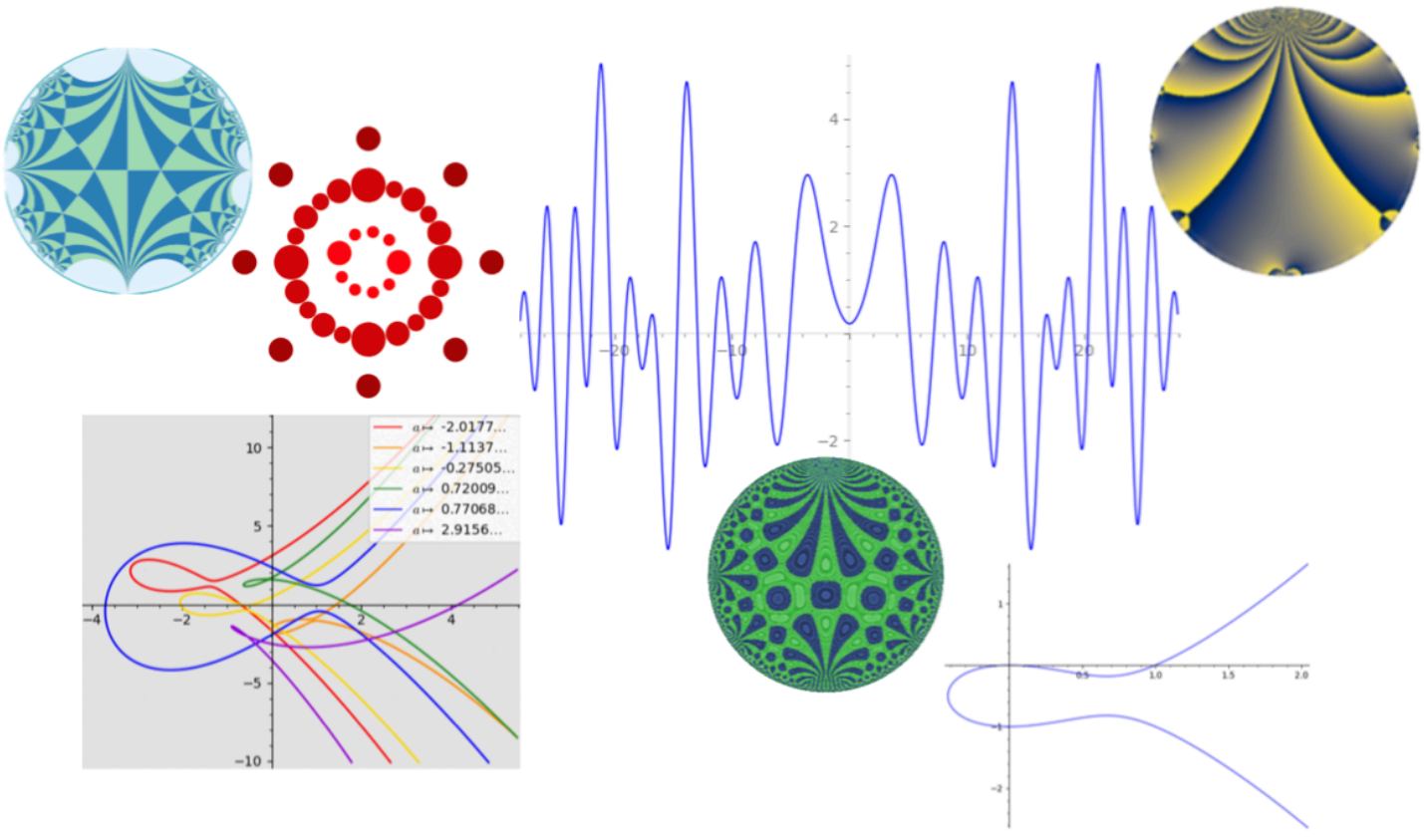
Building mathematical databases

David Roe

Department of Mathematics
MIT

June 24, 2026
Machine Computation in Homotopy Theory

The LMFDB is a collection of databases in number theory. While a few parts of the content may be useful to homotopy theorists, I hope the design choices can inspire your own efforts.



Sections of the LMFDB

- **L-functions** – rational and otherwise
- **Modular forms** – classical, Hilbert, Bianchi, Maass, Siegel
- **Varieties** – elliptic curves over $\mathbb{Q}$ and over number fields, genus 2 curves, modular curves, higher genus families, abelian varieties, Belyi maps
- **Fields** – number fields and p-adic fields
- **Groups** – abstract finite groups, permutation groups, Sato-Tate groups, and lattices
- **Representations** – Artin representations and Dirichlet characters
- **Motives** – hypergeometric

Aims to make deep mathematical connections concrete.

