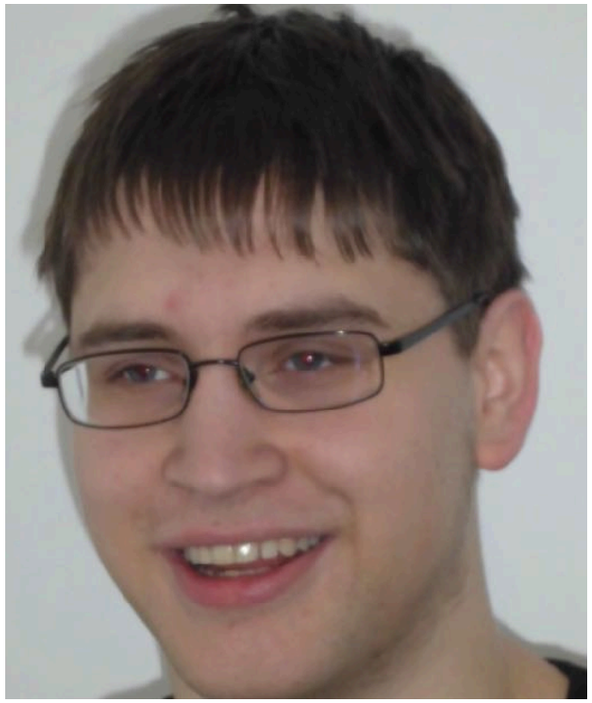
Research data in mathematics

Taking the high road

Katja Berčič
Michael Kohlhase
Florian Rabe
Tom Wiesing

Michael Kohlhase

Florian Rabe

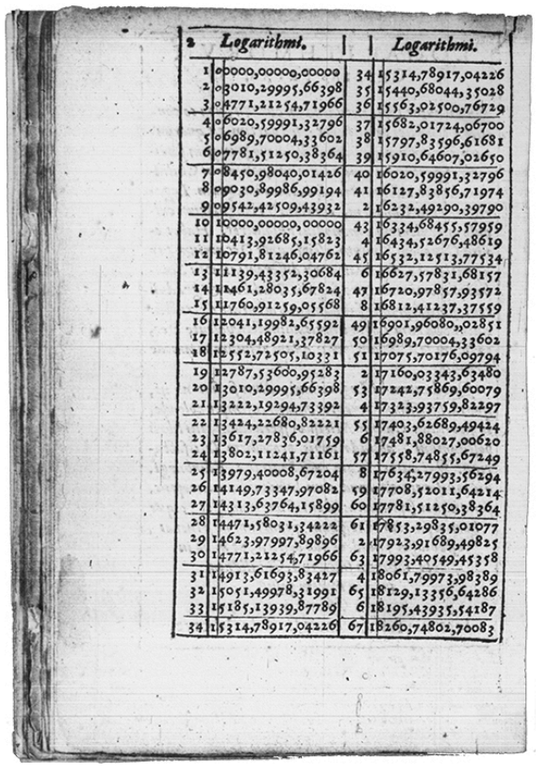


Tom Wiesing

Examples of historical datasets

Plimpton 322

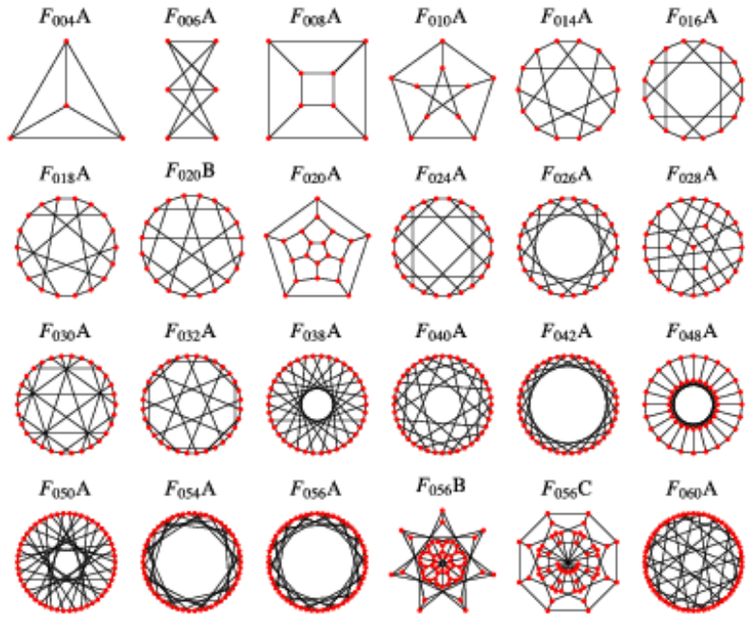
A cuneiform tablet from ca. 1800 BC. with a list of Pythagorean triples $a, b, c \in \mathbb{N}$, $a^2 + b^2 = c^2$.



	Logarithmi.		Logarithmi.
1	0,00000,00000,00000	34	1,53314,78917,04226
2	0,30103,29995,66398	35	1,54406,80443,35028
3	0,47712,12547,19662	36	1,55630,25076,76729
4	0,60205,99913,27324	37	1,56820,17247,06700
5	0,69897,00043,36029	38	1,57978,35961,61681
6	0,77815,12503,83648	39	1,59106,46073,01650
7	0,84509,80401,42619	40	1,60205,99913,27960
8	0,90308,99861,99194	41	1,61278,38561,71974
9	0,95424,25094,03932	42	1,62324,92903,39790
10	1,00000,00000,00000	43	1,63346,84557,57959
11	1,04139,26851,58230	44	1,64345,26761,48619
12	1,07918,12460,47626	45	1,65321,25127,37534
13	1,11394,33523,68489	46	1,66276,92863,68157
14	1,14612,80355,67824	47	1,67209,78571,93372
15	1,17609,12559,05568	48	1,68124,12378,37559
16	1,20411,99824,65592	49	1,69019,60800,02851
17	1,23044,89211,37827	50	1,69897,00043,36029
18	1,25527,25055,10331	51	1,70757,01767,09794
19	1,27875,36000,95283	52	1,71600,33436,63480
20	1,30103,29995,66398	53	1,72427,58693,60079
21	1,32221,92947,33392	54	1,73239,37589,82297
22	1,34242,26801,82221	55	1,74036,26803,49424
23	1,36172,78362,01759	56	1,74818,80274,90620
24	1,38021,12417,11161	57	1,75587,48563,67249
25	1,39794,00086,67204	58	1,76342,79933,56294
26	1,41497,33479,97082	59	1,77085,20116,64214
27	1,43136,37641,15899	60	1,77815,12503,83648
28	1,44715,80311,34222	61	1,78532,98357,01077
29	1,46239,79978,89896	62	1,79239,16890,49825
30	1,47712,12547,19662	63	1,79934,05549,45358
31	1,49136,16934,83427	64	1,80617,99773,98389
32	1,50514,99978,31991	65	1,81291,33356,64286
33	1,51851,93939,87789	66	1,81954,39935,54187
34	1,53147,89170,42260	67	1,82607,48020,70083

