
DEALINGS WITH DATA

PHYSICS, MACHINE LEARNING AND GEOMETRY

AUTHOR

Lorien Xanthe HAYDEN

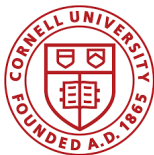
ADVISOR

James Patarasp SETHNA

A Dissertation

*Presented to the Faculty of the Graduate School
of Cornell University*

*in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy*



August 2019



©2019 Lorien Xanthe Hayden

ALL RIGHTS RESERVED

DEALINGS WITH DATA

PHYSICS, MACHINE LEARNING AND GEOMETRY

Lorien Xanthe Hayden, Ph.D.

Cornell University 2019

Collecting and interpreting data is key to developing an understanding of the physical underpinnings of observable events. As such, questions of how to generate, curate and otherwise wrangle data become central as systems of interest become increasingly difficult to access experimentally and the sheer quantity of raw information explodes.

The data explored in this dissertation covers a wide range of sources and methods. On the more traditional end, we explore simulation data of the two-dimensional non-equilibrium random-field Ising model which we treat with a novel analytic normal form theory of the Renormalization Group. Branching out from condensed matter, we explore several machine learning and sampling methods in various contexts.

The machine learning projects in particular include three lines of investigation: an unsupervised machine learning analysis of sectors of the economy extracted from stock return data, an analysis of the computational neural networks successfully applied to experimental ATLAS data in a recent Kaggle challenge, and an exploration of the geometical underpinnings of canonical neural networks using Jeffrey's Prior sampling of trained networks.

BIOGRAPHICAL SKETCH

The author was raised in Ashland, Missouri where she attended high school. Throughout this time, she participated extensively in Musical Theatre and cranked out an absurd amount of pizza, pasta, wings and hoagies at the local Pizza Haus. Beginning in 2007, she started college at the University of Missouri - Columbia where she transitioned from slinging food and making floral arrangements at the HyVee grocery store to teaching physics courses and performing scientific research as part of the McNair Scholars Program. She finished her B.Sc. (H) in Physics and Mathematics from the University of Missouri - Columbia in 2011.

Following a year as a Research Assistant with Professor Giovanni Vignale studying spin-torque effects at oxide interfaces, she moved to Ithaca, NY to pursue her Ph.D. in Physics at Cornell University. In 2015, she received her Masters of Science under the direction of her thesis advisor James Sethna. In 2017, Lorien married Jacob Wesley Boyett and they now have a son, Jiraiya Wesley Boyett who is 1 year old and full of boundless energy.

To my husband and son.

ACKNOWLEDGEMENTS

The work presented in the dissertation would not have been possible without the many individuals who have so greatly enriched my life during this period. I would like to first thank my advisor Jim Sethna whose infectious enthusiasm and steadfast support provided the framework for this time of rapid personal and professional growth. His unwavering passion for all things science and boundless curiosity inspired me to take risks and pursue the breadth of work presented here.

I also want to thank my research group including Danilo Liarte, D Zeb Rocklin, Ashivni Shekhawat, Yan-Juin Chen, Ricky Chachra, Alexander Alemi, Matthew Bierbaum, Archishman Raju, Katherine Quinn, Colin Clement, Jaron Kent-Dobias, Eddie Lee, David Hathcock, Alen Senanian, and Noé Besserman for the countless deep and intriguing conversations we have had. The friendship of this talented and exuberant team of people has meant a great deal to me. Special thanks to Katherine Quinn who has been both a confidant and inspiration. Thank you also to Mark Transtrum and Chris Myers for insightful discussions and to Veit Elser and Peter Wittich for providing the exciting questions which made the stressful period of my A exam a delight.

The community I have found at Cornell has also played an integral role in this process. Thank you to the members of GWiS (Graduate Women in Science), CUKE (Cornell University Klezmer Ensemble), and CEME (Cornell Eastern Music Ensemble) and to Jeffrey Bergfaulk, Andi Kao, Kayla Nguyen, Nikolay Zhelev, Shao Min Tan, Kathryn McGill and Kendra Weaver for enriching conversations.

Thank you to the broader Ithaca community including family friends Isaiah Parker for sharing his love of cooking and Valerie Sykes for rich discussions on personal philosophy. Thank you to Krysten Geddes and Emre Simsek for support and encouragement during my pregnancy and to Hailey and Kim Cote who will

undoubtedly remain lifelong friends.

Finally, thank you to my family. This dissertation would not have been possible without the love and continued support of my husband Jacob Boyett, his parents Shari and Ron Korthuis, and siblings Alaina and James Boyett. I am grateful also for my son, Jiraiya Boyett, whose joy and curiosity is an inspiration.

TABLE OF CONTENTS

ABSTRACT	I
BIOGRAPHICAL SKETCH	II
DEDICATION	III
ACKNOWLEDGEMENTS	IV
TABLE OF CONTENTS	VI
LIST OF FIGURES	IX
LIST OF TABLES	XX
1 INTRODUCTION	1
2 2D NON-EQUILIBRIUM RANDOM-FIELD ISING MODEL	5
2.1 NORMAL FORMS OF THE RG FLOWS	7
2.2 SCALING VARIABLES	10
2.3 SIMULATIONS	10
2.4 SCALING COLLAPSES	12
2.5 PARAMETER VALUES	14
2.6 COMPARISON OF FORMS	16
2.7 DISCUSSION	19
3 MACHINE LEARNING INTRODUCTION	21
4 CANONICAL SECTORS AND EVOLUTION OF FIRMS IN THE US STOCK	
MARKETS	23
4.1 ARCHETYPAL ANALYSIS	25

4.2	COMPANY EVOLUTION IN TIME	29
4.3	NUMBER OF CANONICAL SECTORS	30
4.3.1	SECTOR RELATIONSHIPS	33
4.3.2	TWO AND THREE FACTOR DECOMPOSITIONS	34
4.4	COEFFICIENT OF DETERMINATION	34
4.5	DISCUSSION	40
5	ANALYSIS OF THE KAGGLE HIGGS BOSON MACHINE LEARNING CHALLENGE	45
6	JEFFREY’S PRIOR SAMPLING OF DEEP SIGMOIDAL NETWORKS	57
6.1	SLOPPY MODELS	57
6.2	DEEP NETWORKS	59
6.3	JEFFREY’S PRIOR	62
6.4	RESULTS	65
6.5	DISCUSSION	68
7	SEQUENTIAL IMPORTANCE SAMPLING	77
8	DISCUSSION	85
	BIBLIOGRAPHY	89
	APPENDIX A CORRELATION LENGTH	101
	APPENDIX B INVARIANT SCALING COMBINATIONS	101
B.1	POWER LAW FORM	101
B.2	TRANSCRITICAL FORM	103
B.3	ALTERNATIVE TRANSCRITICAL FORM	104
B.4	PITCHFORK FORM	106

APPENDIX C	UNIVERSAL SCALING FUNCTION FORMS	107
APPENDIX D	FORM COMPARISON	108
APPENDIX E	PARAMETER VALUES	109
APPENDIX F	KAGGLE DATASET DETAILS	114
APPENDIX G	JEFFREY'S PRIOR SAMPLING: ADDITIONAL RESULTS	116
APPENDIX H	SAMPLING OF A VARIATIONAL AUTOENCODER: PRE- LIMINARY RESULTS	123

LIST OF FIGURES

1	Avalanches produced for different values of disorder. $w = 0.5, 1.0, 5.0, 50.0$ from left to right, top to bottom	11
2	Area weighted avalanche size distribution. The avalanche size s times the area weighted avalanche size distribution $A(s w)$ for values of w ranging from 0.8 to 8.0.	12
3	Change in magnetization with field. $\frac{dM}{dh}(h w)$ for values of w ranging from 0.8 to 8.0.	13
4	Scaling collapse of the area weighted avalanche size distribution. w ranges from 0.8 to 8.0. There is a slight bulge at $s/\Sigma(w) \sim 10^{-2}$ for small w	14
5	Scaling collapse of the change in magnetization with respect to the field. w ranges from 0.8 to 8.0.	15
6	Comparison of the best fit of $\Sigma(w)$ and $\eta(w)$ derived with different functional forms of $\frac{dw}{dt}$. We have $w = (r - r_c)/s_s$ such that $\Sigma(r) = \Sigma(w)$ and $\eta(r) = \eta(w)$. ‘NF’ corresponds to Σ and η derived from the transcritical normal form, ‘Power Law’ the hyperbolic (power law) form and ‘Pitchfork’ the pitchfork form. . .	17
7	Comparison of $1/\log \Sigma(w)$ for the best fit of $\Sigma(w)$ derived with different functional forms of $\frac{dw}{dt}$. We have $w = (r - r_c)/s_s$ such that $\Sigma(r) = \Sigma(w)$. ‘NF’ corresponds to Σ derived from the transcritical normal form, ‘Power Law’ the hyperbolic (power law) form and ‘Pitchfork’ the pitchfork form.	18

- 8 **Low-dimensional projection of the stock price returns data.** Stock price returns are projected onto a plane spanned by two stiff vectors from the SVD of the emergent simplex corners [53, Supplement]. Each colored circle corresponds to one of the 705 stocks in the dataset used in the analysis. Colors denote the sectors assigned to companies by Scottrade [97]. The grey corners of the simplex correspond to sector-defining prototype stocks, whereas all other circles are given by a suitably weighted sum of these grey corners. . . . 26
- 9 **Canonical sector decomposition of stocks of selected companies.** A complete set of all 705 stocks is provided on the companion website [113]; the color scheme is shown on the right. Conglomerates like GE decompose roughly into their core business lines. Tech firms such as Apple that sell mass-market consumer goods have an important fraction in *c-cyclical*, whereas IBM has a significant portion of *c-non-cyclical* returns presumably due to its government contracts. Telecom companies like AT&T are generally classified under a separate telecom category by major classification systems, yet analysis shows their returns are described by a combination of *c-non-cyclical* and *c-utility* sectors. Health insurance providers like Aetna are commonly classified as financial services firms, but their returns consist of a major part *c-non-cyclical* and only a minor part of *c-financial*—the healthcare sector is generally less prone to economic downturns. Defense contractors like Lockheed are listed as capital goods companies, but their returns are seen to be majority *c-non-cyclical* and only a smaller share of *c-industrial* sector. . . . 28

10 **Emergent sector time series.** Annualized cumulative log price returns of the eight emergent sectors are shown. The time series capture all important features affecting different sectors: building-up of the dot-com bubble (c. 2000) followed by a burst, the soaring energy valuations (2003–08) followed by a crash, and financial crisis of 2008. We note that the dot-com bubble was confined to the c-tech whereas the financial crisis effects were spread throughout the sectors. Precise definition of the cumulative returns plotted here is given in [53]. 29

11 **Evolving sector participation weights.** Results from the sector decomposition made with rolling two-year Gaussian windows are shown for selected stocks. A complete set of 705 charts is provided on the companion website [113]. Color scheme is as in (Figure 9). For stable and focused companies such as Pacific Gas & Electric or IBM, one sees no significant shifts in sector weights; changes in time agree with errors expected from unresolved fluctuations [113]. Wal-Mart’s returns, on the other hand, have moved significantly from *c-cyclical* to *c-non-cyclicals* (consumer staples) in the post-financial crises years as shown; this is also true of other low-price consumer commodities retailers such as Costco, but not true of higher price retailers such as Whole Foods, Macy’s, etc. Corning, previously an *industrial* firm with a huge presence in optical fiber, suffered in the aftermath of the dot-com crisis and now is classified as a *tech* firm presumably due to its Gorilla[®] glass used in cellphones, laptop displays, and tablets. Berry Petroleum grew within its home state of California in the early 1990s through development on properties that were purchased in the earlier part of 20th century. In 2003, the company embarked on a transformation [10] by direct acquisition of light oil and natural gas production facilities outside California. The figure shows a clear shift in the distribution of sector weights as the company has moved toward *c-energy* and away from *c-real estate*. Similarly, as Plum Creek Timber converted to a real estate investment trust (REIT) in the late 1990s [85], its sector weights have significantly shifted toward *c-real estate* sector. 31

12	<p>Comparison between flow diagrams presented in Figure 11 with simulated data. The simulated data is created from the dot product of the weight vector of the company with the corner time series. This yields a version of the company with constant weights in time. To this we add gaussian noise with standard deviation one and repeat the analysis to generate the flows in time. In the left column are the actual flows for companies, on the right is their constant in time counterpart with added noise. We see that key features noted are in fact signal while small fluctuations correspond to noise.</p>	32
13	<p>Changes in the decomposition with dimensionality. A Sankey diagram (generated using D3 [16]) displaying the relationships between sector decompositions with $n = N + 1$ and $n = N$ Relative node sizes correspond roughly to the amount of the market participating in the sector. Connection width depicts how strongly the sectors for decompositions with different n relate.</p>	41

- 14 **Pie charts depicting sectors as linear combinations of other sector decompositions having a different value of the dimensionality n .** (a) Two sector decomposition with respect to the eight sector version (b) Three with respect to eight (c) eight with respect to two (d) eight with respect to three. For (a) and (b) the color scheme is the same as used throughout for the eight sector decomposition. For (c) and (d) colors correspond to those in Figure 9 for the two and three sector nodes. Through these charts it is evident that the two sector decompositions corresponds to an *c-assets* sector containing *c-finance* and *c-real estate*. and a *c-goods* sector containing companies which provide goods and services. In (c) and (d) we see *c-industrial*, *c-cyclical* and *c-non-cyclical* which merge by $n = 5$ split between the two and three factor decompositions respectively, consistent with Figure 9. 42
- 15 **Normalized distribution of singular values.** Filled blue histogram corresponds to distribution of singular values of returns from the dataset R_{ts} —one notices a clear separation of the hump-shaped bulk of singular values, and about 20 stiff singular values (the largest singular value ~ 952 , corresponding to the *market mode* is not shown). Pink line histogram outline shows the distribution of singular values of a matrix of the same shape as R but containing purely random Gaussian entries. 43

16	3 Factor Model vs. Fama and French 2D projections of the weights for each company in the SP500 with current tickers and data in the date range we consider. Red denotes companies with large market caps (market cap >10 billion), blue denotes medium (market cap 2-10 billion) and green denotes small (market cap < 2 billion). For our decomposition (a), there is no separation distinguishable by size of company. In comparison, for the Fama and French decomposition (b), there appears a gradation from large to small companies consistent with a factor of the model being related to size. (This is natural, since one of Fama and French’s factors explicitly is the difference between large and small-cap returns). Thus our unsupervised 3-factor decomposition appears quite distinct from Fama and French’s hand-created one.	44
17	Sample decision tree from scikit-learn [81]. A single variable along with its decision rule is chosen at each level in the tree such that the cost function, such as gini or cross-entropy, is minimized. .	50
18	Diagram of the decoding map learned by a generic 3-layer neural network. Once trained, a DBN or SdA provides a function f such that, for any given neural activation in the top hidden layer $\vec{\theta}$, f provides a corresponding image in reconstruction space $\vec{y} = f(\vec{\theta})$.	62
19	Widths of the reconstructed manifold for each network along PCA directions.	66

