Big Data and the Landscape:

A Case Study in F-theory

Jim Halverson
Northeastern University

Based on:

1706.02299 with **Cody Long** and Ben Sung
1707.00655 with Jon Carifio, Dima Krioukov, and Brent Nelson
1710.09374 with **Cody Long** and Ben Sung
1711.06685 with Jon Carifio, Will Cunningham, Dima Krioukov, **Cody Long**, and Brent Nelson

Outline

- Big Data and the String Landscape.

what do we mean by “**big**”?
is the landscape big?
if yes, how do we deal with it?

broad proposal: literally in any way that makes progress.

technical proposal: algorithmic universality, supervised machine learning, deep reinforcement learning, network science, . . .

- A Big Network of F-theory Geometries.

exact lower bound on landscape of geometries.
form graph representing topological transitions.

- Universality in a Big Network of String Geometries.

enormous gauge sectors, no weak coupling limits. (algorithmic universality)
rate of E6 appearance. (supervised machine learning)

Big Data and the String Landscape

What do we mean by big?

- **In data science:** many different usages.

  (bigger than previous data,
  or can just barely read into memory and / or process with current tech,
  or too big for current memory / processor and motivating new tech.)

- Another usage: so **big** that no conceivable computer will ever be able to directly store or process the set.

- **A worry?** You can't do anything with a big set?

- Computer scientists, e.g., make progress with big sets.

  Machine-learning / AI *in some domains* achieve superhuman progress.

AlphaGo Zero

"Mastering the game of Go without human knowledge."

Silver et al. (Google DeepMind), Nature Oct. 2017.

A long-standing goal of artificial intelligence is an algorithm that learns, tabula rasa, superhuman proficiency in challenging domains. Recently, AlphaGo became the first program to defeat a world champion in the game of Go. The tree search in AlphaGo evaluated positions and selected moves using deep neural networks. These neural networks were trained by supervised learning from human expert moves, and by reinforcement learning from self-play. **Here we introduce an algorithm based solely on reinforcement learning, without human data, guidance or domain knowledge beyond game rules.** AlphaGo becomes its own teacher: a neural network is trained to predict AlphaGo's own move selections and also the winner of AlphaGo's games. This neural network improves the strength of the tree search, resulting in higher quality move selection and stronger self-play in the next iteration. **Starting tabula rasa, our new program AlphaGo Zero achieved superhuman performance, winning 100–0 against the previously published, champion-defeating AlphaGo.**

Point: Go has 10^{172} states, therefore big, and for the task of playing excellently, superhuman progress achieved tabula rasa.

Is the String Landscape big?

- **Previous big landscape:**

  IIB flux vacua. Fix geometry, turn on fluxes.

  Flux estimates: $O(10^{500})$ Ashok, Denef, Douglas . . . $O(10^{272,000})$ Taylor, Wang

- **Emerging (?) big landscape:**

  Of topologically distinct geometries.

  Geometries: $4/3 \times 2.96 \times 10^{755}$ JH, Long, Sung $O(10^{3000})$ Taylor, Wang

- Logistical memory realities: $\text{Memory required for string geometries} \geq 10^{745}\ \text{GB}$

  $\text{Memory required for flux vacua} \geq O(10^{272,000})\ \text{GB}.$

- Logistical processing realities: (streaming algorithms?)

  $\text{Time required for streaming string geometries} \geq 10^{684}\ T_{\text{univ}}$

  $\text{Time required for streaming flux vacua} \geq O(10^{272,000})\ T_{\text{univ}},$

How to handle a big landscape?

- **Algorithmic universality:** *universality derived not from a constructed set, but instead detailed knowledge of a concrete construction algorithm.*

- **Techniques from data science / AI for strings:**

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

  (simple algs, neural nets, “predict”)

  reinforcement learning / genetic algorithms: RL: [JH, Ruehle, Nelson] to appear. Genetic: [Abel, Rizos], [Ruehle]

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

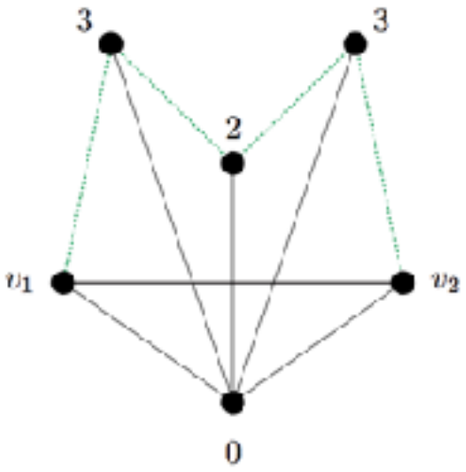
  network science: (“connect”) [Carifio, Cunningham, JH, Krioukov, Long, Nelson] [Taylor, Wang]

  topological data analysis: (“shape” of data) [Cole, Shiu] (for non-gaussianity)

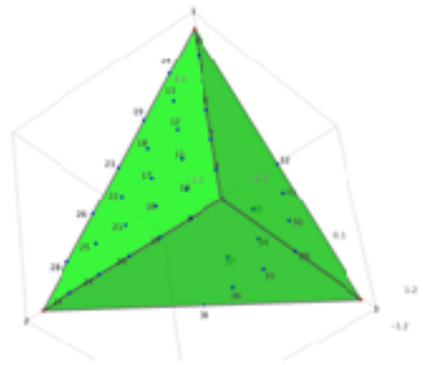
  conjecture generation / intelligible AI: [Carifio,JH, Krioukov,Nelson]

  (use ML to generate conjectures, prove theorems. “rigorify”.)

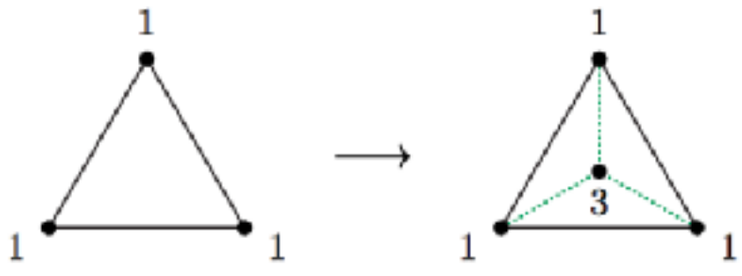
- **Vacuum selection:** maybe once we fully understand string theory, cosmological dynamics will allow us to ignore vast swaths of the landscape. (too hopeful?).

