UN SEMINAR BIG DATA

Arnaud Pincet - Policy and Strategy Team
OECD Development Cooperation Directorate

BETTER POLICIES FOR BETTER LIVES

Preliminary questions

- How many of you are statisticians?
- How many of you use R or Python in your work?
- How many of you are doing or have done regressions (time series, logit, multivariate,…)?
- How many of you have tried successfully or unsuccessfully to train and deploy a ML algorithm?
- Who knows StackOverflow? StackExchange?

Are public and private actors focusing on the right SDGs? How would we know?

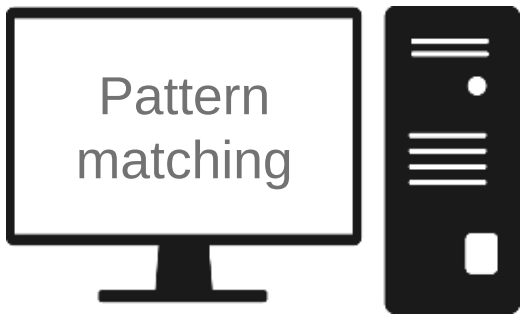
What is the issue?

- Most of the “trillions” in sustainable finance are already here but need to be better targeted
  - Are we focusing on the right areas?
  - What are the SDG darlings and orphans?
- The problem is not that there is not enough information but too much information

The OECD develops a new tool that helps managing these large flows of information and provide mapping of both private and public sector contribution to the SDGs

The OECD is using Machine Learning and Natural Language Processing to map descriptions of aid projects