20	<p>Two dimensional PCA projections of the MNIST digit data for ones and the Jeffrey’s Prior sample of the model manifold. The $(i, j)^{th}$ figure in the grid is a plane spanned by singular vectors i and $j + 1$ of the centered data. Image reconstructions lie on the edges of the manifold. Width of the network decreases slightly along each principal vector; however the aspect ratio is much closer to unity than that observed in other models [106].</p>	69
21	<p>Two dimensional PCA projections of the MNIST digit data for ones, the Jeffrey’s Prior sample of the model manifold and a few selected images from the dataset. The top row corresponds to the top row in Figure 20. Ones with differing distinct characteristics such as tilt or bases have been labeled to display how the manifold is roughly arranged according to digit behavior. The images corresponding to these digits are shown on the bottom row. For each, the original digit has been plotted in pink with the reconstruction overlaid in black.</p>	70
22	<p>Two dimensional PCA projections of the MNIST digit data and the Jeffrey’s Prior sample of the model manifold for a 4 layer DBN. The $(i, j)^{th}$ figure in the grid is a plane spanned by singular vectors i and $j + 1$ of the centered data.</p>	71
23	<p>Examples of MNIST images, their reconstructions, and images sampled using Jeffrey’s Prior for the DBN. For the sampled ‘digits’, each snapshot corresponds to the same walker. The final row corresponds to the top five ‘eigen-digits’ of the dataset.</p>	72

24	PCA projection of the Jeffrey's Prior sampling with MNIST digit data for the DBN along the two largest principal component vectors.	In the middle column 'Transparent' the transparency of the MNIST points have been enhanced to show the position of the sampling. On the right under 'Corners' the digits and sampling have again been plotted in blue and black respectively with the corners added in pink. Corners correspond to activations in the top hidden layer ($\vec{\theta}$) for which $\theta_i \in \{-\infty, \infty\} \sim \{-10^6, 10^6\}$. The axes are shared along each row. In order to deal with the large values of the corners and sampling, the top layer is shown with a sigmoid applied. Although the sampling spans the parameter space (Top Layer), it is subsequently mapped to the interior of the manifold in reconstruction space ('Layer 1'). Note that the radius of the sampling is roughly $\frac{2}{3}$ that of the digits - resulting in a 30-D volume $(\frac{2}{3})^{30} = 5.2 \times 10^{-6}$ of the total volume. Notice also that the digits lie outside the corners.	73
25	One of Knuth's original samples on a 10x10 grid [61].	Numbers refer to the number of choices available at each vertex.	78
26	Images of the first few samples on a 100 by 100 grid.	80
27	λ as a function of the number of walks sampled.	(red: k=4, orange: k=5 ...purple: k=10)	81
28	β as a function of the number of walks sampled.	(red: k=4, orange: k=5 ...purple: k=10)	82
29	κ as a function of the number of walks sampled.	(red: k=4, orange: k=5 ...purple: k=10)	82
30	Fit comparisons $\Sigma_{th}(w)$, transcritical form.	109
31	Fit comparisons $\eta_{th}(w)$, transcritical form.	110

32	Fit comparisons $\Sigma_{\text{alt}}(w)$, alternative transcritical form.	. . . 110
33	Fit comparisons $\eta_{\text{alt}}(w)$, alternative transcritical form.	. . . 113
34	PCA projection of the Jeffrey's Prior sampling with MNIST '1' digit data for the single-digit network along the two largest principal component vectors. In the middle column 'Transparent' the transparency of the MNIST points have been enhanced to show the position of the sampling. On the right under 'Corners' the '1's and sampling have again been plotted in blue and black respectively with the corners added in pink. Corners correspond to activations in the top hidden layer ($\vec{\theta}$) for which $\theta_i \in \{0, 1\}$.	117
35	Examples of MNIST '1' images, their reconstructions, and images sampled using Jeffrey's Prior for the single-digit network. For the sampled '1's, each snapshot corresponds to the same walker. The final row corresponds to the top five 'eigen-digits' of the '1's dataset. 118

36	<p>PCA projection of the Jeffrey’s Prior sampling with MNIST digit data for the SdA along the two largest principal component vectors. In the middle column ‘Transparent’ the transparency of the MNIST points have been enhanced to show the position of the sampling. On the right under ‘Corners’ the digits and sampling have again been plotted in blue and black respectively with the corners added in pink. Corners correspond to activations in the top hidden layer ($\vec{\theta}$) for which $\theta_i \in \{-\infty, \infty\} \sim \{-10^6, 10^6\}$. The axes are shared along each row. In order to deal with the large values of the corners and sampling, the top layer is shown with a sigmoid applied. Although the sampling spans the parameter space (Top Layer), it is subsequently mapped to the interior of the manifold in reconstruction space (‘Layer 1’).</p>	119
37	<p>Examples of MNIST images, their reconstructions, and images sampled using Jeffrey’s Prior for the SdA. For the sampled ‘digits’, each snapshot corresponds to the same walker. The final row corresponds to the top five ‘eigen-digits’ of the dataset. .</p>	120
38	<p>PCA projection of the Jeffrey’s Prior sampling for the 3D manifold (Pink) and 30D manifold (Black) in reconstruction space for the DBN.</p>	121
39	<p>PCA projection of the Jeffrey’s Prior sampling for the 3D manifold (Pink) and 30D manifold (Black) in reconstruction space for the SdA.</p>	122
40	<p>PCA projection of the sampling of a variational autoencoder and the MNIST digits.</p>	123
41	<p>PCA projection of the sampling of a variational autoencoder.</p>	124

LIST OF TABLES

1 **Table of the parameter values determined through a joint fit of $\Sigma(w)$ and $\eta(w)$.** *NF* corresponds to the transcritical form and *NF_{alt}* to the alternative transcritical form described in Appendix B.3. *NF₀* corresponds to the transcritical form with $r_c = 0$ and *NF_{Harris}* to $\lambda_h = 1$, the Harris criteria. To compute the error bars, we performed the collapses and subsequent fits of the nonlinear forms using subsets of the disorders for which we have data [11 out of 13 points]. The errors given are the standard deviation of the values determined in this way. Values in bold were fixed in the corresponding fit. 16

2 **Canonical sectors and major business lines of primary constituent firms.** The eight canonical sectors identified by the analysis described here are listed in the column on the left; these were named in accord with the business lines of firms that show strong association with these sectors. Some examples are provided in the right column; a full list is available on companion website [113]. . . 27

3 **Top 20 contributing companies to each sector in the two sector decomposition.** Ranking is determined by the matrix $C_{s,f}$ which describes each sector as a linear combination of stocks. Labels are those given by Scottrade and percentage describes the percentage of the sector attributable to the company. 35

4	<p>Top 20 contributing companies to each sector in the three sector decomposition. Ranking is determined by the matrix $C_{s,f}$ which describes each sector as a linear combination of stocks. Labels are those given by Scottrade and percentage describes the percentage of the sector attributable to the company.</p>	36
5	<p>Top 20 contributing companies to each sector in the three sector decomposition. Ranking is determined by the matrix $C_{s,f}$ which describes each sector as a linear combination of stocks. (Accompaniment to Table 4.)</p>	37
6	<p>Percentage of the Explainable Variance captured by our model compared with the Fama and French factor model. Regression is done on the normalized dataset of 705 stocks without the market mode removed. To capture this, we add the market mode to factors obtained by our decomposition.</p>	39
7	<p>Scores obtained after 1000 epochs of training with 108 neural networks on each dataset. Dimension denotes the the number of entries for each event in the given dataset. AMS was evaluated by the Kaggle site on the unlabeled test.csv data provided. Percent Error was evaluated on the 20% of the training data held out. . . .</p>	55

8	Training characteristics of the neural networks we study.	
	Classification accuracy on the MNIST dataset [67] was evaluated by training a support vector machine to classify the data given the top layer of features [81]. The network and support vector machine were then applied in tandem to the test set in order to calculate the error. Training of the networks were achieved using Theano [6, 9] and MATLAB code provided by the Hinton group [55]. The 4-layer SdA trained with theano was trained with a linear mapping at the top layer. This choice was made to ease comparison between the SdA and the Hinton group’s DBN which has this characteristic. For each of the 4-layer networks the layer sizes were chosen to correspond with Hinton’s original network. The top hidden layer in each case had a dimension of 30.	61
9	Numerical values of the constants obtained from sampling.	81
10	Comparison of the known number of walks with the estimates given by sampling.	83
11	Table of the best fit values corresponding to Figures 30 and 31. Values in bold correspond to values fixed in the fit.	111
12	Table of the best fit values corresponding to Figures 32 and 33. Values in bold correspond to values fixed in the fit.	112

1 INTRODUCTION

As computing power continues to increase, so does the development of techniques which harness it efficiently. These come in a variety of flavors including novel approaches to simulations, sampling, visualization and machine learning. In this work, we touch on each of these subjects informed by the Sethna group expertise in dealing with high-dimensional geometry and canonical materials science models.

This dissertation is divided roughly into two sections. In the first, I discuss our work on the non-equilibrium random-field Ising model. This model is well studied, yet there are outstanding questions. In two dimensions, power law scaling approaches fail and the critical disorder is difficult to pin down. Additionally, the presence of faceting on the square lattice creates avalanches that are lattice dependent at small scales. We propose two methods which we find solve these issues. First, we perform large scale simulations on a Voronoi lattice to mitigate the effects of faceting. Secondly, the invariant arguments of the universal scaling functions necessary to perform scaling collapses can be directly determined using our recent normal form theory of the Renormalization Group. This method has proven useful in cleanly capturing the complex behavior which occurs in both the lower and upper critical dimensions of systems and here captures the 2D NE-RFIM behavior well. The obtained scaling collapses span over a range of a factor of ten in the disorder and a factor of 10^4 in avalanche cutoff. They are consistent with a critical disorder at zero and with a lower critical dimension for the model equal to two.

These results are exciting for two reasons. One, the NE-RFIM is an interesting model in its own right. Barkhausen noise in magnets [11] decision making in socio-economics [17], absorption and desorption in superfluids [41, 69] as well as

the effects of nematicity in high T_c superconductors [15, 25, 83] can each be understood in terms of 'crackling noise' naturally described by the NE-RFIM. Secondly, these results provide a nice example of the effectiveness of considering the underlying Renormalization Group structure when dealing with cases where traditional power law scaling fails. Power law scaling collapses are ubiquitous in a number of fields, however, the mathematical structure of the corresponding critical point is often overlooked. Not all critical points correspond to hyperbolic fixed points and recognizing this provides a straightforward prescription for dealing with 'difficult' cases.

While the first portion of this dissertation details a standard approach to dealing with large amounts of simulation data, namely to think hard and come up with an appropriate mathematical/physical description, the second focuses on tools that attempt to find a description with as little human input as possible. While providing little to no intuition of underlying physical truths, machine learning models are useful in their heavy handedness. It is not necessary for the practitioner to have any particularly deep knowledge of the dataset in question to extract useful information although domain knowledge may be leveraged to enhance the results.

One such exploration we have undertaken is to extract information about the sectors of the economy from raw stock return data. Imaging using principle component analysis reveals that stock data lies in a high-dimensional tetrahedron. Using this geometrical information only, it is possible to choose an appropriate machine learning approach using no properties of economics. This is remarkable and allows us to extract a number of interesting results in an unsupervised way. These include elucidating sectors of the economy, company composition in terms of how strongly they participate in a given sector, changes in company composition with time and relationships between the sectors themselves.

In 2014, the Higgs Boson Machine Learning Challenge was hosted on Kaggle,

a platform for predictive modelling and analytics competitions. The challenge was to use a subset of ATLAS data to train the most effective classifier of Higgs events. Surprisingly, despite domain knowledge, the best team of physicists came in 8th behind data scientists. This seems to be in part due to underestimation of the power of machine learning combined with a lack of experience with certain pitfalls in implementing these algorithms. To look at the question of whether domain knowledge could in principle be useful, we consider another type of machine learning algorithm: computational neural networks. Training an ensemble of such networks on subsets of the data provided for the Kaggle challenge suggests that features derived using physics knowledge do in fact show an advantage over raw data.

In the context of neural networks, we also explore another common tool and its potential pitfalls, Monte Carlo sampling. Both computational neural networks and Monte Carlo sampling algorithms have been around for some time [76, 91], however, recent advances have proven them to be much more powerful than their predecessors [12, 96]. Neural networks such as those used in the Kaggle challenge have been shown to have a remarkable ability to uncover low dimensional structure in data. In this next portion of this dissertation, we explore this idea directly by analysing the manifold learned by Deep Belief Networks and Stacked Denoising Autoencoders using Monte Carlo sampling. Previously, the Sethna group has studied manifolds generated by models from a variety of fields. What has been found repeatedly is that the manifold composed of all possible predictions of the model forms a hyperribbon in data space. What this corresponds to conceptually is that certain combinations of parameters dominate the behavior of the model while others barely have an effect [112]. In studying neural networks, however, we observe a very different behavior. The model manifold forms an only slightly elongated hypersphere and, more curiously, the actual data appears only on the

boundaries of the manifold. The shape of the manifold suggests that the network is doing well in weighting each parameter equally, however, the vast regions which do not correspond to data indicate that the network may waste a considerable amount of descriptive ability. On the other hand, the prior used for the sampling, known as Jeffrey's Prior, places equal weight per equal volume in data space. Recent work by Transtrum et. al. [107] suggests that the boundaries of a model manifold are the most important for describing the behavior of a model. Rather than pointing to a deficit of the network, the behavior observed may point to a sickness with Jeffrey's Prior.

Sampling with Jeffrey's Prior in the case of the neural networks appears to be flawed. In fact, although sampling is a very important tool when all elements of a distribution cannot be enumerated, it can also be very problematic. Choice of prior, specific algorithm, as well as a host of hyperparameters can strongly affect the validity of the results. As an example, consider the convergence of sequential importance sampling in a problem originally proposed by Knuth [61]. Exploring this case, the results suggest that to get an accurate approximation with the method one would need approximately 10^{212} times the age of the universe with the naive approach. This highlights that although sampling is very useful for studying problems where analyticity is prohibited, it is important to understand the limitations and to assess carefully the results.

Results from each of our explorations into machine learning and sampling approaches highlight two key and contrasting ideas which can be summed up concisely. These types of algorithms can be extremely useful if used with care. As physicists implement these algorithms, it is very important to understand the pitfalls in order to gain an accurate understanding of what the results really convey.

2 2D NON-EQUILIBRIUM RANDOM-FIELD ISING MODEL

The non-equilibrium random-field Ising model (NE-RFIM) is a model of long-standing interest. This model, albeit simple, contains the necessary ingredients to describe hysteretic and avalanche behaviors in a diverse set of systems. Barkhausen noise in magnets [11] decision making in socio-economics [17], absorption and desorption in superfluids [41, 69] as well as the effects of nematicity in high T_c superconductors [15, 25, 83] can each be understood in terms of ‘crackling noise’ naturally described by the NE-RFIM.

Although the NE-RFIM itself has been around in various forms since the 1970s [57], there are still a number of open questions:

- Is it in the same universality class as the equilibrium model?
- Is the lower critical dimension two?
- What is the value of the critical disorder?
- Is power law scaling sufficient to capture the behavior?

It has long been debated whether the equilibrium and non-equilibrium versions of the model are in the same universality class. This question of universality has been approached in a number of ways which have suggested the same class for the two models [4, 34, 70, 71, 74, 82]. Recently, however, evidence has been provided that this is not the case [3]. Our findings pretty clearly imply this latter result.

Another open question concerns the lower critical dimension (LCD) of the non-equilibrium model. For the equilibrium case, the LCD is accepted to be two [20] and there is evidence to believe the same is true of the front-propagation model [44]. For the nucleated model, there have been conflicting analysis including work suggesting that the LCD is two [29, 30], that power-laws are indeed able to capture the

behavior and no crossover occurs in 2D [27, 28], and even that a lower critical dimension does not exist for this model [65, 99, 100, 104]. Here, we find the success of our results to be consistent with a LCD of two.

Yet another open question is what value r_c takes in two dimensions for the nucleated model. In the nucleated model, the critical disorder appears to decrease with dimension, going from 5.96 ± 0.02 in 5D to 2.16 ± 0.03 in 3D [98]. This behavior in conjunction with the observation that for both the equilibrium and front-propagation problems, r_c is found to be zero [44] suggests that r_c may be quite small. The results we present here are consistent with $r_c = 0$.