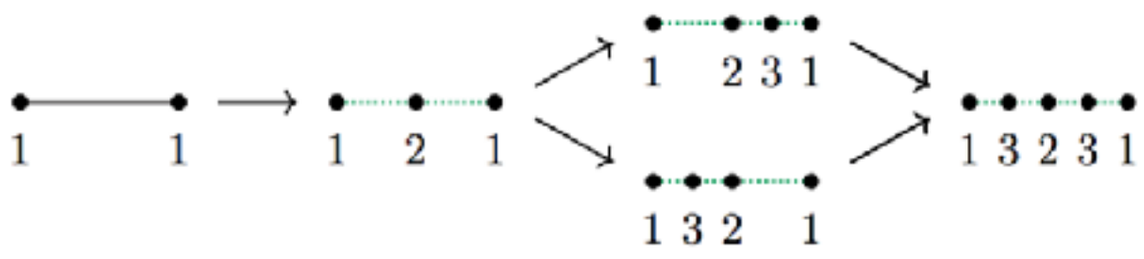


A Big Network of String Geometries



29623152130665884767352987834249686608037802178207672224146242 5
15006529850884562115569564407600452463538913299714609706711172 8
16338803139605678208334761486656697665587902198401279376915721 3
77916981675332533597832585618652359806540480037550724846434238 8
51104723500665582211995572458694879689651884757702603641702684 0
94643379095601672444383055842508595943837154889760954003267052 0
28291799835039979311595823142349439965584910156501341310099487 0
23066141830666224173372736021192320842584597278827051675591361 7
37179257019926973119075478386400898933595790023262079596527939 5
19024066607607335973871447210947571693449780071901582629967340 5
21121413917651351072732370916229999166974674117171609600000000 0
00000000000000000000000000000000000000000000000000000000000000 0





F-theory

- IIb with generalized 7-branes, varying axiodilaton, strong coupling.

- Mathematically described by a Calabi-Yau elliptic fibration over base B, where B is the internal space. Use Weierstrass form:

$$y^2 = x^3 + fx + g$$

- Seven-branes live in the base on the discriminant locus:

$$\Delta = 4f^3 + 27g^2 = 0$$

- The network is a network of bases B and Calabi-Yau elliptic fibrations over them.

Strategy for generating bases

- **Want:** large ensemble of bases, and then to understand its physics, and ideally universal features (if they exist).

- **Strategy:**

  1. Start with some “minimal” base geometry.

  2. Perform topological transitions to other bases.

  3. Satisfy “(4,6)” condition, ensures finite distance movement in CY moduli.

- For simplicity: **toric threefolds.**

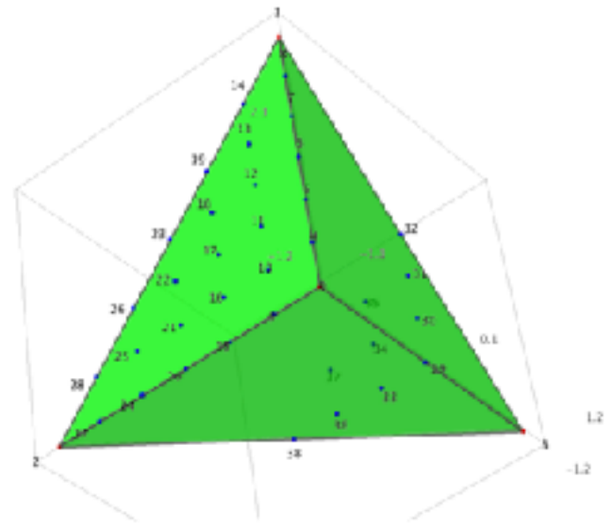
Setting some language.

- We'll refer to a sequence of blowups as a "tree", exceptional ray in fan from blowup as a "leaf"
- Trees over edges = "edge trees"
- Trees over faces = "face trees"
- Points on polytope = "roots"
- Need to classify all trees with h ≤ 6 for all leaves.
- Do so by exhaustively constructing the toric blowups.

The Tree Ensemble

JH, Long, Sung

1. **Minimal geometry:** weak Fano toric 3-fold base, corresponding to a triangulated 3d reflexive polytope. Defines a **Fan** with rays v_i.



2. **Blowup:** along subvarieties to reach a new toric base. Combinatorially described by adding new ray v_e to the fan, corresponding to a new exceptional divisor D_e.

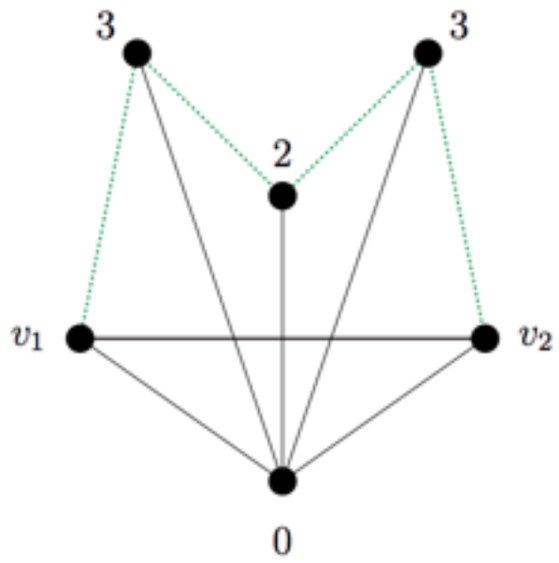
$$v_e = \sum_i a_i v_i$$

The Tree Ensemble

$$v_e = \sum_i a_i v_i$$

- Define the height of a blowup as $h = \sum_i a_i$
- In general, can blow up along

