Some Experiments with Machine Learning, Reasoning and Discovery

Josef Urban

Czech Technical University in Prague

Deep-Learning Models for Mathematics and Type Theory
April 23, 2025, Goteborg

European Research Council
Established by the European Commission

How Do We Automate Math and Science?

- What is mathematical and scientific thinking/intelligence?
- Pattern-matching, analogy, induction/learning from examples
- Learning both fuzzy hunches and crisp algos/heuristics/procedures
- Deductive reasoning and proving (*btw, computation is reasoning*)
- intelligently guided search, exploration and guessing/conjecturing
- Complicated feedback loops and interplays between all these
- Using a lot of previous knowledge - both for induction and deduction
- Examples from physics (Popper?): deduction pushing us to Relativity, QM

- We need to develop such methods on computers
- Are there any large corpora suitable for nontrivial deduction?
- Yes! Large libraries of formal proofs and theories
- So let's try to develop strong AI on them!

- In my case done for 25-30 years. Recent overviews:
- *Learning Guided Automated Reasoning: A Brief Survey. 2024*
- Zar Goertzel's Phd thesis: *Learning Inference Guidance in ATP. 2023*
- *AI4REASON*: `http://ai4reason.org/PR_CORE_SCIENTIFIC_4.pdf`

Quotes: Learning vs. Reasoning vs. Guessing

“C’est par la logique qu’on démontre, c’est par l’intuition qu’on invente.”
(It is by logic that we prove, but by intuition that we discover.)

– Henri Poincaré, Mathematical Definitions and Education.

“Hypothesen sind Netze; nur der fängt, wer auswirft.”
(Hypotheses are nets: only he who casts will catch.)

– Novalis, quoted by Popper – The Logic of Scientific Discovery

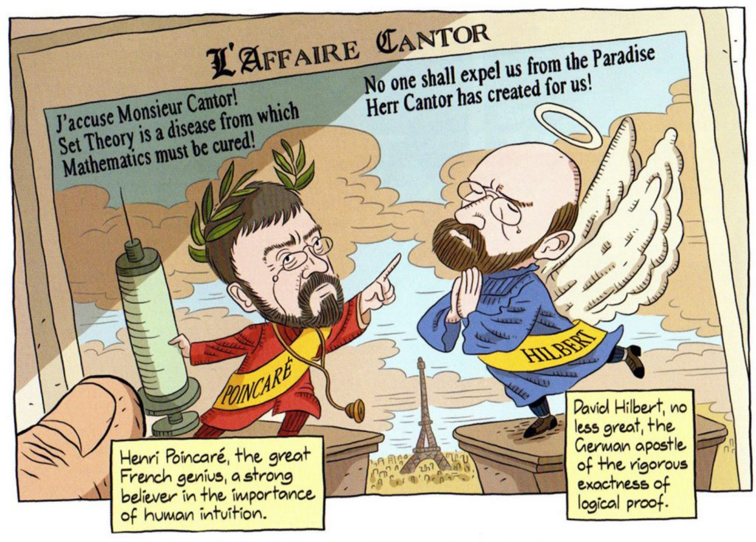
Certainly, let us learn proving, but also let us learn guessing.

– G. Polya - Mathematics and Plausible Reasoning

Galileo once said, "Mathematics is the language of Science." Hence, facing the same laws of the physical world, alien mathematics must have a good deal of similarity to ours.

– R. Hamming - Mathematics on a Distant Planet

Intuition vs Formal Reasoning – Poincaré vs Hilbert



[Adapted from: *Logicomix: An Epic Search for Truth* by A. Doxiadis]

Leibniz's/Hilbert's/Russell's Dream: Let Us Calculate!

Solve all (math, physics, law, economics, society, ...) problems by
reduction to logic/computation

[Adapted from: *Logicomix: An Epic Search for Truth* by A. Doxiadis]

Quick intro: *Prove/Learn feedback loop* on formal math

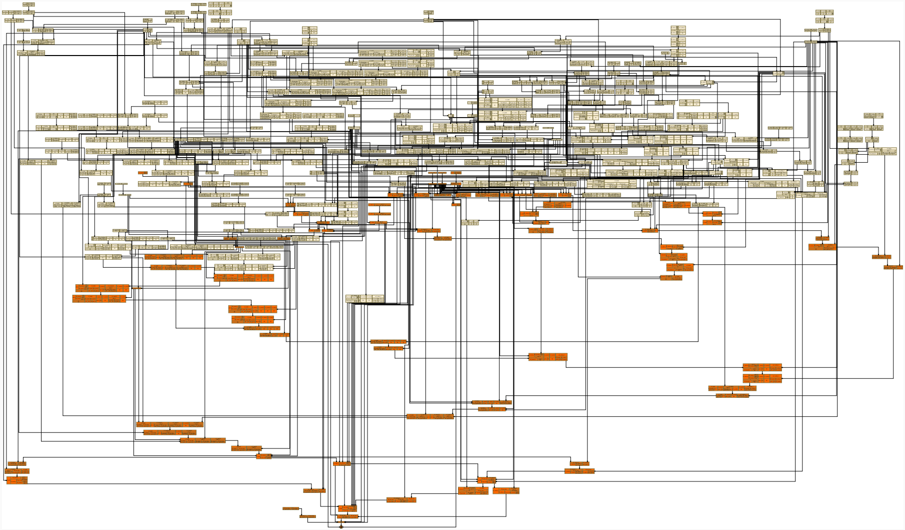
- Done on 57880 Mizar Mathematical Library formal math problems in 2019
- Efficient ML-guidance inside the best ATPs like E prover (ENIGMA)
- *Training* of the ML-guidance is *interleaved* with *proving* harder problems
- Ultimately a 70% improvement over the original E strategy:
- ... from 14933 proofs to 25397 proofs (all in 10s CPU - no cheating)

	$\mathcal{S}$	$\mathcal{S}\odot\mathcal{M}_9^0$	$\mathcal{S}\oplus\mathcal{M}_9^0$	$\mathcal{S}\odot\mathcal{M}_9^1$	$\mathcal{S}\oplus\mathcal{M}_9^1$	$\mathcal{S}\odot\mathcal{M}_9^2$	$\mathcal{S}\oplus\mathcal{M}_9^2$	$\mathcal{S}\odot\mathcal{M}_9^3$	$\mathcal{S}\oplus\mathcal{M}_9^3$
solved	**14933**	16574	20366	21564	22839	22413	23467	22910	23753
$\mathcal{S}\%$	+0%	+10.5%	+35.8%	+43.8%	+52.3%	+49.4%	+56.5%	+52.8%	+58.4
$\mathcal{S}+$	+0	+4364	+6215	+7774	+8414	+8407	+8964	+8822	+9274
$\mathcal{S}-$	-0	-2723	-782	-1143	-508	-927	-430	-845	-454

	$\mathcal{S}\odot\mathcal{M}_{12}^3$	$\mathcal{S}\oplus\mathcal{M}_{12}^3$	$\mathcal{S}\odot\mathcal{M}_{16}^3$	$\mathcal{S}\oplus\mathcal{M}_{16}^3$
solved	24159	24701	25100	**25397**
$\mathcal{S}\%$	+61.1%	+64.8%	+68.0%	**+70.0%**
$\mathcal{S}+$	+9761	+10063	+10476	+10647
$\mathcal{S}-$	-535	-295	-309	-183