Finally, it has yet to be resolved whether a power law form is sufficient to describe the behavior in 2D. Fitting assuming a power law form, Vives *et al.* found r_c to take the value 0.75 ± 0.03 [110]. More recently, on a much larger square lattice, Spasojevic *et al.* find $r_c = 0.54 \pm 0.02$ [27, 28] collapsing over a range of $r \in [0.64, 0.70]$. We expect this discrepancy to be replicated in simulations on a larger scale with r_c taking a yet smaller value and suggest this type of inconsistency in power law scaling points to a deficit in its ability to accurately capture the critical behavior in 2D.

Power law scaling collapses have long been a preferred method for demonstrating that the behavior of a critical system is well understood. That this type of heuristic procedure can work so well in such a widespread number of applications is initially surprising and leads naturally to the question of when and why this approach fails. For example, in the two-dimensional non-equilibrium random-field Ising model (2D NE-RFIM), attempted collapses assuming power law scaling perform in very limited ranges of disorder [27, 28, 30, 63], which we argue is a symptom of non-power law scaling. This failure can also be observed in a number of other systems, particularly at their lower and upper critical dimension. Recently, Raju *et al.* [89] have been successful in describing the non-linearities that

arise in renormalization group flows from the perspective of normal form theory. This formulation provides a systematic method to perform scaling collapses. In the cases for which power laws work well, the dynamics can be described simply by the presence of a hyperbolic fixed point; the eigenvalues are non-zero and there is no qualitative change in the stability of the fixed points. We propose it is the presence of a transcritical bifurcation in the disorder flow equation that corresponds to the rise in complexity needed to describe simulation data of the 2D NE-RFIM. By considering the form the flow equations should take, we are able to provide concrete non-linear scaling variables which enable collapse of our data over a range of a factor of ten in the disorder.

In addition to the application of our normal form theory of the Renormalization Group, another key component to the success of our collapses is an approach to dealing with the faceting. Running simulations on a square lattice leads to distortions in the shape of the distributions of interest due to lattice effects as the critical point is approached. Long, unnaturally straight avalanche boundaries for small disorder arise which serve to effectively decrease the simulation size. To combat this, we run our simulations on a Voronoi lattice. Although this could in principle introduce an amount of intrinsic disorder, we find the Voronoi lattice to be effective in combating faceting effects, enabling clean collapses over a range of a factor of ten in the disorder, a significantly larger range than the current available collapses which use data in a range $\approx 10\%$.

2.1 NORMAL FORMS OF THE RG FLOWS

The model considered is an avalanche model with nearest neighbor coupling J and a randomized bias r under the influence of an adiabatically increasing field h . Avalanche size is denoted s . Following the convention of Bray and Moore [20] for

the equilibrium model, we define a parameter w which corresponds to the ratio of the disorder r over the coupling J and determine its RG flow equation through symmetry considerations.

In the equilibrium model, the flow equation is found to be $dw/d\ell = -(\epsilon/2)w + Aw^3 + h.o.t.$ where $\epsilon = D - 2$ and $w = r/J$ [20]. For the NE-RFIM, however, r has the symmetry $r \leftrightarrow -r$ while J lacks this symmetry due to the external field. This implies $w \leftrightarrow -w$ and suggests that the RG flow for w in the NE-RFIM must include a squared order term.

In principle, there are an infinite series of terms. Assuming the lower critical dimension $D = 2$, we have $\epsilon = 0$, and may choose a scale for the disorder r_s such that the prefactor of the squared order term in the flow equation of w is equal to one. Taking $J = 1$, the choice we make for w is $w = (r - r_c)/r_s$ where r_c defines the critical disorder. The generic form for the flow equation of w is given by

$$\frac{dw}{d\ell} = w^2 + B_1 w^3 + B_2 w^4 + \dots \quad (1)$$

Using only polynomial changes of variables, it is possible to remove all terms of $O(4)$ or higher without removing any universal behavior. To demonstrate, consider the change of variables $w = \tilde{w} + b_1 \tilde{w}^2 + b_2 \tilde{w}^3 + b_3 \tilde{w}^4 + \dots$. The resulting flow equation takes the form:

$$\frac{d\tilde{w}}{d\ell} = \tilde{w}^2 + B_1 \tilde{w}^3 + (B_1 b_2 + b_2^2 + B_2 - b_3) \tilde{w}^4 + \dots \quad (2)$$

With an appropriate choice of b_2 in terms of b_3 , the coefficient of \tilde{w}^4 may easily be set to zero. Likewise, all higher order terms may be systematically removed. Dropping the tildes and subscripts for clarity, the final form of the flow equation

is given by

$$\frac{dw}{d\ell} = w^2 + Bw^3 \quad (3)$$

which corresponds to the normal form of a transcritical bifurcation ¹. We may directly solve for the correlation length $\xi \sim (1/w + B)^{-B} \exp(1/w)$ in the normal form variables (Appendix A).

Next consider the flow equations for s and h . The eigenvalues for these are given by $\lambda_s = d_f$ and λ_h respectively where d_f denotes the fractal dimension. In each case, the zero eigenvalue of w gives rise to cross terms between s and w and h and w . Again, in principle, we have an infinite number of possible terms but most all terms may be removed with a polynomial change of variables. The flow equations for s and h are hence given by

$$\begin{aligned} \frac{ds}{d\ell} &= -d_f s - Csw, \\ \frac{dh}{d\ell} &= \lambda_h h + Fhw \end{aligned} \quad (4)$$

where in higher dimensions $d_f = 1/\sigma\nu$ and $\lambda_h = \beta\delta/\nu$. In two dimensions, the individual exponents $\sigma \rightarrow 0$ and ν and $\beta\delta \rightarrow \infty$, keeping the combinations we use finite. The coefficients B , C , and F are universal. Just as the linear terms at ordinary (hyperbolic) fixed points yield universal critical exponents, these terms control universal dependences of physical behavior with changes in the control parameters. Note that, while they cannot be set to zero by a coordinate change, they may have universal values equal to zero, especially in special cases like the lower critical dimension.

¹The traditional transcritical bifurcation normal form [103] $dw/d\ell = w^2$ is derived using the implicit function theorem, but involves changes of variables that alter critical properties in singular ways. Eq. 3 is the simplest form that can be reached by successive polynomial changes of variables.

2.2 SCALING VARIABLES

The appropriate scaling variables to collapse the data can be directly calculated from the flow equations as detailed in Appendix B.2. The invariant scaling combination obtained takes the form $s/\Sigma(w)$ where $\Sigma(w)$ is a nonlinear function of w . We allow for an undetermined scale factor Σ_s . The resulting form is given by

$$\Sigma(w) = \Sigma_s \left(B + \frac{1}{w} \right)^{-Bd_f + C} \exp \left(\frac{d_f}{w} \right). \quad (5)$$

Likewise for h , we obtain:

$$\eta(w) = \eta_s \left(B + \frac{1}{w} \right)^{B\lambda_h - F} \exp \left(-\frac{\lambda_h}{w} \right) \quad (6)$$

where $(h - h_{max})/\eta(w)$ is invariant under the RG, and η_s is another scale factor.

2.3 SIMULATIONS

Experience simulating the RFIM on a square lattice has revealed a propensity for faceting in which the shape of the avalanche size distribution becomes dependent on properties of the lattice for small avalanche sizes. To mitigate this effect, we perform simulations on a periodic Voronoi lattice where, for each value of r , we consider 100 distinct lattices of size 1000x1000. Voronoi cells were chosen by generating random coordinates between 0 and 1 and constructing the cells with a 2D implementation of Voropp [93] provided by C. H. Rycroft. Examples of the avalanche behavior for different values of r are shown in Figure 1.

We note that much larger simulations have been done on the square lattice, including a thorough analysis of results from a $131,072^2$ lattice [27, 28]. In analysis of in house simulations on a square lattice, however, we encountered long, unnaturally straight avalanche boundaries. We found these distortions strongly affected

the shape of the size distribution for small disorders and served to effectively decrease the system size, a difficulty which became dramatically more pronounced as the disorder decreased. In addition to lattice dependent effects infecting the distributions for larger and larger avalanche sizes approaching the critical point, this effective reduction of system size encouraged the use of a Voronoi lattice.

From the simulations we extract two quantities of interest: the area weighted avalanche size distribution $A(s|w)$ [33] and the change in magnetization of the sample with respect to the field $\frac{dM}{dh}(h|w)$. The data collected is shown in Figures 2 and 3

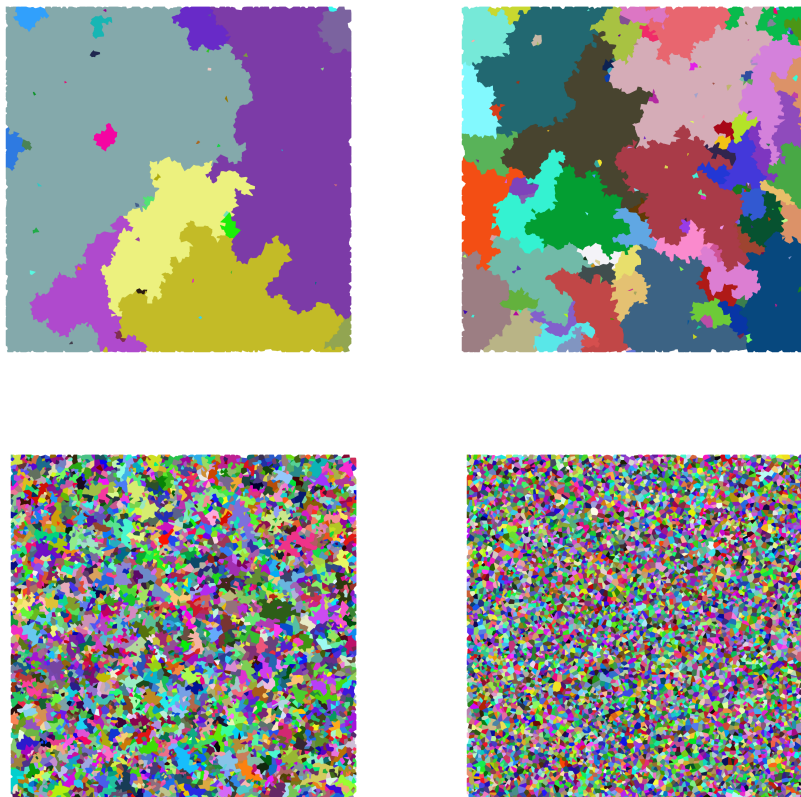


Figure 1: **Avalanches produced for different values of disorder.** $w = 0.5, 1.0, 5.0, 50.0$ from left to right, top to bottom

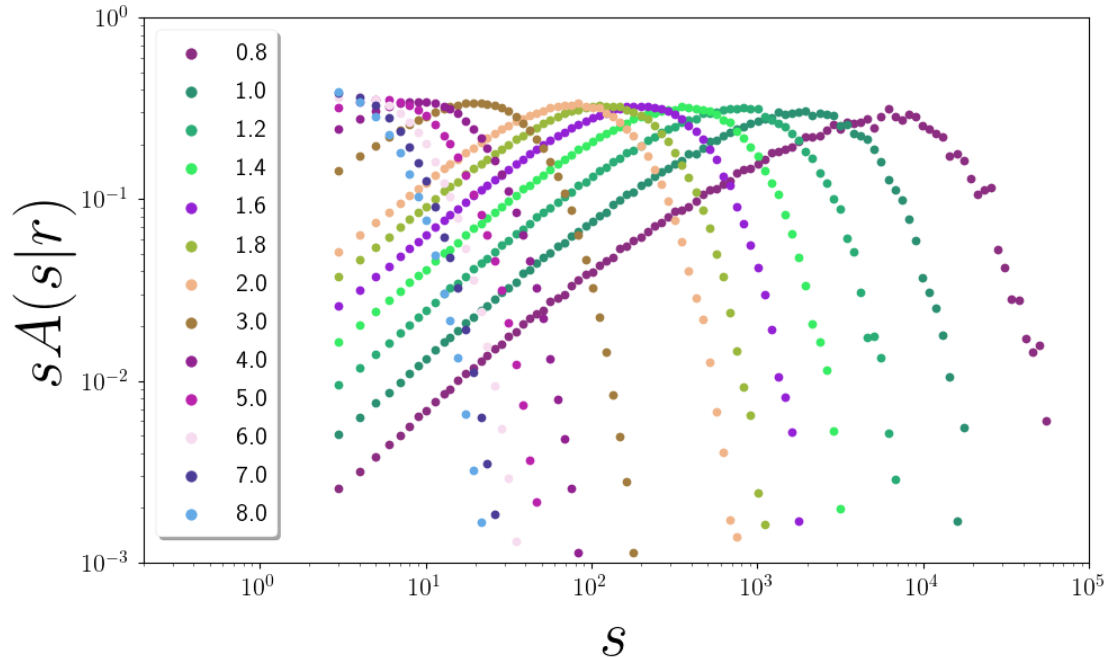


Figure 2: **Area weighted avalanche size distribution.** The avalanche size s times the area weighted avalanche size distribution $A(s|w)$ for values of w ranging from 0.8 to 8.0.

2.4 SCALING COLLAPSES

First consider the area weighted size distribution $A(s|w)$. In analogy with three dimensions, we take $A(s|w) = s^{-1}v_s^x\mathcal{A}(v_s^y)$ where v_s is the scaling variable and the prefactor of s^{-1} arises from normalization constraints with $v_s = s/\Sigma(w)$ from Equation 5. The avalanche size distribution also depends on an unknown universal scaling function, \mathcal{A} . To perform scaling collapses via a fit, we assume a form for this equation given in Appendix C. The associated collapse is shown in Figure 4.

Next consider dM/dh . The scaling form for the magnetization as a function of field in three dimensions would be

$$M_{3D}(h|w) \sim w^\beta \mathcal{M}((h - h_c)/w^{\beta\delta}) \quad (7)$$

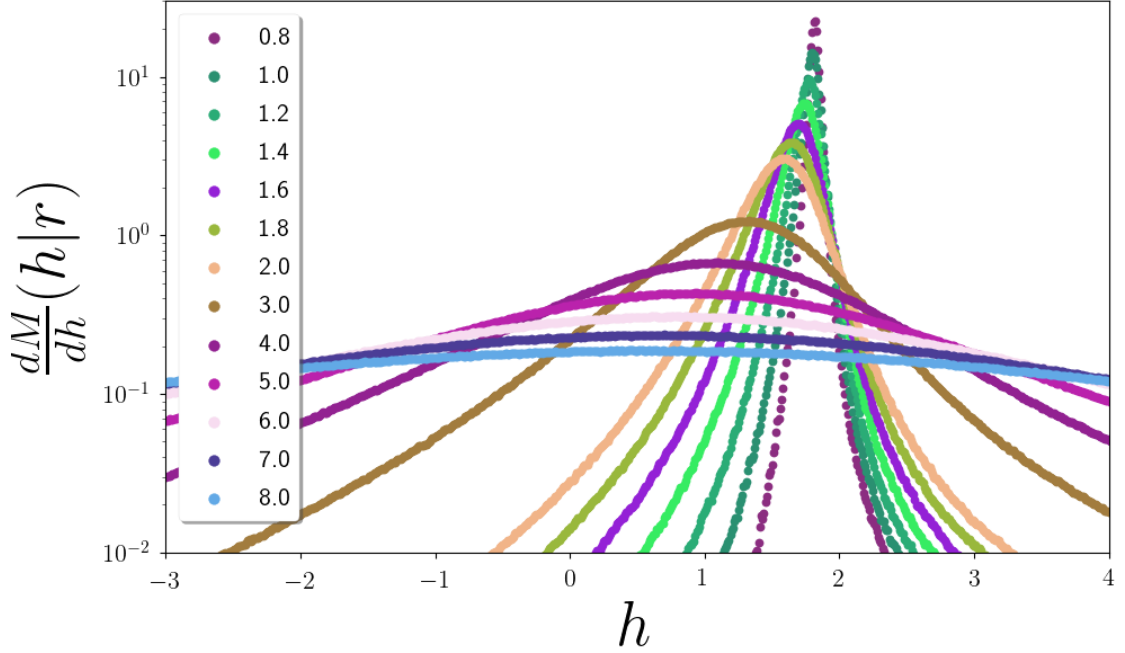


Figure 3: **Change in magnetization with field.** $\frac{dM}{dh}(h|w)$ for values of w ranging from 0.8 to 8.0.

yielding a 3D scaling form

$$\frac{dM_{3D}}{dh}(h|w) = w^{\beta-\beta\delta} \frac{d\mathcal{M}_{3D}}{dh}((h-h_c)/w^{\beta\delta}) \quad (8)$$

In two dimensions, $w^{\beta\delta}$ is simply replaced by $\eta(w)$ from Eq. 6.