1. Toric curves <—> edges in the triangulated polytope.

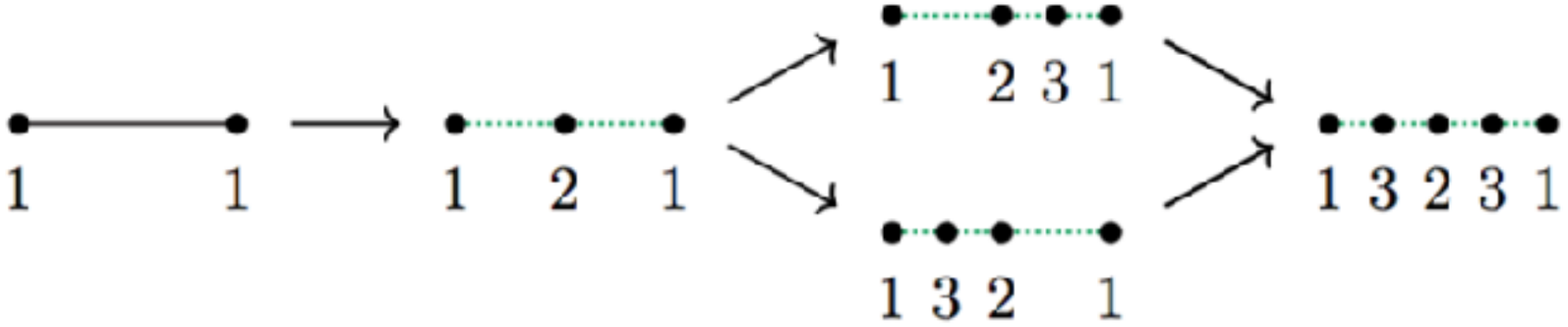


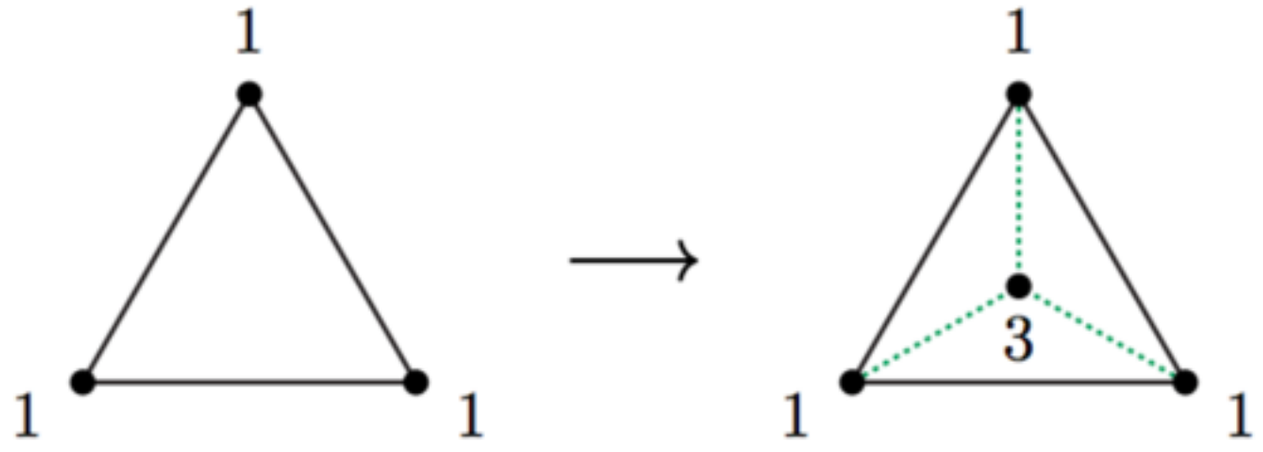
Growing a tree above the edge!
Disclaimer: not a graph theory tree.

2. Toric points <—> faces (triangles) in the triangulated polytope.

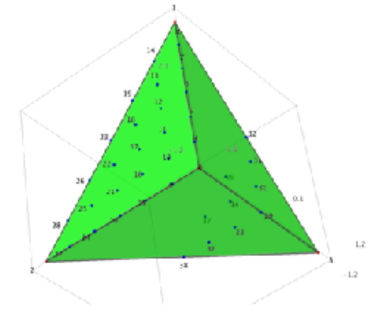
The Tree Ensemble

- To visualize, it's easier to project all rays back onto the polytope, so 'growing a tree' corresponds to subdividing edges and faces.





The Tree Ensemble



- Calabi-Yau elliptic fibrations over these bases form a connected moduli space, related by topological transitions, under certain technical conditions. A necessary one is

$$MOV_{D_e}(g) < 6 \text{ or } MOV_{D_e}(f) < 4$$ Hayakawa, Wang

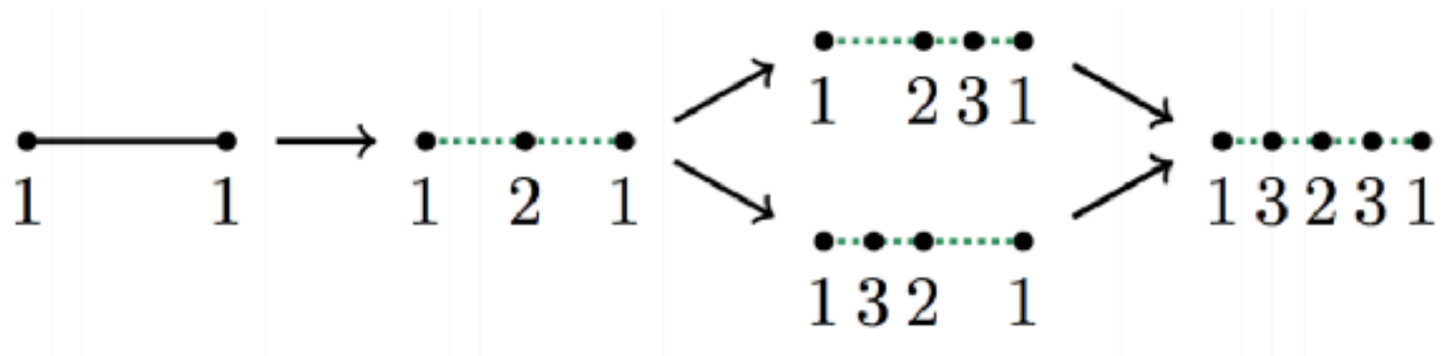
- A **sufficient condition** to ensure that each Calabi-Yau is connected in moduli space limits the possible blowups in a given local patch to a finite set, rendering the ensemble finite.

$$MOV_{D_e}(g) < 6 \Leftarrow h(v_e) \leq 6 \text{ for all } v_e$$ JH, Long, Sung

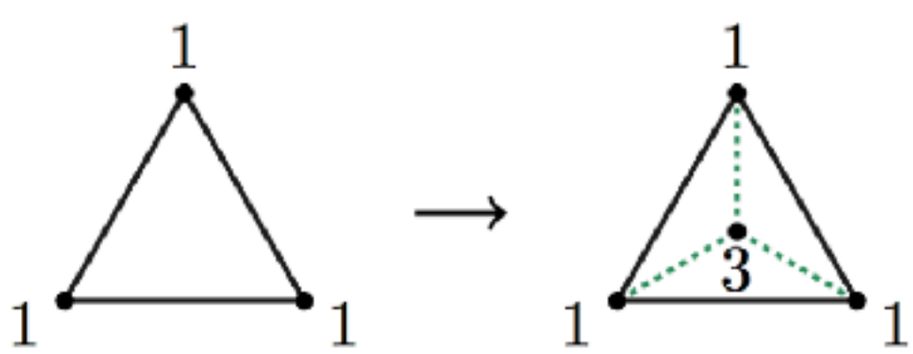
- The topological transitions give this ensemble a network structure: **geometries are nodes**, and **topological transitions are edges**.

Classification of Trees

- All 5 $h \leq 3$ edge trees.



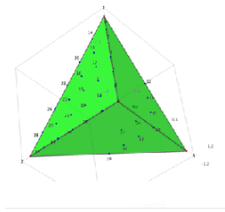
- Both $h \leq 3$ face trees.



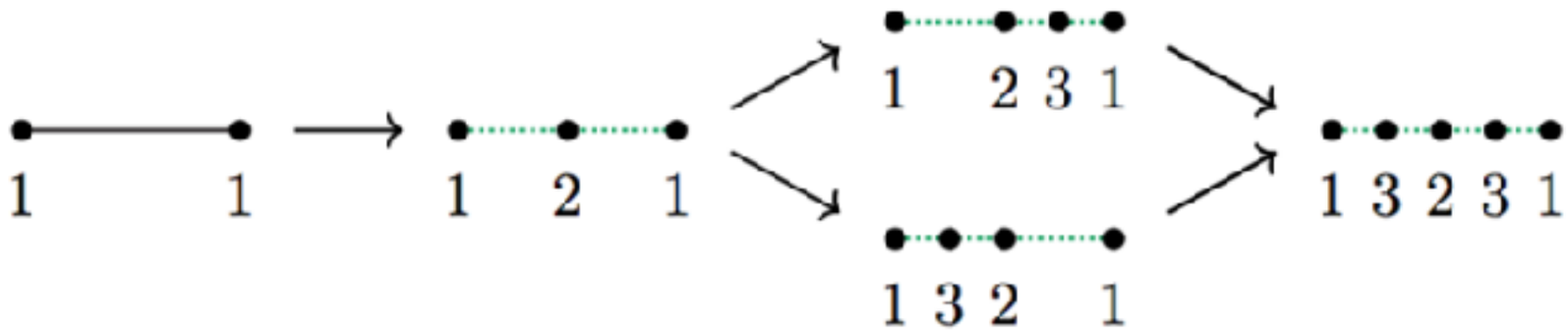
- # for $h \leq N$:

