Universes as Bigdata:

Strings, Manifolds and Machine-Learning

YANG-HUI HE

London Institute, *Royal Institution of GB*

Dept of Mathematics, City, *University of London*

Merton College, *University of Oxford*

School of Physics, *NanKai University*

Beyond Standard Model: From Theory to Experiment (BSM- 2021)

1984: $10 = 4 + 3 \times 2$

- First String Revolution [Green-Schwarz] anomaly cancellation;

  Heterotic string [Gross-Harvey-Martinec-Rohm]: $E_8 \times E_8$ or $SO(32)$, 1984 - 5
- String Phenomenology [Candelas-Horowitz-Strominger-Witten]: 1985
  - $SU(3) \times SU(2) \times U(1) \subset SU(5) \subset SO(10) \subset E_6 \subset E_8$
  - Standard Solution (MANY more since): $\mathbb{R}^{3,1} \times X$, X Ricci-flat, Kähler
- *mathematicians were independently thinking of the same problem*:
  - Riemann Uniformization Theorem in $\dim_{\mathbb{C}} = 1$: Trichotomy $R < 0, = 0, > 0$
  - Euler, Gauss, Riemann, Bourbaki, Atiyah-Singer ...

$$\chi(\Sigma) = 2 - 2g(\Sigma) = [c_1(\Sigma)] \cdot [\Sigma] = \frac{1}{2\pi} \int_{\Sigma} R = \sum_{i=0}^{2} (-1)^i h^i(\Sigma)$$

Calabi-Yau

1984: $10 = 4 + 3 \times 2$

- First String Revolution [Green-Schwarz] anomaly cancellation;

  Heterotic string [Gross-Harvey-Martinec-Rohm]: $E_8 \times E_8$ or $SO(32)$, 1984 - 5
- String Phenomenology [Candelas-Horowitz-Strominger-Witten]: 1985
  - $SU(3) \times SU(2) \times U(1) \subset SU(5) \subset SO(10) \subset E_6 \subset E_8$
  - Standard Solution (MANY more since): $\mathbb{R}^{3,1} \times X$, X Ricci-flat, Kähler
- *mathematicians were independently thinking of the same problem*:
  - Riemann Uniformization Theorem in $\dim_{\mathbb{C}} = 1$: Trichotomy $R < 0, = 0, > 0$
  - Euler, Gauss, Riemann, Bourbaki, Atiyah-Singer ...

$$\chi(\Sigma) = 2 - 2g(\Sigma) = [c_1(\Sigma)] \cdot [\Sigma] = \frac{1}{2\pi} \int_\Sigma R = \sum_{i=0}^{2} (-1)^i h^i(\Sigma)$$

Calabi-Yau

An Early Physical Challenge to Algebraic Geometry

- CY3 X, tangent bundle $SU(3) \Rightarrow$
  1. E_6 GUT: commutant $E_8 \to SU(3) \times E_6$, then
  2. Wilson-line/discrete symmetry to break E_6-GUT to some SUSY version of Standard Model (generalize later)
  3. Particle Spectrum:

     Generation $\quad n_{27} = h^1(X, TX) = h^{2,1}_{\overline{\partial}}(X)$

     Anti-Generation $\quad n_{\overline{27}} = h^1(X, TX^*) = h^{1,1}_{\overline{\partial}}(X)$
- Net-generation: $\chi = 2(h^{1,1} - h^{2,1}) =$ Euler Number
- 1980s Question: Are there Calabi-Yau threefolds with Euler number ± 6?
- None of obvious ones ☹

  e.g., Quintic Q in $\mathbb{P}^4$ is CY3 $Q^{h^{1,1},h^{2,1}}_{\chi} = Q^{1,101}_{-200}$ so too may generations

  (even with quotient $-200 \notin 3\mathbb{Z}$)

An Early Physical Challenge to Algebraic Geometry

- CY3 X, tangent bundle $SU(3) \Rightarrow$
  1. E_6 GUT: commutant $E_8 \to SU(3) \times E_6$, then
  2. Wilson-line/discrete symmetry to break E_6-GUT to some SUSY version of Standard Model (generalize later)
  3. Particle Spectrum:

     Generation $\quad n_{27} = h^1(X, TX) = h^{2,1}_{\overline{\partial}}(X)$

     Anti-Generation $\quad n_{\overline{27}} = h^1(X, TX^*) = h^{1,1}_{\overline{\partial}}(X)$
- Net-generation: $\chi = 2(h^{1,1} - h^{2,1}) =$ Euler Number
- 1980s Question: Are there Calabi-Yau threefolds with Euler number ± 6?
- None of obvious ones ☹

  e.g., Quintic Q in $\mathbb{P}^4$ is CY3 $Q^{h^{1,1},h^{2,1}}_{\chi} = Q^{1,101}_{-200}$ so too may generations

  (even with quotient $-200 \notin 3\mathbb{Z}$)

An Early Physical Challenge to Algebraic Geometry

- CY3 X, tangent bundle $SU(3) \Rightarrow$
  1. E_6 GUT: commutant $E_8 \to SU(3) \times E_6$, then
  2. Wilson-line/discrete symmetry to break E_6-GUT to some SUSY version of Standard Model (generalize later)
  3. Particle Spectrum:

     Generation $\quad n_{27} = h^1(X, TX) = h^{2,1}_{\overline{\partial}}(X)$

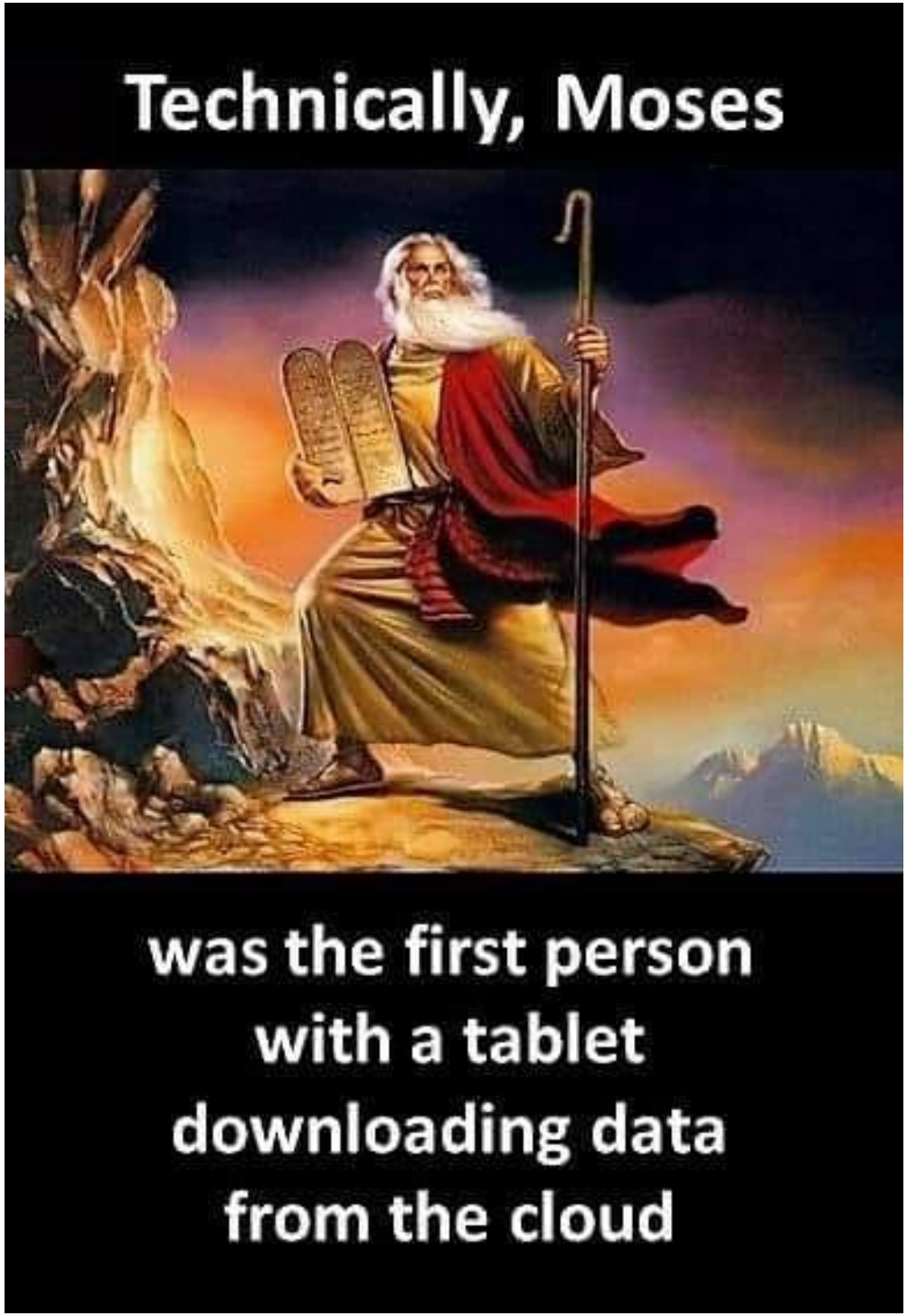
     Anti-Generation $\quad n_{\overline{27}} = h^1(X, TX^*) = h^{1,1}_{\overline{\partial}}(X)$
- Net-generation: $\chi = 2(h^{1,1} - h^{2,1}) =$ Euler Number
- 1980s Question: Are there Calabi-Yau threefolds with Euler number ± 6?
- None of obvious ones ☹