But what of the term w^β ? It is quite typical for critical exponents to take saturating values in the lower critical dimension. We know that $\beta\delta \rightarrow \infty$ as $d \rightarrow 2$, but that does not tell us how β varies with dimension. Numerical simulations in higher dimensions [30] show β decreasing from its mean-field value $\beta_{MF} = 1/2$ in $d = 6$ down to $\beta_{3D} = 0.035 \pm 0.0280$ in three dimensions. It is natural to expect that $\beta = 0$ in two dimensions, and that the universal scaling function \mathcal{M} varies from -1 to 1 as the field increases (saturating the behavior). This implies that

$$\frac{dM}{dh}(h|w) = \eta(w)^{-1} \frac{d\mathcal{M}}{dh}(v) \quad (9)$$

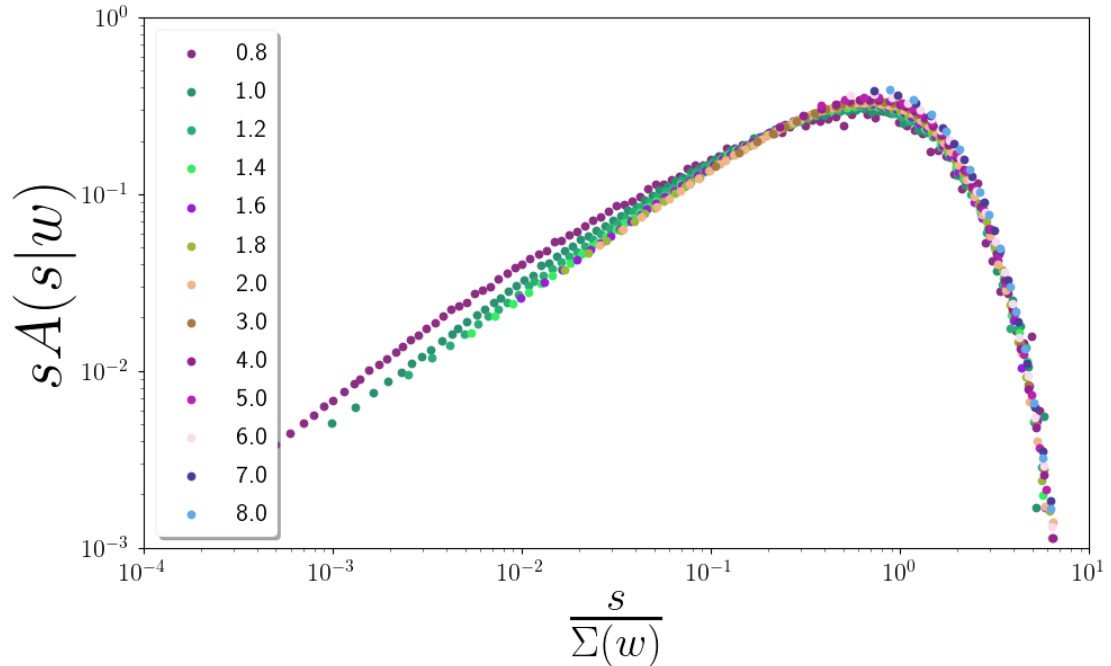


Figure 4: **Scaling collapse of the area weighted avalanche size distribution.** w ranges from 0.8 to 8.0. There is a slight bulge at $s/\Sigma(w) \sim 10^{-2}$ for small w .

where the invariant scaling combination

$$v = (h - h_{max})/\eta(w). \quad (10)$$

It is traditional to scale with $h - h_c$, but since $h_{max} - h_c \propto \eta(w)$, scaling to h_{max} is equivalent. The form chosen for the universal scaling function $\frac{dM}{dh}(h|w)$ is given in Appendix C. The associated collapse is shown in Figure 5.

2.5 PARAMETER VALUES

Through performing the scaling collapses we are provided with values of Σ and η for each value of disorder, r . Using the nonlinear scaling forms for each of these we may then extract values for the associated parameters. An unconstrained fit yields a fractal dimension larger than two, the dimension of the system, which is

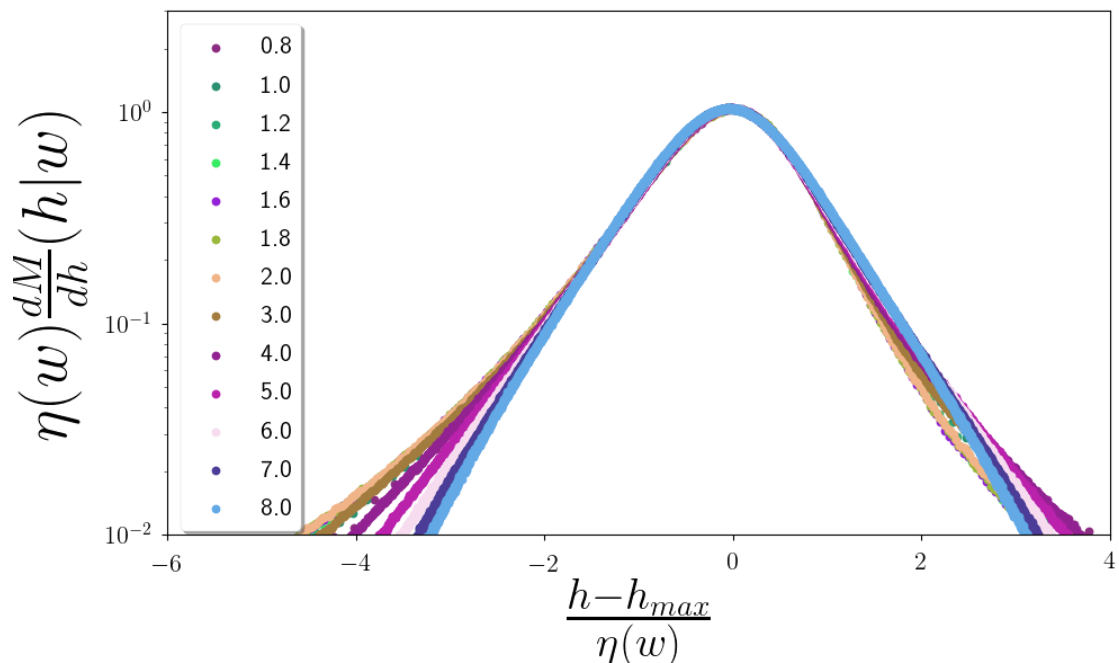


Figure 5: **Scaling collapse of the change in magnetization with respect to the field.** w ranges from 0.8 to 8.0.

unphysical. The 2D avalanches we consider appear compact. This suggests that the fractal dimension should be given by $d_f = 2$ and that the maximum avalanche size should scale as the square of the correlation length. For this reason, we expect also that $\Sigma(w) \sim \xi^2$ and set $C = 0$. Imposing these constraints, the fits obtained are able to describe the data well, as shown in Figure 6.

We expect the statistical errors and dependence on functional forms chosen for the universal scaling functions to be small. It is useful, however, to consider finite size effects at small r and lattice effects for large r . To compute these error bars, we performed the collapses and subsequent fits of the nonlinear forms using subsets of the disorders for which we have data [11 out of 13 points]. The errors given are the standard deviation of the values determined in this way. The best fit values with associated errors are given in Table 1. Likewise, the standard deviation for both Σ and η are provided for points in the overlap of the subsets. The error bars

	NF	NF_0	NF_{alt}	NF_{Harris}	Conjecture
r_c	-0.46 ± 0.06	0	0	-0.46 ± 0.06	$[-0.5, 0.0]$
λ_h	0.52 ± 0.07	0.24 ± 0.08	0.70 ± 0.05	1	1
B	-0.15 ± 0.01	0.039 ± 0.007	-0.76 ± 0.14	-0.25 ± 0.03	$[-0.8, 0.0]$
F	1.33 ± 0.12	2.02 ± 0.13	0.45 ± 0.04	0.45 ± 0.06	$[0.0, 0.5]$
C	0	1.76 ± 0.28	0	0	0
d_f	2	2	2	2	2

Table 1: **Table of the parameter values determined through a joint fit of $\Sigma(w)$ and $\eta(w)$.** NF corresponds to the transcritical form and NF_{alt} to the alternative transcritical form described in Appendix B.3. NF_0 corresponds to the transcritical form with $r_c = 0$ and NF_{Harris} to $\lambda_h = 1$, the Harris criteria. To compute the error bars, we performed the collapses and subsequent fits of the nonlinear forms using subsets of the disorders for which we have data [11 out of 13 points]. The errors given are the standard deviation of the values determined in this way. Values in bold were fixed in the corresponding fit.

for Σ and η are smaller than the datapoints (Figure 6).

Note that the best fit value of r_c is found to be less than zero. There are several possible explanations for this. One, $r_c < 0$ could indicate the Voronoi lattice used introduces an amount of intrinsic disorder. This is certainly plausible as random bond and random field disorder are expected to belong to the same universality class [40, 110]. Alternatively, constraining $r_c = 0$ we obtain a comparable fit by including an alternative normal form, NF_{alt} , differing from $\Sigma(w)$ and by analytic corrections to scaling (expected for the larger disorders considered). This form is described in Appendix B.3. In either case, the results are consistent with $r_c = 0$.

2.6 COMPARISON OF FORMS

As a test of our finding that the 2D NE-RFIM corresponds to a transcritical bifurcation, we may compare the fits obtained to those using different underlying assumptions. In particular, it is straightforward to calculate Σ and η assuming a hyperbolic fixed point (corresponding to power law scaling: Appendix B.1) and a pitchfork bifurcation (Appendix B.4). For each of these cases we can perform a fit

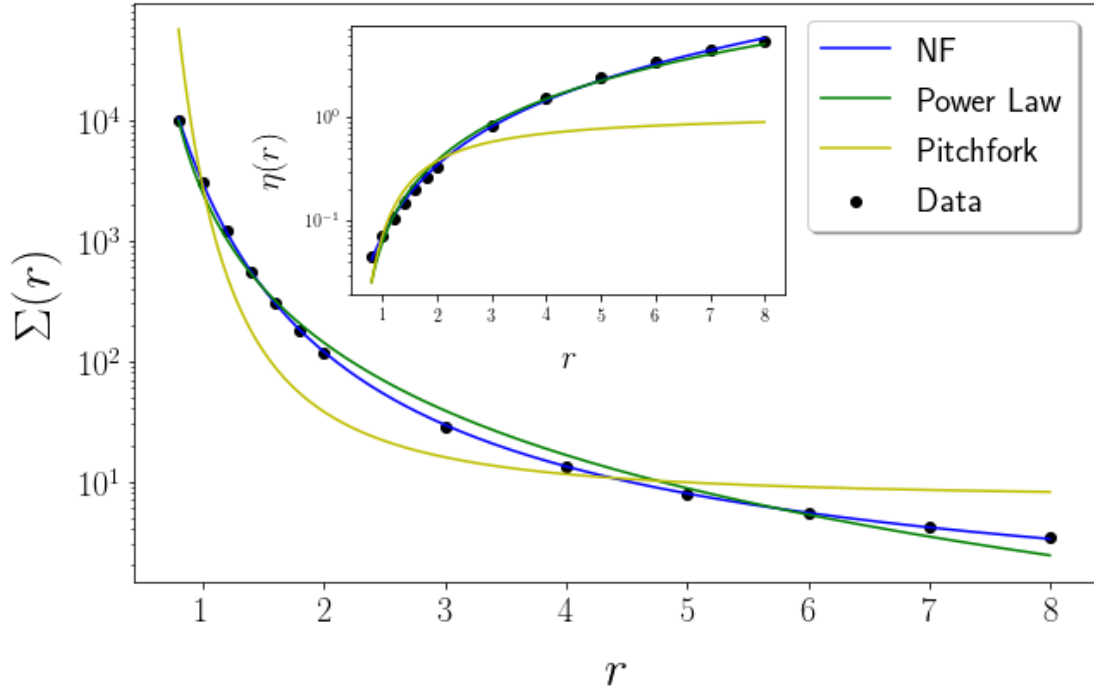


Figure 6: **Comparison of the best fit of $\Sigma(w)$ and $\eta(w)$ derived with different functional forms of $\frac{dw}{dt}$.** We have $w = (r - r_c)/s_s$ such that $\Sigma(r) = \Sigma(w)$ and $\eta(r) = \eta(w)$. ‘NF’ corresponds to Σ and η derived from the transcritical normal form, ‘Power Law’ the hyperbolic (power law) form and ‘Pitchfork’ the pitchfork form.

to the values of $\Sigma(w)$ and $\eta(w)$ extracted from the collapse. The comparison of these fits are shown in Figure 6.

It is particularly illuminating to consider the behavior of $1/\log \Sigma(w)$. For a transcritical bifurcation, the exponential divergence (ignoring B and C in Equation 5) gives

$$\frac{1}{\log \Sigma(w)} \sim \frac{1}{d_f} w. \quad (11)$$

Hence, if the behavior corresponds to a transcritical bifurcation, we would expect a plot of $1/\log \Sigma$ to scale linearly with the disorder. A comparison of the linear fit to $1/\log \Sigma$, along with the plots of $1/\log \Sigma$ for the best fits with a power law and pitchfork form are shown in Figure 7. The results clearly support a transcritical

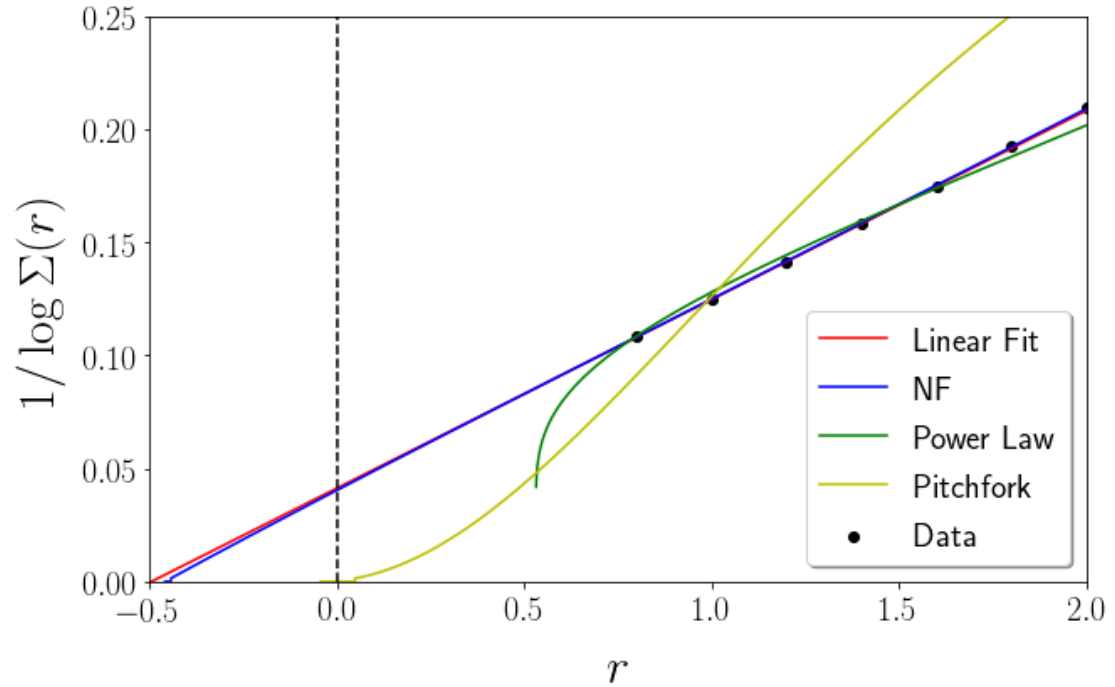


Figure 7: **Comparison of $1/\log \Sigma(w)$ for the best fit of $\Sigma(w)$ derived with different functional forms of $\frac{dw}{dt}$.** We have $w = (r - r_c)/s_s$ such that $\Sigma(r) = \Sigma(w)$. ‘NF’ corresponds to Σ derived from the transcritical normal form, ‘Power Law’ the hyperbolic (power law) form and ‘Pitchfork’ the pitchfork form.

bifurcation, with $r_c < 0$, and challenge the alternative power law and pitchfork assumptions.