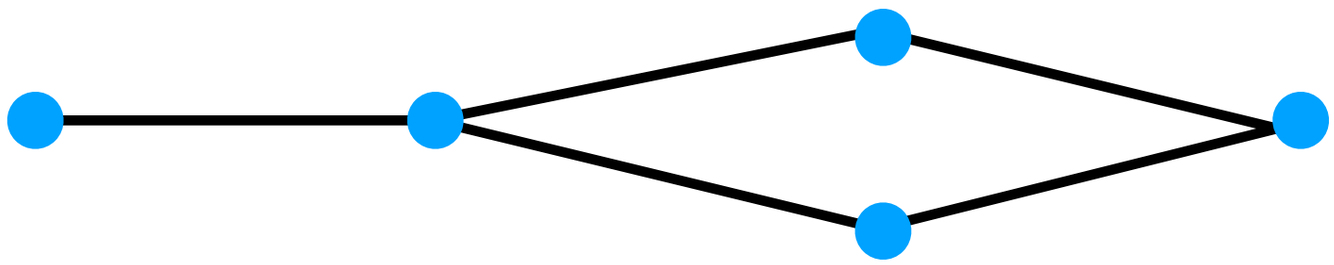
N	# Edge Trees	# Face Trees
3	5	2
4	10	17
5	50	4231
6	82	41,873,645

The Edge Network N_E



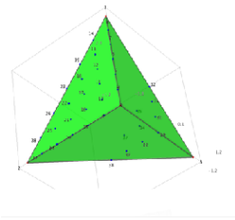
- First consider blowup of curves. Toric curves correspond to edges in the triangulation.



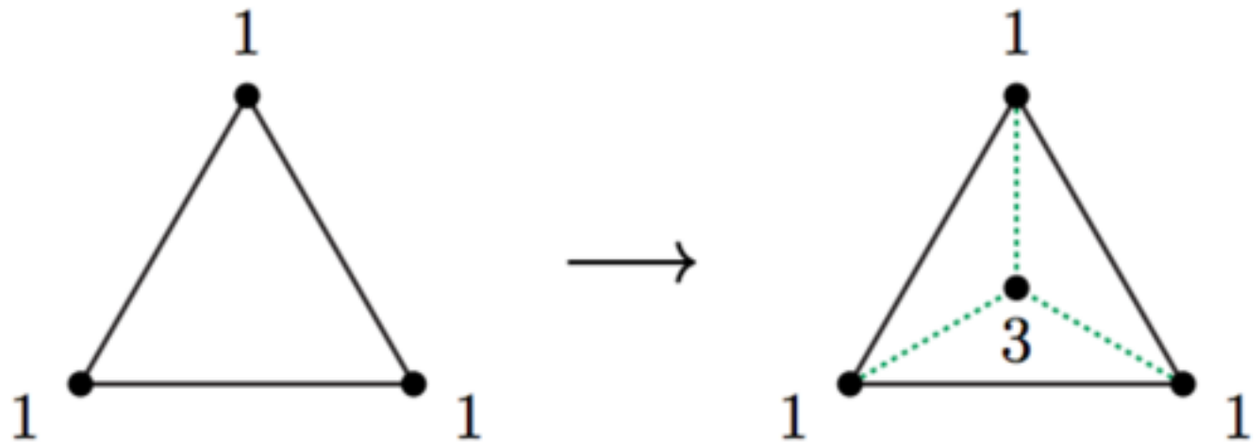


- A single toric curve, corresponding to an edge in the triangulation, admit 82 configurations of blowups.

  These configurations form a network N_E ,
  with **82 nodes** and **1386 edges**.

The Face Network N_F 

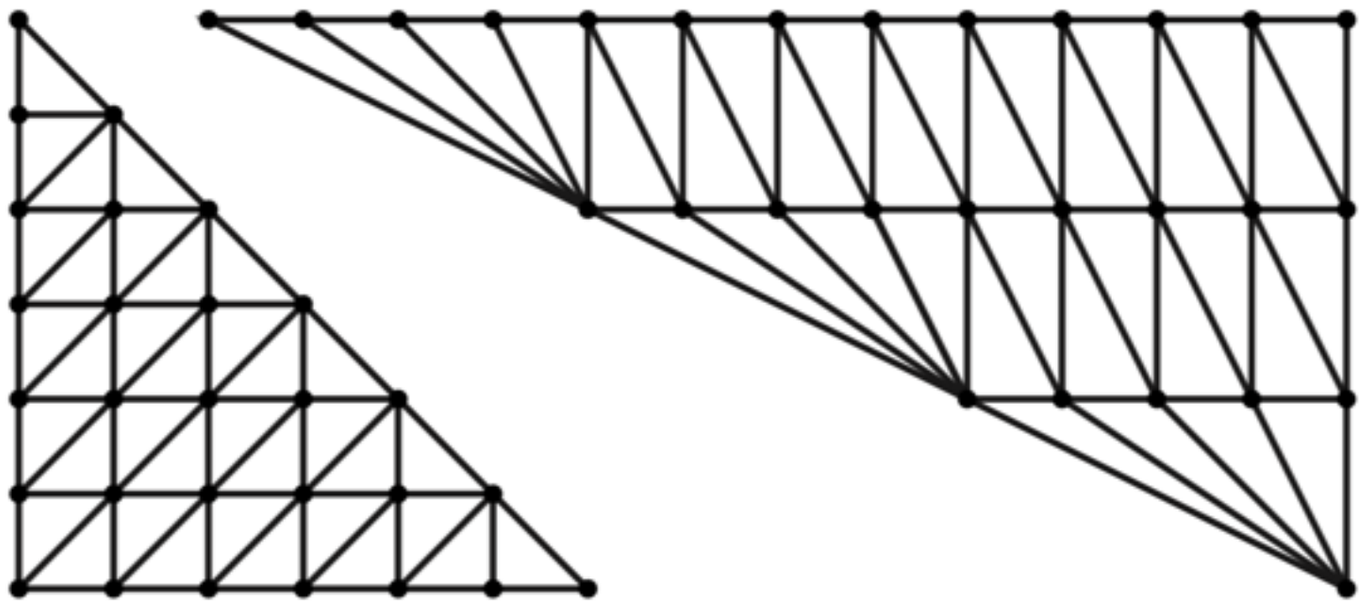
A toric point corresponds to a triangle in the triangulation.



These configurations form a network N_F with **41,873,645 nodes** and **100,036,155 edges**.