- 75% of the Mizar corpus (43414) reached in July 2021 - higher times and many prove/learn cycles: `https://github.com/ai4reason/ATP_Proofs`
- Details in our Mizar60 paper: `https://arxiv.org/abs/2303.06686`

Can you do this in 4 minutes? (359-step ATP proof)



Can you do this in 4 minutes? (human-written code)

```
theorem Th31: :: BORSUK_5:31
 for A being Subset of R^1
 for a, b being real number st a < b & A = RAT (a,b) holds
 Cl A = [.a,b.]
proof
 let A be Subset of R^1; :: thesis:
 let a, b be real number ; :: thesis:
 assume that
 A1: a < b and
 A2: A = RAT (a,b) ; :: thesis:
 reconsider ab = ].a,b.[, RT = RAT as Subset of R^1 by NUMBERS:12, TOPMETR:17;
 reconsider RR = RAT /\ ].a,b.[ as Subset of R^1 by TOPMETR:17;
 A3: the carrier of R^1 /\ (Cl ab) = Cl ab by XBOOLE_1:28;
 A4: Cl RR c= (Cl RT) /\ (Cl ab) by PRE_TOPC:21;
 thus Cl A c= [.a,b.] :: according to XBOOLE_0:def 10 :: thesis:
 proof
  let x be set ; :: according to TARSKI:def 3 :: thesis:
  assume x in Cl A ; :: thesis:
  then x in (Cl RT) /\ (Cl ab) by A2, A4;
  then x in the carrier of R^1 /\ (Cl ab) by Th15;
  hence x in [.a,b.] by A1, A3, Th16; :: thesis:
 end;
 thus [.a,b.] c= Cl A :: thesis:
 proof
  let x be set ; :: according to TARSKI:def 3 :: thesis:
  assume A5: x in [.a,b.] ; :: thesis:
  then reconsider p = x as Element of RealSpace by METRIC_1:def 13;
  A6: a <= p by A5, XXREAL_1:1;
  A7: p <= b by A5, XXREAL_1:1;
  per cases by A7, XXREAL_0:1;
   suppose A8: p < b ; :: thesis:
    now :: thesis:
     let r be real number ; :: thesis:
     reconsider pp = p + r as Element of RealSpace by METRIC_1:def 13, XREAL_0:def 1;
     set pr = min (pp,((p + b) / 2));
     A9: min (pp,((p + b) / 2)) <= (p + b) / 2 by XXREAL_0:17;
     assume A10: r > 0 ; :: thesis:
     p < min (pp,((p + b) / 2))
     proof
      per cases by XXREAL_0:15;
       suppose min (pp,((p + b) / 2)) = pp ; :: thesis:
        hence p < min (pp,((p + b) / 2)) by A10, XREAL_1:29; :: thesis:
       end;
       suppose min (pp,((p + b) / 2)) = (p + b) / 2 ; :: thesis:
        hence p < min (pp,((p + b) / 2)) by A8, XREAL_1:226; :: thesis:
       end;
      end;
     end;
     then consider Q being rational number such that
     A11: p < Q and
     A12: Q < min (pp,((p + b) / 2)) by RAT_1:7;
     (p + b) / 2 < b by A8, XREAL_1:226;
     then min (pp,((p + b) / 2)) < b by A9, XXREAL_0:2;
     then A13: Q < b by A12, XXREAL_0:2;
     min (pp,((p + b) / 2)) <= pp by XXREAL_0:17;
     then A14: (min (pp,((p + b) / 2))) - p <= pp - p by XREAL_1:9;
     reconsider P = Q as Element of RealSpace by METRIC_1:def 13, XREAL_0:def 1;
     P - p < (min (pp,((p + b) / 2))) - p by A12, XREAL_1:9;
     then P - p < r by A14, XXREAL_0:2;
     then dist (p,P) < r by A11, Th14;
     then A15: P in Ball (p,r) by METRIC_1:11;
     a < Q by A6, A11, XXREAL_0:2;
     then A16: Q in ].a,b.[ by A13, XXREAL_1:4;
     Q in RAT by RAT_1:def 2;
     then Q in A by A2, A16, XBOOLE_0:def 4;
     hence Ball (p,r) meets A by A15, XBOOLE_0:3; :: thesis:
    end;
    hence x in Cl A by GOBOARD6:92, TOPMETR:def 6; :: thesis:
```

ENIGMA - The Rise of Computronium (2021)

From: Josef Urban <josef.urban@gmail.com>, Date: Jul 26, 2021 at 9:47
Subject: ENIGMA - The Rise of Computronium

I am happy to announce that the ENIGMA system of the E lineage, helped by its Deepire Vampiric cousin, has reached today (July 26th, 2021) the landmark of 75% automatically proved Mizar top-level problems.
[..]
Many of the proofs show that ENIGMA has autonomously (without any human-programmed decision procedures, tactics, and/or dataset preparation/tuning) learned how to routinely perform common mathematical algorithmic tasks such as numeric calculation, matrix manipulation, boolean algebra, integration and differentiation, sequences of standard rewriting and normalization operations in various algebraic theories, etc., and combine them with other reasoning tasks needed for completing the fully formal proofs.
[..]
This suggests that learned guidance combined with efficient search may in near future lead to a new fully declarative problem-solving computing/reasoning architectures applicable to arbitrary computing/reasoning problems without any human engineering.
[..]
Combined with autoformalization this could lead us to large deployment of reasoning machines in science, following Leibniz’s Calculemus dream.

Intro2: *Search/Check/Learn feedback loop* on OEIS

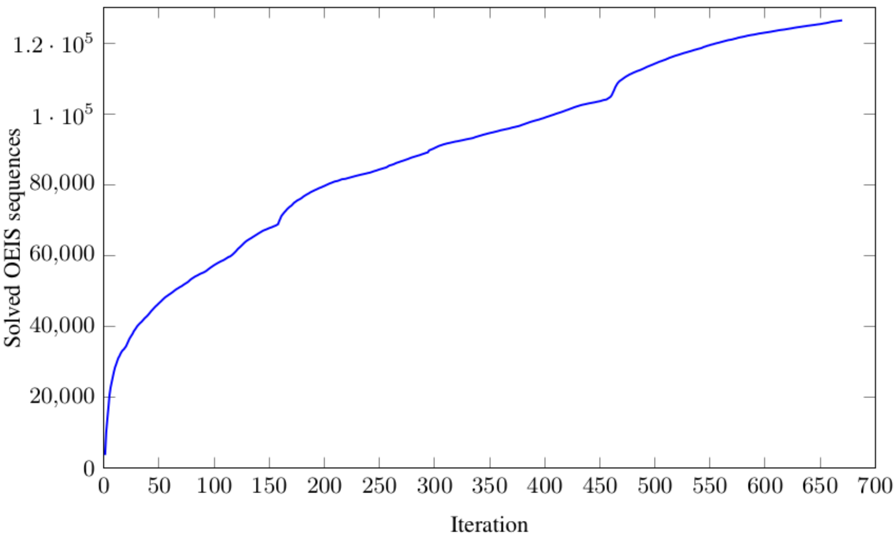
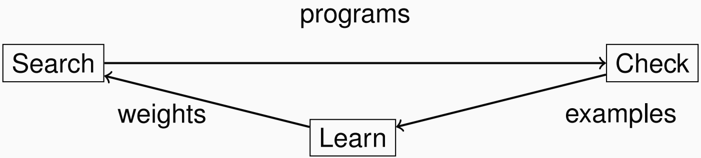