2.7 DISCUSSION

Simulation data of the 2D non-equilibrium random-field Ising model on a lattice which suppresses faceting is explained well by the presence of a transcritical bifurcation, and is incompatible with power law scaling or pitchfork normal forms without large corrections to scaling. This provides evidence that (1) the universality class of the equilibrium and non-equilibrium models are indeed different and that (2) power law scaling (which is governed by a hyperbolic fixed point) is not the correct approach for this system in this regime. The latter conclusion, in turn, is consistent with (3) the LCD of the model being equal to two, or perhaps close to two.

Although the transcritical bifurcation provides the best description of our simulation data, the corresponding parameter values are difficult to pin down. There are a number of restrictions we can make to the parameter values and still obtain a reasonable joint fit of $\Sigma(w)$ and $\eta(w)$. For example, we may require that the Harris criteria saturates, that $r_c = 0$ [30] or that the coefficient of the quintic order term $B = 0$. Each of these provides a good description of our data and are discussed in Appendix E.

In three and higher dimensions [30, 64], measuring a variety of avalanche properties was crucial in pinning down the universal critical exponents and scaling functions. dM/dH and the cumulative avalanche size distribution, measured here, were supplemented by measurements of finite-size scaling, avalanche correlation functions, avalanche sizes binned in H , spanning avalanches, avalanche durations, and average avalanche temporal shapes. Larger system sizes should be possible

with improved Voronoi data structures; Fig. 7 implies that whether $r_c = 0$ or is negative would be definitively answered by a simulation big enough to contain avalanches with $1/\log(\Sigma) = 0.05$, so with $L \sim \sqrt{(\Sigma)} = e^{10} \approx 22,000$.

In summation, performing large scale simulations on a Voronoi lattice and analyzing the RG flow equations yields valuable insight into the behavior of the NE-RFIM in 2D. The obtained scaling collapses span over a range of a factor of ten in the disorder and a factor of 10^4 in avalanche cutoff. They are consistent with a critical disorder at zero and with a lower critical dimension for the model equal to two.

This work was partially supported by NSF grants DMR-1719490 and DGE-1144153. AR acknowledges support from the Simons Foundation. We thank A. Alan Middleton, Gilles Tarjus, and Karin A. Dahmen for helpful discussions.

3 MACHINE LEARNING INTRODUCTION

In the preceding section, data analysis was performed in a very standard way for the scientific community. We developed a theory informed by domain knowledge and worked through the resulting mathematics to make predictions. If the underlying mechanism corresponds to the theory, it is expected that the behavior of the data will be well described by the model.

The following sections deal with a philosophically different approach. The aim of machine learning (ML) is to make these types of predictions in cases where the underlying theory would be too time-consuming or even intractable with current methods. In other words, it attempts to remove the requirement of domain knowledge with the trade off being a lack of insight into the underlying mechanisms that produced the data and a necessary 'domain knowledge' of different ML approaches and their applicability.

The following sections deal with three investigations into ML from the perspective of physics, finance, and information geometry. In Section 4, we present the results of applying an unsupervised algorithm to stock return data. The geometry of the data suggests the appropriate approach and leads to rich predictions about the exposure of companies to different sectors of the economy, changes in company composition with time and relations between sectors of the economy itself ². Following this, we turn our attention to computational neural network algorithms. Section 5 evaluates the subtleties of selecting and implementing an algorithm and includes a discussion of whether physics domain knowledge can be leveraged ef-

²This work was published in Quantitative Finance [53]. Ricky Chachra collected the data, performed the archetypal analysis and wrote the first draft of the manuscript and supplement. Lorien X. Hayden investigated the effects of noise in determining company changes in time, performed the r^2 calculations to add a comparative study of the model with that of Fama and French and analysed the changes in the decomposition with dimension. Lorien X. Hayden revised subsequent drafts and prepared the work for publication. Alexander A. Alemi, Paul H. Ginsparg and James P. Sethna collaborated on all aspects of the project.

fectively when asking questions of a physical nature. In Section 6, we turn our attention to an analysis of these networks from the viewpoint of information geometry. With the lens chosen, neural networks appear not to fall into the class of 'sloppy' models studied extensively by the Sethna group. The results, however, are somewhat ambiguous and suggest a number of potential avenues for understanding the relation between the model these algorithms learn and high dimensional geometry.

4 CANONICAL SECTORS AND EVOLUTION OF FIRMS IN THE US STOCK MARKETS

Classifying companies participating in the US stock markets based upon their participation in various sectors of the economy is important for macroeconomic analysis and for investments into the sector-specific financial indices and exchange traded funds (ETFs). Major industrial classification systems and financial indices have historically been based on expert opinion and developed manually. The extensive amount of data present on the performance of these companies, however, lends itself readily to analysis via machine learning methods. Examining the data, a low-dimensional structure in the space of historical stock price returns emerges which naturally suggests unsupervised ML via Archetypal Analysis (AA) [39]. Implementing this algorithm to determine the convex hull of the dataset automatically identifies 'canonical sectors' in the market, and assigns every stock a participation weight into these sectors. In particular, it provides an unsupervised way to generate a more objective and comprehensive broad-level sector decomposition of stocks.

Stock market performance itself is measured with aggregated quantities called indices that represent a weighted average price of a basket of stocks. Market-wide indices such as Russell 3000 [92] and the S&P 500 [101] consist of stocks from diverse companies reflecting a broad cross-section of the market. Sector-specific indices such as the Dow Jones Financials Index [42], CBOE Oil Index [26] and the Morgan Stanley High-Tech 35 Index [77], etc., are more granular and their composition requires a classification of companies into sectors. The major industrial classification schemes used to perform this classification do so with many ambiguities [80]. It is not clear, for example, how to assign a sector to

conglomerates or diversified companies such as General Electric. Conversely, non-conglomerates with exposure to firms outside their own sector (for example, an investment bank exclusively serving pharmaceutical firms) also blur the boundaries of sector-identification. Moreover, as economic environment or companies evolve, neither the industrial sectors nor the firms' sector association remains static, necessitating updates to sector assignments and addition of new sectors.

A significant number of studies have previously aimed at finding categories of stocks in financial markets with a variety of approaches. Recent numerical techniques have included extensive use of random matrix theory, principal component analysis or associated eigenvalue decomposition of the correlation matrix [36, 37, 46, 48, 59, 84], specialized clustering methods [5, 13, 14, 54, 62, 73, 79] or time series analysis [75, 86], pairwise coupling analysis [24], and even topic-modeling of returns [43]. Indeed, relevant prior work analyzing historical stock price returns [47, 66, 84] elucidated that the high-dimensional space of stock price returns has a low-dimensional representation.

In parallel with this, there is a long tradition of style analysis in finance in which time series can be selected which serve as useful benchmarks for the performance of other stocks or indices. The 3-factor model of Fama and French [47] is one such example. Recently, D. Vistocco and C. Conversano [109] proposed that Archetypal Analysis (AA) [39] could provide these benchmark time series while also providing a way to plot this data in a meaningful way. In particular, they provide a triangular plot for Italian mutual funds and suggest parallel coordinate plots or asymmetric maps for higher dimensional representations.

In contrast to previous studies, the method presented here provides a new holistic way of classifying stocks into industrial sectors by utilizing the emergent structure of price returns in data space. Beyond the proposal of Vistocco and Conversano, we provide an interpretation of the archetypes of AA as sectors of the

economy. This structure is purely contained in the geometry of the time series. Other methods, such as SVD, can discern that there is some such structure but are not well suited to a clean description. AA, on the other hand, determines the convex hull of the dataset making it uniquely suited to creating a quantitative analysis of the data.

4.1 ARCHETYPAL ANALYSIS

To prepare the raw data, we take the log price returns of individual stocks, remove the overall market return, and normalize to zero mean and unit standard deviation. The matrix of daily log returns of a stock s are defined as $r_{ts} = \log P_{ts} - \log P_{(t-1)s}$ where P_{ts} are adjusted closing prices (i.e. corrected for stock splits and dividend issues) and t is in trading days. The normalized returns are then given by $R'_{ts} = (r_{ts} - \langle r_{ts} \rangle_t) / \sigma_s$, where $\sigma_s^2 = \langle r_{ts}^2 \rangle_t - \langle r_{ts} \rangle_t^2$ is the variance (squared volatility). Removing the overall market returns from each stock yields $R_{ts} = R'_{ts} - \langle R'_{ts} \rangle_s$. For details see [53].

The stock returns represented by R_{ts} are well-approximated by a hyper-tetrahedral structure. The hyper-tetrahedron, or simplex, which emerges is a self-organized structure: it has prototypical firms in corners, closely related firms clumped together in each lobe, diversified companies (GE, Walt Disney, 3M, etc.) close to the center, and the number of lobes denoting how many distinct sectors are exhibited by the data. Each lobe of the hyper-tetrahedron is populated by stocks of similar or related businesses. A PCA projection of this hyper-tetrahedron is shown in Figure 8. The lobe-corners, or vertices of the tetrahedron, correspond to *canonical sectors* and approximate the returns of companies that are prototypical of individual sectors. Table 2 details the prototypical companies of each. This suggests a natural way to decompose stocks into canonical sectors: for convex sets, each

interior point is representable as a unique weighted sum of corner points, implying here that every stock’s return is approximated by a weighted sum of returns from the canonical sectors. The weights for a given stock quantify its exposure to the canonical sectors.

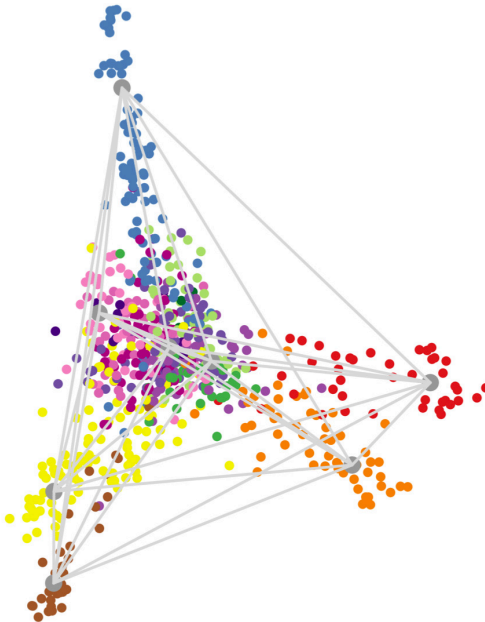


Figure 8: **Low-dimensional projection of the stock price returns data.** Stock price returns are projected onto a plane spanned by two stiff vectors from the SVD of the emergent simplex corners [53, Supplement]. Each colored circle corresponds to one of the 705 stocks in the dataset used in the analysis. Colors denote the sectors assigned to companies by Scottrade [97]. The grey corners of the simplex correspond to sector-defining prototype stocks, whereas all other circles are given by a suitably weighted sum of these grey corners.

To perform the decomposition, we applied an in house python implementation of the AA algorithm described by Mørup and Hansen [78]. The AA factorization is defined as:

$$\begin{aligned}
 R_{ts} &\sim R_{ts'}C_{s'f}W_{fs} \\
 C_{s'f} &\geq 0, \sum_{s'} C_{s'f} = 1, \\
 W_{fs} &\geq 0, \sum_f W_{fs} = 1.
 \end{aligned}
 \tag{12}$$

Columns of $R_{ts}C_{s'f} = E_{tf}$ are the emergent sector time series (basis vectors) rep-

Canonical sector	Prototypical examples
<i>c-cyclical</i>	Gap, Macy’s, Target
<i>c-energy</i>	Halliburton, Schlumberger
<i>c-financial</i>	US Bancorp., Bank of America
<i>c-industrial</i>	Kennametal, Regal-Beloit
<i>c-non-cyclical</i>	Pepsi, Procter & Gamble
<i>c-real estate</i>	Post Properties, Duke Realty
<i>c-technology</i>	Cisco, Texas Instruments
<i>c-utility</i>	Duke Energy, Wisconsin Energy

Table 2: **Canonical sectors and major business lines of primary constituent firms.** The eight canonical sectors identified by the analysis described here are listed in the column on the left; these were named in accord with the business lines of firms that show strong association with these sectors. Some examples are provided in the right column; a full list is available on companion website [113].

representing the n corners of the hyper-tetrahedron, and W_{fs} are the participation weights ($W_{fs} \geq 0$) in sector f so that $\sum_f W_{fs} = 1$ for each stock s . The sector matrix E_{tf} is within the convex hull ($C > 0$, $\sum_s C_{sf} = 1$) of the data R_{ts} . It can be found by either minimizing the squared error with convex constraints in factorization as originally proposed [39], or by making a convex hull in low-dimensions and choosing one or more of its vertices to be basis vectors [105], or by minimizing after initializing with candidate archetypes that are guaranteed to lie in the minimal convex set of the data [78] which is the approach we employ. For details on convergence and other consistency checks see [53].

The dataset consisted of 705 US firms’ stocks with a minimum \$1 billion June 2013 market capitalization and with continuous 20 years (1993–2013) of listing on major exchanges. Analysis of this dataset revealed eight emergent sectors which were named in accordance with the companies they comprised (prefix *c-* denotes “canonical”): *c-cyclical* (including retail), *c-energy* (including oil and gas), *c-industrial* (including capital goods and basic materials), *c-financial*, *c-non-cyclical* (including healthcare and consumer non-cyclical goods), *c-real estate*, *c-technology*, and *c-utility*. Calculated participation weights for a sample of 12 firms in (Fig-

ure 9) show a decomposition of their stocks into the canonical sectors with resulting insights discussed in the caption. Associated with each canonical sector is a time series of returns. As expected, these series show hallmark historical events of individual sectors (Figure 10): the dot-com bubble, the energy crisis, and the financial crisis being the major events in the last two decades.