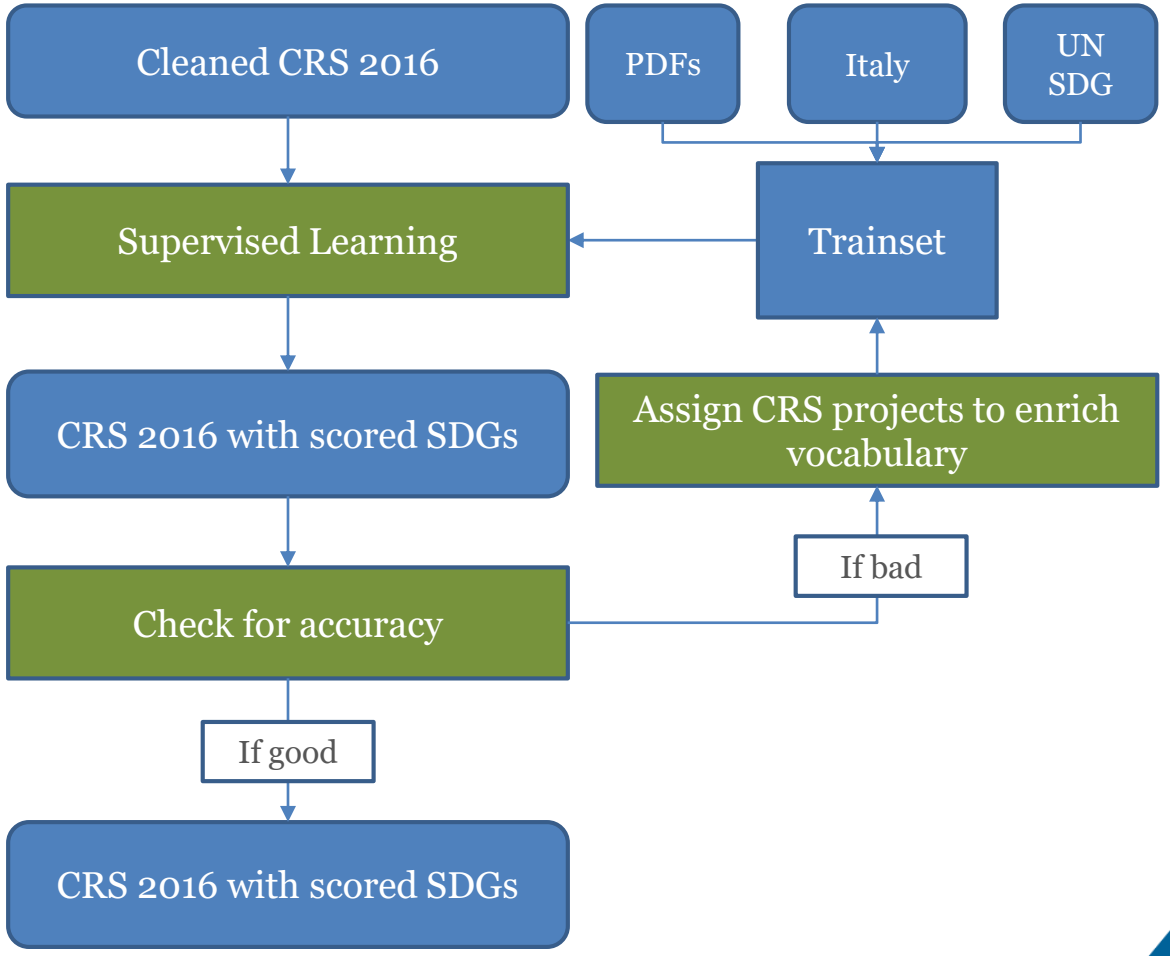
Unstructured Descriptions



Projects Mapped into SDGs

The main algorithm uses supervised learning techniques on sentences converted in vectors

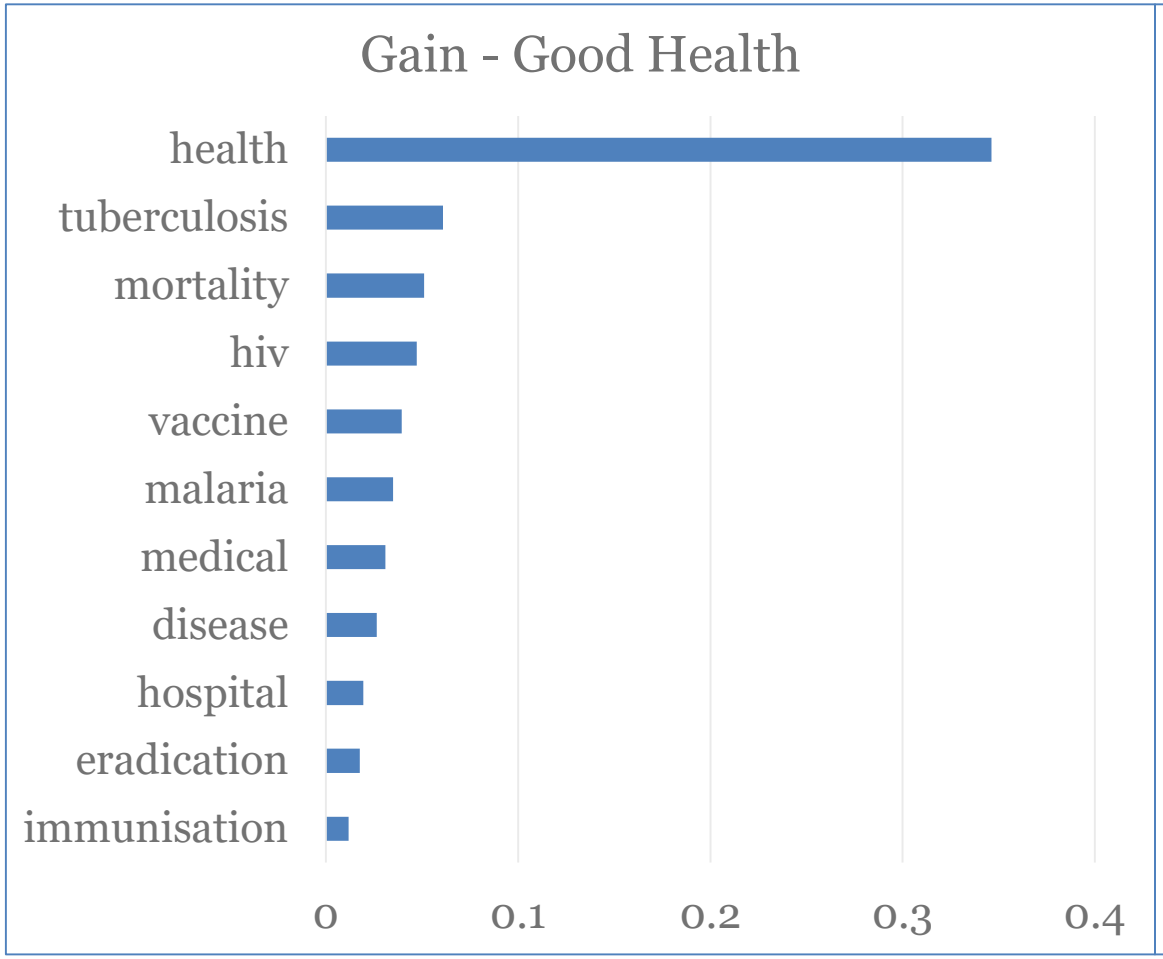
- Created a trainset with PDFs talking about a specific SDG, some projects assigned by Italy, and the UN description of the SDGs
- Asked Policy analyst to manually label 10'000 aid projects to improve accuracy
- Used 2 algorithms that provide predictions on a text to belong to one or many SDGs
- We used purpose codes and other methods to check whether we had some vocabulary missing.

