OUR COMMUNITY TECH · FIELD GUIDE

Build Your Own Data Agent

A free, 30-minute guide for nonprofits, educators, and anyone who assumed this kind of tooling was out of reach.

Author	Kim Wright, Our Community Tech
Published	April 2026
Time to build	~30 minutes
Prerequisites	A web browser. That's it.
Companion code	https://ourcommunity.tech/build-your-own-data-agent

WHY THIS EXISTS

On April 21, 2026, OpenAI's head of ChatGPT engineering, Sulman Choudhry, described **Kepler** — OpenAI's internal data agent, built by two engineers in three months, now serving around 4,000 of OpenAI's 5,000 employees daily. His framing was that the models, the APIs, and the orchestration are all public — anyone can replicate this. This guide takes that invitation seriously, for the organizations that usually assume it isn't meant for them.

WHAT YOU'RE BUILDING

A data scientist in your pocket

By the end of this guide you will have a working tool that lets anyone in your organization type a plain-English question about your data — and get a plain-English answer back. No spreadsheets to open, no pivot tables to build, no one waiting on the one person who knows how to do it.

THE HONEST FRAME

OpenAI built Kepler with two engineers in three months. We built a version your nonprofit can run in thirty minutes. The point isn't that they're the same thing — they aren't. The point is that the gap between 'serious infrastructure' and 'something your team can actually use' is a lot smaller than it used to be.

How it works

Someone on your team asks a question in plain English — for example, *"Which donors gave less this year than last year?"* The agent looks at a summary of your data, writes a small query that answers the question, runs that query against your CSV files, and sends the result back as a short written answer.

Two things are worth saying up front. First: this is a tutorial version, not a production data platform. It is meant to teach the pattern and give you something useful while you decide what to invest in next. Second: it is real. It runs on the same models and the same API that OpenAI uses internally. The difference between this and Kepler is mostly scale, polish, and the number of data sources connected — not the underlying idea.

What the finished product does

- Reads CSV files you drop into a folder.
- Accepts plain-English questions at an interactive prompt.
- Writes a safe, minimal query under the hood.
- Returns a short, written answer with the relevant numbers.
- Optionally shows the query it ran, so you can sanity-check.

What it does not do

- It is not real-time — it only knows what's in the CSVs at load.
- It is not meant for very large datasets (beyond ~50MB).
- It is not a replacement for a data analyst.
- It is not appropriate for protected data (PII, HIPAA, FERPA) — see the **Safety** section.

BEFORE YOU START

What you'll need

Everything below is free to sign up for. You will pay OpenAI for usage — typically a few dollars a month at nonprofit volume. See *Cost expectations* below for what to expect.

1	**A Replit account**	Free. Lets you run Python in the browser with no local install. Sign up at replit.com.
2	**An OpenAI API key**	Free to create. You add a small amount of credit (most nonprofits do $10 to start). Get it at platform.openai.com/api-keys.
3	**Your data as CSV files**	Any spreadsheet can export to CSV: File → Download → CSV. Start with the included sample data for the tutorial.
4	**About 30 minutes**	Most of which is clicking buttons. The reading is the longest part.

What you do not need

- A developer on staff.
- Experience with Python, pandas, SQL, or any other language.
- A local install of anything. Everything runs in your browser.
- A paid OpenAI subscription. ChatGPT Plus is unrelated — the API is billed separately.

THE BUILD

Step 1 — Get your OpenAI API key

The API key is how the agent talks to OpenAI's models. Treat it like a password: one per organization, stored somewhere safe, rotated if it ever leaks.

Create the key

- Go to platform.openai.com/api-keys and sign in (or create a free account).
- Click **+ Create new secret key**.
- Name it something honest — *"ourcommunity-data-agent"* is fine.
- Copy the key immediately. It starts with **sk-**. You will not be shown it again.
- Paste it somewhere safe for the next ten minutes (a password manager is ideal; a sticky note is not).

Add a small amount of credit

A new OpenAI account needs a credit card and a minimum purchase (currently $5). Go to **Billing → Payment methods**, add a card, and add $10 of credit. This will last most nonprofits several months at tutorial volume. See *Cost expectations* later in this guide for specifics.

IF YOUR ORGANIZATION DOESN'T LET YOU EXPENSE AN API

OpenAI for Nonprofits exists and includes usage credits. It takes a few weeks to get approved, but if you're going to keep building, apply early. Start at openai.com/nonprofits.

Step 2 — Fork the Replit template

Replit is a browser-based coding environment. For our purposes, it means you can run the agent without installing anything on your own computer.

- Go to https://ourcommunity.tech/build-your-own-data-agent and click **Clone the Replit template**. This opens the template in Replit.
- Click **Use template** (or **Fork**, depending on the UI). This gives you your own private copy.
- Replit will take a minute to set up — it installs the two Python packages the agent needs (pandas and the OpenAI client). You do not need to do anything during this step.

Add your API key as a secret

Replit has a **Secrets** panel for storing API keys. Never paste your key directly into the code.