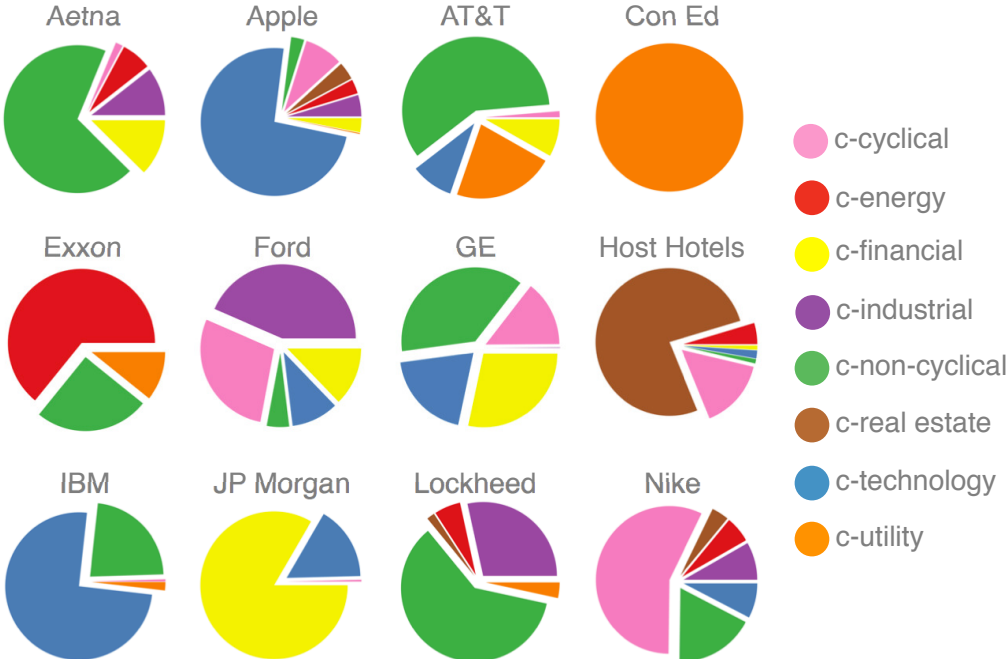


Figure 9: **Canonical sector decomposition of stocks of selected companies.** A complete set of all 705 stocks is provided on the companion website [113]; the color scheme is shown on the right. Conglomerates like GE decompose roughly into their core business lines. Tech firms such as Apple that sell mass-market consumer goods have an important fraction in *c-cyclical*, whereas IBM has a significant portion of *c-non-cyclical* returns presumably due to its government contracts. Telecom companies like AT&T are generally classified under a separate telecom category by major classification systems, yet analysis shows their returns are described by a combination of *c-non-cyclical* and *c-utility* sectors. Health insurance providers like Aetna are commonly classified as financial services firms, but their returns consist of a major part *c-non-cyclical* and only a minor part of *c-financial*—the health-care sector is generally less prone to economic downturns. Defense contractors like Lockheed are listed as capital goods companies, but their returns are seen to be majority *c-non-cyclical* and only a smaller share of *c-industrial* sector.

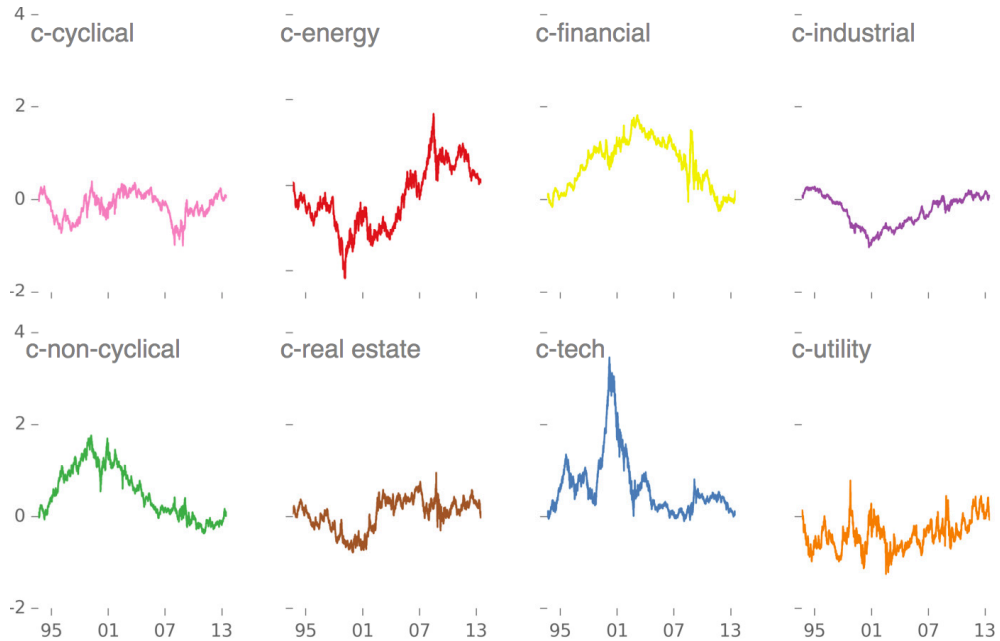


Figure 10: **Emergent sector time series.** Annualized cumulative log price returns of the eight emergent sectors are shown. The time series capture all important features affecting different sectors: building-up of the dot-com bubble (c. 2000) followed by a burst, the soaring energy valuations (2003–08) followed by a crash, and financial crisis of 2008. We note that the dot-com bubble was confined to the c-tech whereas the financial crisis effects were spread throughout the sectors. Precise definition of the cumulative returns plotted here is given in [53].

4.2 COMPANY EVOLUTION IN TIME

The formal framework of AA applied to stock return data lays the foundation for several interesting extensions. In addition to the full data set of $20 \text{ years} \times 705$ firms, we also applied the algorithm to overlapping, two-year Gaussian windows to study to how the sector weights for firms have evolved in time. We decomposed the local normalized log returns for each stock into the canonical sectors determined from the entire time series. Each column (time series) of the returns matrix R_{ts} was multiplied with a Gaussian, $G_{\mu}(\tau) = \exp(-(\tau - \mu)^2/(2 \times 250^2))$ of standard deviation 250 centered at μ to obtain R_{ts}^{μ} . We use $C_{s'f}$ found using the full dataset (Eqn. 12) (corresponding to keeping the sector-defining simplex corners fixed).

R_{ts}^μ is factorized to obtain new weights W_{fs}^μ that describe sector decomposition of stocks in that period focused at $t = \mu$: $R^\mu = R_{ts}^\mu C_{s'f} W_{fs}^\mu$. μ is increased in steps of 50 starting at $\mu = 0$ and ending at $\mu = 5000$, and W^μ is calculated at each μ with the corresponding R^μ . These results are plotted in Figure 11 for a select group of companies; the remainder are available on the companion website [113].

As expected, the sector decomposition of firms is dynamic. Mergers, acquisitions, spin-offs, new products, effect of competitive environments or shifting consumer preferences can change the business foci of firms and hence alter the sector association of firms. External events affecting companies in an idiosyncratic manner also show clear signature in this analysis.

To address the challenge of distinguishing signal from noise in the evolving sector weights, we simulate data to which we add noise and then compare. This was done by repeating the analysis for the flows where the companies from Figure 11 were replaced. For each of these companies, we took its sector weights, $\vec{\omega}_f$, and multiplied by E_{tf} to obtain a time series for the company with weights that are constant in time. We then added gaussian random noise with standard deviation one and replaced these companies by this simulated data. Figure 12 shows the comparison between the real flows from the main text and the simulated constant data with noise added. General features described in the text are shown to be signal while small fluctuations are consistent with noise.

4.3 NUMBER OF CANONICAL SECTORS

Determining the correct number of canonical sectors that appropriately describe the space of stock market returns is akin to the more general issue of selecting a signal-to-noise ratio cutoff, or a truncation threshold in the dimensional-reduction of data. The choice of this threshold is generally sensitive to sampling,

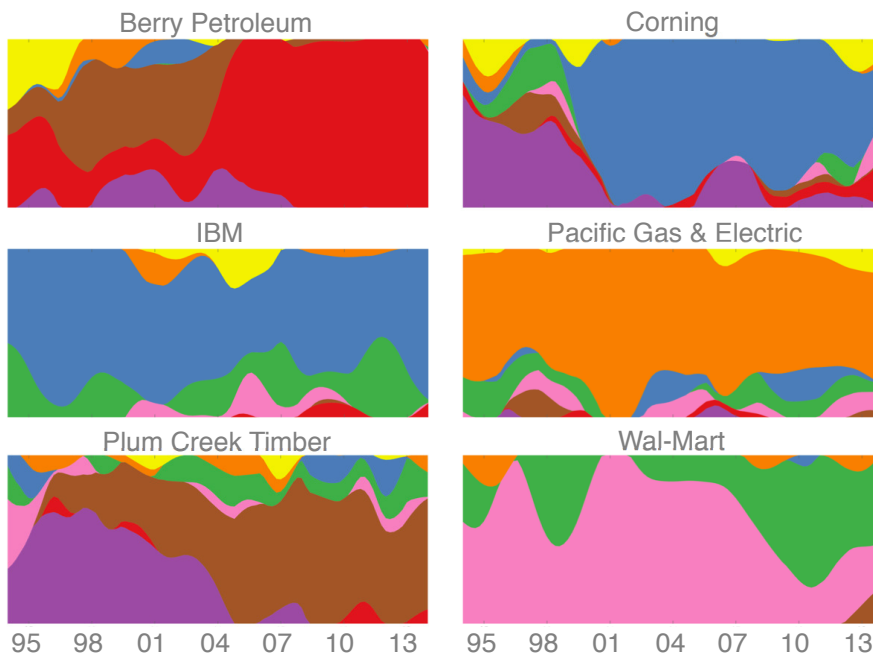


Figure 11: **Evolving sector participation weights.** Results from the sector decomposition made with rolling two-year Gaussian windows are shown for selected stocks. A complete set of 705 charts is provided on the companion website [113]. Color scheme is as in (Figure 9). For stable and focused companies such as Pacific Gas & Electric or IBM, one sees no significant shifts in sector weights; changes in time agree with errors expected from unresolved fluctuations [113]. Wal-Mart’s returns, on the other hand, have moved significantly from *c-cyclical* to *c-non-cyclicals* (consumer staples) in the post-financial crises years as shown; this is also true of other low-price consumer commodities retailers such as Costco, but not true of higher price retailers such as Whole Foods, Macy’s, etc. Corning, previously an *industrial* firm with a huge presence in optical fiber, suffered in the aftermath of the dot-com crisis and now is classified as a *tech* firm presumably due to its Gorilla[®] glass used in cellphones, laptop displays, and tablets. Berry Petroleum grew within its home state of California in the early 1990s through development on properties that were purchased in the earlier part of 20th century. In 2003, the company embarked on a transformation [10] by direct acquisition of light oil and natural gas production facilities outside California. The figure shows a clear shift in the distribution of sector weights as the company has moved toward *c-energy* and away from *c-real estate*. Similarly, as Plum Creek Timber converted to a real estate investment trust (REIT) in the late 1990s [85], its sector weights have significantly shifted toward *c-real estate* sector.

yet the results presented here are reasonably robust with different choices leading to meaningful and similar decompositions. It is an open problem to determine the

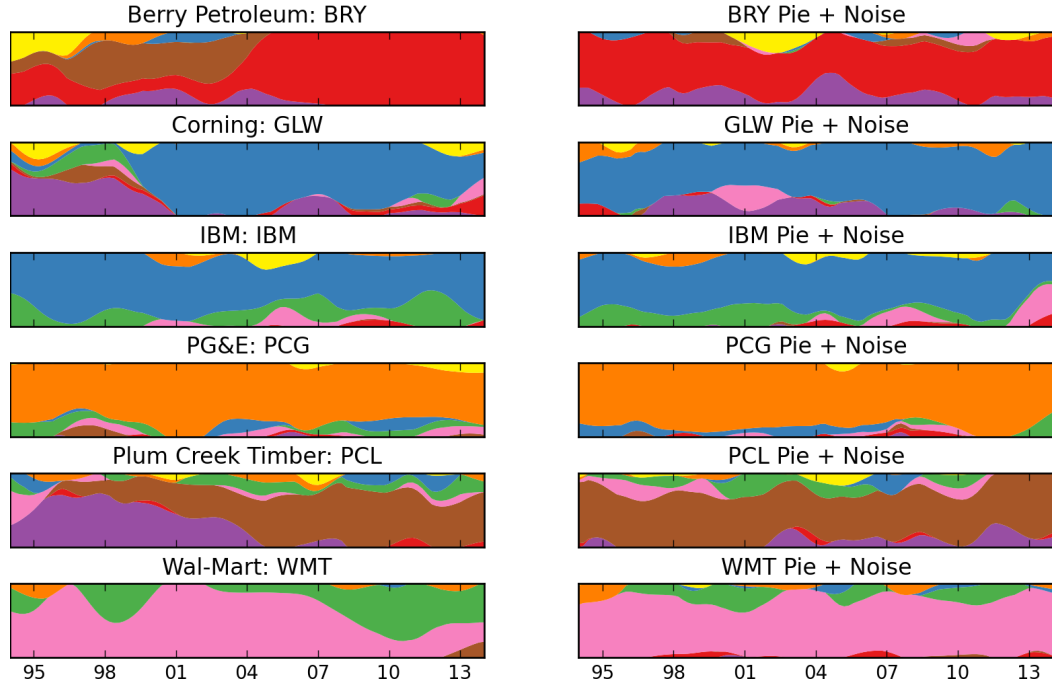


Figure 12: **Comparison between flow diagrams presented in Figure 11 with simulated data.** The simulated data is created from the dot product of the weight vector of the company with the corner time series. This yields a version of the company with constant weights in time. To this we add gaussian noise with standard deviation one and repeat the analysis to generate the flows in time. In the left column are the actual flows for companies, on the right is their constant in time counterpart with added noise. We see that key features noted are in fact signal while small fluctuations correspond to noise.

effective dimensionality (optimal rank) of a general dataset (matrix). One could select among models of different dimensions using statistical tests such as the r^2 discussed below, or information theory based criteria such as Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC), but the choice of the selection criterion is itself generally made on an *ad hoc* basis.

4.3.1 SECTOR RELATIONSHIPS

One approach to investigating how the sector decomposition changes with dimension is to produce a flow diagram. To do this, we performed the fit

$$\|E_{t,f} - E_{t,f'}S_{f',f}\|_F^2 \quad (13)$$

with the constraint $\sum_{f'} S_{f',f} = 1$. Hence the sectors for $n = 9$ can be expressed as a linear combination of sectors for $n = 8$, $n = 8$ as a linear combination of $n = 7$, and so forth. The results of these fits are presented in Figure 13. The figure represents these relationships through connections between the decompositions for $n = N + 1$ and $n = N$ weighted according to the matrix $S^{(N,N+1)}$. More precisely, we create a node corresponding to each of the 9 sectors whose size is proportional to $\sum_s W_{f,s}$ where $W_{f,s}$ is the weight matrix for the 9 sector decomposition. Hence, the relative node sizes represent the amount of the market participating in the sector. Multiplying this vector by $S^{(8,9)}$ gives the approximate size for each node in $n = 8$. Multiplying this vector by $S^{(7,8)}$ gives the approximate size for each node in $n = 7$, and so on. In this way, we generate a Sankey diagram whose node sizes correspond roughly to the amount of the market in the sector and whose connections depict how strongly the sectors for decompositions with different n overlap. In the image, we see that the $n = 9$ decomposition gives the 8 sector version with an additional small sector whose companies are listed in the paper [53, Supplement]. We also see that for $n = 7$, *c-finance* and *c-real estate* merge. At $n = 6$, *c-industrial* and *c-cyclical* merge. For $n = 5$, the new sector containing *c-industrial* and *c-cyclical* merges with *c-non-cyclical*. For $n = 4$, *c-utility* and *c-energy* merge. Finally, for $n = 3$ and $n = 2$, no clear pattern emerges given this image alone.