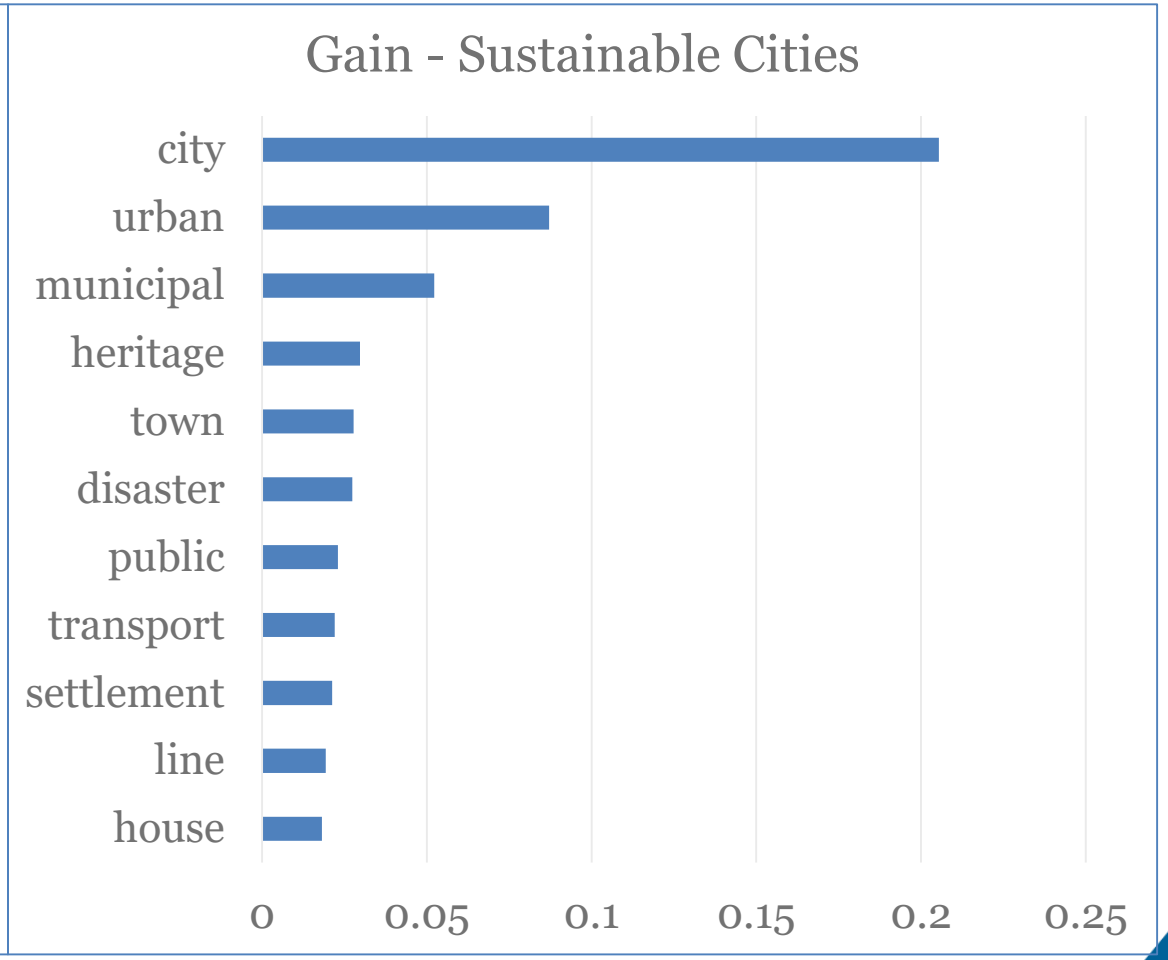


Since SDGs' definitions are interpreted differently even among experts we used different text sources for training