Logarithmorum Chilias Prima (1617)

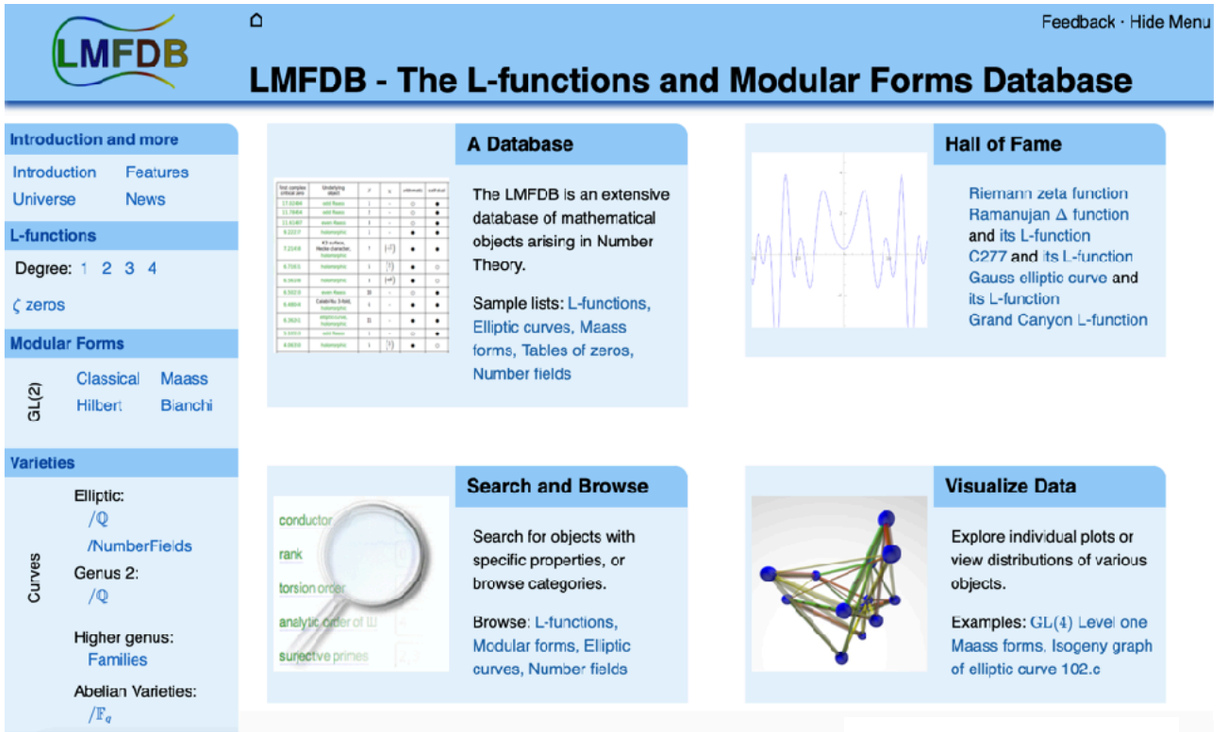
Base-10 logarithm of the integers 0 to 67 to fourteen decimal places.



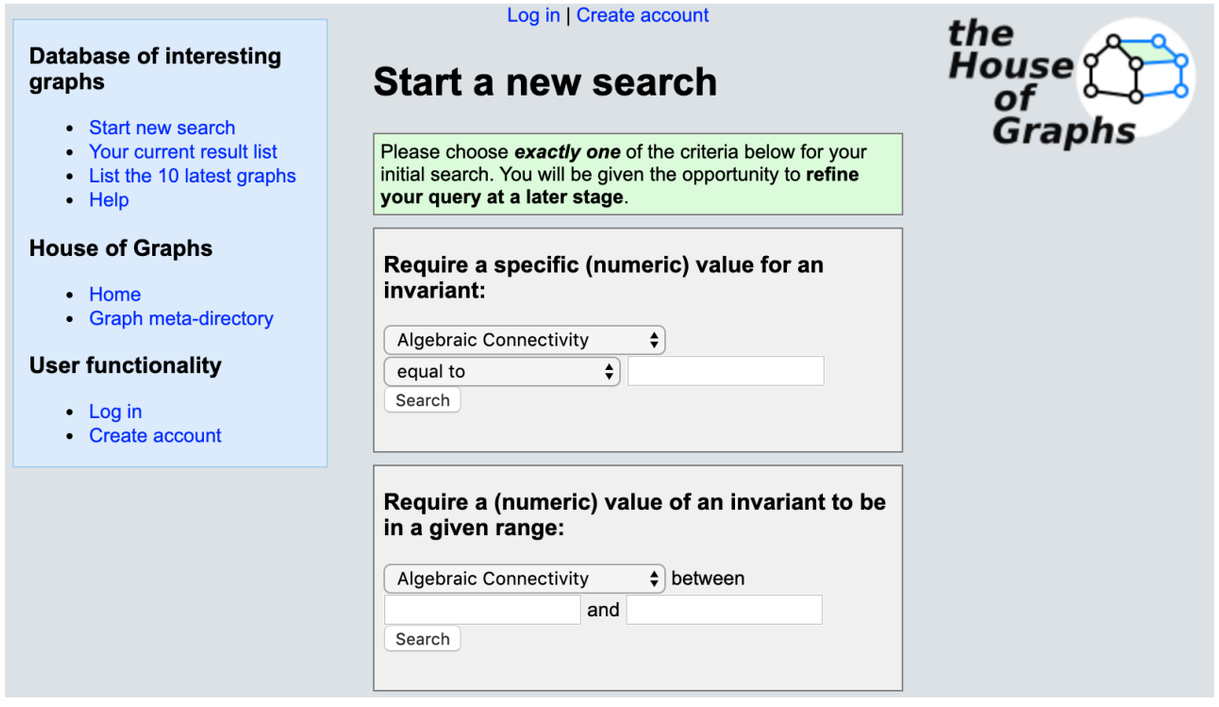
The Foster census (1930)

Cubic symmetric graphs with up to 512 vertices (207 such, Foster only missed a handful).

