String Theory and Data Science

Jim Halverson
Northeastern University

String Pheno 2018

Based on:
1706 with **Long**, Sung
1707 with Carifio, Krioukov, Nelson
1710 with **Long**, Sung

To appear:
1808 with Nelson, Ruehle
18xx with **Long**, Tian, Ruehle
18xx with **Long**, Nelson, Ruehle

“Can data science / ML be useful for problem X?”

“Can data science / ML be useful for problem X?”

- media coverage has a certain flavor . . .

“Can data science / ML be useful for problem X?”

- media coverage has a certain flavor . . .
- but in media because of non-trivial results.

“Can data science / ML be useful for problem X?”

- media coverage has a certain flavor . . .
- but in media because of non-trivial results.
- balanced view: ask a CS colleague or industry data scientist.

  typical Q’s from them:
  - what does your data “look like”? (construction is fine)
    bad answer: 3d toric variety.
    good answer: constrained set of sets of vectors in Z^3.

  - what are you trying to do / understand with it? (helps det. tech.)

The Data Science Zoo

and some string applications of those techniques,
mostly string compactification, but a few AdS / CFT and QFT

supervised machine learning. [He] [Krefl, Song] [Ruehle] [Carifio, JH, Krioukov, Nelson]

(simple algs, neural nets, "predict")

[Liu] [You, Yang, Qi] [Hashimoto, Sugishita, Tanaka, Tomiya]
[Wang, Zhang] [Bull, He, Jejjala, Mishra] [Jinno] [Krippendorf, Mayrhofer]

reinforcement learning (RL) / genetic algorithms:

(DNN + psych, DNN + evolution, agents that learn, move, and "search")

RL: [JH, Ruehle, Nelson] [JH, Long, Ruehle, Tian] [JH, Nilles, Vaudrevange, Ruehle], [Faraggi, Harries et al], [JH, Long, Ruehle, Nelson] Genetic: [Abel, Rizos], [Ruehle]

network science: ("connect") [Taylor, Wang] [Carifio, Cunningham, JH, Krioukov, Long, Nelson]

topological data analysis: (persistent homology, "shape" of data) [Cole, Shiu] (for non-gaussianity) [Cole, Shiu] (for string vacua)

conjecture generation / intelligible AI: [Carifio, JH, Krioukov,Nelson] [JH, Long, Ruehle, Tian]

(use ML to generate conjectures, prove theorems. "make rigorous".)

generative adversarial networks (GANs): [JH, Long, Ruehle]

("generate", produce interesting new examples from noise.) **and many more techniques**

blue = out, black = to appear but presented.

Three Goals

1) data science $\supsetneq$ supervised machine learning

they have suite of techniques. we have many problems.
is there a useful map between the two?

2) use some to tackle physics in landscape, which is both enormous and complex.

Desire better understanding of landscape implications for particle physics and cosmology. Q: requires formal theory progress but will smarter CS techniques also be necessary? Opinion: almost certainly.

3) higher level view: understand the broad ideas and what is possible.

broader string / QFT applications?

Outline

- **Primary Dataset:**
  large ensemble of F-theory geometries,
  physical facts about them.

- **Data Science for Rigor:**
  supervised learning —> conjecture —> gauge sector theorem

- **Data Science for Boundary Detection:**
  deep reinforcement learning the boundary of weak IIB.

- **Data Science for Complexity:** **(!! in progress !!)**
  deep reinforcement learning for Bousso-Polchinski CCs.

Large Dataset

- *topologically distinct, F-theory geometries, connected in moduli space. BP prob on top.*
- *have some universal physical features*

[JH, Long, Sung] x 2, 1706 and 1709

The Mathematics

- **4D F-theory:** 3-fld base B, 7-brane structure at generic CS det'd by B topology, called "non-Higgsable cluster."

  Some selective progress: Anderson, JH, Heckman, Grassi, Morrison, Rudelius, Shaneson, Taylor, Wang, Vafa.

- **Starting point:** B a weak Fano toric threefold, encoded in a fine regular star triangulation of a 3d reflexive polytope.

- **Topological transitions:** systematically perform sequences of toric blowups over toric points, then toric curves.

- **Sequence Bounds:** if all singularities are canonical, geom. is at finite distance from bulk of CS in the Weil-Petersson metric.

  Alg. Geom: [Hayakawa] [Wang] **in F-theory: [Morrison]**

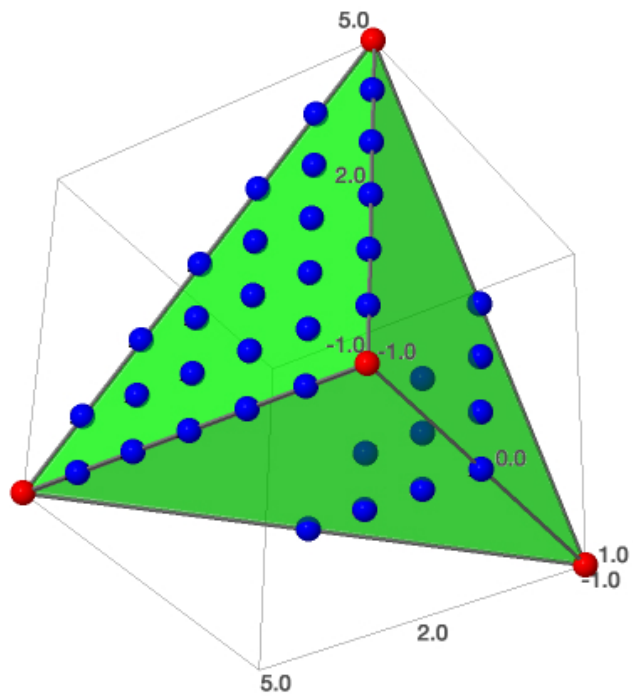
- **Classification:** there are 82 (41,873,645) sequences over curves (points) that satisfy a sufficient condition for canonical singularities.

- **Ensemble:** all ways of performing these sequences of blowups. from an initial, fixed, triangulated polytope.

