (basically it just re-emphasizes unique words in the string without getting rid of the old words)

Um, ya—nevermind. It doesn't really work that well. First of all, you need a large sample to get okay results (i.e. >500) and even then, it depends largely on the assumption that variables (indicators) of similar topics are named similarly, which is rarely the case, even within organizations. All that being said, it does work wonders on categorical variables. For example:

- Meningitis number of reported cases
- Malaria number of reported confirmed cases
- Poliomyelitis number of reported cases
- Yellow Fever number of reported cases

For these, it's going to pull out "Meningitis," "Malaria," etc. just like we'd want it to. However, other "malaria"-oriented variables are not likely to have the string "reported cases," and thus it is likely that "malaria" won't be the only thing pulled out of these other strings. Basically you need superfluous words to be so common and included in variable groups that they easily exclude themselves.

2. Manually classify frequent words as topics or non-topics, delete non-topics, and re-run (i.e. manually create a list of "stop-words")

This works, but honestly it's just quicker to create a list of topics you're looking for—stop-word creation can take a *lot* of iterations and can still miss certain words you may be interested in. For example, the word "meningococcal" may only appear once in your entire dataset, but you'll still want to This then becomes a supervised method (see below).

3. Collect more topical information about indicator strings

Doing this…

B. Supervised (i.e. machine learning) classification for comparison with the Top 100
C. Scraping and comparing the actual data
D. Mapping (incorporating AidData?)

We have researched data from the following sources:

1. * World Bank HealthStats
2. World Bank Microdata Library
3. * World Bank World Development Indicators
4. * WHO Global Health Observatory (GHO)
5. * WHO Mortality Database
6. Global Fund
7. * Global Fund Monitoring and Evaluation Requirements
8. * Demographic Health Survey (DHS)
9. USAID Document Experience Clearinghouse
10. Global Health Data Exchange (GHDx)
11. More organizations here…

Step 1. We need to be able to narrow down comparable sets of projects. To do this we need to figure out how WHO measures HIV vs. how WB measures HIV at an aggregate level—i.e. we need to put each organization's indicators into a category. Then, we can start searching for organization-specific projects that relate to those indicators. When we find one, we will, ideally, already know what topic and category it belongs in and what it should thus be compared to. Basically we're creating a set of tags for each organization and then we can classify project reports based on those tags.

Couldn't we just create the tags based on the language of the report? Yes, actually—then we could compare these tags to the language of the Top 100 indicators and see where the project fits.

We're going to do all of this with the indicator names first in order to show that it's possible—i.e. we will create "tags" for each indicator name and compare it to the Top 100. Then we can show how these indicators could potentially be compared within these subgroups.

What are the advantages? Aren't we really just boosting sample size? I mean, pretty much yes, but the main advantage is boosting sample size *across* organizations instead of just within. This allows a level of understanding and coordination that is impossible now. This is more the case insofar as we can incorporate geographic information.

Note: in the end, results may simply not be comparable—that's okay. Just show what could be done if they were comparable/ what is lost with them not being.