Forests from Trees

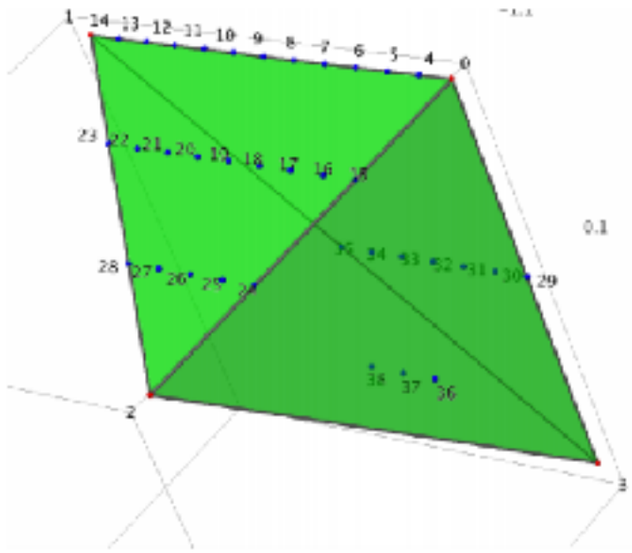
- Each “tree” is data representing a local sequence of blowups.
- **Form “forest”** (threefold base B) from trees by systematically adding trees to FRST of a 3d reflexive polytope. Face trees first, then edge.
- **Count:** polytopes whose FRST’s have the largest number of faces and edges dominate the ensemble.
- **Two polytopes dominate:** have 108 edges and 72 faces, very large facet.

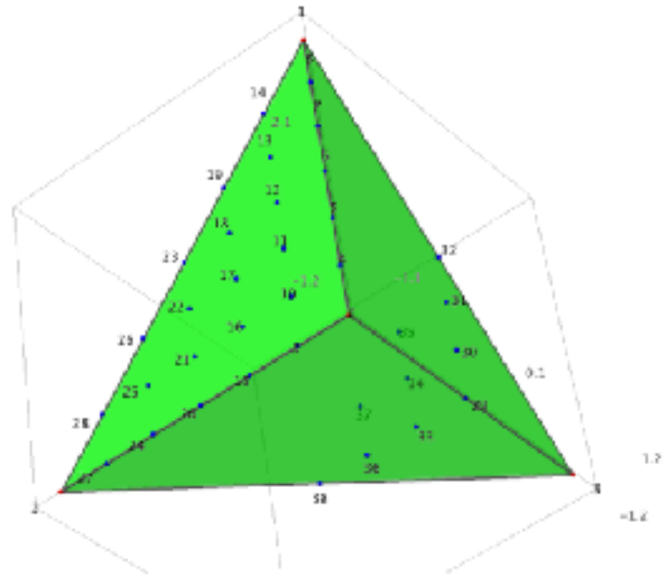


large face of Δ_1° **large face of** Δ_2°

The Tree Network

The dominant polytopes:





Each has 108 toric curves (edges) and 72 toric points (triangles) when triangulated. The number of bases in these ensembles is:

$$|S_{\Delta_1^\circ}| = \frac{2.96}{3} \times 10^{755} \qquad |S_{\Delta_2^\circ}| = 2.96 \times 10^{755}$$

Studied network properties: [Carifio, Cunningham, JH, Krioukov, Long, Nelson]

Recapping the language.

- We'll refer to a sequence of blowups as a "tree", exceptional ray in fan from blowup as a "leaf"
- Trees over edges = "edge trees"
- Trees over faces = "face trees"
- Points on polytope = "roots"
- Need to classify all trees with h ≤ 6 for all leaves.
- Do so by exhaustively constructing the toric blowups.

Universality in a Big Network of String Geometries

Universal: (technique: algorithmic universality)

- Non-Higgsable seven-branes.
- Enormous gauge sectors.
- Strong coupling.

Semi-common: (technique: supervised machine learning)

- E6 on distinguished divisor.

(see upcoming universal results with supervised ML).

Non-Higgsable 7-branes

Some selective progress: Halverson, Grassi, Morrison, Shaneson, Taylor, Wang

F	a	b	c	Sing.	G
I_0	≥ 0	≥ 0	0	none	none
I_n	0	0	$n \geq 2$	A_{n-1}	$SU(n)$ or $Sp(\lfloor n/2 \rfloor)$
II	≥ 1	1	2	none	none
III	1	≥ 2	3	A_1	$SU(2)$
IV	≥ 2	2	4	A_2	$SU(3)$ or $SU(2)$
I_0^*	≥ 2	≥ 3	6	D_4	$SO(8)$ or $SO(7)$ or G_2
I_n^*	2	3	$n \geq 7$	D_{n-2}	$SO(2n-4)$ or $SO(2n-5)$
IV^*	≥ 3	4	8	E_6	E_6 or F_4
III^*	3	≥ 5	9	E_7	E_7
II^*	≥ 4	5	10	E_8	E_8

$$f = \tilde{f} \prod_i x_i^{a_i}$$

$$g = \tilde{g} \prod_i x_i^{b_i}$$

$$\Delta = \tilde{\Delta} \prod_i x_i^{c_i}$$

$$c_i = min(3a_i, 2b_i)$$

- **Non-Higgsable seven-brane:** $c_i > 0$ for some i. (NH7)

$$G \in \{E_8, E_7, E_6, F_4, SO(8), SO(7), G_2, SU(3), SU(2)\}$$

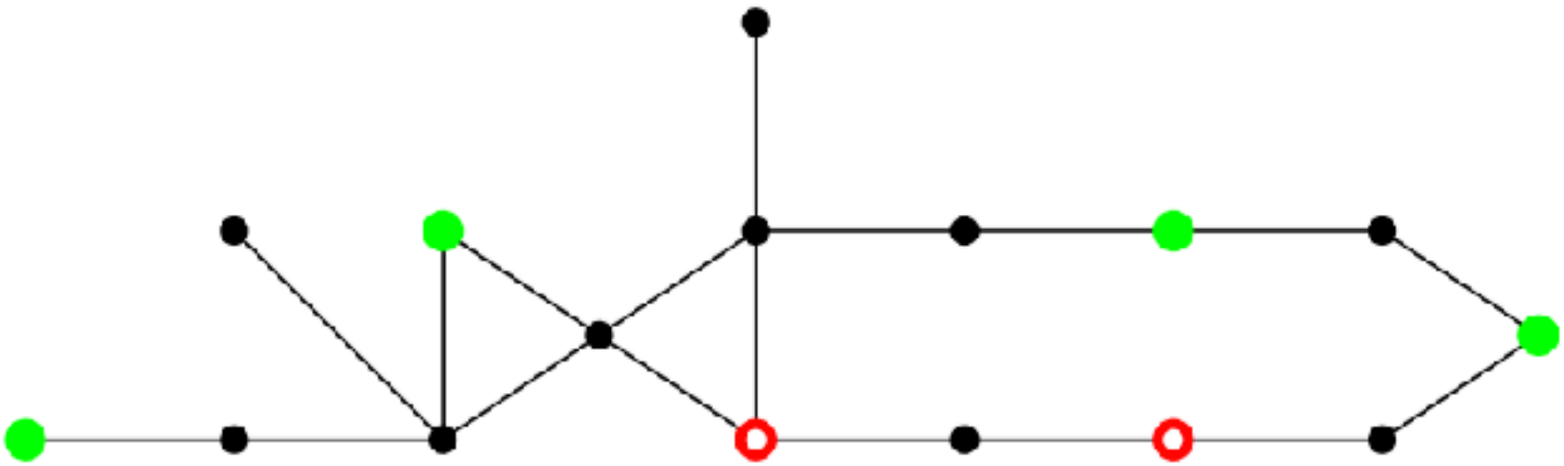
- **Cannot be Higgsed** by a complex structure deformation!