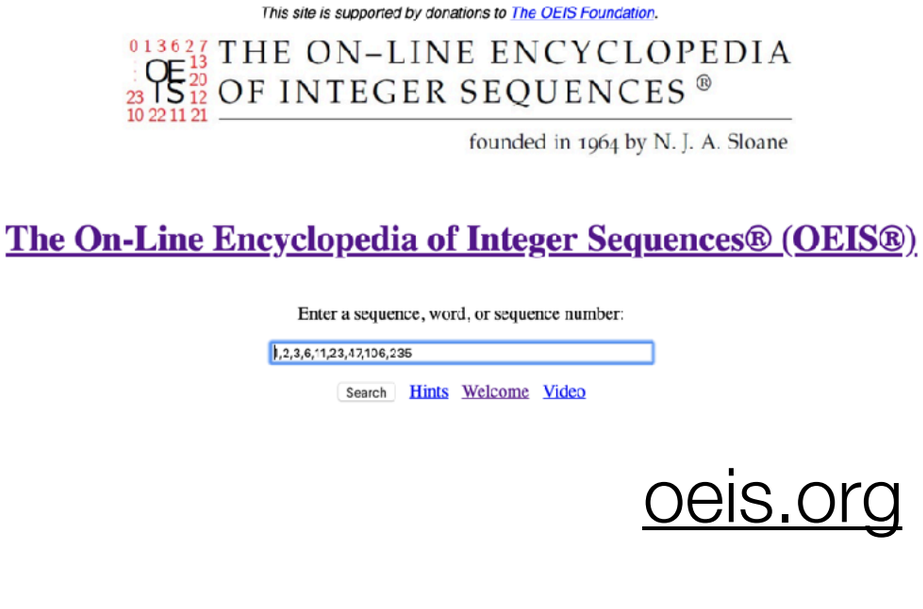
Examples of modern online mathematical databases



lmfdb.org



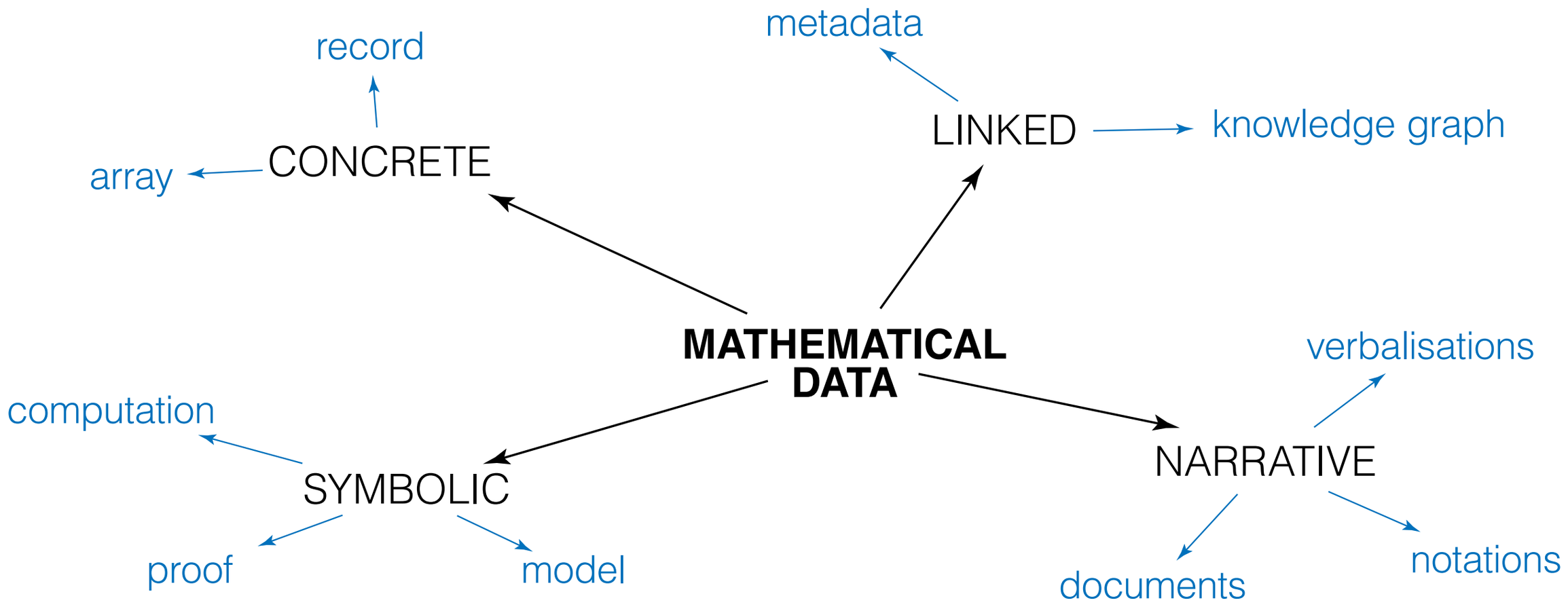
hog.grinvin.org



oeis.org

Examples:
OEIS, LMFDB, ...

Examples:
zbMATH, MathSciNet, MathHub



Examples:
GAP, SageMath,...
Mizar, Coq, PVS, HOL,...

