ATLAS: A spend classification dataset for estimating scope 3 emissions

Andrew Dumit, Krishna Rao, Travis Kwee, Jonathan Glidden, Varsha Gopalakrishnan, Katherine Tsai, Sangwon Suh

3°

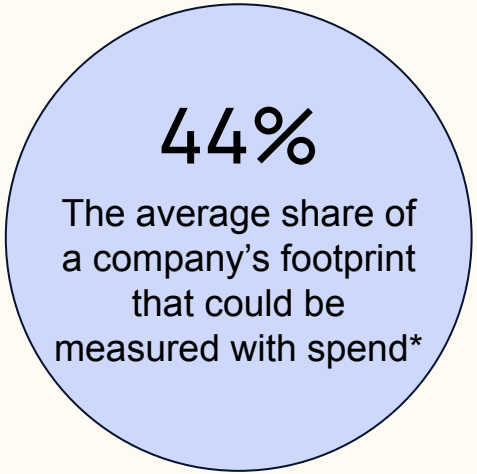
tCO_2e

Why do companies perform spend classification?

Scope 3 emissions are indirect GHG emissions that occur from a company's value chain.

Because they are not from sources owned by the company, they are typically estimated.

Spend classification is a crucial task that enables companies to estimate a large fraction of their emissions.



Watershed

* Estimated by computing the share of Scopes 3.1 + 3.2 divided by total footprint. We used the population of CDP reporters from the 2023 reporting year that reported 3.1 or 3.2
† Estimated from CDP 2023 reporting. Based on the share companies who reported scope 3.1 or 3.2 and reported using a spend-based model.

Company Ledgers

General ledgers are a “source of truth” for all business activity

These are complex datasets covering various hierarchies at differing levels of detail

Mapped to industry category codes

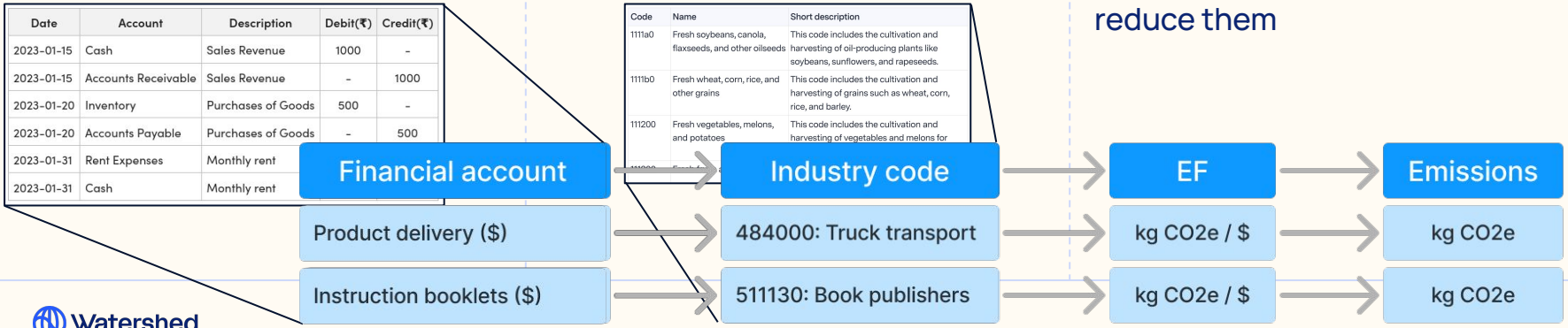
Industry codes classify the primary business activity occurring

There are about 400 industry codes spanning resource extraction, mfg, retail, services

Resulting emissions

Based on the industry code, average emissions per $ from that industry are applied to the spend

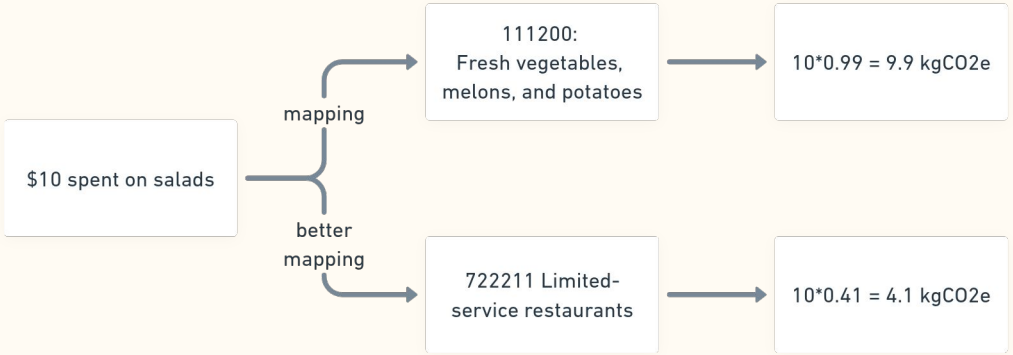
The resulting emissions are used to understand what emissions a company is responsible for and so it can take steps to report on and reduce them



Watershed

Why is it important to get it right?