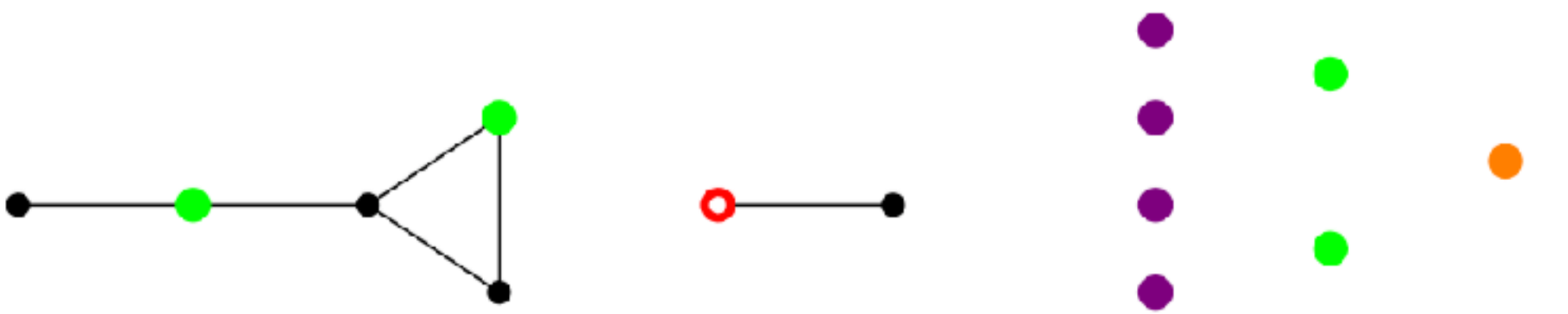
- **Non-Higgsable clusters:** network of intersecting NH7. (NHC).

- **Entirely determined by topology of B!**

A Typical NHC

(a beautiful picture from Taylor-Wang)

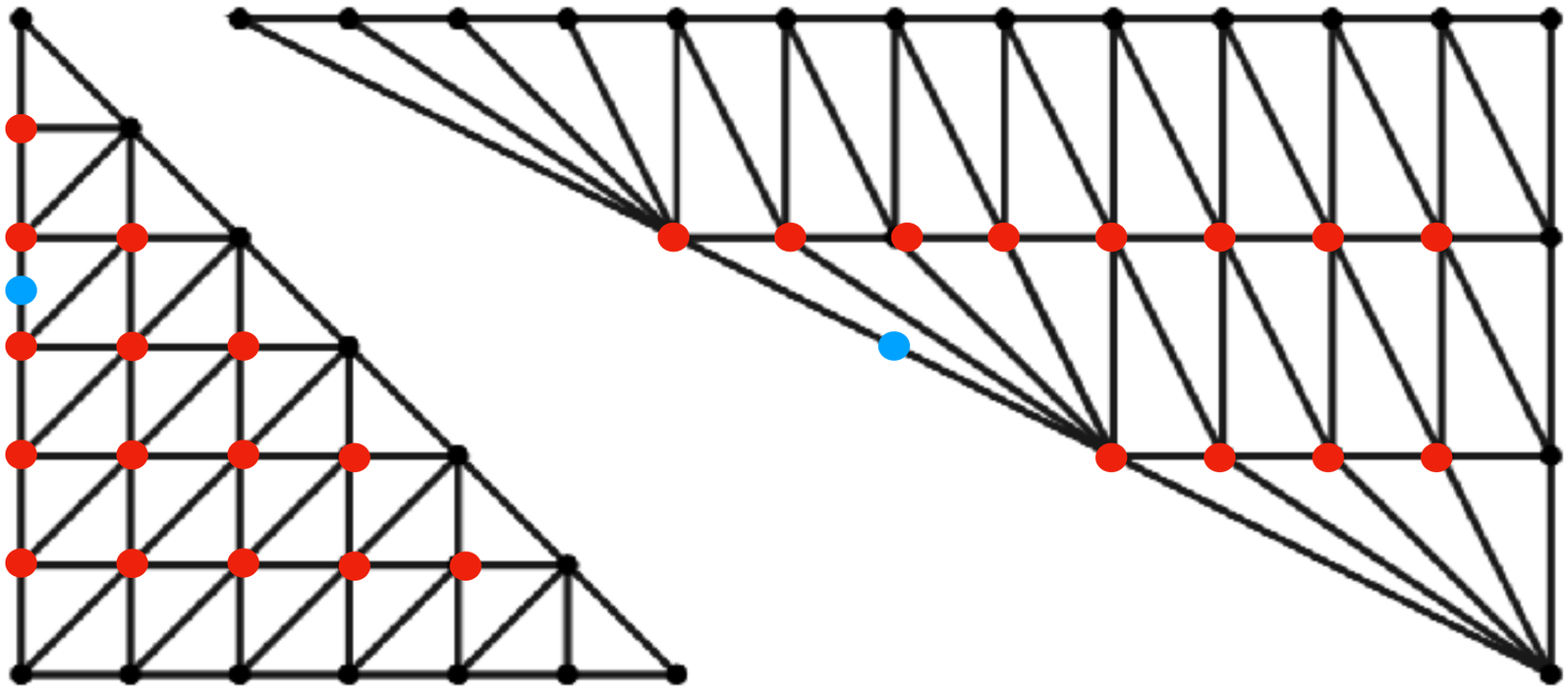




Universality of NH7 JH, Long, Sung

- Consider an edge or facet of a polytope, and perform a height > 2 blowup on that edge or facet.

- This cuts out a special monomial in **f**, **g**, forces type II NH7 on all divisors correspondir



blow up

type II NH7

All face trees (except for one on ground) have a h > 2 leaf.

All but two edge trees have a h > 2 leaf.

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

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

Universality of Large Gauge Sectors

JH, Long, Sung

- E_8 on roots (divisors on facet) are extremely common.

- **Theorem:** A leaf built on E_8 roots with height h = 1,2,3,4,5,6 has Kodaira fiber $F = II^*, IV^*_{ns}, I^*_{0ns}, IV_{ns}, II, -$ and geometric gauge group $E_8, G_2, SU(2), -, -$ respectively.

- Let H_i be number of height i leaves above E_8 roots. Then:

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

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

with probability $\geq .999995$

- Theorem confirmed by “machine learning”.

(fancy technique: linear regression. one of many sklearn defaults.)

Universality of Strong Coupling

(Sen limit almost never exists). **JH, Long, Sung**

- **Sen's limit:** weakly coupled limit in CS moduli space.

- Require having only I_n or I_n^* fibers.

- **Does not exist** if you have NH7 on rigid divisors with too large $MOV_D(f,g)$ (higher than I_0^*).

  (i.e. no seven-branes with exceptional G at weak coupling.)

  Q: how often do you have at least one such NH7?

Universality of Strong Coupling

(Sen limit almost never exists). JH, Long, Sung

A Sen limit is spoiled by:

- A height-2 blowup along two 1-simplices on the same edge, coupled with a height-3 blowup along that same edge.
- A height-3 blowup along a 2-simplex strictly interior to a face.