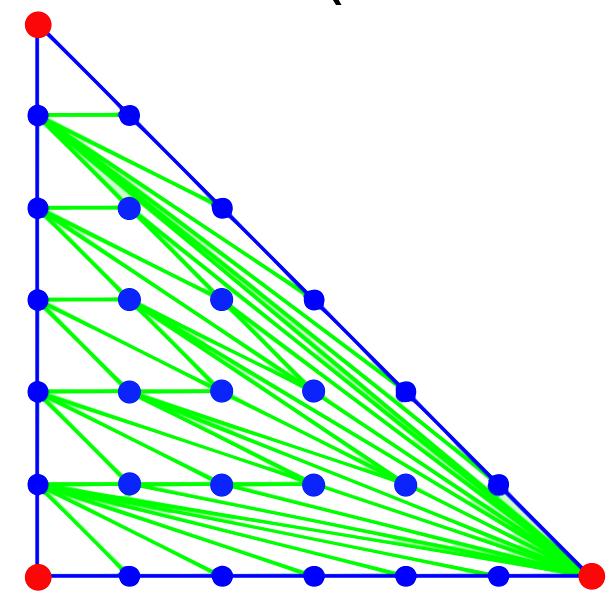
The Combinatoric Picture

The Combinatoric Picture

- **Polytope:**



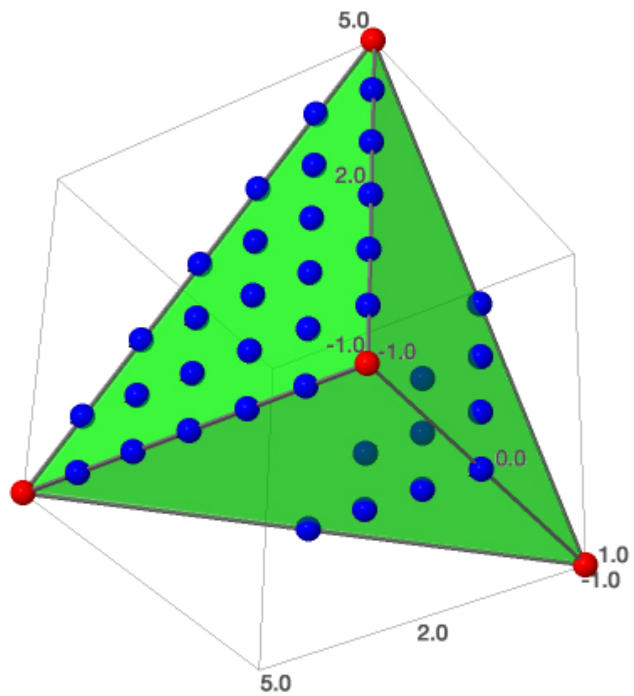
Triangulation: (codim 1 faces)



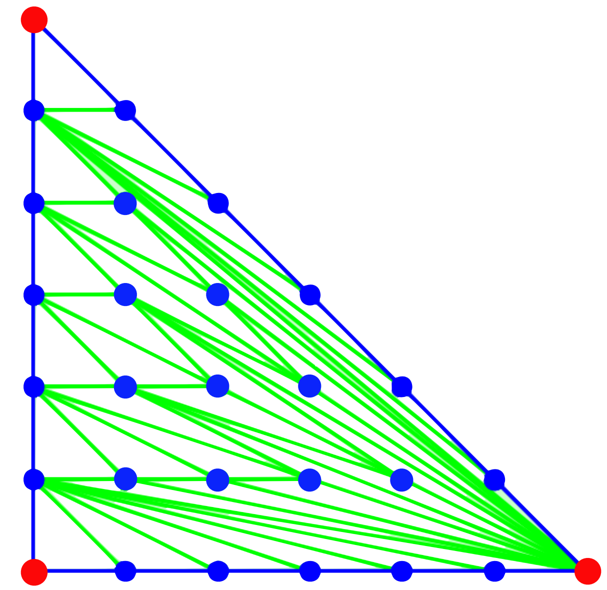
Fact: any FRS triangulation of this has 108 edges, 72 faces.

The Combinatoric Picture

- **Polytope:**

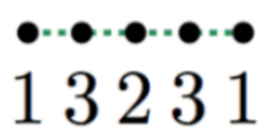


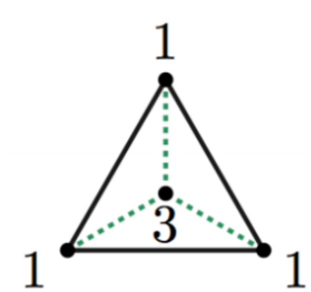
Triangulation: (codim 1 faces)

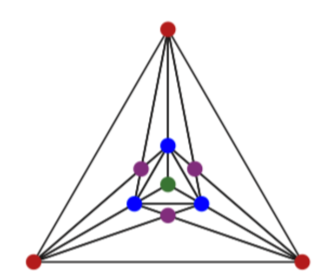


Fact: any FRS triangulation of this has 108 edges, 72 faces.

- **Rep seq. of blowups:** (topological transitions, project into board)