- The PDFs comes from different organisations such as:
  - International organisations (UN, OECD, WB, ADB, …)
  - Private organisations (KPMG, UBS,…)
  - Researcher (universities, think tanks,…)
- 206 PDFs have been selected and assigned to its respective SDG
- Each paragraph of each PDF represents a source of vocabulary that the computer will learn from

Based on the texts, the algo. determines the relative importance of words in order to link a project to a certain SDG





CRS 2016, preliminary results

It assigns one, multiple, or zero SDG based on the score of each text for the 17 categories

Description	Poverty	Hunger	Health	Educ.	Gender	Water	Energy
Department for International Development other **education** sector support programme in Nigeria - procurement of services the planning, financing and delivery of sustainable and replicable basic **education** services in terms of access, equity and quality are improved at federal level and up to six states.	0.01	0.02	0.01	0.84	0.01	0.01	0.01
Federal ministry of finance donor government subsidy for frame ii export credit charges: **solar driven** warm **water supply** and cooling for dr. Alejandro Davila Bolanos **hospital**	0.01	0.02	0.63	0.01	0.01	0.63	0.25
Donor's* international development agency donor government preliminary services rendered abroad - Europe	0.01	0.02	0.02	0.01	0.01	0.01	0.01
French development agency other tskb line of credit focused on **workplace** safety and **women employment**.**	0.02	0.02	0.01	0.01	0.62	0.01	0.01

* Anonymised donor name

**SDG Decent Work and Eco. Growth (SDG 8) also detected for this example

Using this tool on the aid database (OECD CRS) we can map official aid to the SDGs

Distribution of Official Development Assistance (ODA) from Development Assistance Committee (DAC) members

USD disbursement, CRS 2016

http://em-sbx-dev-8.main.oecd.org:8008/

Because the algorithm attributes one or multiple SDGs can also explores the synergies between them

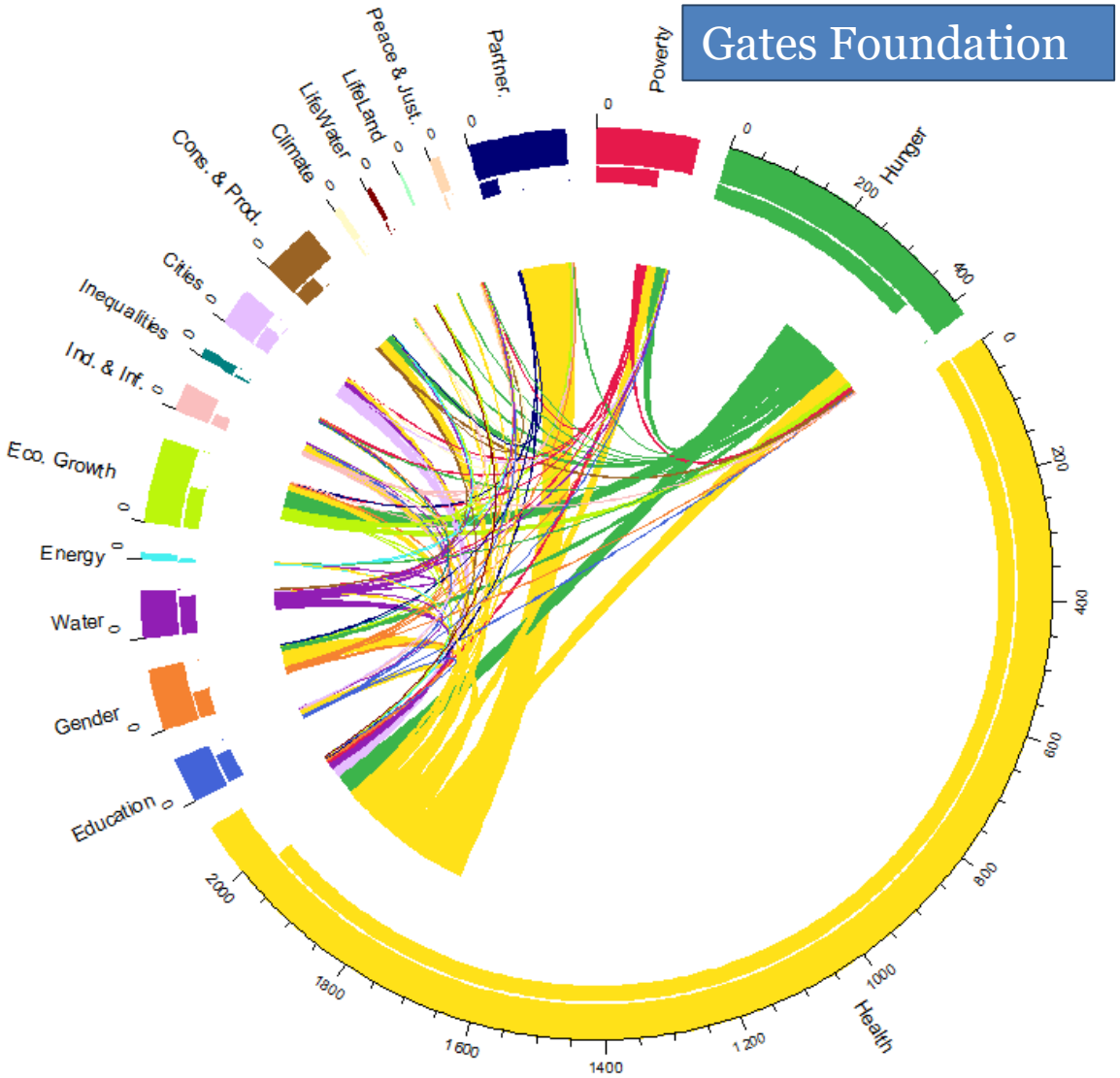
What are the most linked SDGs for the Bill and Melinda Gates Foundation?