Figure 12: Number y of solved OEIS sequences after x iterations

Search-Verify-Train Positive Feedback Loop (OEIS)

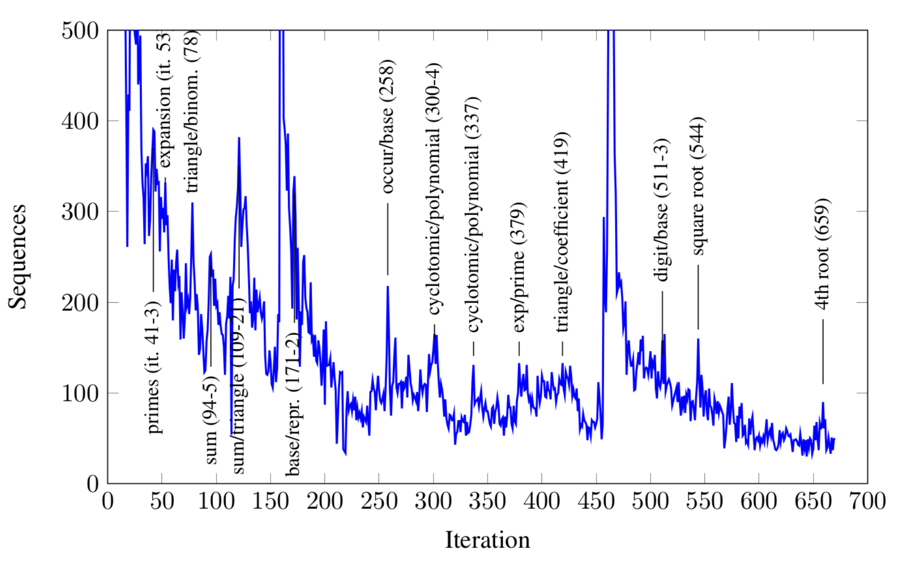


- Small Turing-complete DSL for our programs, e.g.:
  $2^{\mathbf{x}} = \prod_{y=1}^{x} 2 = loop(2 \times x, \mathbf{x}, 1)$
  $\mathbf{x}! = \prod_{y=1}^{x} y = loop(y \times x, \mathbf{x}, 1)$
- Analogous to our Prove/Learn feedback loops in learning-guided proving (since 2006 – Machine Learner for Automated Reasoning – MaLARea))
- However, OEIS allows much faster feedback on *symbolic conjecturing*
- 670 iterations and still refuses to plateau - counters RL wisdom?
- Since it interleaves symbolic breakthroughs and statistical learning?
- Cheap: The electricity bill is only \$1k-\$3k, you can do this at home
- ~4.5M explanations invented: 50+ different characterizations of primes

Some Invented Explanations for OEIS ("Alien Coder")

- `https://oeis.org/A4578`: Expansion of sqrt(8) in base 3:
  loop2(((y * y) div (x + y)) + y, y, x + x, 2, loop((1 + 2) * x, x, 2)) mod (1 + 2)
- `https://oeis.org/A4001`: Hofstadter-Conway $10k seq: a(n) = a(a(n-1)) + a(n-a(n-1)) with a(1) = a(2) = 1:
  loop(push(loop(pop(x), y-x,pop(x)),x) + loop(pop(x), x-1, x), x - 1, 1)
- `https://oeis.org/A40`: prime numbers:
  2 + compr((loop(x * y, x, 2) + x) mod (2 + x), x)
- `https://oeis.org/A30184`: Expand $\eta(q) * \eta(q^3) * \eta(q^5) * \eta(q^{15})$ in powers of q (elliptic curves):
  loop(push(loop((pop(x) * loop(if (pop(x) mod y) <= 0 then (x - loop(if (x mod (1 + (y + y))) <= 0 then (x + x) else x, 2, y)) else x, y, push(0, y))) + x, y, push(0, x)), x) div y, x, 1)
- `https://oeis.org/A51023`: Wolfram's $30k Rule 30 automaton:
  loop2(y, y div 2, x, 1, loop2(loop2((((y div (0 - (2 + 2))) mod 2) + x) + x, y div 2, y, 1, loop2(((y mod 2) + x) + x, y div 2, y, 1, x)), 2 + y, x, 0, 1)) mod 2
- `https://oeis.org/A2580`: $\sqrt[3]{2}$ Hales's blog: `https://t.ly/tHs1d`

Some Automatic Technology Jumps



Serious Math Conjecturing – Elliptic Curves

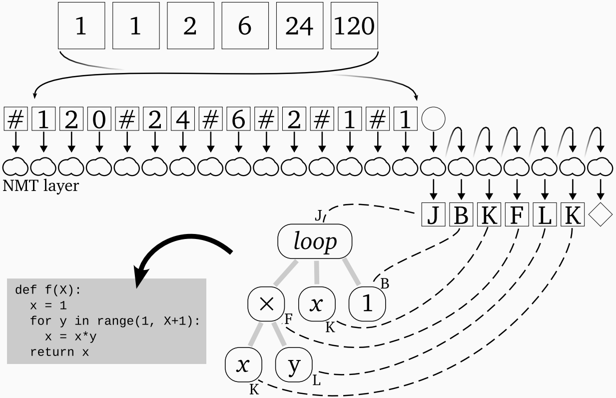
- Sander Dahmen: *Here are some OEIS labels related to elliptic curves (and hence modular forms), ordered by difficulty. It would be interesting to know if some of these appear in your results.*
- A006571 A030187 A030184 A128263 A187096 A251913

- JU: *We have the first three:*
- A6571 : loop((push(loop((pop(x) * loop(if (pop(x) mod y) <= 0 then ((if (y mod loop(1 + (x + x), 2, 2)) <= 0 then (x - y) else x) - y) else x, y, push(0, y))) + x, y, push(0, x)), x) * 2) div y, x, 1)
- A30187 : loop(push(loop((pop(x) * loop(if (pop(x) mod y) <= 0 then (x - loop(if (x mod (((2 + y) * y) - 1)) <= 0 then (x + x) else x, 2, y)) else x, y, push(0, y))) + x, y, push(0, x)), x) div y, x, 1)
- A30184 : loop(push(loop((pop(x) * loop(if (pop(x) mod y) <= 0 then (x - loop(if (x mod (1 + (y + y))) <= 0 then (x + x) else x, 2, y)) else x, y, push(0, y))) + x, y, push(0, x)), x) div y, x, 1)

A6571: Expansion of $q * Product_{k>=1}(1-q^k)^2 * (1-q^{11*k})^2$
A30187: Expansion of $\eta(q) * \eta(q^2) * \eta(q^7) * \eta(q^{14})$ in powers of q.
A30184: Expansion of $\eta(q) * \eta(q^3) * \eta(q^5) * \eta(q^{15})$ in powers of q.