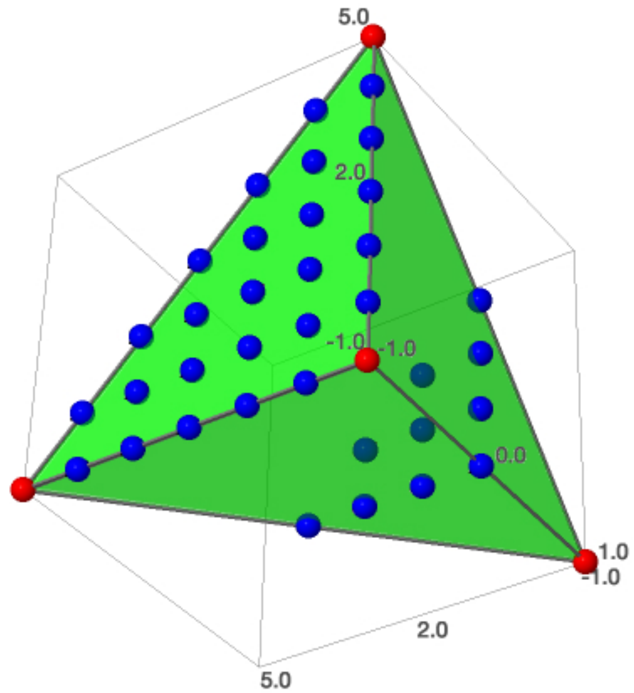




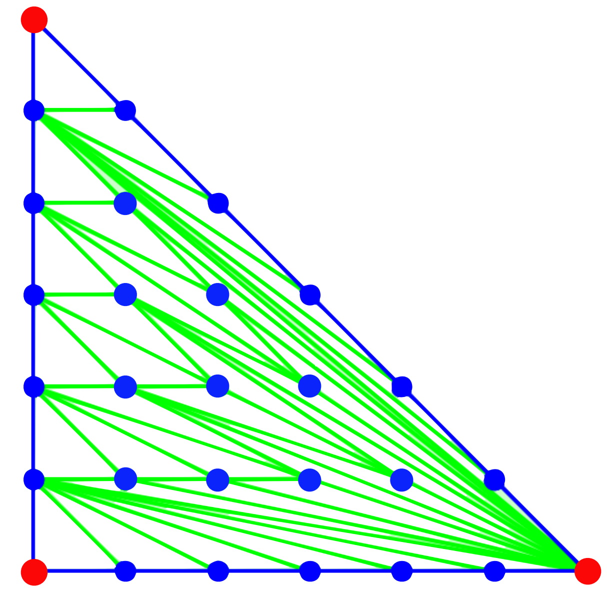


The Combinatoric Picture

- **Polytope:**

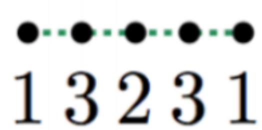


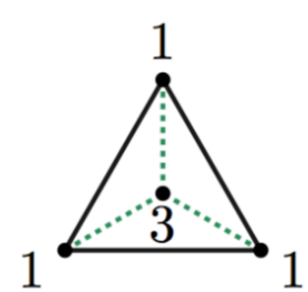
Triangulation: (codim 1 faces)

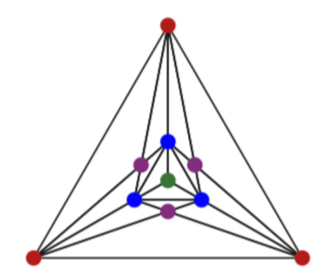


Fact: any FRS triangulation of this has 108 edges, 72 faces.

- **Rep seq. of blowups:** (topological transitions, project into board)







- **Ensemble Size:** (put the widgets on the triangulation)

$$82^{108} \times 4187364\,5^{72} = 2.96 \times 10^{755}$$

The Integer

exact lower bound on topologically distinct F-theory geometries.

296231521306658847673529878342496866080378021782076722241462425150065298508845621155
695644076004524635389132997146097067111728163388031396056782083347614866566976655879
021984012793769157213779169816753325335978325856186523598065404800375507248464342388
511047235006655822119955724586948796896518884757702603641702684094643379095601672444 3
830558425085959438371548897609540032670520282917998350399793115958231423494399655849
101565013413100994870230661418306662241733727360211923208425845972788270516755913617
371792570199269731190754783864008989335957900232620795965279395190240666076073359738
714472109475716934497800719015826299673405211214139176513510727323709162299991669746
741171716096000000000000000000000000000000000000000000000000000000000000000000000000

of 4319 3d reflexive polytopes,
there’s one other polytope that yields this same number of geometries.
they dominate the ensembles from other polytopes
by over 60 orders of magnitude.

Physics Universality

related ensemble of [Taylor, Wang] has similar results

- **Universality from algorithm:** (nice when this possible)
  geometric ansatz with computable high prob. —> physics property

- for any geom., easy to compute geometric 7-brane structure at generic CS

- **Universality of Non-Higgsable Seven-branes:**

$$P(\text{NHC in } S_{\Delta_1^\circ}) \geq 1 - 1.01 \times 10^{-755}$$

$$P(\text{NHC in } S_{\Delta_2^\circ}) \geq 1 - .338 \times 10^{-755}$$

- **Universality of Large Gauge Sectors:** $rk(G) \geq 160$

$$G \geq E_8^{10} \times F_4^{18} \times U^9 \times F_4^{H_2} \times G_2^{H_3} \times A_1^{H_4} \qquad U \in \{G_2, F_4, E_6\}$$

$$rk(G) \geq 160 + 4H_2 + 2H_3 + H_4$$

- **Cosmology Suggestion: Dark Glueballs**
  A Problem: [JH, Nelson, Ruehle] If solved, ultralight axions: [JH, Nelson, Ruehle, Salinas]

- **Universality of Strong Coupling:** $\frac{N_{\text{Sen}}}{N_{\text{Total}}} \leq 3.0 \times 10^{-391}$

Rigor

- *data science:*
  supervised ML —> conjecture —> theorem
- *this physics application: E6 in ensemble*

[Carifio, JH, Krioukov, Nelson] 1707

An E6 Puzzle

- **Gauge group result:** dominated by $G_i \in \{E_8, F_4, G_2, A_1\}$

  (interesting: groups with only self-conjugate reps!)

- **Something SM-useful?** E6 and SU(3) allowed for generic CS.

  - Simple conditions / probabilities for them not known.
  - in random samples, prob(E6) ~ 1/2000.
  - when E6 arises in RS, on a distinguished four-cycle T.

- Q: Can we train a ML model to accurately predict yes or no for E6 on T?

  Q: If so, can we learn how it makes its decision?

  in our paper: called **conjecture generation**.
  as a CS buzzword: **intelligible AI**.