Examples:
Wikipedia, Arxiv

K. Berčič, M. Kohlhase, and F. Rabe *(Deep) fair mathematics*. (Submitted)
K. Berčič *Towards a Census of Relational Data in Mathematics*. mathdb.mathhub.info

The Other 80% (or more)

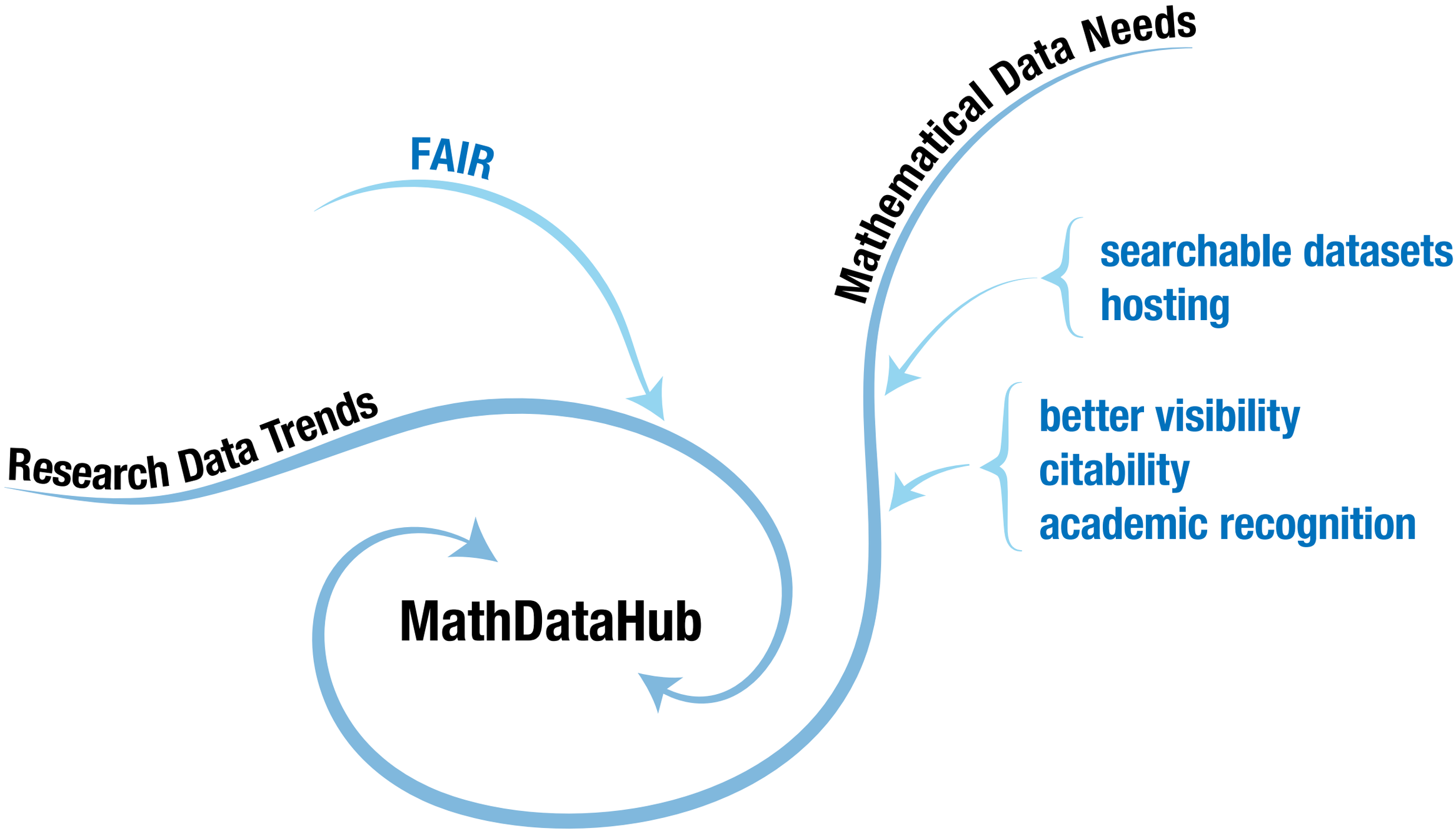
[N,i]	V	E	Tr	W?	B?	|AG|
C4[5, 1]	5	10	DT	W	NB	120
C4[6, 1]	6	12	DT	U	NB	48
C4[8, 1]	8	16	DT	U	Bip	(2^7)(3^2)
C4[9, 1]	9	18	DT	W	NB	72
C4[10, 1]	10	20	DT	U	NB	320
C4[10, 2]	10	20	DT	W	Bip	240
C4[12, 1]	12	24	DT	U	Bip	768
C4[12, 2]	12	24	DT	W	NB	48
C4[13, 1]	13	26	DT	W	NB	52
C4[14, 1]	14	28	DT	U	NB	(2^8)(7^1)
C4[14, 2]	14	28	DT	W	Bip	336
C4[15, 1]	15	30	DT	W	NB	60
C4[15, 2]	15	30	DT	W	NB	120
C4[16, 1]	16	32	DT	U	Bip	(2^12)
C4[16, 2]	16	32	DT	W	Bip	384
C4[17, 1]	17	34	DT	W	NB	68
C4[18, 1]	18	36	DT	U	NB	(2^10)(3^2)
C4[18, 2]	18	36	DT	W	Bip	144
C4[20, 1]	20	40	DT	U	Bip	(2^12)(5^1)
C4[20, 2]	20	40	DT	W	Bip	80
C4[20, 3]	20	40	DT	W	NB	320
C4[20, 4]	20	40	SS	U	Bip	(2^8)(3^1)(5^1)
C4[21, 1]	21	42	DT	W	NB	84
C4[21, 2]	21	42	DT	W	NB	336

Wilson, Potočnik; A Census of edge-transitive tetravalent graphs

- Graphs of order 4 to 300 (18 MB)
- Graphs of order 302 to 500 (66 MB)
- Graphs of order 502 to 600 (69 MB)
- Graphs of order 602 to 700 (84 MB)
- Graphs of order 702 to 800 (114 MB)
- Graphs of order 802 to 900 (147 MB)
- Graphs of order 902 to 1000 (183 MB)
- Graphs of order 1002 to 1050 (164 MB)
- Graphs of order 1052 to 1100 (113 MB)
- Graphs of order 1102 to 1150 (103 MB)
- Graphs of order 1152 to 1200 (234 MB)
- Graphs of order 1202 to 1250 (137 MB)
- Graphs of order 1252 to 1280 (131 MB)