Encoding OEIS for Language Models

- Input sequence is a series of digits
- Separated by an additional token # at the integer boundaries
- Output program is a sequence of tokens in Polish notation
- Parsed by us to a syntax tree and translatable to Python
- Example: $a(n) = n!$




```
def f(X):
  x = 1
  for y in range(1, X+1):
    x = x*y
  return x
```

Search-Verify-Train Feedback Loop for OEIS

- **search phase**: LM synthesizes many programs for input sequences
- typically 240 candidate programs for each input using beam search
- 84M programs for OEIS in several hours on the GPU (depends on model)
- **checking phase**: the millions of programs efficiently evaluated
- resource limits used, fast indexing structures for OEIS sequences
- check if the program generates *any* OEIS sequence (hindsight replay)
- we keep the shortest (Occams's razor) and fastest program (efficiency)
- from iter. 501, we also keep the program with the best speed/length ratio
- **learning phase**: LM trains to translate the "solved" OEIS sequences into the best program(s) generating them
- from iter. 336: train LMs to transform (generalization, optimization)
- our learned version of human-coded methods like ILP and compilation

Search-Verify-Train Feedback Loop

- The weights of the LM either trained from scratch or continuously updated
- This yields *new minds* vs *seasoned experts* (who have seen it all)
- We also train experts on varied selections of data, in varied ways
- Orthogonality: common in theorem proving - different experts help
- Each iteration of the self-learning loop discovers more solutions
- ... also improves/optimizes existing solutions
- The alien mathematician thus self-evolves
- Occam's razor and efficiency are used for its weak supervision
- Quite different from today's LLM approaches:
- LLMs do one-time training on everything human-invented
- Our alien instead starts from zero knowledge
- Evolves increasingly nontrivial skills, may diverge from humans
- Turing complete (unlike Go/Chess) – arbitrary complex algorithms

History, Motivation, AI/TP/ML

- Intuition vs Formal Reasoning – Poincaré vs Hilbert, Science & Method
- Turing's 1950 paper: Learning Machines, learn Chess?, undecidability??
- 50s-60s: Beginnings of ATP and ITP – Davis, Simon, Robinson, de Bruijn
- Lenat, Langley: AM, manually-written heuristics, learn Kepler laws,...
- Denzinger, Schulz, Goller, Fuchs – late 90's, ATP-focused: *Learning from Previous Proof Experience* (Tree NNs for ATP, E prover, ...)
- My MSc (1998): Try ILP to learn rules and heuristics from IMPS/Mizar
- Since: Use large formal math corpora: Mizar, Isabelle, HOL, Coq, Lean
  ... to combine/develop symbolic/statistical deductive/inductive ML/TP/AI
  ... hammer-style methods, internal guidance, feedback loops, ...
- Buzzword bingo timeline: AI vs ML vs NNs vs DL vs LLMs vs AGI vs ...?
  See Ben Goertzel's 2018 Prague talk: `https://youtu.be/Zt2HSTuGBn8`

Why Combine Learning and Reasoning Today?

1. Practically Useful for Verification of Complex HW/SW and Math
   - Formal Proof of the Kepler Conjecture (2014 – Hales et al – 20k lemmas)
   - Formal Proof of the Feit-Thompson Theorem (2012 – Gonthier et al)
   - Verification of several math textbooks and CS algorithms
   - Verification of compilers (CompCert)
   - Verification of OS microkernels (seL4), HW chips (Intel), transport, finance,
   - Verification of cryptographic protocols (Amazon), etc.
2. Blue Sky AI Visions:
   - Get strong AI by learning/reasoning over large KBs of human thought?
   - Big formal theories: good semantic approximation of such thinking KBs?
   - Deep non-contradictory semantics – better than scanning books?
   - Gradually try learning math/science
   - automate/verify them, include law, etc. (Leibniz, McCarthy, ..)
     - What are the components (inductive/deductive thinking)?
     - How to combine them together?
   - As of 2022/23: Overlaps/analogies/differences with LLMs?

Sample of Learning Approaches

- **neural networks** (statistical ML, old!) – backprop, SGD, deep learning, convolutional, recurrent, attention/transformers, tree NNs, graph NNs, etc.
- **decision trees, random forests, gradient boosted trees** – find good classifying attributes (and/or their values); more explainable, often SoTA
- **support vector machines** – find a good classifying hyperplane, possibly after non-linear transformation of the data (*kernel methods*)
- **k-nearest neighbor** – find the k nearest neighbors to the query, combine their solutions, good for *online learning* (important in ITP)
- **naive Bayes** – compute probabilities of outcomes assuming complete (naive) independence of characterizing features, i.e., just multiplying probabilities: $P(y|\mathbf{x}) = P(x_1|y) * P(x_2|y) * \ldots * P(x_n|y) * P(y)/P(\mathbf{x})$
- **inductive logic programming** (symbolic ML) – generate logical explanation (program) from a set of ground clauses by generalization
- **genetic algorithms** – evolve large population by crossover and mutation
- various **combinations** of statistical and symbolic approaches
- **supervised, unsupervised, online/incremental, reinforcement** learning (actions, explore/exploit, cumulative reward)

Learning – Features and Data Preprocessing

- **Extremely important** - if irrelevant, there is no way to learn the function from input to output ("garbage in garbage out")
- **Feature discovery/engineering** – a big field, a bit overshadowed by DL
- **Deep Learning (DL)** – deep neural nets that automatically find important high-level features for a task, can be structured (tree/graph NNs)
- **Data Augmentation and Selection** – how do we generate/select more/better data to learn on?
- **Latent Semantics, PCA, dimensionality reduction**: use linear algebra (eigenvector decomposition) to discover the most similar features, make approximate equivalence classes from them; or just use *hashing*
- **word2vec and related/neural methods**: represent words/sentences by *embeddings* (in a high-dimensional real vector space) learned by predicting the next word on a large corpus like Wikipedia
- **math and theorem proving**: syntactic/semantic/computational patterns/abstractions/programs
- How do we **represent** math data (formulas, proofs, models) in our mind?

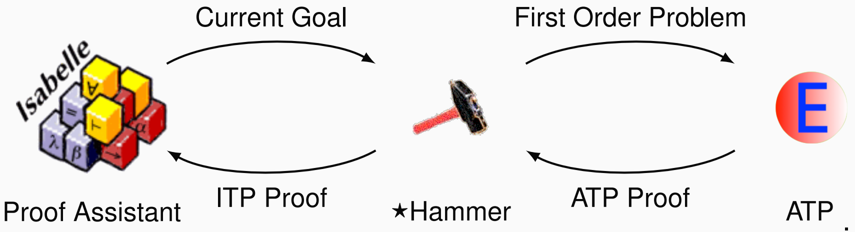
Using Learning to Guide Theorem Proving

- **high-level**: pre-select lemmas from a large library, give them to ATPs
- **high-level**: pre-select a good ATP strategy/portfolio for a problem
- **low-level**: guide every inference step of ATPs (tableau, superposition)
- **low-level**: guide every kernel step of LCF-style ITPs
- **mid-level**: guide application of tactics in ITPs, learn new tactics
- **mid-level**: invent suitable strategies/procedures for classes of problems
- **mid-level**: invent suitable conjectures for a problem
- **mid-level**: invent suitable concepts/models for problems/theories
- **proof sketches**: explore stronger/related theories to get proof ideas
- **theory exploration**: develop interesting theories by conjecturing/proving
- **feedback loops**: (dis)prove, learn from it, (dis)prove more, learn more, ...
- **autoformalization**: (semi-)automate translation from LaTeX to formal
- ...