Demo

beta.lmfdb.org

Statistics

Size

About 3TB static (Riemann zeta zeros and class groups of imaginary quadratic fields) and 2TB searchable data, 450GB of search indexes (more details).

Number of objects

24M L-functions, 850k modular forms, 7.5M varieties (with 34M modular curves coming soon), 21M fields, 1M groups, 30M Dirichlet characters, 650k Artin representations, 60k families of hypergeometric motives.

Statistics for each section

Most sections of the LMFDB have a statistics page.

Web interface

- Three versions of the website:
  1. lmfdb.org – hosted on google cloud,
  2. beta – hosted at MIT, includes more preliminary sections like modular curves, hypergeometric motives and lattices.
  3. alpha – hosted at MIT, unfinished sections like Shimura curves and mod-ℓ Galois representations, and in-progress updates to existing sections.
- Runs an open source Flask frontend, allowing for templated webpages that are filled in from the database.
- The backend uses PostgreSQL, with data access also available through the API or Python.

Features

- Knowls – often used for defining terms, expand on the same page. Editable online.
- Aim at both experts and novices. Both search and browse modes, interesting examples.
- Completeness – describe regimes in which data is complete. Search results now show if complete.
- Source and reliability – describe how data was computed, any assumptions used and verifications made
- Results can be further refined, resorted, with displayed columns customized.
- Properties box (often with a picture) and related objects.
- Aim to have downloads in various formats.
- Each object has a mathematically meaningful label and homepage.

AI and mathematical data

- LLMs are getting good at many things: writing code, searching the literature, creating proofs
- Still bound by computational complexity, and people want results quickly. Value remains in **collecting** reusable computations that take a long time.
- Collecting computational results also allows machine-assisted **search** for interesting examples.
- At a technical level, MCP servers (Model Context Protocol) provide ways for LLMs to interact with computational resources.
- LLMs make formalization more accessible; I think there is a lot of room for improvement on reasons to trust data.

Database of databases

MathBases (mathbases.org)

- started at an AIM workshop in December 2023,
- grew out of Katja Berčič's MathDB (mathdb.mathhub.info),
- has an editorial board: Katja Berčič, Edgar Costa, Benjamin Hutz, David Lowry-Duda, David Roe, Ben Spitz, Adam Towsley,
- is a database that does not contain itself.

Mission

1. provide a searchable index of existing mathematical databases,
2. support mathematicians interested in creating new mathematical databases and improving existing ones.

Scope

MathBases currently indexes 130 databases; there are about 25 more on the to-add list.

Criteria for inclusion of a dataset

1. publicly available,
2. mathematical and of interest to research mathematicians
3. consisting of examples of mathematical objects, rather than exposition or formal statements

Database topics

52. Combinatorics
26. Number theory
22. Group theory
19. Metric geometry
13. Algebraic geometry
6. Rings and algebras
5. Geometric topology
4. Functional analysis, linear algebra
3. Dynamical systems, mathematical physics, representation theory
2. Algebraic topology, category theory, classical analysis and ODEs, commutative algebra, differential geometry
1. Analysis of PDEs, complex variables, general mathematics, general topology, logic, numerical analysis, operator algebras
0. K-theory and homology, optimization and control, probability, quantum algebra, symplectic geometry, spectral theory, statistics

Expository articles

In addition to serving as an index, MathBases aims to collect articles aimed at mathematicians who are creating databases.

- The Create page outlines the process of creating a new database, with links to more details
- At the moment, all the articles are stubs, which we hope to expand.
- We have a process for submitting articles. This could be
  - fleshing out one of the stubs;
  - writing something new, like a handbook for running computations on google cloud or a guide on accepting contributions from anonymous users;
  - an improvement to an existing article.
- Please submit things so that I don't have to write it all!
- We have not yet wrestled with how this part of MathBases should change now that LLMs exist.

Starting a new database

In addition to expository articles, we offer several additional resources if you are interested in making a database:

- A Zulip server (part of `code4math`) where you talk to others about mathematical databases. Feel free to join, even if you don't want to make a new database!
- We plan to add templates based on existing databases (the LMFDB and MathBases to start), since modifying a template is easier than starting from scratch.
- We hope to create a community around mathematical databases; conferences like this one are part of that.

Contributing

We'd love your help! Here are some ways to contribute:

- Add an existing database to the index, especially in sparse areas,
- Tell your colleagues about the site,
- Contribute an article,
- Help improve the project infrastructure (improved search interface, testing, handling dead links, displaying when a database record was last updated...),
- Fill in more metadata for databases in the index,
- Report a problem,
- Help us make a logo,
- Tell us about a database-related conference or add an idea for a new database.