4.3.2 TWO AND THREE FACTOR DECOMPOSITIONS

We further explore the two and three sector decompositions by examining their constituent companies and looking at pie charts describing the relationship between our 8 sector decomposition and those with $n = 2$ and $n = 3$ respectively. Recall that each archetype is constrained to be a linear combination of companies, or in other words to lie in the convex hull of the data. Using this information, we list the 20 companies which contribute the most to each sector in the two and three factor decompositions (Tables 3, 4, and 5). For the two sector decomposition, we find the sectors divide roughly into *c-assets* (e.g. financial and real estate companies) and *c-goods* (e.g. companies which provide goods and services). For $n = 3$, the division is less clear. Another way to look at the constituents of these sectors is by examining pie chart representations of these decompositions. Again consider the fit $\|E_{t,f} - E_{t,f'}S_{f',f}\|_F^2$ with the constraint $\sum_{f'} S_{f',f} = 1$. Applying this, we can express the two sector archetypes as linear combinations of the 8 sector archetypes and vice versa. Additionally, we can do the same for the three factor decomposition. The pie charts these fits produce are shown in Figure 14. The results are consistent with the sector breakdowns described from examining the constituent companies.

4.4 COEFFICIENT OF DETERMINATION

Another method to examine the 'goodness' of the returns decomposition $R = EW$ can be given by measuring the coefficient of determination (r^2) as follows:

$$r^2 = 1 - SSE/SST \tag{14}$$

c-assets	label	percent	full name	c-goods	label	percent	full name
DDR	real estate	1.77%	DDR Corp.	HON	tech	0.53%	Honeywell International Inc.
ONB	financial	1.7%	Old National Bankcorp.	TMO	health	0.51%	Thermo Fisher Scientific Inc.
BRE	real estate	1.66%	Brookfield Real Estate Serv.	NAV	cyclical	0.49%	Navistar International Corp.
PEI	real estate	1.54%	Pennsylvania RIT	CSL	basic	0.47%	Carlisle Companies Inc.
FMBI	financial	1.5%	First Midwest Bancorp. Inc.	IRF	tech	0.47%	International Rectifier Corp.
PRK	financial	1.5%	Park National Corp.	APD	basic	0.46%	Air Products & Chemicals Inc.
BAC	financial	1.42%	Bank of America Corp.	PCP	basic	0.43%	Precision Castparts Corp.
STI	financial	1.41%	SunTrust Banks Inc.	OMC	misc services	0.43%	Omnicom Group Inc.
DRE	real estate	1.29%	Duke Realty Corp.	MXIM	tech	0.43%	Maxim Integrated Products, Inc.
UBSI	financial	1.28%	United Bankshares Inc.	TFX	health	0.41%	Teleflex Inc.
CPT	real estate	1.28%	Camden Property Trust	NSC	transport	0.41%	Norfolk Southern Corp.
PPS	real estate	1.28%	Post Properties Inc.	NBL	energy	0.4%	Noble Energy Inc.
WABC	financial	1.26%	Westamerica Bancorp.	SM	energy	0.4%	SM Energy Company
FMER	financial	1.26%	FirstMerit Corp.	WMT	retail	0.39%	Wal-Mart Stores Inc.
CNA	financial	1.26%	CNA Financial Corp.	CR	basic	0.38%	Crane Co.
VLY	financial	1.25%	Valley National Bancorp.	ADI	tech	0.38%	Analog Devices Inc.
MTB	financial	1.24%	M&T Bankcorp.	ITW	cyclical	0.38%	Illinois Tool Works Inc.
WRI	real estate	1.23%	Weingarten Realty Investors	PPG	basic	0.38%	PPG Industries Inc.
BDN	real estate	1.21%	Brandywine Realty Trust	BA	capital	0.38%	The Boeing Company
ZION	financial	1.2%	Zions Bancorp.	AME	tech	0.38%	Ametek Inc.
Total		27.54%		Total		8.53%	

Table 3: **Top 20 contributing companies to each sector in the two sector decomposition.** Ranking is determined by the matrix $C_{s,f}$ which describes each sector as a linear combination of stocks. Labels are those given by Scottrade and percentage describes the percentage of the sector attributable to the company.

sector 1	label	percent	sector 2	label	percent	sector 3	label	percent
XOM	energy	1.29%	BRE	real estate	2.16%	IRF	tech	1.29%
HP	energy	1.22%	PEI	real estate	2.08%	EMC	tech	1.22%
CVX	energy	1.21%	BWS	retail	1.99%	ADI	tech	1.21%
ETR	utility	1.2%	CNA	financial	1.79%	CSCO	tech	1.2%
APD	basic	1.2%	ONB	financial	1.73%	TXN	tech	1.2%
OXY	energy	1.19%	DDR	real estate	1.63%	BMC	tech	1.19%
NFG	utility	1.18%	PRK	financial	1.59%	SNPS	tech	1.18%
PX	basic	1.17%	CBSH	financial	1.59%	PLXS	tech	1.17%
CL	non-cyclical	1.16%	BC	cyclical	1.56%	CPWR	tech	1.16%
NBL	energy	1.15%	FMER	financial	1.55%	AVT	tech	1.15%
OII	energy	1.11%	RDN	financial	1.54%	SWKS	tech	1.11%
LNT	utility	1.11%	MAS	capital	1.54%	HPQ	tech	1.11%
D	utility	1.08%	DDS	retail	1.47%	PMCS	tech	1.08%
DTE	utility	1.07%	FMBI	financial	1.47%	MXIM	tech	1.07%
SCG	utility	1.06%	ALK	transport	1.46%	ARW	tech	1.06%
WEC	utility	1.04%	WABC	financial	1.43%	TER	tech	1.04%
APA	energy	0.99%	PCH	real estate	1.42%	ATML	tech	0.99%
BAX	health	0.98%	VLY	financial	1.41%	MCHP	tech	0.98%
MUR	energy	0.98%	BAC	financial	1.41%	LRCX	tech	0.98%
CPB	non-cyclical	0.98%	STI	financial	1.37%	CGNX	tech	0.98%
Total		22.38%	Total		19.14%	Total		32.18%

Table 4: **Top 20 contributing companies to each sector in the three sector decomposition.** Ranking is determined by the matrix $C_{s,f}$ which describes each sector as a linear combination of stocks. Labels are those given by Scottrade and percentage describes the percentage of the sector attributable to the company.

sector 1	full name	sector 2	full name	sector 3	full name
XOM	Exxon Mobil Corp.	BRE	Brookfield Real Estate Serv.	IRF	International Rectifier Corp.
HP	Helmerich & Payne Inc.	PEI	Pennsylvania RIT	EMC	EMC Corp.
CVX	Chevron Corp.	BWS	Brown Shoe Co. Inc.	ADI	Analog Devices Inc.
ETR	Energy Corp.	CNA	CNA Financial Corp.	CSCO	Cisco Systems Inc.
APD	Air Products & Chemicals Inc.	ONB	Old National Bancorp.	TXN	Texas Instruments Inc.
OXY	Occidental Petroleum	DDR	DDR Corp.	BMC	BMC Software Inc.
NFG	National Fuel Gas Company	PRK	Park National Corp.	SNPS	Synopsys Inc.
PX	Praxair Inc.	CBSH	Commerce Bancshares Inc.	PLXS	Plexus Corp.
CL	Colgate-Palmolive Co.	BC	Brunswick Corp.	CPWR	Compuware Corp.
NBL	Noble Energy Inc.	FMER	FirstMerit Corp.	AVT	Avnet Inc.
OII	Oceanenergy International Inc.	RDN	Radian Group Inc.	SWKS	Skyworks Solutions Inc.
LNT	Alliant Energy Corp.	MAS	Masco Corp.	HPQ	Hewlett-Packard Company
D	Dominion Resources Inc.	DDS	Dillard's Inc.	PMCS	PMC-Sierra Inc.
DTE	DTE Energy Corp.	FMBI	First Midwest Bancorp. Inc.	MXIM	Maxim Integrated Products Inc.
SCG	SCANA Corp.	ALK	Alaska Air Group Inc.	ARW	Arrow Electronics Inc.
WEC	Wisconsin Energy Corp.	WABC	Westamerica Bancorp.	TER	Teradyne Inc.
APA	Apache Corp.	PCH	Potlatch Corp.	ATML	Atmel Corp.
BAX	Baxter International Inc.	VLY	Valley National Bancorp.	MCHP	Microchip Technology Inc.
MUR	Murphy Oil Corp.	BAC	Bank of America Corp.	LRCX	Lam Research Corp.
CPB	Campbell Soup Company	STI	SunTrust Banks Inc.	CGNX	Cognex Corp.

Table 5: **Top 20 contributing companies to each sector in the three sector decomposition.** Ranking is determined by the matrix $C'_{s,f}$ which describes each sector as a linear combination of stocks. (Accompaniment to Table 4.)

Here, SSE is denotes the sum of square errors $\|R - EW\|_F^2$, and SST is the total sum of squares $\|R\|_F^2$. This is also known as the *proportion of variance explained* (PVE). For the factorization of the full dataset, normalized with the market mode removed, the calculated r^2 value is 11.1%. The SVD of R with singular values shown in (Figure 15) provides a convenient way to put this number in context for the returns dataset. Only 20 singular values (excluding the market mode) were above the cut-off that was predicted by random matrix theory for a matrix of purely random Gaussian entries. For any matrix M with elements m_{ij} , the norm $\|M\|_F^2 = \sum_{i,j} m_{ij}^2 = \sum_i s_i^2$, where s_i are the singular values [87]. Thus, the fraction of intrinsic variation in R above the cutoff is the sum of squares of the 20 singular values (not including market mode) divided by SST, $\sum_{i=1}^{i=20} s_i^2 / \|R\|_F^2 = 19.8\%$. Therefore, as a first approximation, the factorization explains $11.1/19.8 = 56\%$ of the *random matrix theory (RMT) explainable variation*. For reference we provide the RMT explainable variation for the factor decomposition of Fama and French, the classification by Scottrade, and the top 8 singular vectors given by SVD. The percentage of the RMT explainable variation for different numbers of factors compared to the 3 factor decomposition of Fama and French is shown in (Table 6). Fama and French have the benefit of allowing factors to have positive or negative weights. In order to compare with another non-negative decomposition, we fix the weight matrix according to the Scottrade labels and run archetypal analysis for this $n = 14$ factor version. The r^2 value for this decomposition is 10.7% with a corresponding RMT explainable variance of 54.2% compared to 56% for our 8 factors. For completeness, we also note that if R is rank-reduced to the eight stiffest components found by SVD (not including market mode), then the factorization explains 85% of the the RMT explainable variation in R with overall results in good accord with the analysis presented here. This implies that sector decomposition information was already contained in the stiff modes from the SVD

Bulk Variation	80.2%
Explainable Variation	19.8%
Factors	Percent of Explainable Variation
Market Mode (MM)	8.0%
2 factors + MM	26.0%
3 factors + MM	36.1%
4 factors + MM	42.8%
5 factors + MM	48.9%
6 factors + MM	55.3%
7 factors + MM	59.4%
8 factors + MM	63.7%
9 factors + MM	68.1%
Fama and French	24.0%

Table 6: **Percentage of the Explainable Variance captured by our model compared with the Fama and French factor model.** Regression is done on the normalized dataset of 705 stocks without the market mode removed. To capture this, we add the market mode to factors obtained by our decomposition.

of R , however SVD is not the appropriate tool for the decomposition.

The eight-factor decomposition presented here explains 11.1% of the total variation (r^2) in the normalized returns with the market mode removed, and 56% of the random matrix theory explainable variation defined in [113]. For comparison, the classic three-factor decomposition portfolio returns by Fama and French [47] into market mode, market capitalization, and growth versus value yields an r^2 value of only 4.75%. Indeed, if only three factors are used instead of the eight for the decomposition presented here, the regression yields a comparable r^2 value (5.61%) but there appears to be no correspondence between three factors found by our unsupervised model, and those of Fama and French (Figure 16). Carrying out a similar comparison with Fama and French’s analysis applied to model portfolio returns, the regression on the S&P500 yields an r^2 value of 99.4% for Fama and French compared to 93.5% for our eight-factor decomposition (market mode reintroduced). Our decomposition was optimized without concern for market cap-

italization, which appears to be the key difference: For an equal weighted index of the 338 stocks in the S&P500 with current tickers and a complete data series in our time of interest, we obtain an r^2 value of 99.0% (97.0% for 3 factors) compared to 95.8% for Fama and French.

4.5 DISCUSSION

The emergent hyper-tetrahedral structure of stock returns suggests a natural unsupervised machine learning approach as we have detailed. This approach allows for the determination of sectors of the economy and the corresponding exposure of companies to each sector as described. Furthermore, it provides a basis for further exploration of how companies evolve in time, the relationship between sectors of the economy and a comparison to widely accepted benchmarks.

Future work remains to address survivorship bias, effects of sampling at different frequencies, and incorporating market capitalization. Investors, analysts, and governments alike would benefit from the development of new investable sector indices [53] that measure the health of our industrial sectors just like the macroeconomic indicators (GDP, housing starts, unemployment rate, etc.) measure the health of our broader economy. Tracing the sectors back in time [ArchetypalEvolution] could elucidate the incorporation of science and technology into our economic system. Finally, our unsupervised decomposition could provide data suitable for quantitative modeling of the internal and external dynamics of our economic system.

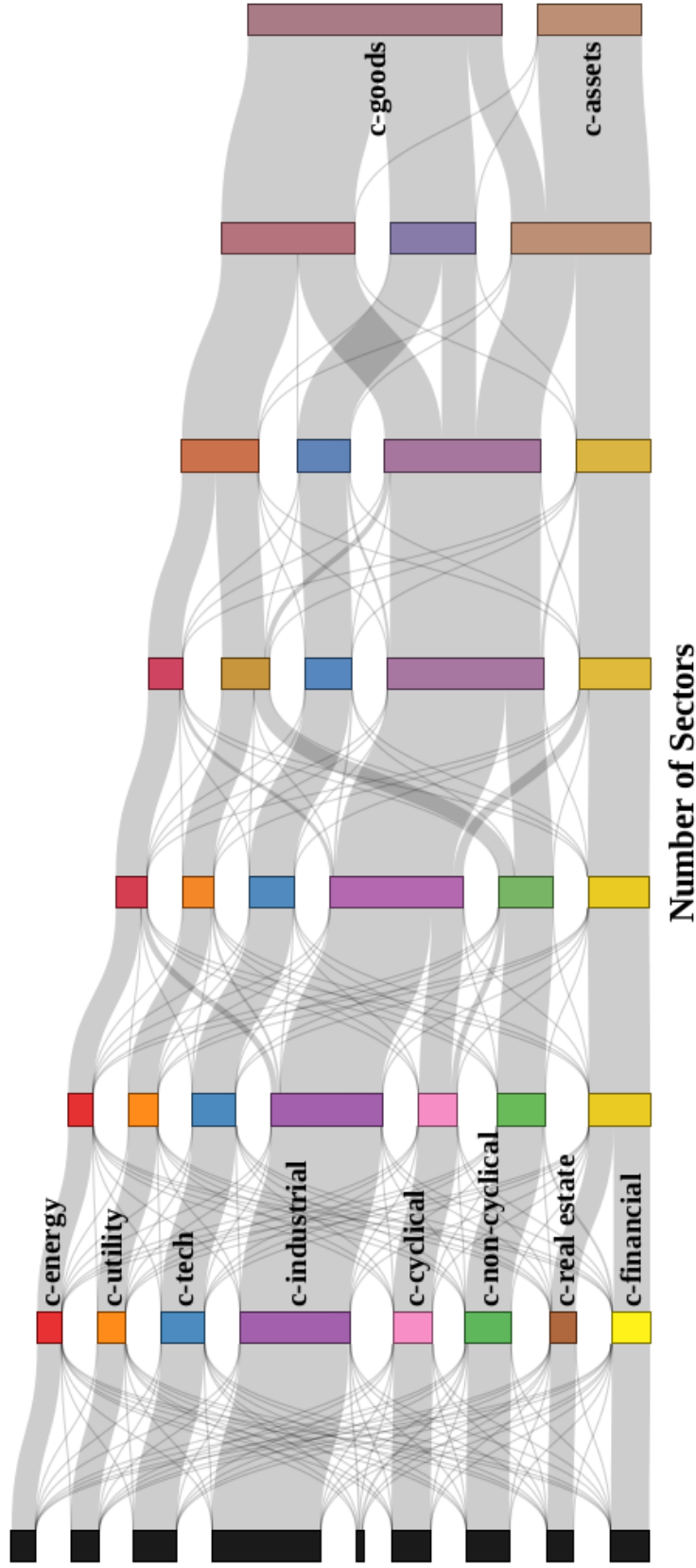


Figure 13: **Changes in the decomposition with dimensionality.** A Sankey diagram (generated using D3 [16]) displaying the relationships between sector decompositions with $n = N + 1$ and $n = N$. Relative node sizes correspond roughly to the amount of the market participating in the sector. Connection width depicts how strongly the sectors for decompositions with different n relate.