Point: ML -> conjecture -> theorem means numerical -> rigorous

Supervised Machine Learning

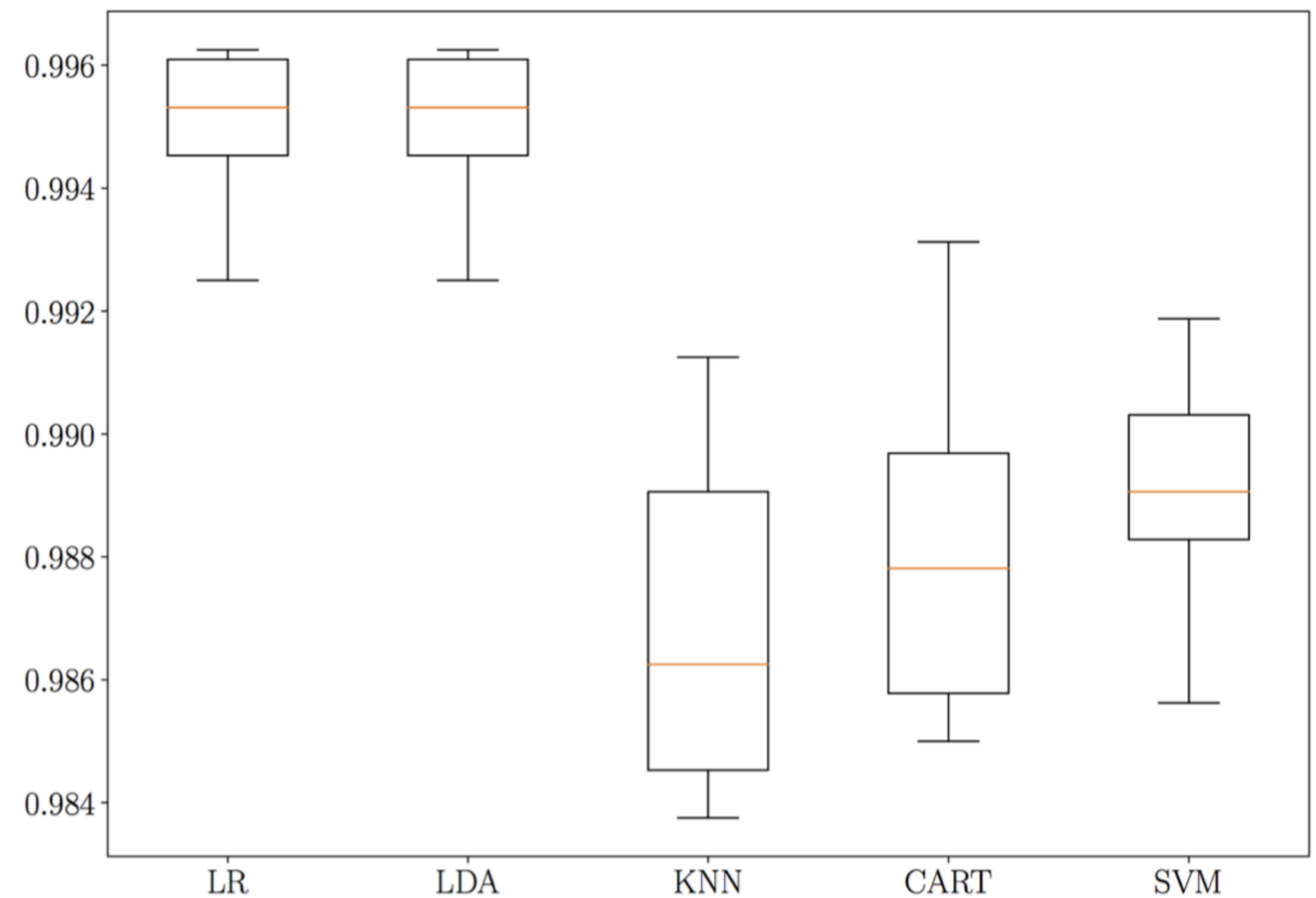
- given (input,output) pairs,
  learns to predict output
  test on unseen data,
  see how well the model does.

- **Training data:**
  in: blowup height data
  out: E6 or not.

  10000 random samples w/ E6,
  10000 w/o

- **Displayed:**
  whisker plots of % accuracy
  with 10-fold cross validation.

- >99% accuracy common.



	LR	LDA	KNN	CART	SVM
50/50 Validation Set	.994	.994	.982	.987	.989
Unenriched Set	.988	.988	.981	.988	.983.

training < 5 minutes per model,
easy to implement using sklearn (python).

note: simple techniques work well here,
no need for neural nets.

ML -> Conjecture -> Theorem

- supervised ML -> one variable was linchpin.
- that fact -> conjecture -> theorem (E6 iff).
- theorem -> probability computation.

Theorem: Suppose that with high probability the group G on v_{E_6} is $G \in \{E_6, E_7, E_8\}$ and that E_6 may only arise with $\tilde{m} = (-2, 0, 0)$. Given these assumptions, there are three cases that determine whether or not G is E_6.

a) If $a_{max} \geq 5$, $\tilde{m}$ cannot exist in Δ_g and the group on v_{E_6} is above E_6.

b) Consider $a_{max} = 4$. Let $v_i = a_i v_{E_6} + b_i v_2 + c_i v_3$ be a leaf built above v_{E_6}, and $B = \tilde{m} \cdot v_2$ and $C = \tilde{m} \cdot v_3$. Then G is E_6 if and only if $(B, b_i) > 0$ or $(C, c_i) > 0$ $\forall i$. Depending on the case, G may or may not be E_6.

c) If $a_{max} \leq 3$, $\tilde{m} \in \Delta_g$ and the group is E_6.

$$P(E_6 \,\text{on}\, v_{E_6} \,\text{in}\, T) = \left(1 - \frac{36}{82}\right)^9 \left(1 - \frac{18}{82}\right)^9 \simeq .00059128$$

$$\text{Number of } E_6 \text{ Models on } T = .00059128 \times \frac{1}{3} \times 2.96 \times 10^{755} = 5.83 \times 10^{751}.$$

- probability checks: 5 batches, 2m random samples each.