  e.g., Quintic Q in $\mathbb{P}^4$ is CY3 $Q^{h^{1,1},h^{2,1}}_{\chi} = Q^{1,101}_{-200}$ so too may generations

  (even with quotient $-200 \notin 3\mathbb{Z}$)

The First Data-sets in Mathematical Physics/Geometry

- [Candelas-A. He-Hübsch-Lutken-Schimmrigk-Berglund] (1986-1990)
  - CICYs (complete intersection CYs) multi-deg polys in products of $\mathbb{CP}^{n_i}$ CICYs
  - Problem: *classify all configuration matrices*; employed the best computers at the time (**CERN supercomputer**); **q.v. magnetic tape and dot-matrix printout in Philip's office**
  - 7890 matrices, 266 Hodge pairs $(h^{1,1}, h^{2,1})$, 70 Euler $\chi \in [-200, 0]$
- [Candelas-Lynker-Schimmrigk, 1990]
  - Hypersurfaces in Weighted P4
  - 7555 inequivalent 5-vectors w_i, 2780 Hodge pairs, $\chi \in [-960, 960]$
- [Kreuzer-Skarke, mid-1990s - 2000]
  - Hypersurfaces in (Reflexive, Gorenstein Fano) Toric 4-folds
  - 6-month running time on dual Pentium SGI machine
  - at least 473,800,776, with 30,108 distinct Hodge pairs, $\chi \in [-960, 960]$



The age of data science in mathematical physics/string theory not as recent as you might think

of course, experimental physics had been decades ahead in data-science/machine-learning

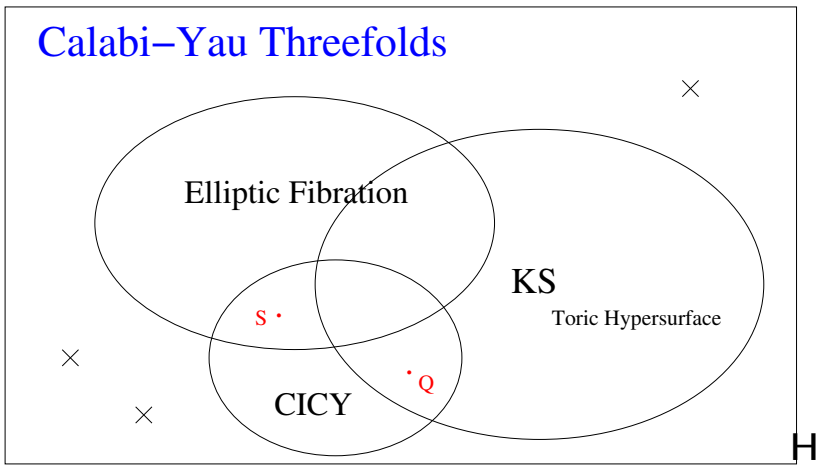
After 40 years of research by mathematicians and physicists

.....