- Scope 3 is frequently the largest portion of the carbon footprint for a company!
- Similar sounding codes have different emissions factors. For example, mapping health insurance to 'insurance agencies and brokers' (the people who sell insurance) vs to 'insurance carriers' (the people providing insurance services) would estimate emissions to be about 16% higher than they should be.
- A more extreme example shows how for "salads", a user could realize >100% difference by wrongly selecting fresh vegetables instead of limited service restaurants.



The ATLAS Dataset

A few examples from ATLAS dataset

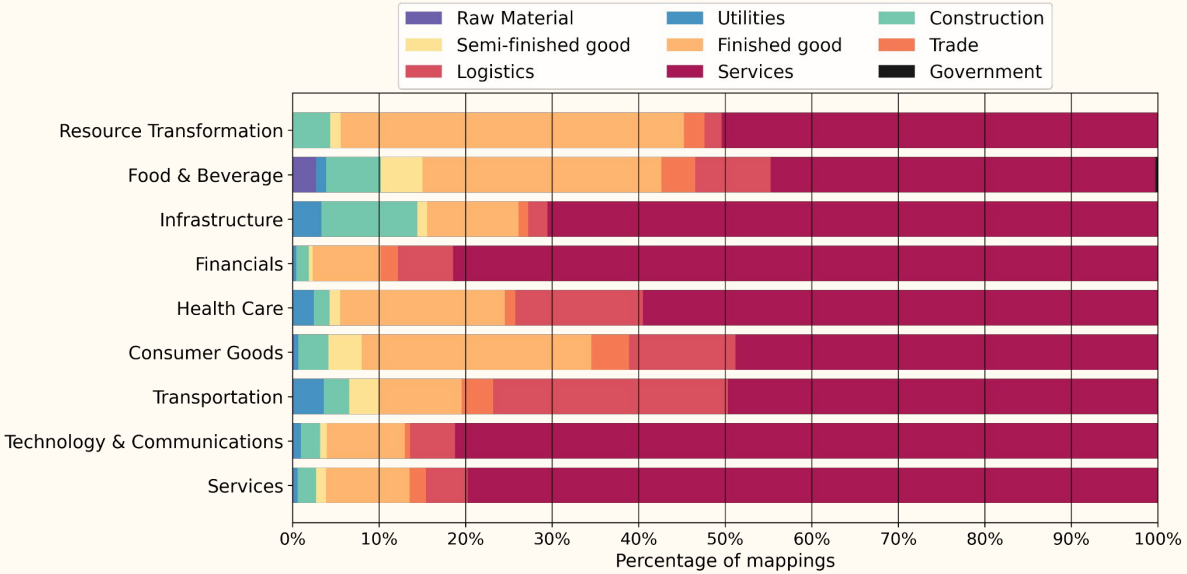
Real spend description	Synthetic spend description	Mapped industry category
Hidden to protect privacy	100A1: video Production. Expense	541800: Advertising, public relations, and related services
	Payment & Commercial Pressure Washer (misc.)	230301: Nonresidential maintenance and repair
	long-distance auto exp GHX	484000: Truck transportation

A few examples from ATLAS dataset

Input	Label
Synthetic spend description	Mapped industry category
100A1: video Production Expense	541800: Advertising, public relations, and related services
Payment & Commercial Pressure Washer (misc.)	230301: Nonresidential maintenance and repair
long-distance auto exp GHX	484000: Truck transportation

The ATLAS dataset

- 10,000 synthetic mappings with labels
- Represents 73% of the types of global economic activities
- Matches the distribution of corporate activities from 9 macro industries across 8 countries

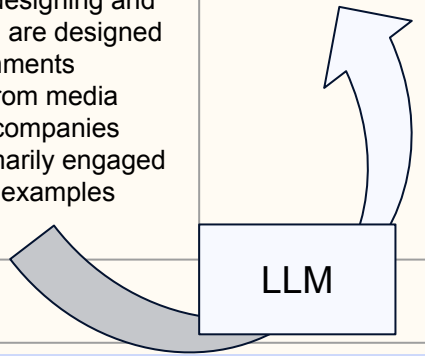


The ATLAS dataset protects privacy while mimicking real world distribution of corporate expenses: examples