From Theorem	:	$.00059128 \times 2 \times 10^6 = 1182.56$
From Random Samples	:	$1183, 1181, 1194, 1125, 1195$

- **the point:** intelligible AI / conjecture generation can yield rigor.
  simpler the ML -> easier to conjecture. hard with ANNs?

Boundary Detection

- *data science:*
  reinforcement learning for AI game play.

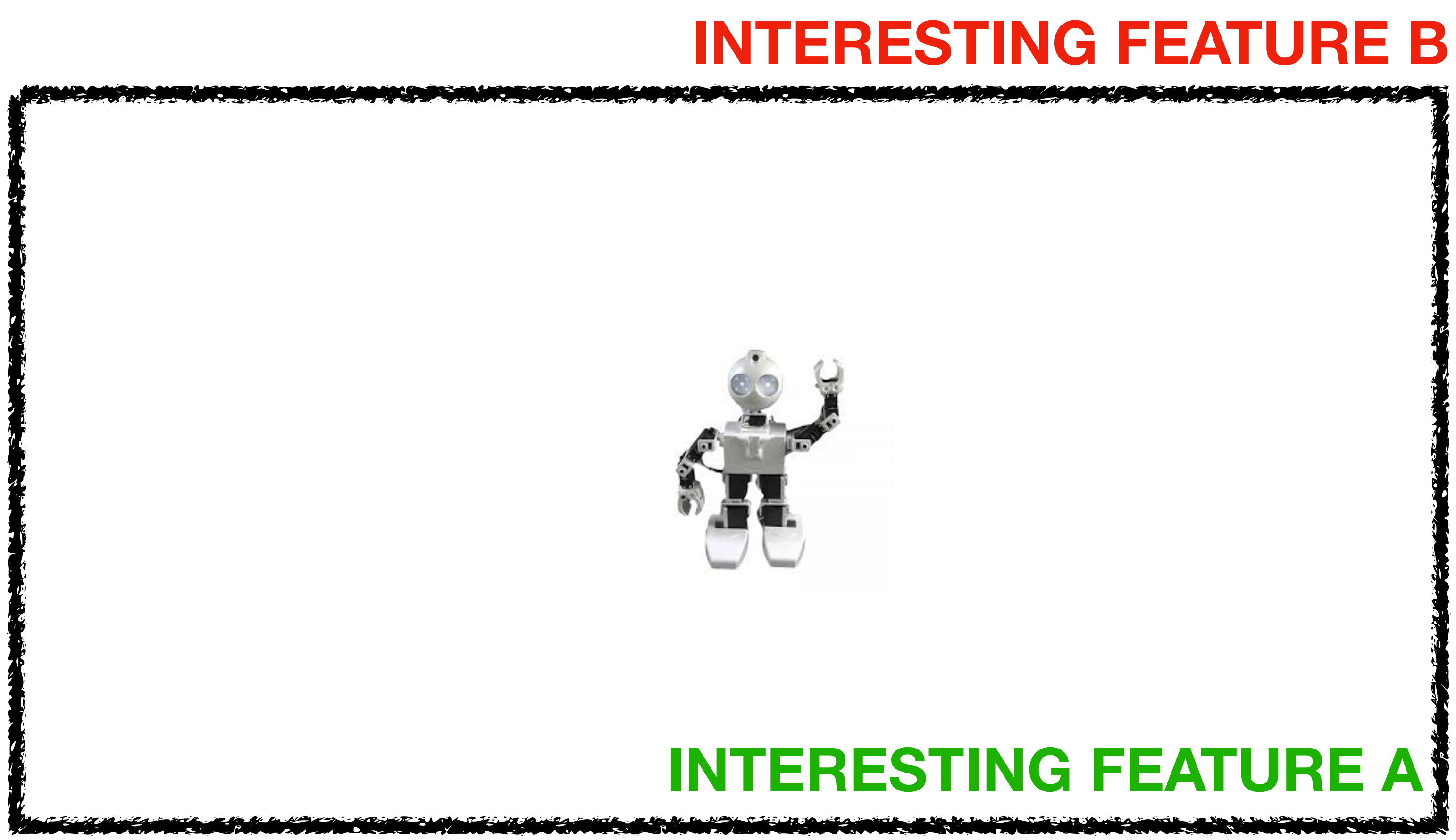
- *physics application:*
  what does weak IIB "look like" inside of F-theory?

[JH, Nelson, Ruehle] to appear, 1808
[JH, Long, Ruehle, Tian] to appear, 18xx

Picture: Boundary Detection

suppose you have a robot in large, complex space that wants to determine the boundary between feature A and B.

it doesn't know the global structure of the space, but it does know how to determine in vs. out.



in some cases, random walking and checking in vs. out isn't so inefficient, see above.

Picture: Boundary Detection

other case: random walk would not be so good, e.g. hard to discover deep crevices.

Q: can we reward robot so it learns how to not go out of bounds? explore space more intelligently?

Reinforcement Learning

supervised ML **predicts**, RL **explores / searches**
famous examples: AlphaGo & AlphaGo Zero

- an ***agent*** interacts in an ***environment***.
- it perceives a ***state*** from ***state space***.
- its ***policy*** picks and executes an action, given the state.
- agent arrives in new state, receives a ***reward.***
- successive rewards accumulate into ***return.***
- return may penalize future rewards via ***discount factor***.
- policy optimized to maximize reward, i.e. **agent learns how to act!**