AI/TP Examples and Demos

- ENIGMA/hammer proofs of Pythagoras : `https://bit.ly/2MVPAn7` (more at `http://grid01.ciirc.cvut.cz/~mptp/enigma-ex.pdf`) and simplified Carmichael `https://bit.ly/3oGBdRz`,
- 3-phase ENIGMA: `https://bit.ly/3C0Lwa8`, `https://bit.ly/3BWqR6K`
- Long trig proof from 1k axioms: `https://bit.ly/2YZ0OgX`
- Extreme Deepire/AVATAR proof of $\epsilon_0 = \omega^{\omega^{\omega^{\cdot^{\cdot^{\cdot}}}}}$ `https://bit.ly/3Ne4WNX`
- Hammering demo: `http://grid01.ciirc.cvut.cz/~mptp/out4.ogv`
- TacticToe on HOL4: `http://grid01.ciirc.cvut.cz/~mptp/tactictoe_demo.ogv`
- TacticToe longer: `https://www.youtube.com/watch?v=BO4Y8ynwT6Y`
- Tactician for Coq: `https://blaauwbroek.eu/papers/cicm2020/demo.mp4`, `https://coq-tactician.github.io/demo.html`
- Inf2formal over HOL Light: `http://grid01.ciirc.cvut.cz/~mptp/demo.ogv`
- QSynt: AI rediscovers the Fermat primality test: `https://www.youtube.com/watch?v=24oejR9wsXs`

Today's AI-ATP systems (⋆-Hammers)



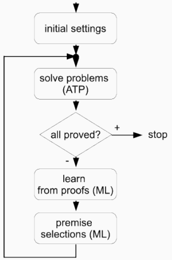
How much can it do?

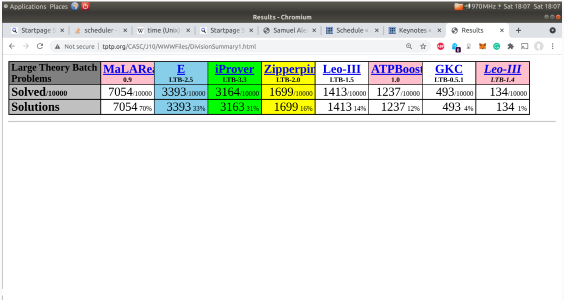
- Mizar / MML – MizAR
- Isabelle (Auth, Jinja) – Sledgehammer
- Flyspeck (including core HOL Light and Multivariate) – HOL(y)Hammer
- HOL4 (Gauthier and Kaliszyk)
- CoqHammer (Czajka and Kaliszyk) - about 40% on Coq standard library

$\approx$ 40-45% success by 2016, 60% on Mizar as of 2021

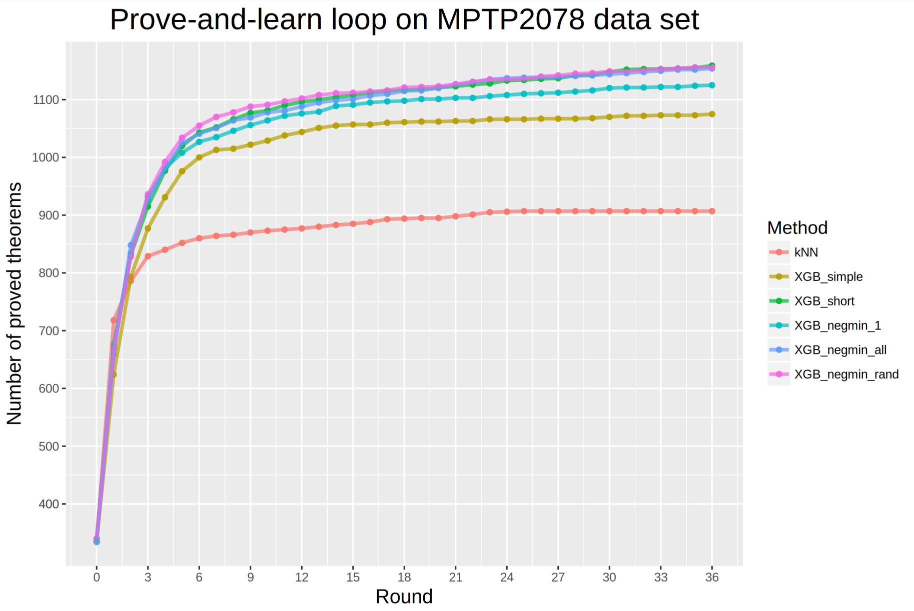
High-level feedback loops – MALARea, ATPBoost

- Machine Learner for Autom. Reasoning (2006) – infinite hammering
- feedback loop interleaving ATP with learning premise selection
- both syntactic and semantic features for characterizing formulas:
- evolving set of finite (counter)models in which formulas evaluated
- winning AI/ATP benchmarks (MPTPChallenge, CASC 08/12/13/18/20)
- ATPBoost (Piotrowski) - recent incarnation focusing on multiple proofs

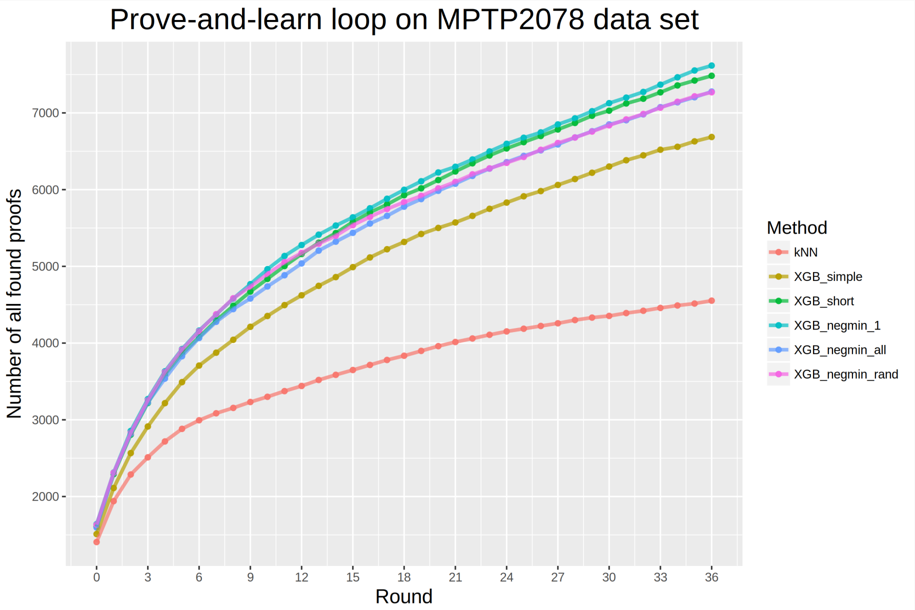




Large Theory Batch Problems	MaLARe 0.9	E LTB-2.5	iProver LTB-3.3	Zipperpin LTB-2.0	Leo-III LTB-1.5	ATPBoost 1.0	GKC LTB-0.5.1	Leo-III LTB-1.4
Solved/10000	7054/10000	3393/10000	3164/10000	1699/10000	1413/10000	1237/10000	493/10000	134/10000
Solutions	7054 70%	3393 33%	3163 31%	1699 16%	1413 14%	1237 12%	493 4%	134 1%



Prove-and-learn loop on MPTP2078 data set
Number of proved theorems
Round
Method
kNN
XGB_simple
XGB_short
XGB_negmin_1
XGB_negmin_all
XGB_negmin_rand