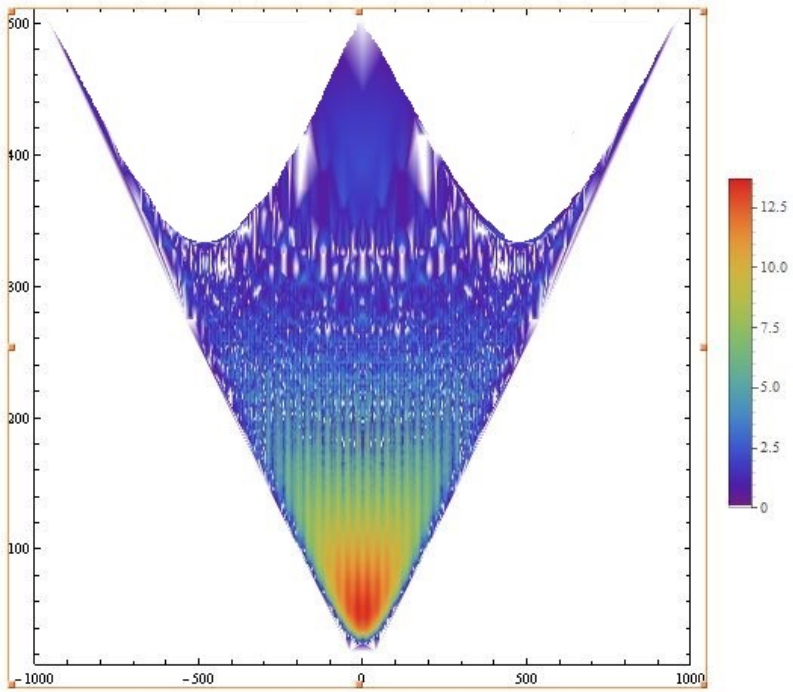
The Compact CY3 Landscape

cf. YHH, *The Calabi-Yau Landscape: from Geometry, to Physics, to Machine-Learning,* 1812.02893, [Springer, to appear]

Vienna (KS, Knapp,...), Penn (Ovrut, Cvetic, Donagi, Pantev ...), Oxford/London (Candelas, Constantin, Lukas, Mishra, YHH, ...), MIT (Taylor, Johnson, Wang, ...), Northeastern/Wits (Halverson, Long, Nelson, Jejjala, YHH), Virginia Tech (Anderson, Gray, SJ Lee, ...), Utrecht (Grimm ...), CERN (Weigand, ...), Cornell (MacAllister, Stillman), Munich (Luest, Vaudravange), Uppsala (Larfors, Seong) ...



Georgia O'Keefe on Kreuzer-Skarke



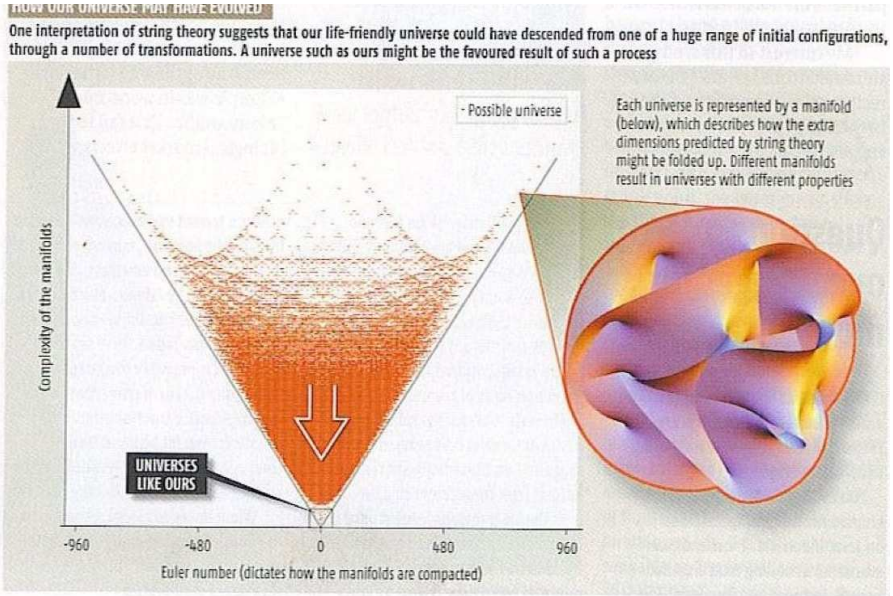
Horizontal $\chi = 2(h^{1,1} - h^{2,1})$ vs. Vertical $h^{1,1} + h^{2,1}$