Weak Coupling RL Game

- **state space:** 10^{755} F-theory geometries

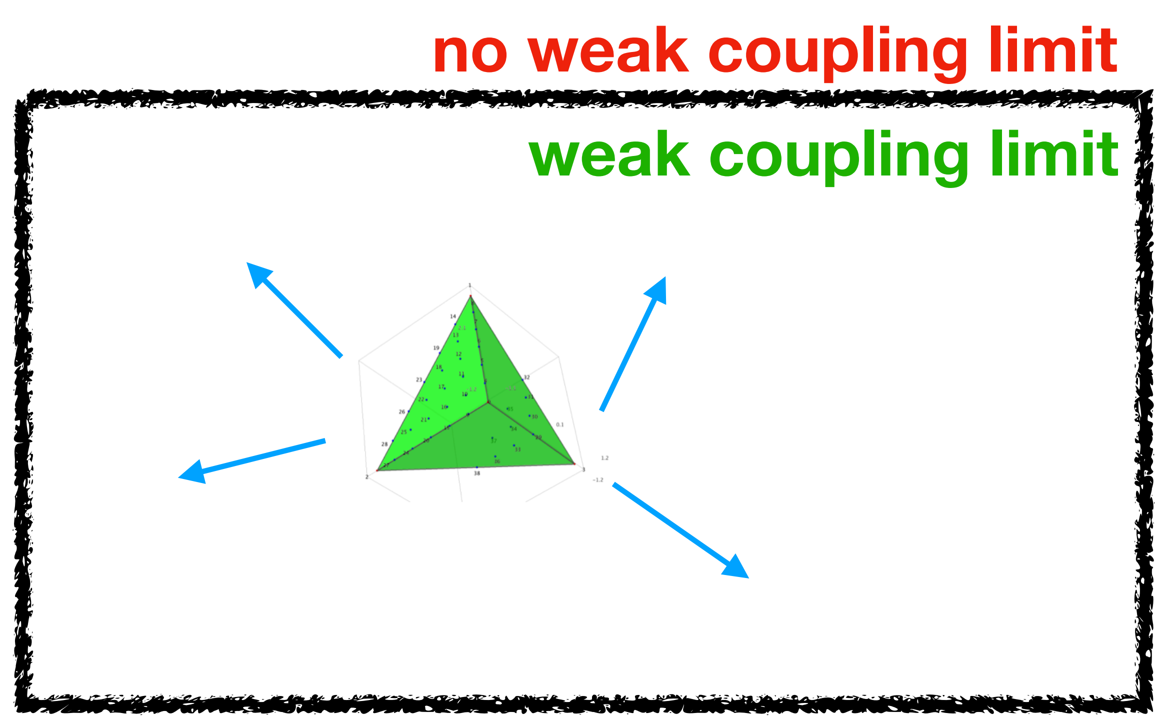
- **action space:** sequences of pt. or curve blowups that don't immediately rule out weak coupling limit.

- **the game:** start with weak Fano (i.e. no blow-ups). peform sequence of blowups. if: no weak coupling limit, out of bounds, end game.
  else: weak coupling limit possible, reward = 100 points, repeat.

- **RL algorithm:** A3C, an **A**synchronous **A**dvantage **A**ctor-**C**ritic
  [Mnih et al] Google DeepMind, 2016.

- **Implementation:**

  OpenAI Gym (RL framework)
  + ChainerRL (provides A3C)
  + physicist-provided game environment.



RL Game Results

recall: a "step" is performing a sequence of blowups.

- learning in under 1m steps.

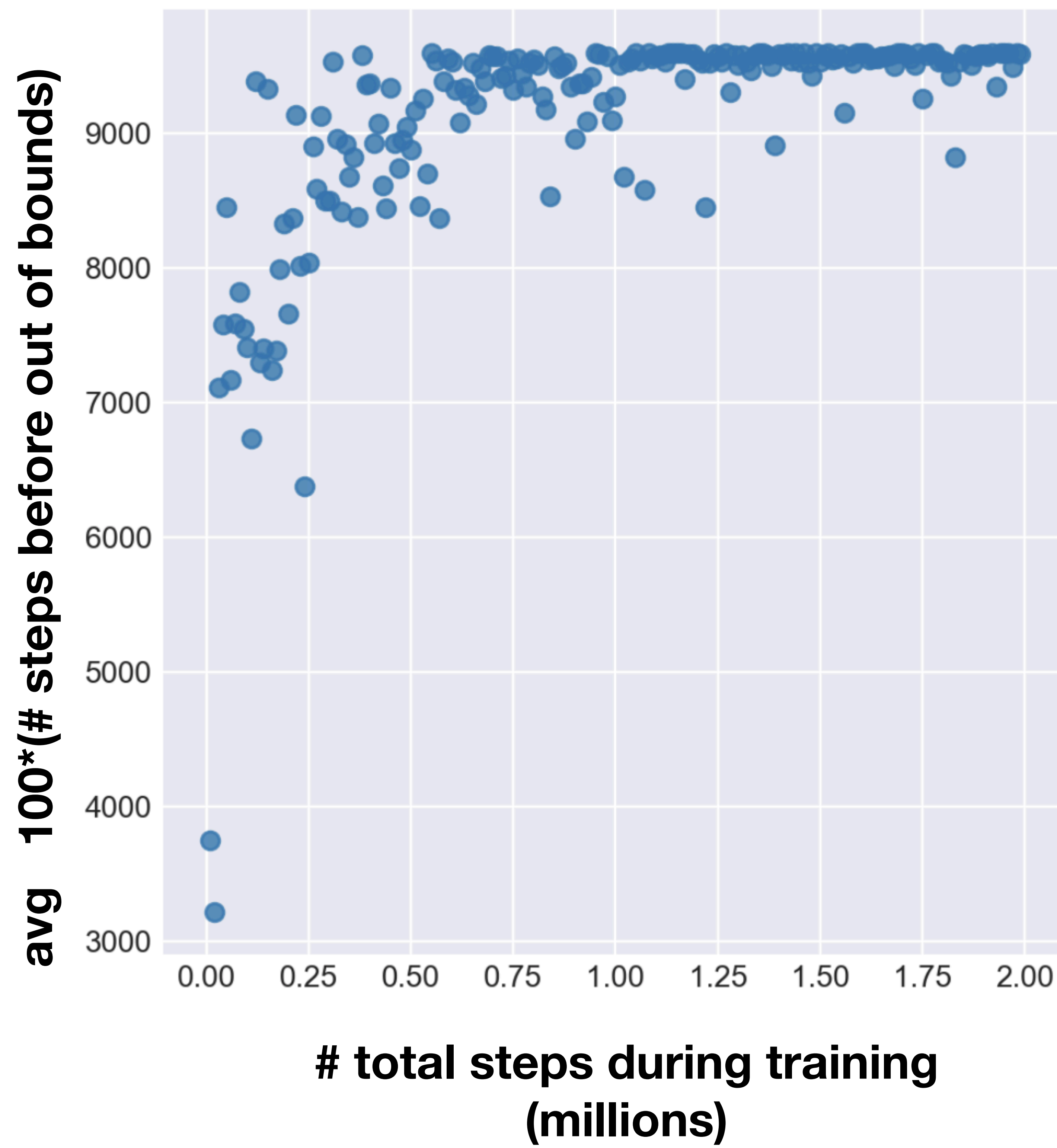
- score ~95k means can perform 95 sequences of blowups.