Potočnik, Spiga, Verret; A census of small connected cubic vertex-transitive graphs

```
CubicVT:=[ [] : i in [1..1280]];

CubicVT[4,1] := Graph<4 | {{1,3}, {1,4}, {2,4}, {2,3}, {1,2}, {3,4}}>;

CubicVT[6,1] := Graph<6 | {{2,5}, {1,3}, {2,6}, {1,4}, {3,5}, {4,6}, {2,3},
{1,6}, {4,5}}>;

CubicVT[6,2] := Graph<6 | {{1,3}, {1,5}, {2,6}, {5,6}, {4,5}, {2,4}, {1,2},
{3,4}, {3,6}}>;

CubicVT[8,1] := Graph<8 | {{2,8}, {1,5}, {1,7}, {7,8}, {4,8}, {5,6}, {6,7},
{4,5}, {1,2}, {2,3}, {3,4}, {3,6}}>;

CubicVT[8,2] := Graph<8 | {{1,8}, {2,6}, {6,8}, {4,7}, {1,4}, {4,5}, {5,8},
{1,2}, {2,7}, {3,7}, {3,5}, {3,6}}>;

CubicVT[10,1] := Graph<10 | {{4,6}, {3,5}, {2,6}, {4,8}, {5,6}, {3,4}, {1,5},
{1,10}, {2,10}, {7,9}, {3,7}, {9,10}, {1,7}, {2,8}, {8,9}}>;

CubicVT[10,2] := Graph<10 | {{4,6}, {3,5}, {3,6}, {4,5}, {8,10}, {1,3}, {6,8},
{1,9}, {5,7}, {7,10}, {9,10}, {2,4}, {1,7}, {2,8}, {2,9}}>;

CubicVT[10,3] := Graph<10 | {{2,6}, {6,7}, {4,8}, {3,9}, {1,3}, {4,10}, {6,8},
{5,9}, {1,4}, {1,2}, {2,5}, {7,10}, {3,7}, {8,9}, {5,10}}>;

CubicVT[12,1] := Graph<12 | {{12,10}, {11,7}, {3,9}, {3,7}, {11,9}, {2,4},
{6,10}, {1,9}, {12,5}, {1,5}, {11,6}, {7,8}, {6,8}, {3,5}, {4,10}, {12,2},
{4,8}, {1,2}}>;
```


FAIR
Research Data Trends
MathDataHub
Mathematical Data Needs
searchable datasets
hosting
better visibility
citability
academic recognition

FAIR

- **Findable**: easy to identify and find for *both humans and computers*, e.g. with metadata that facilitate searching for specific datasets.

- **Accessible**: stored for long term so that they can easily be accessed and/or downloaded with well-defined access conditions, whether at the level of metadata, or at the level of the actual data.

- **Interoperable**: ready to be combined with other datasets by *humans or computers*, without ambiguities in the meanings of terms and values.

- **Reusable**: ready to be used for future research and to be further processed using computational methods.

MathHub Data - your dataset, but FAIR

About GitHub

A census of small connected cubic vertex-transitive graphs

All connected cubic vertex-transitive graphs of order at most **1280** .

This dataset has 111360 objects.

Matches found: *26*

More about this dataset

Display results

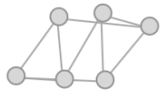
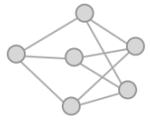
Choose columns

Available conditions

- Order
- CVT Index
- Graph
- Name
- Clique Number

Active conditions

- Is Partial Cube
- Order<100

	Order	Graph	Name	Is Arc-Transitive	Is Partial Cube	Is Prism
	4		Tetrahedron	true	false	false
	6		3-Prism	false	false	true
	6		Utility graph	true	false	false

Number of vertices in the graph.