- Fraction of geometries that admit a Sen limit: $< 3 \times 10^{-391}$

- These geometries are inherently strongly-coupled F-theory geometries that do not admit a weakly coupled string theory description. There is no such limit in moduli space, for fixed base.

An E_6 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? SU(3)?

  - Simple conditions / probabilities for then not known. **JH, Long, Sung**
  - In random samples, prob ~ 1/1000.
  - When E6 arises in RS, on a distinguished vertex: (1,-1,-1).

- **Machine Learning:** **Carifio, Halverson, Krioukov, Nelson**

  Q: Can we train an ML model to accurately predict yes or no for E6 on (1,-1,-1)?

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

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

Point: by using machine learning to generate conjectures, we may be able to take its numerical / empirical results and turn it into rigorous results.

Training the Model

- **Supervised machine learning:** given a large number of (input,output) pairs, **learn to predict** output given input, and then **test on unseen data**, see how well the model does.

- **Training data:**

  Input: (max height above v, # of such rays) for all v in polytope.
  Output: E6 on (1,-1,-1) or not.

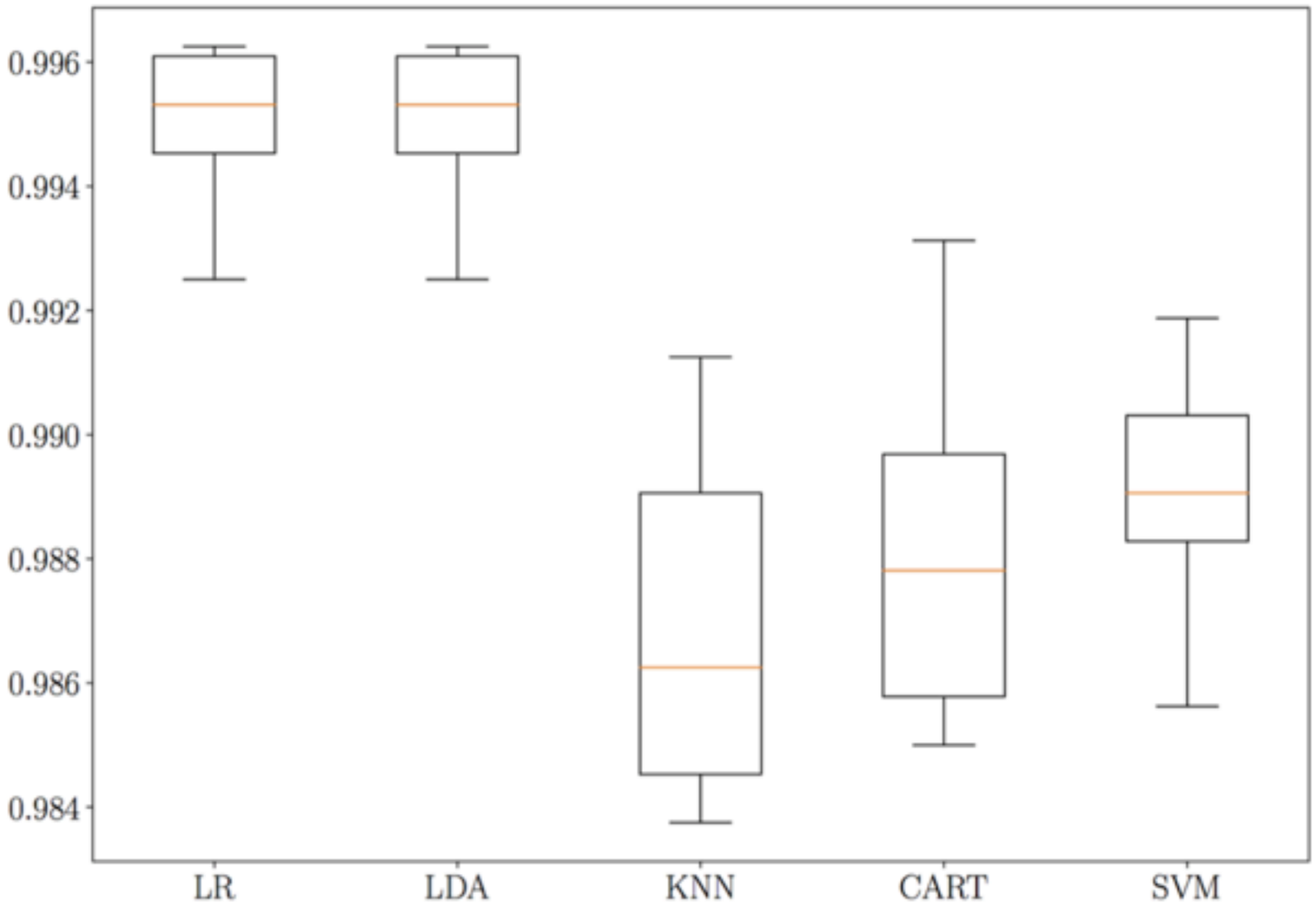
$$S_{a,v_1} := \{v \in V | v = av_1 + bv_2 + cv_3, \;\; a,b,c \geq 0\}$$

$$(a_{max}, |S_{a_{max},v}|) \quad \forall v \in \Delta_1^\circ \qquad \xrightarrow{A} \qquad E_6 \text{ on } v_{E_6} \text{ or not}$$

- **sklearn:** a very nice free Python package.

- **Training sample:** 10000 random with no E6, 10000 random with E6.

Evaluating the Model on Unseen Data



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

- **Displayed:** whisker plots of % accuracy with 10-fold cross validation.
- **Gold bar:** mean % accuracy.
- **Factor analysis:** only two of the variables really matter:

$$(a_{max}, |S_{a_{max}, v_{E6}}|)$$

Conjecture Generation

- **Organizing principle?** See what it gets right and wrong! (using the model trained with logistic regression.)

- **Observation:**
  amax = 5: always no
  amax = 4: usually no.

- **Initial Conjecture:**

Conjecture: If $a_{max} = 5$ for v_{E6}, then v_{E6} does not carry E_6. If $a_{max} = 4$ for v_{E6} it may or may not carry E_6, though it is more likely that it does.

a_{max}	$\lvert S_{a_{max},v_{E6}} \rvert$	Pred. for E_6 on v_{E_6}	Hyperplane Distance
4	5	No	0.88
4	6	No	0.29
4	7	Yes	−0.31
4	8	Yes	−0.90
4	9	Yes	−1.50
4	10	Yes	−2.09
4	11	Yes	−2.69
4	12	Yes	−3.28
4	13	Yes	−3.88
4	14	Yes	−4.47
4	15	Yes	−5.07
4	16	Yes	−5.67
4	17	Yes	−6.26
4	18	Yes	−6.85
4	19	Yes	−7.45
4	20	Yes	−8.04
4	21	Yes	−8.64
4	22	Yes	−9.23
4	23	Yes	−9.83
4	24	Yes	−10.42
5	1	No	7.34
5	2	No	6.75
5	3	No	6.15
5	4	No	5.56
5	5	No	4.96
5	6	No	4.37
5	7	No	3.78
5	8	No	3.18
5	9	No	2.59
5	10	No	1.99
5	11	No	1.40
5	12	No	0.80

Conjecture Refinement and Theorem

- **Use info from ML, think a bit, write down conjecture.**

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.