- In the left sidebar of your Replit, click the padlock icon (**Secrets**).
- Click **New secret**.
- Key: **OPENAI_API_KEY** (exactly, all caps, with the underscore).
- Value: paste the **sk-...** key from Step 1.
- Click **Add secret**.

WHY THIS MATTERS

If you paste your key into the code directly and later share the Replit with a colleague (or make it public), you've just given them your billing key. The Secrets panel keeps the key out of the code and out of any copy of the project.

Step 3 — Run the sample tutorial

Before you point this at your own data, run it on the synthetic nonprofit dataset that ships with the template. Three CSV files live in **/data**:

`donors.csv`	60 fictional donors with giving history — first gift date, lifetime giving, last year's gift, this year's gift.
`grants.csv`	24 fictional grants across programs and funders, with amounts, status, and reporting dates.
`program_budget.csv`	7 program budgets for FY2025 with budgeted vs. actual spending and variance.

Click Run

Press the big green **Run** button at the top of the Replit. You should see a banner that looks like this:

```
──────────────────────────────────────────────────
 Your Nonprofit's Data Agent  —  minimum viable version
──────────────────────────────────────────────────
 Loaded: donors, grants, program_budget

 Ask a question in plain English.
 Type  quit  (or press Ctrl-C) to exit.
 Type  /code  to toggle showing the generated query.
──────────────────────────────────────────────────

You ▸
```

Ask your first question

At the **You** prompt, try any of these:

- *Which donors gave less this year than last year, and by how much?*
- *Who lapsed this year (gave last year, nothing this year)?*
- *What was our average grant size by program?*
- *Which programs are over budget, and by how much?*
- *Which funders gave us more than one grant?*

If you want to see the query the agent wrote, type `/code` and press Enter. The next time you ask a question, it will print the generated code before running it. This is the single most useful trick for building trust in the answers: you can see exactly what it did.

Step 4 — Swap in your own data

Once the sample is working, replacing it with your own data is a three-minute task. Before you do this, read the **Safety** section on the next page. There are kinds of data this minimum version is **not** appropriate for.

Export your data to CSV

- Open your source (Excel, Google Sheets, Numbers, or a report from your donor database).
- For each table you want the agent to see, use **File → Download → CSV** (or equivalent).
- Give each file a short, clean name: **donors.csv**, **grants.csv**, **programs.csv**. No spaces, lowercase.

Upload to Replit

- In your Replit, find the **data/** folder in the file tree on the left.
- Drag your CSV files into that folder. (Or right-click → **Upload file**.)
- Delete the sample files once your own files are in place — keeps things tidy.

Tell the agent which files to load

Open **agent.py** in the editor. Near the top, find the `DATA_FILES` dictionary. Edit it to match your filenames:

```
DATA_FILES = {
    "donors":   "data/donors.csv",
    "grants":   "data/grants.csv",
    "programs": "data/programs.csv",
}
```

The *keys* on the left are short names the agent will use internally — keep them lowercase and plural (**donors**, **grants**, **volunteers**). The *values* on the right are the paths to your CSV files. Save the file (Ctrl-S / Cmd-S) and click **Run** again.

Ask a real question

Start with something you already know the answer to. That's the cheapest way to build calibration on whether the agent is reliable for your specific data. A good opening question is usually a count or a sum you can verify by hand: *"How many active donors do we have?"* or *"What was our total grant revenue in 2024?"*

CREDIBILITY ANCHOR

Safety — read this before using real data

Every question you ask sends a summary of your data schema and a small result set to OpenAI's API. The full dataset stays on the Replit server, but the column names, types, a few sample rows, and any result the agent computes do cross the internet to OpenAI. That has consequences.

Do not put through this tool

- PII your donors, students, or clients didn't consent to being sent to a third-party AI provider.
- Social Security numbers, tax IDs, bank account numbers. Ever.
- HIPAA-protected health data (medical records, clinical notes).
- FERPA-protected student records (grades, discipline, IEPs).
- Anything covered by a grant confidentiality clause or NDA.
- Case notes, client services records, legal matters.

Fine to put through it

- Aggregated or de-identified data (totals, counts, averages).
- Grant tracking that is not confidential.
- Public-facing financial data (the kind that ends up on your 990).
- Board-reportable metrics.
- The synthetic sample data included in the template.

A good rule of thumb: *"Would I be comfortable pasting this into an email to an outside consultant?"* If yes, it's probably fine. If no, it isn't.

How OpenAI handles data from the API

- **API requests (what this tool uses) are not used to train OpenAI's models by default.** That is stated policy for the paid API.
- **ChatGPT is different.** Do not conflate the two. This tool talks to the API, not to ChatGPT.
- OpenAI retains API data briefly for abuse monitoring (typically up to 30 days) unless you sign a Zero Data Retention agreement.
- "Not used for training" is not the same as "never seen." The data still leaves your environment.