MathData

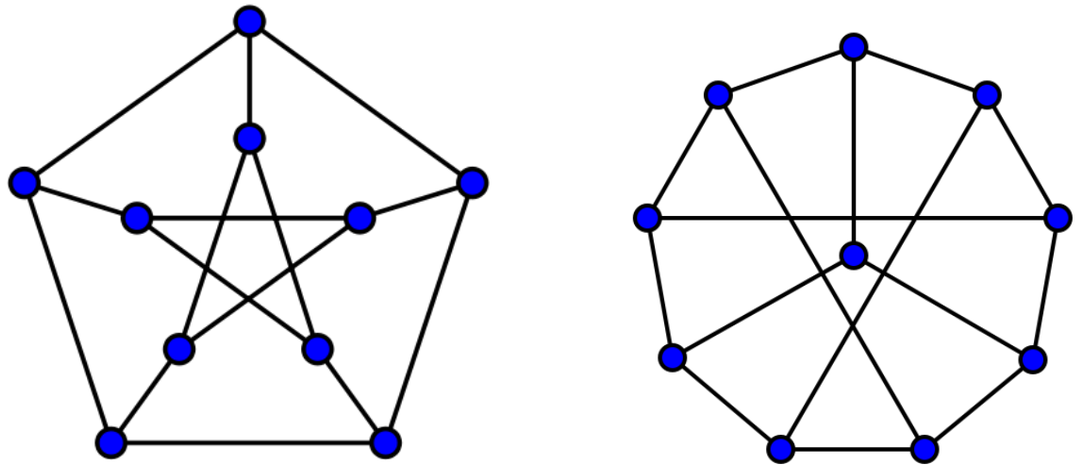
data.mathhub.info

Isomorphism and representation

Humans do everything **up to isomorphism**

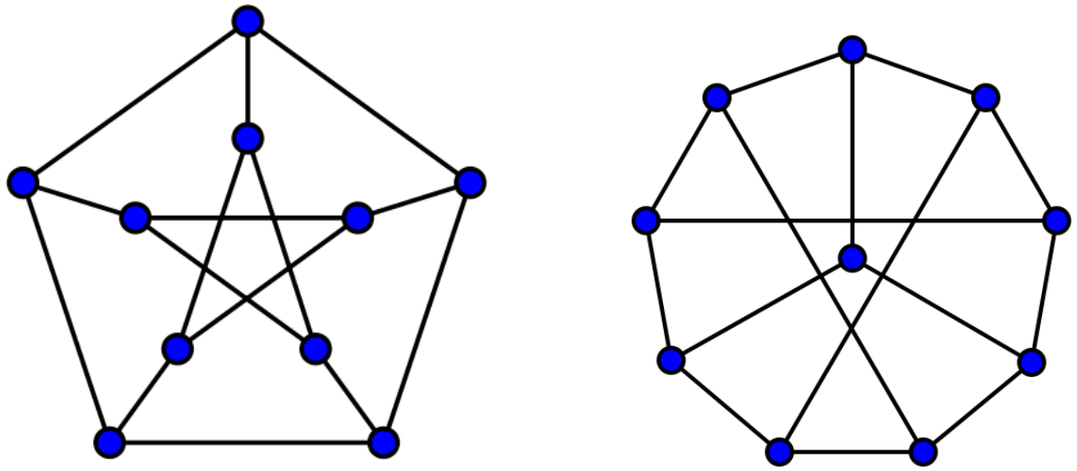
but

Computer systems are terrible at it



Mathematical objects usually don't come in one of the standard database types.

Storing complex objects



- Ideally: store canonically* chosen members of an isomorphism class

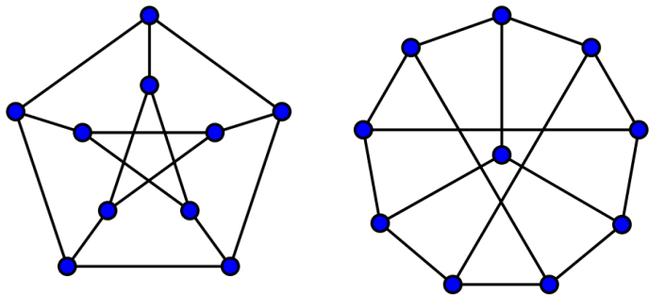
- Machine friendly format

Representation Theorem (1)

representation theorem

mathematical type
(graph)

database type
(string)



:I`ES@obGkqegW~

graph6

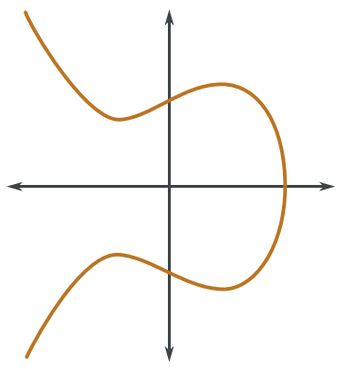
canonical labeling + edge list as string

Representation Theorem (2)

representation theorem

mathematical type
(elliptic curve)

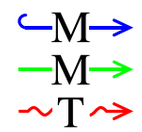
database type
(integer 5-tuple)