- **Key point:** ML-inspired focus on one particular variable, led quickly (< 24 hours) to a theorem once identified.

  “Back and forth” process, could be of broad applicability.

Probability and Checks

- **Probability computation:**

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

computed using # appropriate edge trees relative theorem.

Result:

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

- **Check:** with 5 batches, 2 million random samples each.

$$\begin{aligned} \text{From Theorem} \quad &: \quad .00059128 \times 2 \times 10^6 = 1182.56 \\ \text{From Random Samples} \quad &: \quad 1183,\ 1181,\ 1194,\ 1125,\ 1195 \end{aligned}$$

Concluding Thoughts

Better understanding the landscape likely requires
both formal progress and progress on dealing with its size.

It is big, in the sense that no conceivable computer
will ever be able to store or process it.

Given this, in addition to formal progress,
universality from construction algorithm and
techniques from data science seem like
promising directions for understanding it.

I used both to understand a huge ensemble of geometries.
Non-Higgsable seven-branes, large gauge sectors,
and strong coupling are all universal in the ensemble.

I also showed how machine learning can be used to
generate a conjecture and subsequently prove a theorem,
in this case related to the prevalence of E6.

Thank you!

296231521306658847673529878342496866080378021782076722241462425
150065298508845621155695644076004524635389132997146097067111728
163388031396056782083347614866566976655879021984012793769157213
779169816753325335978325856186523598065404800375507248464342388
511047235006655822119955724586948796896518847577026036417026840
946433790956016724443830558425085959438371548897609540032670520
282917998350399793115958231423494399655849101565013413100994870
230661418306662241733727360211923208425845972788270516755913617
371792570199269731190754783864008989335957900232620795965279395
190240666076073359738714472109475716934497800719015826299673405
211214139176513510727323709162299991669746741171716096000000000
000000000000000000000000000000000000000000000000000000000000000

Extra Slides

Base transitions

- Starting with an elliptically fibered Calabi-Yau **X -> B**, one can crepantly pass to another elliptically fibered Calabi-Yau **X" -> B'** by a base-change, and pass to a minimal Weierstrass model.

- This procedure is

  1. Perform a blowup **B' -> B** in the base along a subvariety **C** and perform a base change

$$X^{'} = X \times_B B^{'} \rightarrow B^{'}$$

  2. Perform a change of coordinates and pass to a minimal Weierstrass model **X" -> B'**.

Candelas, Diaconescu, Florea, Morrison, Rajesh

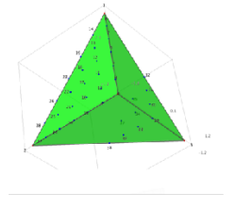
- For this procedure to be crepant we need

$$MOV_C(f, g) \geq (4, 6) \qquad \textbf{if C is a curve in B}$$

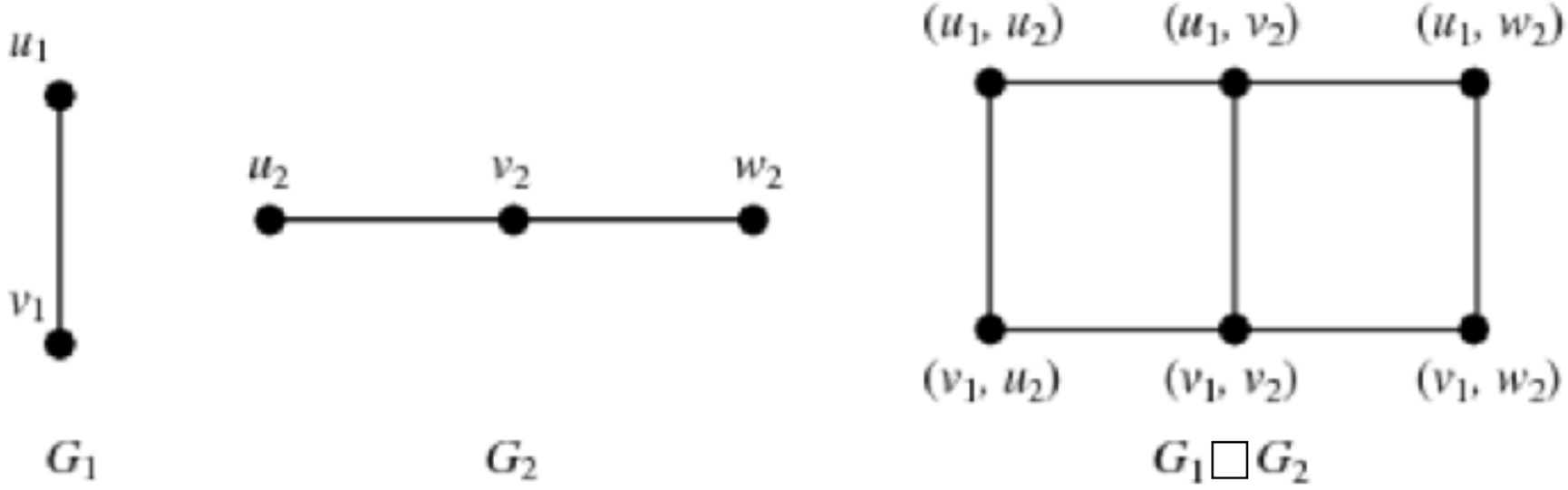
$$MOV_C(f, g) \geq (8, 12) \qquad \textbf{if C is a point in B}$$

- This produces a new elliptic Calabi-Yau **X" -> B'**, with a new base **B'** which is a blowup of **B**.

The Tree Network



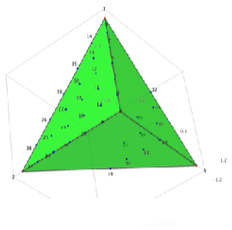
- A generic network with 10^{755} nodes would be completely intractable, but this network factorizes into a cartesian product of graphs:



Wolfram

Cartesian product = $\square$

The Tree Network



- The tree network N_{tree} factorizes as

$$N_{\text{tree}} = N_E^{\square\, 108} \square\, N_F^{\square\, 72}$$

- Simply put, two geometries in the Cartesian product are adjacent if they are related by a single blowup in a single local patch.

By understanding N_E and N_F we can learn about N_{tree} !

Beyond toric bases

- Can we generalize beyond toric bases? Observation: the minimal geometries we've considered (WFTV) can be viewed as patching together crepant resolutions of orbifold singularities of $\mathbb{C}^3$ of the form:

  Roan

  1. Isolated singularities. **Degeratu, Yau**

  2. A_n singularities fibered over curves.

- A natural generalization to move beyond toric threefolds is to consider crepant resolutions of other orbifold singularities. What's left are

  1. D_n singularities fibered over a curve.

  Joyce

  2. E_n singularities fibered over a curve.

  Facchini, Gonzalez-Alonso, Lason

- By looking at the Cox ring of the resolutions, we find that building any trees above these geometries forces non-Higgsable clusters on rigid divisors arising in the crepant resolution, and spoils the existence of a Sen's limit. These geometries produce inherently strongly coupled physics as well!

  JH, Long, Sung