Do all donors have the same interlinkage or is it due to donor's specific policies?

Disbursement, CRS 2016, USD Millions
Links represents projects that targets two SDGs

We can identify cross-cutting projects such as

- Gender-based health policies
- Water and hygiene
- Nutrition and health



For senior managements, the tool provides some insights on alignment of fund disbursed compare to their policy priorities

Kathryn, Senior Policy Advisor,
Development Agency HQ

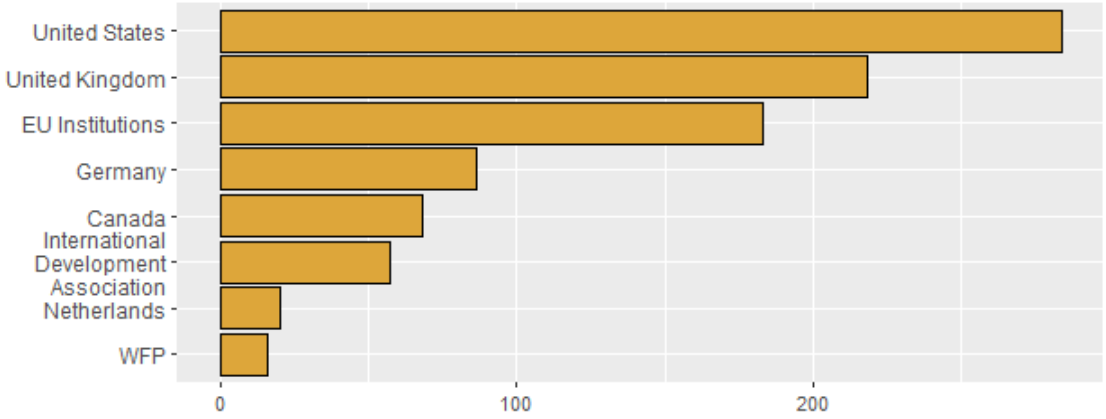
"With a tool that tracks the SDGs I can quickly understand where our aid portfolio stands and see where we specialise at the global and regional level."