- random walker: can only perform 3-4 sequences of blowups before out of bounds (strong coupling).

- preliminary physics results:

  1) weak coupling very rare: $10^{30} < N_{weak} < 10^{80}$ in 10^{755} ensemble

  2) typical weakly coupled model has at least 30 SO(8) seven-brane stacks that can typically be Higgsed in CS.



Complexity

- *RL progress on NP-hard problems?*
- *first attempts at RL for Bousso-Polchinski.*

[JH, Long, Nelson, Ruehle], to appear

CCs and Complexity

- Bousso-Polchinski:

$$\Lambda = \Lambda_0 + g_{ij} N_i N_j \qquad N \in \mathbb{Z}^k$$

- Douglas-Denef:
for general metric, whether or not there is a lattice point with small CC in above model is NP-hard. (see DD for toy model caveats)

- Complexity vs. Practicality? in real world, concrete parameters, and it can pay it have "good enough" solutions to NP-hard problems. (Amazon?)

- CS for CCs in another complex model: **[Arkani-Hamed, Dimopoulos, Kachru]**

  - optimization via Karmarkar-Karp @ 10^6 - 10^9 moduli. lattice sieve @ lower, e.g. 10^4
  [Bao, Bousso, Jordan, Lackey]
  - **model-free** reinforcement learning (sim to A3C) @ 200 moduli. (KNAP200)
  [Bello et al.] Google Brain, 2016.

- **gen for complexity:** optimization? human-derived strategy, model-dependent.
RL? teach the game, machine learns the strategy.

trade-offs, not a priori clear which wins. should try both. OTOH, but model-free is good, and there there are famous cases where RL wins (AlphaGo).

Bousso-Polchinski RL game

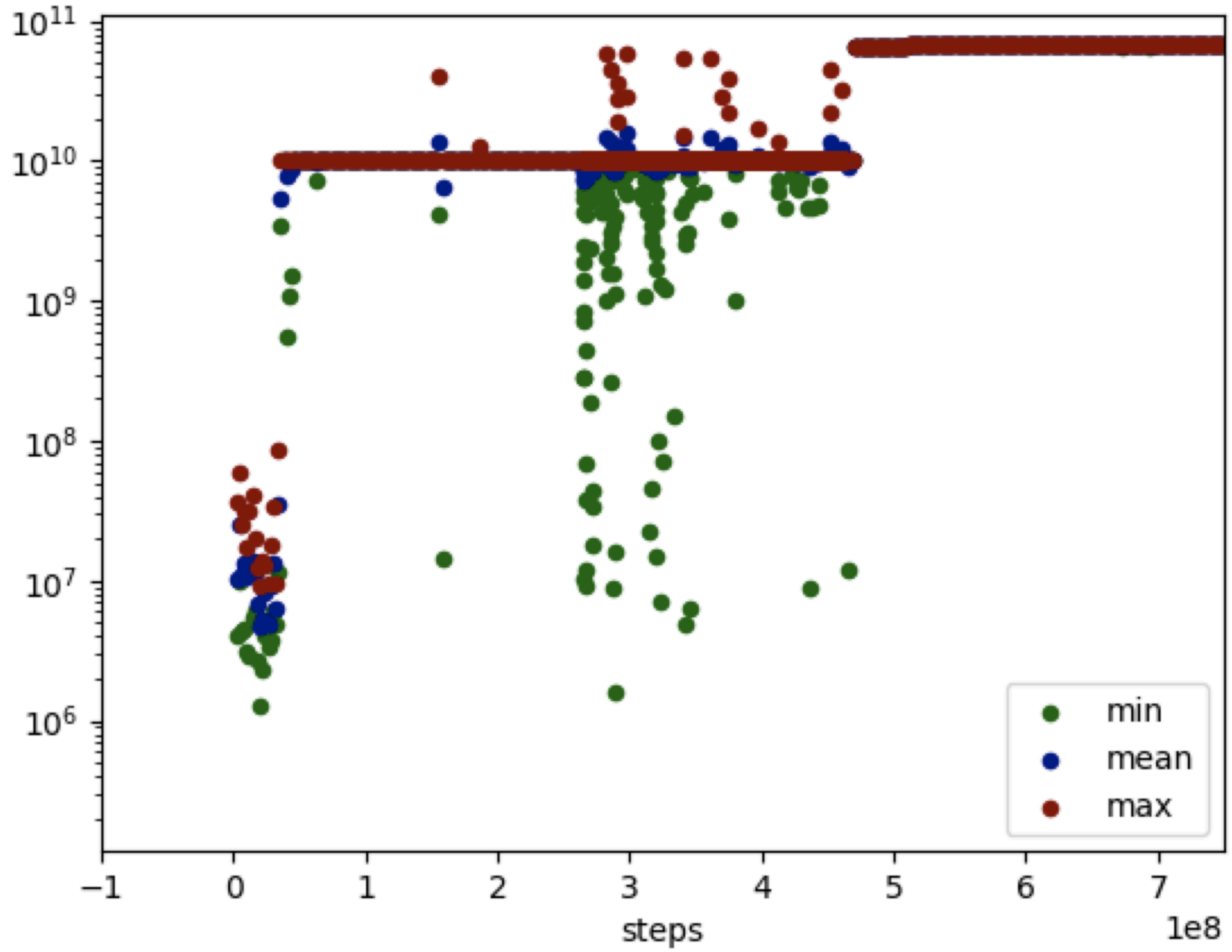
- metrics from Wishart ensemble, O shift to shortest eigenvector.
- state: a vector $N \in \mathbb{Z}^k$
- action: ++ or - - on any vector entry.
- CC formula with choice Λ_0 = -1.