Prove-and-learn loop on MPTP2078 data set
Number of all found proofs
2000
3000
4000
5000
6000
7000
0
3
6
9
12
15
18
21
24
27
30
33
36
Round
Method
kNN
XGB_simple
XGB_short
XGB_negmin_1
XGB_negmin_all
XGB_negmin_rand

Various Improvements and Additions

- Model-based features for semantic similarity [IJCAR'08]
- Features encoding term matching/unification [IJCAI'15]
- Various learners: weighted k-NN, boosted trees (LightGBM,XGBoost)
- Matching and transferring concepts and theorems between libraries (Gauthier & Kaliszyk) – allows "superhammers", conjecturing, and more
- Lemmatization – extracting and considering millions of low-level lemmas
- LSI, word2vec, neural models, definitional embeddings (with Google)
- Learning in binary setting from many alternative proofs
- Negative/positive mining (ATPBoost - Piotrowski & JU, 2018)
- Stateful neural methods: RNNs and Transformers (Piotrowski & JU, 2020) (smooth transition from fact selection to conjecturing – Jakubuv & JU 2020)
- Currently strongest: Name-independent graph neural nets (Olsak, 2020) (generalize very well to new terminology/lemmas)

Low-level: Statistical Guidance of Connection Tableau

- learn guidance of every clausal inference in connection tableau (leanCoP)
- set of first-order clauses, *extension* and *reduction* steps
- proof finished when all branches are closed
- a lot of nondeterminism, requires backtracking
- *Iterative deepening* used in leanCoP to ensure completeness
- good for learning – the tableau compactly represents the proof state

Clauses:

$c_1 : P(x)$

$c_2 : R(x, y) \lor \neg P(x) \lor Q(y)$

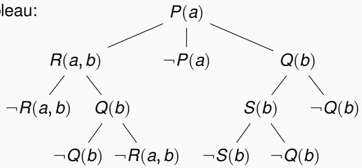
$c_3 : S(x) \lor \neg Q(b)$

$c_4 : \neg S(x) \lor \neg Q(x)$

$c_5 : \neg Q(x) \lor \neg R(a, x)$

$c_6 : \neg R(a, x) \lor Q(x)$

Closed Connection Tableau:



Statistical Guidance of Connection Tableau – rlCoP

- 2018: strong learners via C interface to OCAML (boosted trees)
- remove iterative deepening, the prover can go arbitrarily deep
- added Monte-Carlo Tree Search (MCTS) (inspired by AlphaGo/Zero)
- MCTS search nodes are sequences of clause application
- a good heuristic to explore new vs exploit good nodes:

$$UCT(i) = \frac{w_i}{n_i} + c \cdot p_i \cdot \sqrt{\frac{\ln N}{n_i}} \qquad \text{(UCT - Kocsis, Szepesvari 2006)}$$

- learning both policy (p) (clause selection) and value (w) (state evaluation)
- clauses represented not by names but also by features (generalize!)
- binary learning setting used: | proof state | clause features |
- mostly term walks of length 3 (trigrams), hashed into small integers
- many iterations of proving and learning
- More recently also with GNNs (Olsak, Rawson, Zombori, ...)

Statistical Guidance of Connection Tableau – rlCoP

- On 32k Mizar40 problems using 200k inference limit
- nonlearning CoPs:

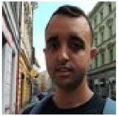
System	leanCoP	bare prover	rlCoP no policy/value (UCT only)
Training problems proved	10438	4184	7348
Testing problems proved	**1143**	431	804
Total problems proved	11581	4615	8152

- rlCoP with policy/value after 5 proving/learning iters on the training data
- $1624/1143 = 42.1$% improvement over leanCoP on the testing problems

Iteration	1	2	3	4	5	6	7	8
Training proved	12325	13749	14155	14363	14403	14431	14342	**14498**
Testing proved	1354	1519	1566	1595	**1624**	1586	1582	1591

ENIGMA (2017): Guiding the Best ATPs like E Prover

- ENIGMA (Jan Jakubuv, Zar Goertzel, Karel Chvalovsky, others)

- The proof state are two large heaps of clauses *processed/unprocessed*
- learn on E's proof search traces, put classifier in E
- positive examples: clauses (lemmas) used in the proof
- negative examples: clauses (lemmas) not used in the proof
- 2021 multi-phase architecture (combination of different methods):
    - fast gradient-boosted decision trees (GBDTs) used in 2 ways
    - fast logic-aware graph neural network (GNN - Olsak) run on a GPU server
    - logic-based subsumption using fast indexing (discrimination trees - Schulz)
- The GNN scores many clauses (context/query) together in a large graph
- Sparse - vastly more efficient than transformers ($\sim$100k symbols)
- 2021: leapfrogging and Split&Merge:
- aiming at learning reasoning/algo components

Feedback prove/learn loop for ENIGMA on Mizar data

- Done on 57880 Mizar problems recently
- Serious ML-guidance breakthrough applied to the best ATPs
- Ultimately a 70% improvement over the original strategy in 2019
- From 14933 proofs to 25397 proofs (all 10s CPU - no cheating)
- Went up to 40k in more iterations and 60s time in 2020
- 75% of the Mizar corpus reached in July 2021 - higher times and many runs: `https://github.com/ai4reason/ATP_Proofs`

	$\mathcal{S}$	$\mathcal{S} \odot \mathcal{M}_9^0$	$\mathcal{S} \oplus \mathcal{M}_9^0$	$\mathcal{S} \odot \mathcal{M}_9^1$	$\mathcal{S} \oplus \mathcal{M}_9^1$	$\mathcal{S} \odot \mathcal{M}_9^2$	$\mathcal{S} \oplus \mathcal{M}_9^2$	$\mathcal{S} \odot \mathcal{M}_9^3$	$\mathcal{S} \oplus \mathcal{M}_9^3$
solved	**14933**	16574	20366	21564	22839	22413	23467	22910	23753
$\mathcal{S}\%$	+0%	+10.5%	+35.8%	+43.8%	+52.3%	+49.4%	+56.5%	+52.8%	+58.4
$\mathcal{S}+$	+0	+4364	+6215	+7774	+8414	+8407	+8964	+8822	+9274
$\mathcal{S}-$	-0	-2723	-782	-1143	-508	-927	-430	-845	-454

	$\mathcal{S} \odot \mathcal{M}_{12}^3$	$\mathcal{S} \oplus \mathcal{M}_{12}^3$	$\mathcal{S} \odot \mathcal{M}_{16}^3$	$\mathcal{S} \oplus \mathcal{M}_{16}^3$
solved	24159	24701	25100	**25397**
$\mathcal{S}\%$	+61.1%	+64.8%	+68.0%	**+70.0%**
$\mathcal{S}+$	+9761	+10063	+10476	+10647
$\mathcal{S}-$	-535	-295	-309	-183

ENIGMA Anonymous: Learning from patterns only

- The GNN and GBDTs only learn from formula structure, not symbols
- Not from symbols like $+$ and $*$ as Transformer & Co.
- E.g., learning on additive groups thus transfers to multiplicative groups
- **Evaluation** of old-Mizar ENIGMA on 242 new Mizar articles:
- 13370 new theorems, $>$ 50% of them with new terminology:
- The 3-phase ENIGMA is **58%** better on them than unguided E
- While **53.5%** on the old Mizar (where this ENIGMA was trained)
- Generalizing, analogizing and transfer abilities unusual in the large transformer models