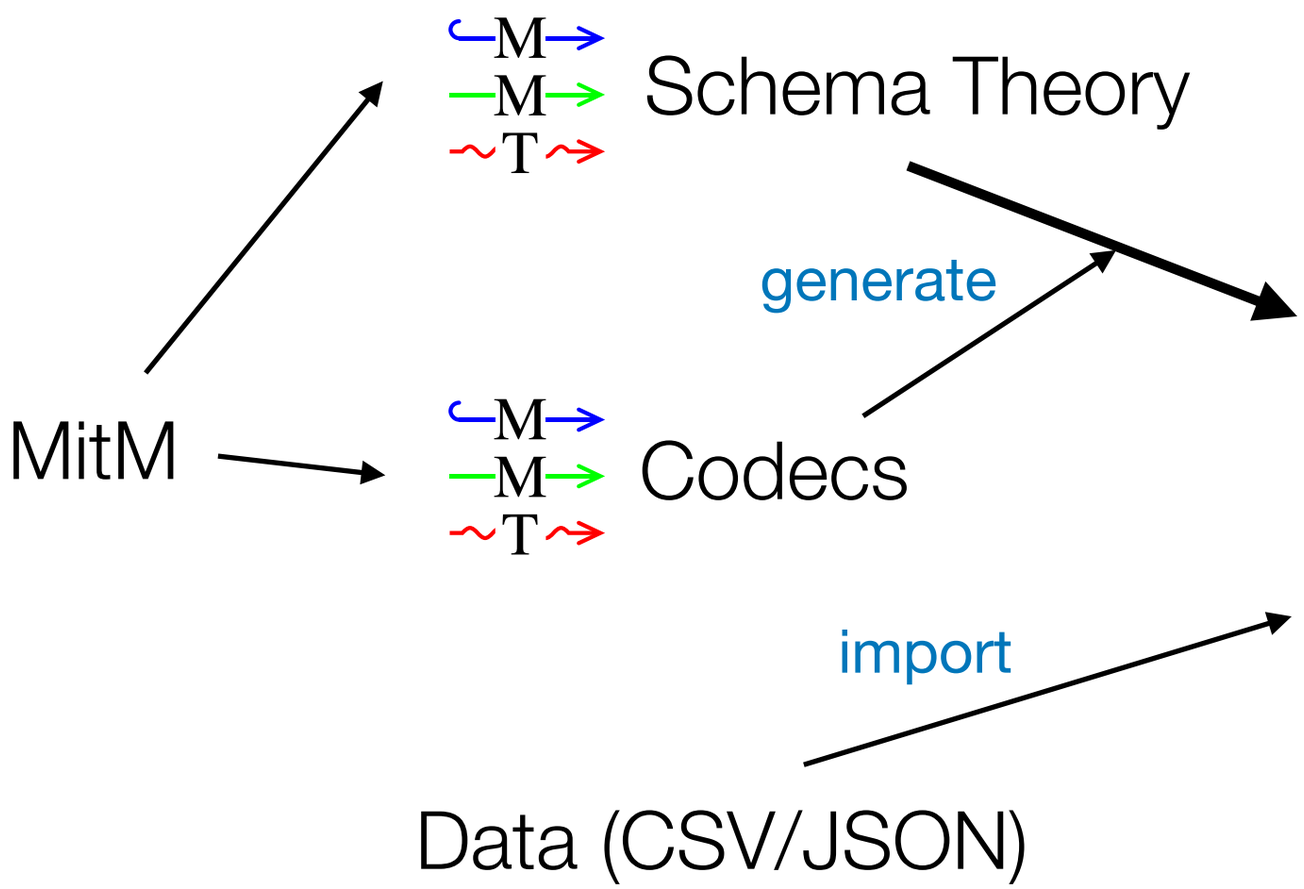
```
(0, 0, 0, -1, 1)
```

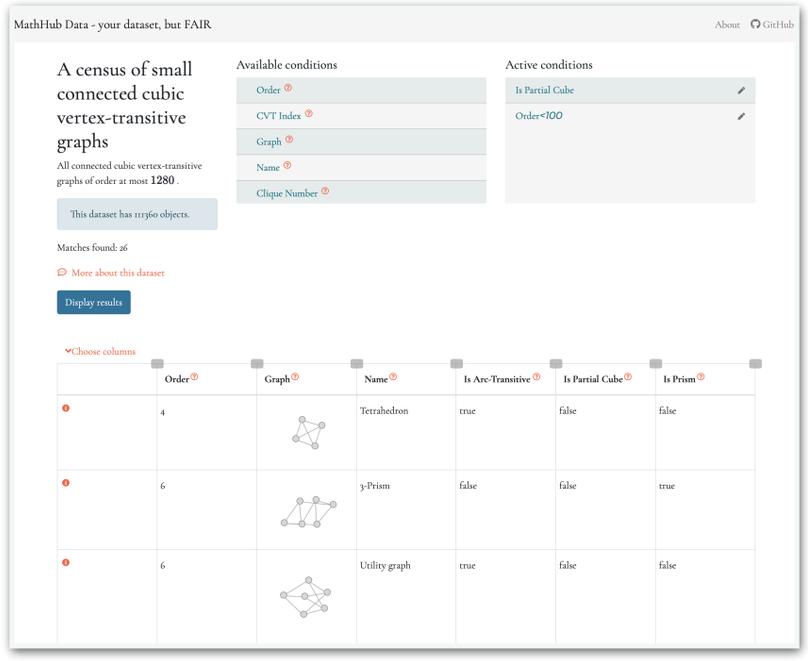
Weierstrass coefficients

Machine actionable specification

- Describe/specify data in a formal language ()
  - **Codecs** connect the mathematical specification and the database specification
  - A **schema theory** describes the dataset structure
- Formal specifications are expensive: collect and share
- Math in the Middle (MitM) first used to connect CAS

MathDataHub





MitM, schema theory

```
theory Graphs : base:?Logic =
  include ☞ts:?AllSets |
  include ☞ts:?Relations |
  include ☞ts:?Union |
  include ☞num:?NaturalArithmetics |
  include ☞ts:?SetQuantifiers |

  nodetype : type |

theory dgraph_theory : base:?Logic =
    nodes : set nodetype |
    edges : relation nodetype nodetype |

    successors : nodetype ⟶ set nodetype |
      = [a] ⦇ nodes | ([x] edges a x) ⦈ |

    predecessors : nodetype ⟶ set nodetype |
      = [a] ⦇ nodes | ([x] edges x a) ⦈ |

    neighbours : nodetype ⟶ set nodetype |
      = [a] (successors a) ∪ (predecessors a)  |
```

```
theory CubicVertexTransitiveGraphsS : ?MDDL =
  link /?Metadata?implements MitM:?Graphs?finite_ugraph |
  meta ?MDDL?schemaGroup "CVT" |

  order: int |
    meta ?Codecs?codec StandardInt |
    link /?Metadata?implements MitM:?Graphs?order |
    @_description "Number of vertices in the graph." |
  |

  cvt_index: int |
    meta ?Codecs?codec StandardInt |
    @_description "Id of the graph in the dataset. A graph
      is uniquely identified by the id and the order." |
  |

  graph: MitM:?Graphs?finite_ugraph |
    meta ?Codecs?codec GraphAsSparse6 |
    tag ?MDDL?opaque |
    @_description "The graph encoded in the sparse6 format." |
  |
```