Triadophilia

Exact (MS)SM Particle Content from String Compactification

- [Braun-YHH-Ovrut-Pantev, Bouchard-Cvetic-Donagi 2005] first exact MSSM
- [Anderson-Gray-YHH-Lukas, 2007-] use alg./comp. algebraic geo & sift
- Anderson-Gray-Lukas-Ovrut-Palti ~ 200 in 10^{10} MSSM Stable Sum of Line Bundles over CICYs (Oxford-Penn-Virginia 2012-)

  Constantin-YHH-Lukas '19: 10^{23} exact MSSMs (by extrapolation on above set)?



A Special Corner

[New Scientist, Jan, 5, 2008 feature]

P. Candelas, X. de la Ossa, YHH, and B. Szendroi

"Triadophilia: A Special Corner of the Landscape" ATMP, 2008

The Landscape Explosion & Vacuum Degeneracy Problem

meanwhile . . . LANDSCAPE grew rapidly where Each Geometry is a Universe

- Orbifold Models/CFT: cf. Nilles, Faraggi, et al. at this conference
- D-branes Polchinski 1995
- M-Theory/G_2 Witten, 1995
- F-Theory/4-folds Katz-Morrison-Vafa, 1996
- AdS/CFT Maldacena 1998
- Flux-compactification Kachru-Kallosh-Linde-Trivedi, 2003, Denef-Douglas 2005-6: $10^{\gg 500}$ possibilities . . .

String theory trades one hard-problem [quantization of gravity] by another [looking for the right compactification] (in many ways a richer and more interesting problem, especially for the string/maths community)

Where we stand . . .

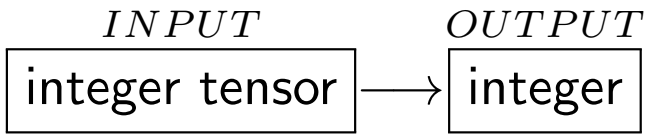
The Good Last 10-15 years: large collaborations of physicists, computational mathematicians (cf. SageMATH, GAP, Bertini, MAGMA, Macaulay2, Singular) have bitten the bullet computed many geometrical/physical quantities and compiled them into various databases **Landscape Data** ($10^{9\sim10}$ entries typically)

The Bad Generic computation HARD: dual cone algorithm (exponential), triangulation (exponential), Gröbner basis (double-exponential) . . . e.g., how to construct stable bundles over the $\gg 473$ million KS CY3? Sifting through for SM computationally impossible . . .

The ??? **Borrow new techniques from "Big Data" revolution**

A Wild Question

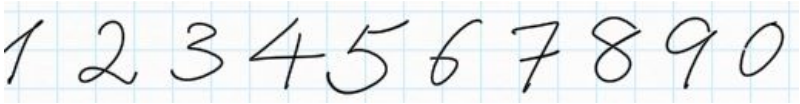
- Typical Problem in String Theory/Algebraic Geometry:



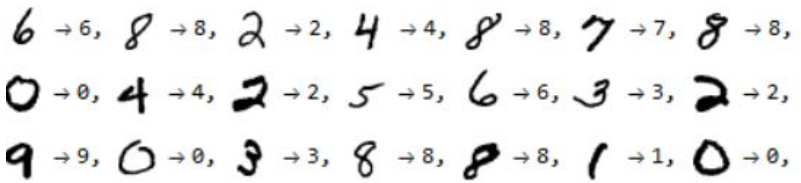
- Q: Can (classes of problems in computational) Algebraic Geometry be "learned" by AI ? , i.e., can we "machine-learn the landscape?"
- [YHH 1706.02714] Deep-Learning the Landscape, *PLB 774, 2017* (*Science*, Aug, vol 365 issue 6452, 2019): Experimentally, it seems so for many situations in geometry and beyond.
- Q: Can we ML (supervised + unsupervised) Mathematical Structure?