Real spend description	Mapped industry category	Mapped industry description
Hidden to protect privacy	541800: Advertising, public relations, and related services	This industry comprises: establishments primarily engaged in creating advertising campaigns and placing such advertising in periodicals, newspapers, radio and television, or other media. These establishments are organized to provide a full range of services, including advice, creative services, account management, production of advertising material, media planning, and buying; establishments primarily engaged in designing and implementing public relations campaigns. These campaigns are designed to promote the interests and image of their clients. Establishments primarily engaged in purchasing advertising time or space from media outlets and reselling it to advertising agencies or individual companies directly; establishments of independent representatives primarily engaged in selling media time or space for media owners. Illustrative examples include …
		… 9,999 more

The ATLAS dataset protects privacy while mimicking real world distribution of corporate expenses: examples

Real spend description	Mapped industry category	Mapped industry description	Synthetic spend description
Hidden to protect privacy	541800: Advertising, public relations, and related services	This industry comprises: establishments primarily engaged in creating advertising campaigns and placing such advertising in periodicals, newspapers, radio and television, or other media. These establishments are organized to provide a full range of services, including advice, creative services, account management, production of advertising material, media planning, and buying; establishments primarily engaged in designing and implementing public relations campaigns. These campaigns are designed to promote the interests and image of their clients. Establishments primarily engaged in purchasing advertising time or space from media outlets and reselling it to advertising agencies or individual companies directly; establishments of independent representatives primarily engaged in selling media time or space for media owners. Illustrative examples include …	100A1: video Production Expense
		… 9,999 more	



Baseline results

Baseline models

We evaluate four baseline models on the ATLAS dataset

1. **Text embedding:** we embed the spend description using OpenAI's text-embedding-large model and compare to embeddings of metadata about each industry code and sort by cosine similarity.

2. **Logistic regression:** we train a logistic regression model on the bag of words representation of the spend description.

3. **LLM with prompt engineering:** construct a prompt with instructions and CoT for an LLM (Claude 3.5 Sonnet). All categories are provided in the prompt.

4. **LLM with prompt eng + fewshot examples:** add 50 top examples to the prompt. Top examples are retrieved from the training set based on cosine similarity of embeddings from OpenAI's text-embedding-large.

Results on the benchmark

Model	Top 1 accuracy	Top 3 accuracy	Top 1 DoF* accuracy
Text embedding	37.1%	57.3%	68.7%
Logistic regression	48.5%	60.2%	73.8%
LLM	40.6%	57.3%	75.9%
LLM with Fewshot	**57.3%**	**72.2%**	**82.8%**

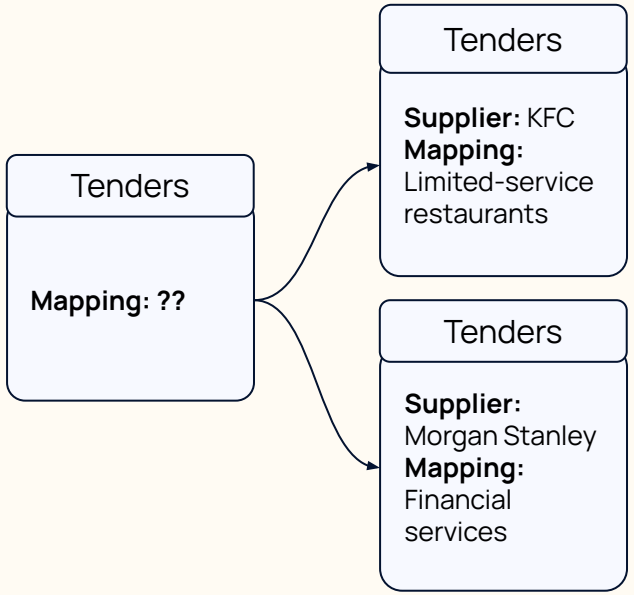
Watershed

* DoF = Degree of Fabrication. The DoF, from input-output analysis, describes where an activity lies in the value chain, between raw materials (low DoF) and finished products (high DoF)

Future directions

- Model improvements
  - Prompt improvements
  - Better retrieval strategies
  - Fine-tuning
  - Determining when to ask for more information

- Expanding the contextual information
  - Financial accounts can have more detailed information underlying them



Recap

Scope 3 is typically the largest source of emissions for companies measuring their emissions inventory.

Spend classification is a crucial part of measuring scope 3.

An incorrect or inconsistent spend classification model can lead to serious over- or underestimation of emissions.

We introduced ATLAS, a benchmark for spend classification, and provided 4 baseline models.

The enterprise
sustainability platform