Math Data Matters

- Exist, are useful* and fun

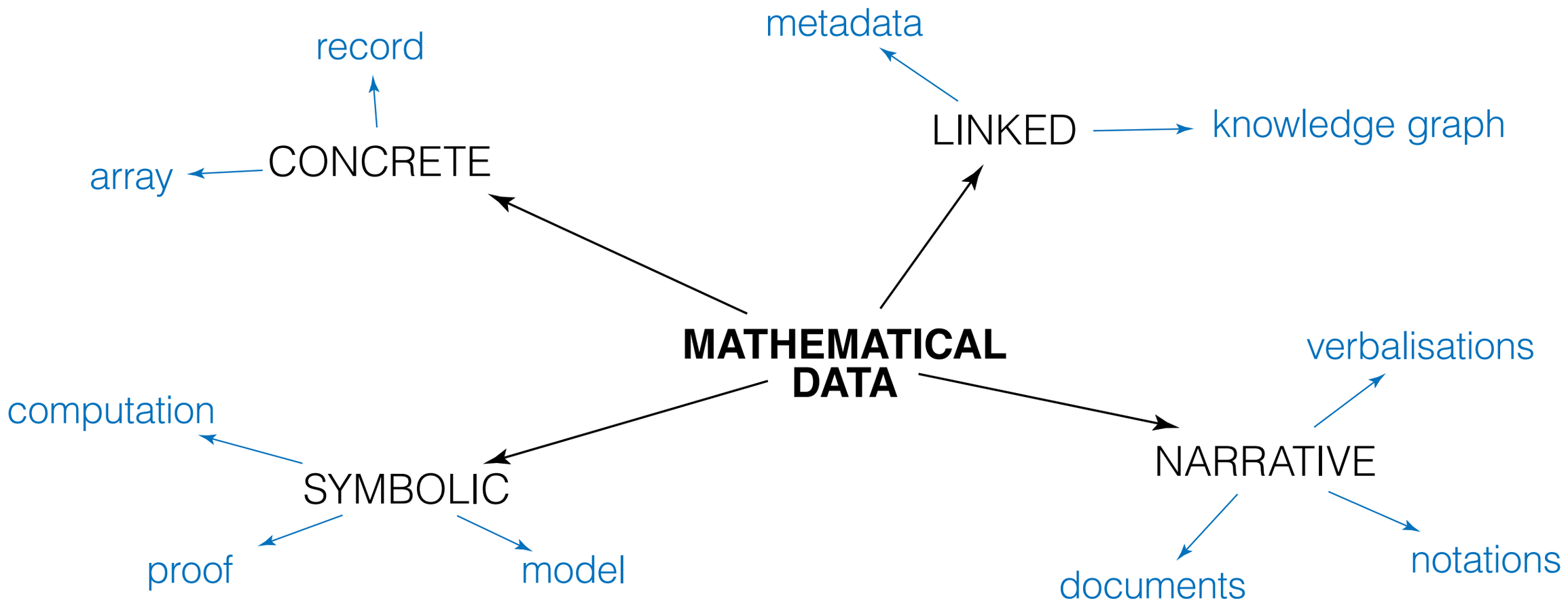
  * but could be more useful

- We want your datasets for MathDataHub
- MaRDI (Germany, 30 million for math research data)
- Demo?

Thank you!

Examples:
OEIS, LMFDB, ...

Examples:
zbMATH, MathSciNet, MathHub



Examples:
GAP, SageMath,...
Mizar, Coq, PVS, HOL,...

Examples:
Wikipedia, Arxiv

K. Berčič, M. Kohlhase, and F. Rabe *(Deep) fair mathematics*. (Submitted)
K. Berčič *Towards a Census of Relational Data in Mathematics*.

(Deep) FAIR mathematics

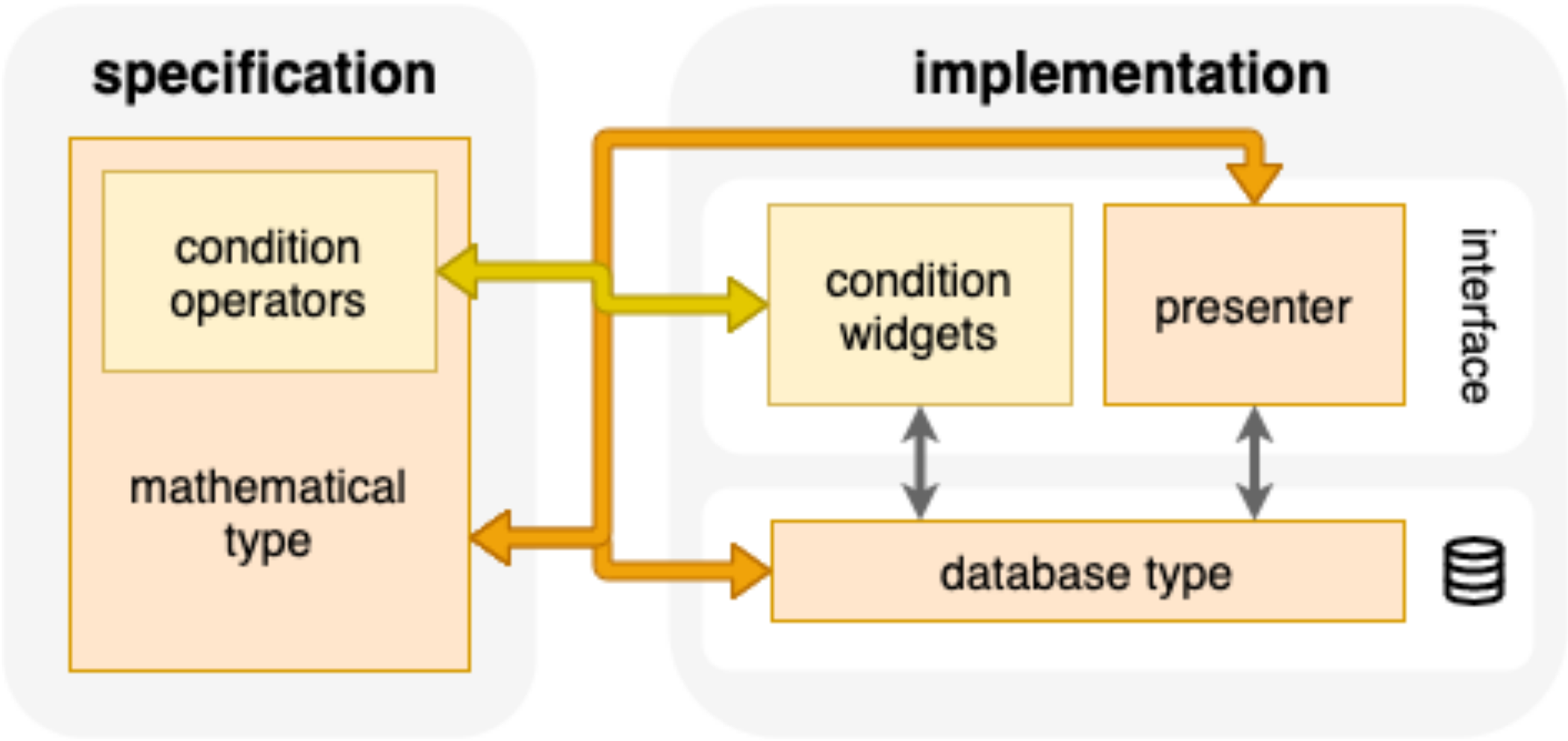
- **Findable.** The internal structure of the dataset is represented and indexed in a way that allows searching for individual entries by *humans or computers*.

- **Accessible.** The long term storage supports access to individual dataset fragments, their metadata and semantics.

- **Interoperable.** Each fragment is ready to be combined with other datasets by *humans or computers* and -- where applicable -- includes qualified references to other fragments.

- **Reusable.** The individual dataset fragments used for future research and to be further processed using computational methods.

Codecs


specification
implementation
condition operators
mathematical type
condition widgets
presenter
interface
database type