A Prototypical Question

- Hand-writing Recognition, e.g., my 0 to 9 is different from yours:



- How to set up a bijection that takes these to $\{1, 2, \ldots, 9, 0\}$? Find a clever Morse function? Compute persistent homology? Find topological invariants? ALL are inefficient and too sensitive to variation.
- What does your iPhone/tablet do? What does Google do? Machine-Learn
  - Take large sample, take a few hundred thousand (e.g. NIST database)



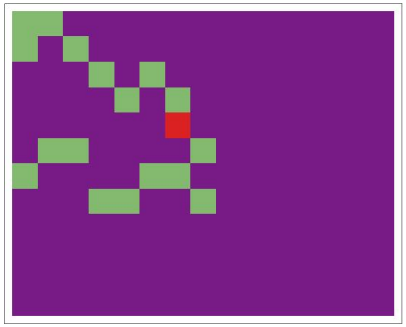
$\ldots$

$28 \times 28 \times (RGB)$

NN Doesn't Care/Know about Alg. Geo (YHH 1706.02714)

- Hodge Number of a Complete Intersection CY is the association rule, e.g.

$$X = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 \end{pmatrix}, \qquad h^{1,1}(X) = 8 \quad \rightsquigarrow$$



$\longrightarrow 8$

CICY is 12×15 integer matrix with entries $\in [0, 5]$ is simply represented as a 12×15 pixel image of 6 colours Proper Way ; ML in matter of seconds/minutes

- Cross-Validation:
  - Take samples of $X \to h^{1,1}$
  - train a NN, or SVM
  - Validation on *unseen* $X \to h^{1,1}$

Progress in String Theory

Major International Annual Conference Series

1986- First "Strings" Conference

2002- First "StringPheno" Conference

2006 - 2010 String Vacuum Project (NSF)

2011- First "String-Math" Conference

2014- First String/Theoretical Physics Session in SIAM Conference

2017- First "String-Data" Conference

- YHH (1706.02714), Seong-Krefl (1706.03346), Ruehle (1706.07024), Carifio-Halverson-Krioukov-Nelson (1707.00655)
- ML+string/alg-geo is now a community (~ 100 papers since): q.v. review YHH 2011.14442 "Universes as Big Data"
- ML various structures/data in pure mathematics, q.v. review YHH 2101.06317 "ML Mathematical Structures"

Summary and Outlook

PHYSICS

- Use AI (Neural Networks, SVMs, Regressor . . .) as
  1. Classifier deep-learn and categorize landscape data
  2. Predictor estimate results beyond computational power

MATHS

- how is AI doing maths w/o knowing any maths? (Alg Geo/$\mathbb{C}$, combinatorics, RT = integer matrices, NT ??)
  1. Predictor form new conjectures/formulae
  2. Classifier stochastically do NP-hard problems
- Hierarchy of Difficulty ML struggles with:

  numerical $<$ **algebraic geometry over** $\mathbb{C}$ $<$ **combinatorics/algebra** $<$ **number theory**

Please Submit papers to . . .

- Y.-H. He, P.-P. Dechant, A. Lukas, A. Kasprzyk, Ed. **"Machine-Learning Mathematical Structures"**, Topical Collection in Adv. Clifford Alg. & Applications, Springer-Birkhäuser, `https://www.springer.com/journal/6/updates/18581430`
- J. Hauenstein, Y.-H. He, I. Kotsireas D. Mehta, T. Tang, Ed. **"Algebraic Geometry and Machine Learning"**, Special Issue in J. Symbolic Computation, Elsevier, `https://www.journals.elsevier.com/journal-of-symbolic-computation/call-for-papers/algebraic-geometry-and-machine-learning`
- D. X. Gu, Y. Wang, S.-T. Yau, Y.-H. He, M. Douglas, et al. Ed. Mathematics, Computation and Geometry of Data, International Press `https://www.intlpress.com/site/pub/pages/journals/items/mcgd/_home/_main/`
- Y.-H. He, et al. Ed. Mathematical Data, World Scientific, To Appear 2021.