The tool allows to:

- Manage the aid portfolio according to the SDG at the national level
- Check for policy coherence and fit with current donor's strategy
- Better communicate donors' action to its multiple stakeholders by linking the CRS statistics to the political agenda

For program managers, the tool helps greater reporting and provides some strategic insights on donors' positions

Distribution of Official Development Assistance (ODA) disbursed in Ethiopia by all donors, *USD disbursement, CRS 2016*



Top 8 donors of ODA in Ethiopia for Zero Hunger, preliminary results, *USD disbursement, CRS 2016*

Mark, Program Manager - Food,

Addis Ababa Office, Ethiopia

"*To start a new initiative in Ethiopia, I can assess the aid profile of Ethiopia and identify the key partners that are already working in the field of Zero Hunger.*"

The tool allows to:

- Assess the SDGs darlings and orphans at the recipient levels
- Assess donor's specialisation and comparative advantage
- Map aid gaps to increase aid effectiveness
- Reduce reporting burden by automatically suggesting SDG to the manager

For statisticians, the tracker alleviate the reporting burden and ensure greater quality controls

James, Statistician,
Development Agency HQ

"With more than 1'000 projects to control per year, I cannot read all reports from the grant managers. The tool allows me to verify the accuracy of information reported and to engage with the grant managers only on the identified issues"

The tool allows to:

- Spot errors more easily: possibility to highlights projects where potential inconsistencies emerges between fields
- Better communicate donors' action to its multiple stakeholders
- Keep CRS stable, and avoid the "political flavour of the month"

Project 1: "Nutrition and agricultural development in the Afar region"

Algorithm also identifies it as SDG2:
OK, no need for human control

Project 2: "Food security and agricultural development in Afar region"

Algorithm also identifies it as SDG2
but does not SDG 5, flagged to James

"Hi Mark, can you provide more details on how this project contributes to reducing the gender gap?"

Using the tools on other datasets: Corporate Social Reports (CSR) can provide valuable information

51 766 reports in GRI Sustainability Disclosure Database

PEPSICO

TARGET

aetna

"Ensure the safety of our **workplaces** to sustain of **forced labour**, **child labour**, and **human trafficking** thus threatening freedom, superior performance and shattering dignity."

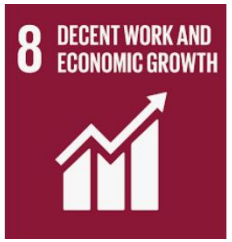
PepsiCo CSR Report 2016

"Divert 70 percent of our retail **waste** from **landfills** through **reuse** or **recycle** programmes by 2020."

Target CSR Report 2016

"Stretching to **Fight Poverty** in Georgia Aetna was the Presenting Sponsor of the second annual Atlanta **Women's** Foundation Yogathon which raises funds for **women** and **girls** fighting to overcome **poverty** in the Atlanta metro area"

Aetna Inc. CSR Report 2016



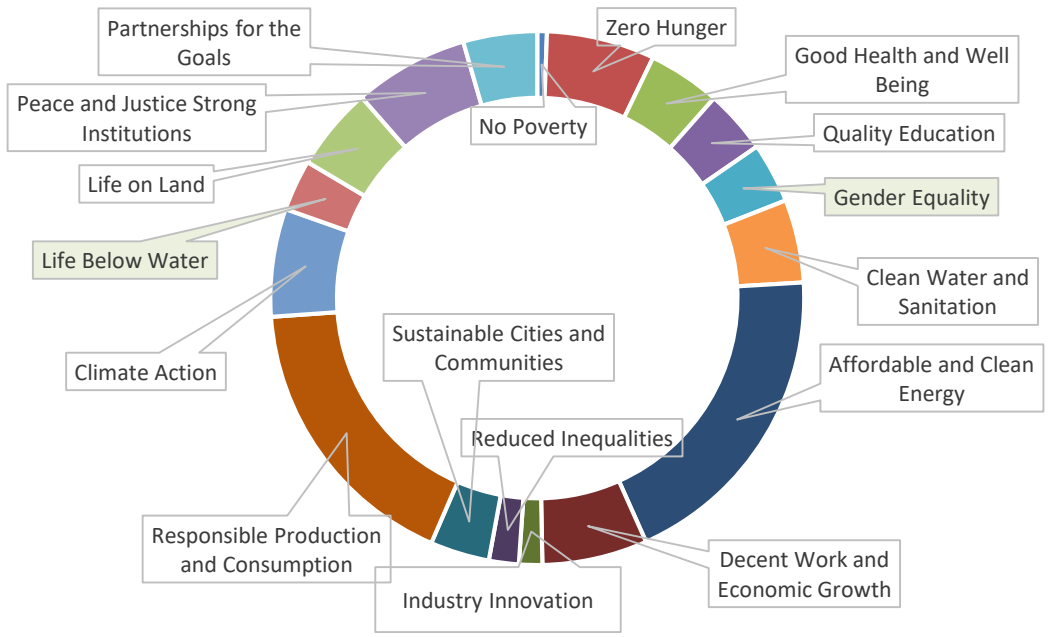
See: GRI, *Integrating the SDGs into Corporate Reporting: A Practical Guide*