More Low-Level Guidance of Various Creatures

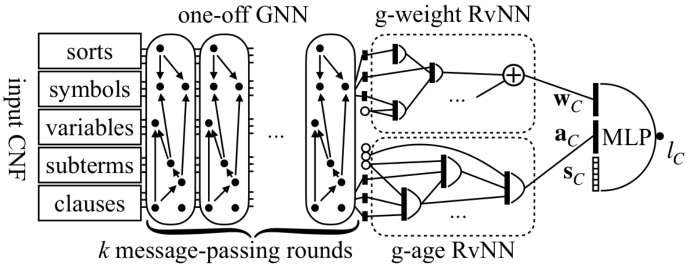
- Neural (TNN) clause selection in Vampire (Deepire - M. Suda):
  Learn from clause *derivation trees only*
  Not looking at what it says, just who its ancestors were.
- Fast and surprisingly good: Extreme Deepire/AVATAR proof of
  $\epsilon_0 = \omega^{\omega^{\omega^{\cdot^{\cdot^{\cdot}}}}}$ `https://bit.ly/3Ne4WNX`
- 1193-long proof takes *about the same resources as one GPT-3/4 reply*
- GNN-based guidance in iProver (Chvalovsky, Korovin, Piepenbrock)
- New (*dynamic data*) way of training
- Led to doubled real-time performance of iProver's instantiation mode
- CVC5: neural & GBDT instantiation guidance (Piepenbrock, Jakubuv)
- very recently 20% improvement on Mizar
- Hints method for Otter/Prover9 (Veroff):
- boost inferences on clauses that match a lemma used in a related proof
- symbolic ML - can be combined with statistical - proof completion vectors

Behold: 1-CPU vampiric “GenAI” proving $\epsilon_0 = \omega^{\omega^{\omega^{\cdot^{\cdot^{\cdot}}}}}$

```
% Refutation found. Thanks to Tanya!
% SZS status Theorem for t36_ordinal5
% SZS output start Proof for t36_ordinal5
fof(f2863755,plain,( $false), inference(avatar_sat_refutation,
[... 2500 lines of proof ...]
% SZS output end Proof for t36_ordinal5
% ------------------------------
% Version: Vampire 4.5.1 (commit 110f4142 on 2020-10-16 16:55:15 +0200)
% Termination reason: Refutation
% Input formulas: 73
% Proof axioms: 49
% Proof steps: 1193
% Main loop iterations started: 38065
% Generated clauses: 2392519
% SAT solver time: 181.936 s
% congruence closure: 167.665 s ( own 156.041 s )
% neural model evaluation: 18.493 s
% other: 503.976 s ( own 26.185 s )
```

A Fresh From the Oven Deepire Slide

- March’25: **20% improvement** of Vampire: *Efficient Neural Clause Selection Reinforcement* (`https://arxiv.org/abs/2503.07792`)
- I.e. the strongest ATP, on the most used TPTP benchmark
- After improving it by 27% in CASC last year!
- Everything is done in real time on a single CPU, anonymous GNN/RNN
- Solved 130 problems with Rating 1 of which 49 never solved before



. Clause evaluation neural architecture. Information flows from left to right.

TacticToe: mid-level ITP Guidance (Gauthier'17,18)

- TTT learns from human and its own tactical HOL4 proofs
- No translation or reconstruction needed - native tactical proofs
- Fully integrated with HOL4 and easy to use
- Similar to rlCoP: policy/value learning for applying tactics in a state
- Demo: `http://grid01.ciirc.cvut.cz/~mptp/tactictoe_demo.ogv`
- However much more technically challenging - a real breakthrough:
  - tactic and goal state recording
  - tactic argument abstraction
  - absolutization of tactic names
  - nontrivial evaluation issues
  - these issues have often more impact than adding better learners
- policy: which tactic/parameters to choose for a current goal?
- value: how likely is this proof state succeed?
- 66% of HOL4 toplevel proofs in 60s (better than a hammer!)
- similar followup work for HOL Light (Google), Coq, Lean, ...

Tactician: Tactical Guidance for Coq (Blaauwbroek'20)

- Tactical guidance of Coq proofs
- Technically very challenging to do right - the Coq internals again nontrivial
- 39.3% on the Coq standard library, 56.7% in a union with CoqHammer (orthogonal)
- Fast approximate hashing for k-NN makes a lot of difference
- Fast re-learning more important than "cooler"/slower learners
- Fully integrated with Coq, should work for any development
- User friendly, installation friendly, integration/maintenance friendly
- Demo: `https://blaauwbroek.eu/papers/cicm2020/demo.mp4`, `https://coq-tactician.github.io/demo.html`
- Took several years, but could become a common tool for Coq formalizers
- Recently GNNs added, a major comparison of k-NN, GNN and LMs (Graph2Tac - `https://arxiv.org/abs/2401.02949`)

Neural Autoformalization (Wang et al., 2018)

- generate about 1M Latex - Mizar pairs synthetically (quite advanced)
- train neural seq-to-seq translation models (Luong – NMT)
- evaluate on about 100k examples
- many architectures tested, some work much better than others
- very important latest invention: attention in the seq-to-seq models
- more data crucial for neural training
- Recent addition: unsupervised MT methods (Lample et all 2018) – no need for aligned data, improving a lot!
- Type-checking not yet internal (boosting well-typed data externally)

Neural Autoformalization data

Rendered LaTeX	If $X \subseteq Y \subseteq Z$, then $X \subseteq Z$.
Mizar	`X c= Y & Y c= Z implies X c= Z;`
Tokenized Mizar	`X c= Y & Y c= Z implies X c= Z ;`
LaTeX	`If $X \subseteq Y \subseteq Z$, then $X \subseteq Z$.`
Tokenized LaTeX	`If $ X \subseteq Y \subseteq Z $ , then $ X \subseteq Z $ .`

Neural Fun – Performance after Some Training

Rendered LaTeX	Suppose s_8 is convergent and s_7 is convergent . Then $\lim(s_8+s_7) = \lim s_8 + \lim s_7$
Input LaTeX	`Suppose $ { s _ { 8 } } $ is convergent and $ { s _ { 7 } } $ is convergent . Then $ \mathop { \rm lim } ( { s _ { 8 } } { + } { s _ { 7 } } ) \mathrel { = } \mathop { \rm lim } { s _ { 8 } } { + } \mathop { \rm lim } { s _ { 7 } } $ .`
Correct	`seq1 is convergent & seq2 is convergent implies lim ( seq1 + seq2 ) = ( lim seq1 ) + ( lim seq2 ) ;`
Snapshot-1000	`x in dom f implies ( x * y ) * ( f | ( x | ( y | ( y | y ) ) ) ) = ( x | ( y | ( y | ( y | y ) ) ) ) ) ;`
Snapshot-2000	`seq is summable implies seq is summable ;`
Snapshot-3000	`seq is convergent & lim seq = 0c implies seq = seq ;`
Snapshot-4000	`seq is convergent & lim seq = lim seq implies seq1 + seq2 is convergent ;`
Snapshot-5000	`seq1 is convergent & lim seq2 = lim seq2 implies lim_inf seq1 = lim_inf seq2 ;`
Snapshot-6000	`seq is convergent & lim seq = lim seq implies seq1 + seq2 is convergent ;`
Snapshot-7000	`seq is convergent & seq9 is convergent implies lim ( seq + seq9 ) = ( lim seq ) + ( lim seq9 ) ;`