WHEN TO GRADUATE OFF THIS VERSION

If your data is sensitive enough that it shouldn't leave your environment, move to **Azure OpenAI Service** (enterprise contract, Zero Data Retention available, data residency controls) or to a private deployment of an open-weights model (Llama 3, Qwen, Mistral) running on your own infrastructure. We help nonprofits make that call all the time — reach out.

Cost expectations

Nonprofits, rightly, hate surprise bills. Here is what this actually costs.

Usage pattern	Monthly cost (est.)	What that looks like
Tutorial only	Under $1	You run the sample questions a few times while following this guide.
Light — one user	$2 – $5	One person asks 5–20 questions a day against a typical nonprofit dataset.
Team — 3 to 5 users	$10 – $25	A few staff members use it as part of weekly check-ins and board prep.
Heavy — team dashboard	$25 – $75	Embedded in a weekly process, with scheduled reports and larger data.

How the pricing works

- OpenAI charges per **token** — a token is roughly three-quarters of a word. Every question uses some tokens going in (your question + data schema) and some coming out (the query and the written answer).
- This template uses **gpt-4o-mini** by default, which is inexpensive and more than capable for this pattern.
- You can set a hard monthly spending cap in your OpenAI billing dashboard. Do this. Even $20/month is enough to keep a stray loop from becoming a news story.

Set a usage cap

Go to platform.openai.com/settings/organization/limits and set a hard monthly limit. We recommend $25 for single-user pilots and $100 for small-team pilots. You will not hit these caps, but if something goes wrong, the cap stops it.

Troubleshooting

The most common issues and how to fix them.

"Missing OPENAI_API_KEY"

You skipped the Secrets step, or named the secret something different. Open Secrets in the left sidebar. The key must be named exactly **OPENAI_API_KEY** and the value must start with **sk-**.

"Could not find data/....csv"

Either the file is in a different folder, or you edited **DATA_FILES** to point at a name that doesn't exist. Check that the filenames in your left sidebar match the paths in **agent.py** exactly, including capitalization.

The agent gave a weird answer

Type **/code** at the prompt and ask the question again. The generated query is almost always the answer to what went wrong — most often it read a column differently than you expected. Renaming the column to something obvious usually fixes it.

It answered in the wrong currency or units

Be explicit: "Which donors gave less **in USD** this year than last year?" Column names like **gave_last_year_usd** help; column names like **amt** hurt.

Replit is slow

The first run after you fork is slow because Replit is installing Python packages. Subsequent runs should start in under ten seconds.

Billing rejected / insufficient credit

You used up your initial credit. Add more in the OpenAI billing dashboard. If you burned through $10 doing the tutorial, something is wrong — contact us.

Where this version breaks

Being honest about the limits is part of the point. This is not the finish line — it's a starting point that teaches the pattern while doing useful work.

- **Not real-time.** The agent only knows what's in the CSVs at the moment you loaded them. If your data changes daily and you need up-to-the-minute answers, you will need a scheduled refresh (see *Next tiers* below).
- **Not for very large data.** This version comfortably handles a few hundred thousand rows. Beyond that, you want a real database, a real BI tool, or both.
- **One question at a time.** There is no follow-up memory in this version — each question is independent. Conversational follow-up is a small upgrade away, but we left it out of the minimum.
- **No joins you can't describe.** If your data requires complex relationships to answer a question, you may need to help the agent by wording the question more explicitly.
- **Not a replacement for a data analyst.** It is a companion for organizations that do not have one.

Where to go from here

Once you have this running on your own data, the next tiers each solve one real problem. We are writing follow-up guides on each.

Donor database integration	Connect directly to Salesforce NPSP, Bloomerang, or Little Green Light instead of exporting CSVs. The agent always sees live data.
Multi-source agent	Combine donor data, grants, program outcomes, and accounting in one agent session — so you can ask questions that cross the boundaries.
Scheduled reports	Run a fixed set of questions every Monday morning and email the team a brief. Good for board packets and weekly standups.

AN INVITATION

If your nonprofit builds this, we want to hear about it — what worked, what didn't, what you wish the next version did. The point of publishing the pattern is to widen the set of organizations that get to use it. Email us, tag us, send a screenshot. We'll share the best of what comes back in the next guide.

Sources & further reading

OpenAI — **Inside our in-house data agent** (2026). openai.com/index/inside-our-in-house-data-agent

VentureBeat — **OpenAI's AI data agent built by two engineers now serves 4,000 employees** (2026). venturebeat.com/orchestration/openais-ai-data-agent-built-by-two-engineers

Our Community Tech — **Notes from MIT: What OpenAI's Head of ChatGPT Engineering Says Comes Next** (April 2026). ourcommunity.tech/insights/emtech-ai-2026-choudhry.html

OpenAI — **API data usage policies**. openai.com/policies/api-data-usage-policies

OpenAI — **Nonprofits program**. openai.com/nonprofits

Replit — replit.com

Kim Wright is a technologist and educator. She teaches at SMU and UTD, leads technology initiatives at Uplift Education, and founded Our Community Tech to help nonprofits build and deploy the tools they need.