Thank you!

Syntax		Semantics
Alpha Go	$\to$	Alpha Zero
ML Patterns	$\to$	Auto Thm Pf&Chk
- Renner et al., PRL/Nature News, 2019: ML (*SciNet*, *autoencoder*)
- Lample-Charton, 2019: ML Symolic manipulations in mathematics
- Tegmark et al., 2019 AI Feynman, symb regressor
- Raayoni et al. 2020 Ramanujan-Machine
- Barbaresco-Nielson 2021 Infor Geom/ML

Sophia (Hanson Robotics, HK)

1st non-human citizen (2017, Saudi)

1st non-human with UN title (2017)

1st String Data Conference (2017)

$\chi(\Sigma)=2$	$\chi(\Sigma)=0$	$\chi(\Sigma)<0$
Spherical	Ricci-Flat	Hyperbolic
$+$ curvature	0 curvature	$-$ curvature
Fano	Calabi-Yau	General Type

- Euler, Gauss, Riemann, Bourbaki, Atiyah-Singer $\ldots \rightsquigarrow$ generalize

$$\chi(\Sigma) = 2 - 2g(\Sigma) = [c_1(\Sigma)] \cdot [\Sigma] = \frac{1}{2\pi}\int_\Sigma R = \sum_{i=0}^{2}(-1)^i h^i(\Sigma)$$

- CONJECTURE [E. Calabi, 1954, 1957] / Thm [ST. Yau, 1977-8] M compact Kähler manifold (g,ω) and $([R]=[c_1(M)])_{H^{1,1}(M)}$. Then $\exists!(\tilde{g},\tilde{\omega})$ such that $([\omega]=[\tilde{\omega}])_{H^2(M;\mathbb{R})}$ and $Ricci(\tilde{\omega})=R$.
- Strominger & Yau were neighbours at IAS in 1985: CHSW named Ricci-Flat Kähler as Calabi-Yau Back

CICYs

$$M = \left[\begin{array}{c|cccc} n_1 & q_1^1 & q_1^2 & \dots & q_1^K \\ n_2 & q_2^1 & q_2^2 & \dots & q_2^K \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ n_m & q_m^1 & q_m^2 & \dots & q_m^K \end{array}\right]_{m\times K}$$

- Complete Intersection Calabi-Yau (CICY) 3-folds
- K eqns of multi-degree $q_j^i \in \mathbb{Z}_{\geq 0}$ embedded in $\mathbb{P}^{n_1} \times \ldots \times \mathbb{P}^{n_m}$
- $c_1(X) = 0 \rightsquigarrow \sum_{j=1}^{K} q_r^j = n_r + 1$
- M^T also CICY

- The Quintic $Q = [4|5]^{1,101}_{-200}$ (or simply $[5]$);
- CICYs Central to string pheno in the 1st decade [Distler, Greene, Ross, et al.]

  E_6 GUTS unfavoured; Many exotics: e.g. 6 entire anti-generations

Back to CICYs

Computing Hodge Numbers $\mathcal{O}(e^{e^d})$

- Recall Hodge decomposition $H^{p,q}(X) \simeq H^q(X, \wedge^p T^\star X) \rightsquigarrow$

$$H^{1,1}(X) = H^1(X, T_X^\star), \qquad H^{2,1}(X) \simeq H^{1,2} = H^2(X, T_X^\star) \simeq H^1(X, T_X)$$

- Euler Sequence for subvariety $X \subset A$ is short exact:

$$0 \to T_X \to T_M|_X \to N_X \to 0$$

- Induces long exact sequence in cohomology:

$$\begin{array}{ccccccc} 0 & \to & \cancel{H^0(X, T_X)}^{\,0} & \to & H^0(X, T_A|_X) & \to & H^0(X, N_X) \quad \to \\ & \to & \boxed{H^1(X, T_X)} & \stackrel{d}{\to} & H^1(X, T_A|_X) & \to & H^1(X, N_X) \quad \to \\ & \to & H^2(X, T_X) & \to & \cdots & & \end{array}$$

- Need to compute Rk(d), cohomology and $H^i(X, T_A|_X)$ (Cf. Hübsch)

Back to ML