More on Conjecturing in Mathematics

- Targeted: generate intermediate lemmas (cuts) for a harder conjecture
- Unrestricted (theory exploration):
- Creation of interesting conjectures based on the previous theory
- One of the most interesting activities mathematicians do (how?)
- Higher-level AI/reasoning task - can we learn it?
- If so, we have solved math:
- ... just (recursively) divide Fermat into many subtasks ...
- ... and conquer (I mean: hammer) them away

Conjecturing and Proof Synthesis by Neural Methods

- Karpathy’15 - RNN experiments with generating fake Math over Stacks
- I have tried to use that for formal math in 2016 but it looked weak
- GPT (-2,3) looks stronger
- Renewed experiments in 2020 (JU & J. Jakubuv: First Neural Conjecturing Datasets and Experiments. CICM’20) on:
  - All Mizar articles, stripped of comments and concatenated together (78M)
  - Articles with added context/disambiguation (156M) (types, names, thesis)
  - TPTP proofs of 28271 Mizar/MPTP theorems by E/ENIGMA (658M)
  - Just the conjecture and premises needed for the 28271 proofs printed in prefix notation
- Quite interesting results, server for Mizar authors
- Quickly taken up by others on HOL, Isabelle, MetaMath ...
- Caveat: Watch for "model pretraining" on undisclosed corpora - often GitHub/math repos that may contain (translations of) the testing data

Can you find the flaw(s) in this fake GPT-2 proof?

Applications Places   4,71 GHz   Wed 15:02   Wed 15:02

emacs@dell

File Edit Options Buffers Tools Index Mizar Hide/Show Help

Save Undo

```
:: generated theorem with "proof"
theorem Th23: :: STIRL2_1:23
for X, Y being finite set st not X is empty & X c= Y
& card X = card Y holds X = Y
proof
  let X, Y be finite set ;
:: thesis: not X is empty & X c= Y & card X = card Y implies X = Y
  assume that
  A1: not X is empty and A2: X c= Y and A3: card X = card Y ;
:: thesis: X = Y
  card (Y \ X) = (card Y) - (card X) by A1, A3, CARD_2:44;
  then A4: card (Y \ X) = ((card Y) - 1) - (card X) by CARD_1:30;
  X = Y \ X by A2, A3, Th22;
  hence X = Y by A4, XBOOLE_0:def_10;
:: thesis: verum
end;
```

-:--- **card_tst.miz** 99% L2131 (Mizar Errors:13 hs Undo-Tree)

Figure: Fake full declarative GPT-2 “Mizar proof” - typechecks!

Gibberish Generator Provoking Algebraists

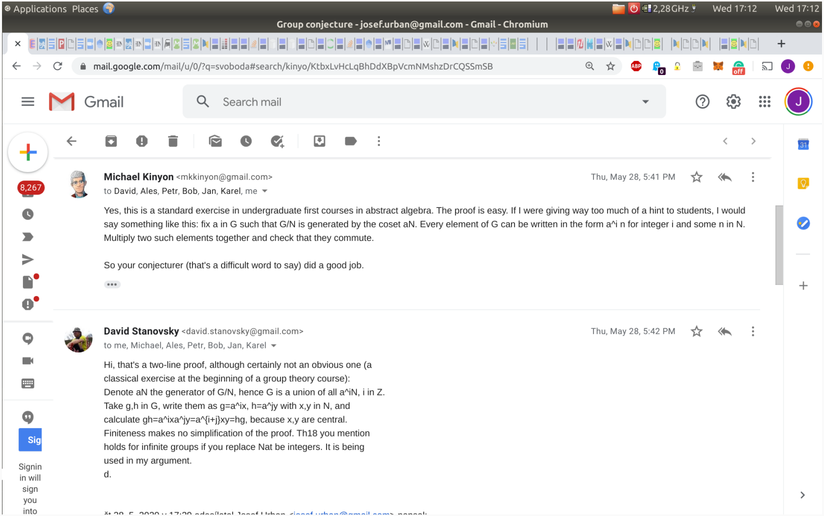


Figure: First successes in making mathematicians comment on AI.

A correct conjecture that was too hard to prove

Kinyon and Stanovsky (algebraists) confirmed that this cut is valid:

```
theorem Th10: :: GROUPP_1:10
for G being finite Group
for N being normal Subgroup of G st
N is Subgroup of center G & G ./. N is cyclic
holds G is commutative
```

The generalization that avoids finiteness:

```
for G being Group
for N being normal Subgroup of G st
N is Subgroup of center G & G ./. N is cyclic
holds G is commutative
```

More cuts

- In total 33100 in this experiment
- Ca 9k proved by trained ENIGMA
- Some are clearly false, yet quite natural to ask:

```
theorem :: SINCOS10:17

sec is increasing on [0, pi/2)
```

leads to conjecturing the following:

```
Every differentiable function is increasing.
```

Thanks and Advertisement

- Thanks for your attention!
- To push AI methods in math and theorem proving, we organize:
- AITP – Artificial Intelligence and Theorem Proving
- 10th issue: September 2025, Aussois, France, `aitp-conference.org/2025`
- ATP/ITP/Math vs AI/ML/AGI people, Computational linguists
- Discussion-oriented and experimental

Caveats, Comments and Dark Sides of ML

- Training on testing data – used to be a big taboo
- Useful as a sanity check: if the trained predictor doesn't predict well on the training data, there is a problem (garbage in garbage out)
- For simple ML systems training on testing data is less troublesome (e.g. naive Bayes with few features)
- Even for the GNN its performance on train/test split is similar
- For large systems like LLMs, this may however be a big issue - memorization vs generalization
- "pretraining": e.g. word embeddings useful for many NLP applications
- capture a lot of "latent semantics" of words, similar for pretrained LLMs
- However, they can also store/memorize any dataset they have seen on internet/GitHub
- E.g. formal proofs written in Lean instead of Coq, Mizar instead of Isabelle, HOL4 instead of HOL light
- Proofwiki/Arxiv versions of the proofs and the important lemmas in them (later used e.g. for conjecturing in the formal setting)

Caveats, Comments and Dark Sides of ML

- Use of synthetic data
- Can be very useful
- In general, there is a large field of data augmentation related to unsupervised and semi supervised learning
- e.g., try to generate many related easier problems, solve those, and learn on them to attack a harder problem
- a major issue is training and testing on an oversaturated dataset:
- even though the test data are formally disjoint from the training set, they can be very close to many training examples
- the synthetic data (and their parameterized generators) may in various hidden/clandestine ways target the test set
- a serious issue: unreleased synthetic data/generators are today used to claim major "breakthroughs"

Caveats, Comments and Dark Sides of ML

- Leaks from the future:
- chronological eval vs random split where the split contains the names of lemmas not yet seen in the chronological evaluation
- chronological evaluation is obviously the best but many ML systems are not easily updatable
- weaker learners like naive Bayes behaved similarly under the random split and in the chronological evaluation
- anonymous settings (typically not LLMs) largely mitigate this - eg the symbol independent GNN, or the binary setting when using GBDTs with anonymous features
- in nonanonymous settings, in-context learning can perhaps be used, this however so far requires a lot of resources
- also, there could be just a single evaluation - e.g. on new articles in a new version of the library (our Mizar 60 eval quite surprising)