The tracker opens multiple area for new policy research based on the many text sources available

Private sector involvement:

- OECD DAC partnered with GRI to investigate which SDGs are mentioned in Corporate Social Responsibility reports
- Preliminary results indicates that gender Equality or Life below Ocean is rarely mentioned

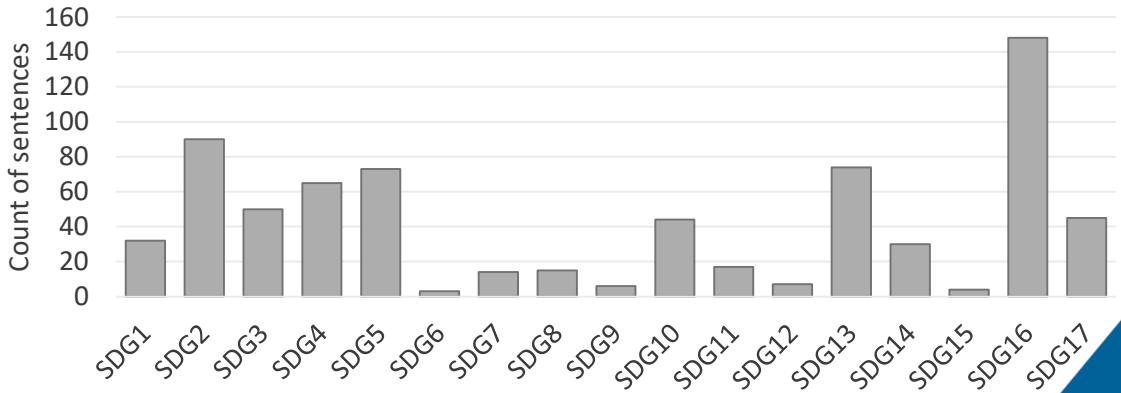
Count of sentences mentioning SDGs in selected Fortune 500 CSR reports, preliminary results



The DAC or other institution could conduct research on many other topics:

- Assess national contribution to the SDGs by "scanning" the laws passed in a particular country
- Explore what are the most SDG related topics mentioned in the news and check whether these are addressed at the international level
- Check for policy coherence by assessing countries reports/speeches/tweets and compare them to their current SDG progress

A Better World: Ireland's Policy for International Development



http://em-sbx-dev-8.main.oecd.org:8009/

The current tracker has some limits that we hope to tackle in the future

- The tracker can identify whether a SDG is mentioned in a sentence, it cannot however assess the impact:
  - *Training 100 doctors* will be categorised similarly to *Training 1'000 doctors*.
- For CRS a project linked with multiple SDGs is counted equally in all SDGs, we do not differentiate allocation.
- Projects without specific enough descriptions cannot be assigned
- Current algorithm cannot differentiate between positive and negative impact. This issue can be addressed in the future depending on resources available
- Current algorithm struggles with very specific words. This issue can be addressed in the future depending on resources available:
  - *Neuroplasticity* or *endocrinology*, might be hard to identify for SDG3

The tool is still in its development phase; we are looking for inputs and comments from policy experts

In the upcoming months the secretariat will:

- **Make the algorithm available & open source**
- **Design workshops** to train statisticians and researchers in capitals and development agencies
- **Publish current results** on a dedicated website
- **Open a competition** for best use case in collaboration with DAC members universities

Additional resources are needed to:

- **Improve the algorithm** and **host it** on a robust platform
- Engage with policy makers, statisticians, and universities to promote the tool as well as **co-organise a conference on SDG investment**

Lessons learned & community engagement

At the stakeholder level

- It is hard to communicate the distinction between statistics and estimates
- Policy makers do not reason in probabilistic ways: “it works 90% of the time” is not accepted

At the SDG level

- Because SDGs are still a blurry concept, even for humans, NLP can help to focus the debate

At the project level

- ML is not understood well by senior management requiring a lot of communication and stakeholders engagement
- If applied to meaningful datasets it can provide high impact and large results
- Barriers to entry are low and there are plenty of resources available online

The OECD will organise a webinar on May 7th 3pm CET on how to use the tool
If you wish to participate please send us an email: arnaud.pincet@oecd.org

Pincet, A., S. Okabe and M. Pawelczyk (2019), "Linking Aid to the Sustainable Development Goals – a machine learning approach", *OECD Development Co-operation Working Papers*, No. 52, OECD Publishing, Paris, https://doi.org/10.1787/4bdaeb8c-en.