$$\Lambda = -1 + N_i\, g_{ij}\, N_j$$

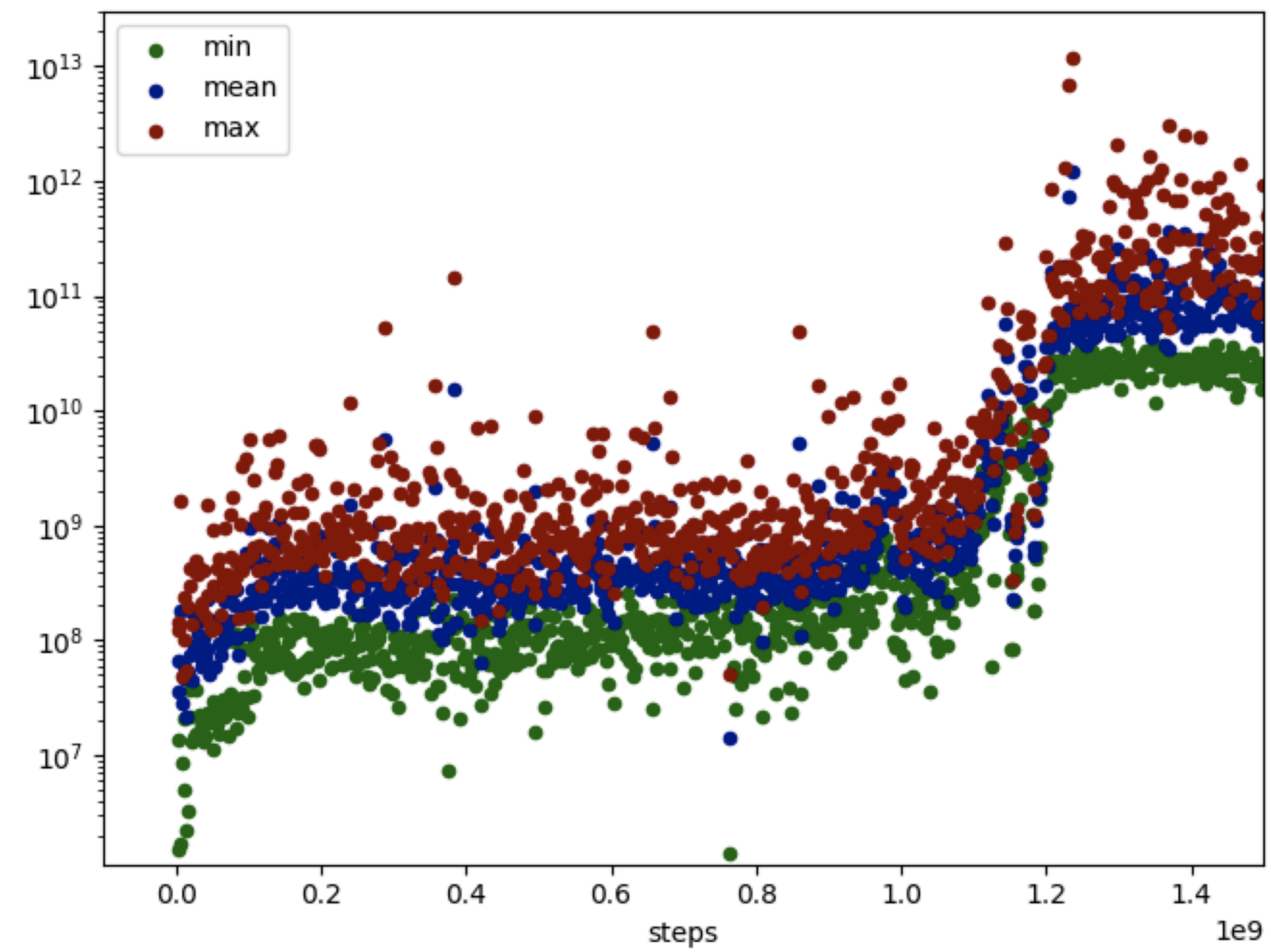
- distance from target ε: $d = |\Lambda - \epsilon|$
- reward as function of power p: $r = d^{-p}$
- episode over if hit ε or max_steps in {10k, 100k}

Very Preliminary Results

tweaking / training code that is O(2) weeks old



LSTM, $N_{mod} = 10,\ \sigma = .0001,\ \gamma = 0.99,\ \beta = .1$



$N_{mod} = 25,\ \sigma = .0001,\ \gamma = 0.9,\ \beta = 1.0$

- note: N_{mod} = 10, 25 here.
- learned 5-6 OOM in evaluation runs. overall best so far: $\Lambda = 10^{-16}$
- tried genetic algorithms, too. both hit a wall — increase moduli? BP is better for > 100.

Will learning stop here or
continue to smaller CCs?

Can we get improvements at higher moduli,
as expected for BP?

Stay tuned.

String Theory and Data Science

- ***for rigor:***
  supervised ML -> conjecture -> theorem. E6.

- ***for boundary detection:***
  RL to stay in bounds. Boundary of weak IIB.

- ***for complexity:*** *model-free RL on NP-hard landscape problems, such as BP CCs.*

standard supervised machine learning is quite useful,
but I wanted to emphasize there is a much broader suite of techniques.

Finish: A Brain Teaser


Finish: A Brain Teaser

- Q: in what 2015 movie did this pair co-star?

Finish: A Brain Teaser

- Q: in what 2015 movie did this pair co-star?
- A: they didn't, these people don't exist.

generated by generated adversarial network. (GAN).

[Karras et al, 2017]

Thanks for listening!