Visualising the aid flows: http://em-sbx-dev-8.main.oecd.org:8008/

Demonstration of the algorithm: http://em-sbx-dev-8.main.oecd.org:8009/

Methodology – Enhancing the descriptions

- Creation of the Description column, containing the Agency Name (e.g. Green Climate Fund or Ministry of Foreign Affairs) and the Channel Name, the English Short Description and Long Description.

- Replacing frequent abbreviations (e.g. SME) and acronyms (e.g. UNHCR).

- For identical descriptions, we only considered them once for our work.

- Removing “stopwords” (e.g. “the”, “and”) but also the name of the DAC countries, main donors and related organisations (e.g. USAID), to avoid biases for the policy analysis.

Methodology – From words to vector

- Lemmatisation of the words, so that the same word in another form (plural, conjugated) are considered as the same word.
  Example: “working”, “worked”, “works” and “work” become “work”
- Transformation of the data frame into a Document-Term Matrix (DTM): a matrix of the frequencies of each word in each document.
- Unigrams and bigrams (e.g. health care) have been considered for the DTM.
- Suppression of too rare unigrams or bigrams: if they are appearing in less than 30 documents in the training dataset, it is not in the DTM.

	Health	Programme	Vulnerable	Child
American health programme for children	1	1	0	1
European programme for vulnerable child	0	1	1	1

The two main types of Machine Learning

Supervised learning

- You know how to classify the input data and the type of behavior you want to predict
- Human driven
- ➢ Example: Nicolas' classification of migration projects in CRS

Unsupervised learning

- You do not know how to find the data and you use the algorithm to find it for you
- Algorithm infers structure from the data, identifies group with similar behaviour
- ➢ Example: finding the themes in different chapters of a book

Algorithms tested

- Supervised learning
  - Binomial/Multinomial Lasso
  - Support Vector Machine
  - Simple Trees
  - More Advanced Trees:
    - Random Forest
    - XG Boost
- Unsupervised learning
  - K-Means Clustering
  - Hierarchical Clustering using various linkage specifications
    - Complete
    - Average, …. Etc.

XGBoost explained

- XGBoost (eXtreme Gradient Boosting) is a widely used algorithm and often wins Machine Learning competitions.
- Advanced version of boosting.
- Weak learners (slightly better predictions than random guess) are weighted according to their performance in order to create a strong learner (good predictions). For XGBoost, the weak learner is a decision tree.

Random Forest Explained

- Classical Tree Method
  - Select a predictor X_j and a cutpoint s such that splitting the predictor space into non overlapping regions results in the greatest possible reduction in our objective function
  - At each step of the tree building process: make the best split at the particular step
  - Suffer from **high variance**. The problem is that slightly different observations can lead to a different split in the beginning of the tree growing process

- Random Forest
  - Now: sample B times from the data with replacement and for every sampled data set we build a separate tree model and record the predicted class for every observation → Variance Reduction
  - We can do another Trick: since all trees are formed using the same predictors with similar observations and therefore the predicted values for each observation will be highly correlated among the B predictions (results from Variance formula)
  - At every step of the tree splitting process, a random sub sample, say m, of the predictors are chosen. Produced trees are now more different. → Less correlated trees → Further Variance Reduction

Limitations – at the SDG level: Lack of Common Definition

- If SDGs are interlinked, some goals are quite confusing or overlapping:

	8.9: By 2030, devise and implement policies to promote sustainable tourism that creates jobs and promotes local culture and products
	12.b: Develop and implement tools to monitor sustainable development impacts for sustainable tourism which creates jobs, promotes local culture and products

	1.b: Create sound policy frameworks […], based on pro-poor and gender-sensitive development strategies, to support […] poverty eradication actions
	5.c: Adopt and strengthen sound policies and enforceable legislation for the promotion of gender equality and the empowerment of all women and girls at all levels