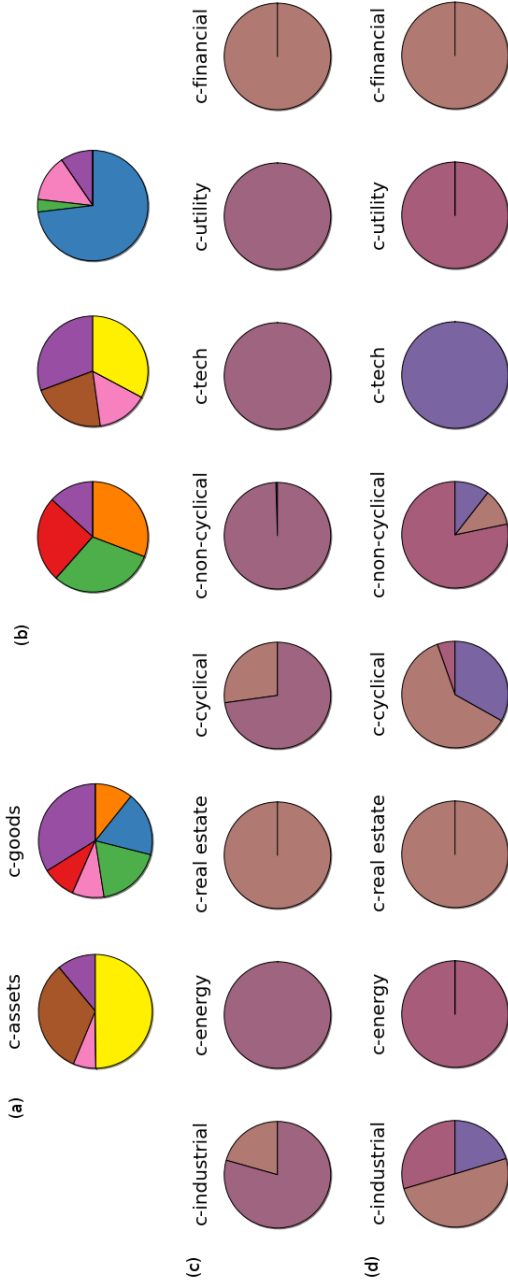


Figure 14: Pie charts depicting sectors as linear combinations of other sector decompositions having a different value of the dimensionality n . (a) Two sector decomposition with respect to the eight sector version (b) Three with respect to eight (c) eight with respect to two (d) eight with respect to three. For (a) and (b) the color scheme is the same as used throughout for the eight sector decomposition. For (c) and (d) colors correspond to those in Figure 9 for the two and three sector nodes. Through these charts it is evident that the two sector decompositions corresponds to an c -assets sector containing c -finance and c -real estate. and a c -goods sector containing companies which provide goods and services. In (c) and (d) we see c -industrial, c -cyclical and c -non-cyclical which merge by $n = 5$ split between the two and three factor decompositions respectively, consistent with Figure 9.

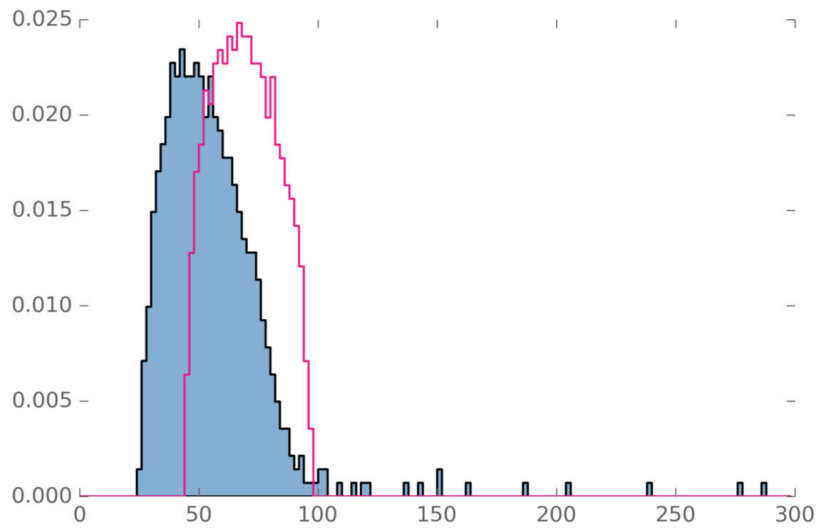


Figure 15: **Normalized distribution of singular values.** Filled blue histogram corresponds to distribution of singular values of returns from the dataset R_{ts} —one notices a clear separation of the hump-shaped bulk of singular values, and about 20 stiff singular values (the largest singular value ~ 952 , corresponding to the *market mode* is not shown). Pink line histogram outline shows the distribution of singular values of a matrix of the same shape as R but containing purely random Gaussian entries.

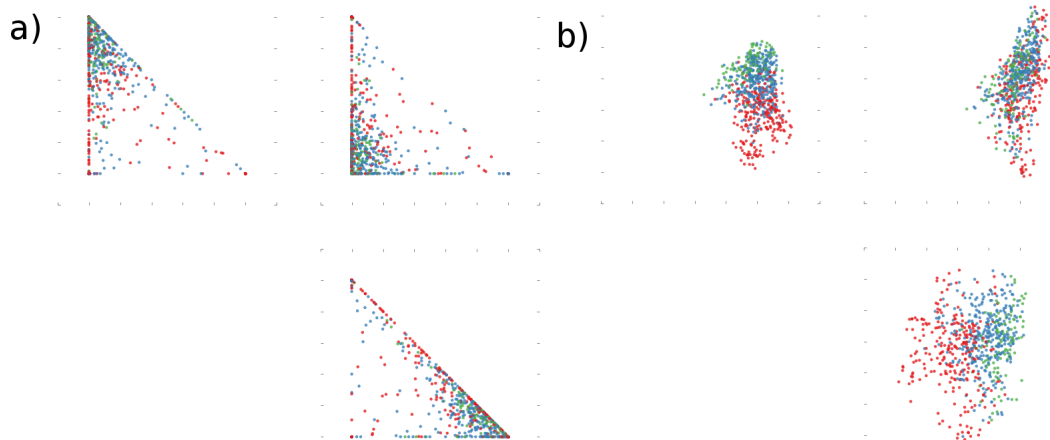


Figure 16: **3 Factor Model vs. Fama and French** 2D projections of the weights for each company in the SP500 with current tickers and data in the date range we consider. Red denotes companies with large market caps (market cap >10 billion), blue denotes medium (market cap 2-10 billion) and green denotes small (market cap < 2 billion). For our decomposition (a), there is no separation distinguishable by size of company. In comparison, for the Fama and French decomposition (b), there appears a gradation from large to small companies consistent with a factor of the model being related to size. (This is natural, since one of Fama and French's factors explicitly is the difference between large and small-cap returns). Thus our unsupervised 3-factor decomposition appears quite distinct from Fama and French's hand-created one.

5 ANALYSIS OF THE KAGGLE HIGGS BOSON MACHINE LEARNING CHALLENGE

Machine Learning has become a very important tool for data processing in recent years. There are a variety of algorithms including support vector machines, neural networks and decision forests but the basic concept for each remains the same. The goal is to train a computer to perform a task which typically would require human insight. Often, these problems are those that are fairly simple yet would be tedious for a human to do in practice. For example, consider a collection of photos such as those accumulated by google street view which are associated with a gps location. For a human to sort through all of this data and create an accurate mapping between address and gps location would be extremely labor intensive. With the development of tailored machine learning algorithms, however, this task becomes computationally straightforward [50]. In general, each of these algorithms comes down to selecting a function with parameters to update and a method with which to update them. The update is designed so that the best accuracy possible on the appropriate task is achieved.

Although ML is extremely useful in practice, its current form employs a bit of the dark arts. In the field of ML, many different kinds of algorithms are implemented in varying combinations. Some algorithms work well for certain tasks while others don't. Experts seem to have good intuition as to what may work and how to tune hyperparameters but ultimately a lot of results are empirical. The combination of these hurdles has spurred research in the direction of understanding these algorithms better, however the field is still very open. One striking example of the empirical nature of this field manifested itself in the recent Kaggle competition: Higgs Boson Machine Learning Challenge. The aim of this challenge was

to inspire data scientists to use their machine learning know-how on ATLAS data to determine whether an event signified the presence of the Higgs or belonged to a background process. Although physicists suspected their considerable domain knowledge would give them an edge in this competition, the best physics group came in 8th. This appears to be due to several causes. One, there are indications that physicists tended overlook some of the common problems when training computers to do these tasks; namely overfitting. While scores remained high on the public board, when tested on the full test data scores were lower. Additionally, the tools used by the top two contestants were very recently shown to perform better than their contemporaries on notable datasets in computer science including CIFAR-10 and 100, STL-10, MNIST, Amazon Sentiment Analysis, CT slices, California Houses, YearPredictionMSD... [56, 58, 102, 111]. The hypotheses that presented here is that features created by domain experts, while beneficial, were not enough in the Kaggle competition to overcome better algorithms and empirical techniques.

Approaches based on computational neural networks took both first and third place in the Kaggle competition. In both of these submissions a large number of deep neural networks were employed and the results averaged to make a prediction. The basic structure of any deep neural network is a sequence of layers. In each layer, the input \vec{x} is transformed linearly to give an output \vec{y} :

$$\vec{y}_i = W_{ij}\vec{x}_j + \vec{b}_i \tag{15}$$

Next, an activation function, f , is applied so that the final output of the layer is $\vec{z} = f(\vec{y})$. This final output of the layer becomes the input to the next layer. In each of the winning networks, the activation function used for the final layer was a softmax, defined below.

$$f(z_j) = \frac{\exp(z_j)}{\sum_i \exp(z_i)} \quad (16)$$

The network output is hence a normalized 2D vector giving the probability that the event is 'signal' or 'background' respectively. Averaging the output of all of the networks yielded the final prediction. To further tune the results, the cutoff value for signal vs background was determined by optimizing the approximate median significance (AMS) metric provided by the challenge organizers.

$$AMS = \sqrt{2 \left((s + b + b_r) \log \left(1 + \frac{s}{b + b_r} \right) - s \right)} \quad (17)$$

where s/b are the unnormalized true positive/false positive rates and $b_r = 10$ is a constant regularization term. More precisely, if the label prediction of the submission is denoted \vec{z} and the true labels by \vec{l} then

$$s = \sum_i w_i \delta(l_i - s) \delta(z_i - s) \quad (18)$$

$$b = \sum_i w_i \delta(l_i - b) \delta(z_i - s) \quad (19)$$

where w_i are the weights given in the data set. As many pointed out on the Kaggle discussion boards, the AMS metric is somewhat noisy and unstable. However, as cited in the documentation for the challenge, this metric is used frequently by high-energy physicists for optimizing the selection region for discovery significance. This makes it a natural choice. In the original paper which introduced AMS as a criteria, b_r was set to zero [38]. In the challenge, b_r was increased to 10 to help alleviate the issues with noise. In addition to using ensembles of deep networks, Gábor Melis (1st) and Courtiol Pierre (3rd) each transformed the data by removing the mean and rescaling the variance to one. Additionally, they both employed

backpropagation with a cross-entropy cost defined below.

$$L(\vec{l}, \vec{z}) = -\frac{1}{N} \sum_{n=1}^N \left[l_n \log(z_n) + (1 - l_n) \log(1 - z_n) \right] \quad (20)$$

where the labels \vec{l} have been converted to integers ($s = 1, b = 0$) and the output of the network (\vec{z}) in this case is the probability that the event is signal. This metric is commonly used and is sometimes called 'logistic loss' or simply 'log loss'. With the exception of the activation function used by the first place contestant, the ensembling approach and training practices used were standard. What seems to have set these contenders apart was the large size of the ensembles, the pre-processing of the data and the use of cross-validation. The combination of these things seems to have helped not only the expressive capabilities of the network but also to have alleviated overfitting.

Although the first and third submissions were very similar in a number of ways there were a couple of things which distinguished them. For one, Gábor Melis used much larger networks (albeit fewer). Additionally, he added small L1 and L2 regularization terms to the cross-entropy loss function. Perhaps the most significant difference was his use of a novel activation function known as channel out [102, 111]. In channel out networks, only a subset of parameters are used to evaluate a given example. This, in turn, allows for more specialization than traditional sigmoidal units. This may explain the increase in performance over the third place contestant who used sigmoidal gates. Channel out is particularly interesting due to its relation to the way biological neural networks function. In biological networks, local competition between neurons has been shown to play a role in gain control and noise reduction. In channel out, sets of n -neurons are forced to compete in a similar fashion. Only the neuron giving the largest output sends on its signal to the next layer. This results in better performance on classification

tasks as well as better performance on implicit long term memory [102].

Second place in the Kaggle challenge went to Tim Salimans who employed a Regularized Greedy Forest (RGF) [58]. This approach is based on a collection of decision trees. Decision trees have become very popular as an 'off the shelf' method for data analysis. Although their performance is somewhat less than neural networks, they require less tuning of hyper-parameters which makes them more approachable. The standard method is to train a large number of decision trees and then apply a wrapper to their output. This results in a decision forest and is known as 'boosting'. Two common methods which fall into this class are Leo Breiman's bootstrap aggregation (bagging) [21] and random forests [22]. The task of a single decision tree at root (pun intended) is, given parameters corresponding to a data point, label the point by applying decision rules which are layed out in a tree-like structure. A simple example of a decision tree is shown in Figure 17. Adding layers of branches and increasing the 'depth' of the tree increases expressive capability, however, it also adds complexity. Hence when training a tree, there is a cost associated with incorrectly labeling data along with a cost proportional to the depth. Common costs associated with labeling include cross-entropy and the gini impurity:

$$C(Z) \propto \sum_n \sum_{k=1}^K (Z_{nk})(1 - Z_{nk}) \quad (21)$$

where Z_{nk} is the fraction of events which are labeled k in node n as given by the output of the tree and K is the number of classes. As with neural networks, there are vast numbers of algorithms that fall under heading of 'decision tree' or 'decision forest'. The fundamental algorithms and their origins are explained nicely in 'Learning Classification Trees' by Wray Buntine [23].

The RGF approach is different from simple boosting in that it tries to make use of the underlying tree structure rather than treating the learned trees as a black

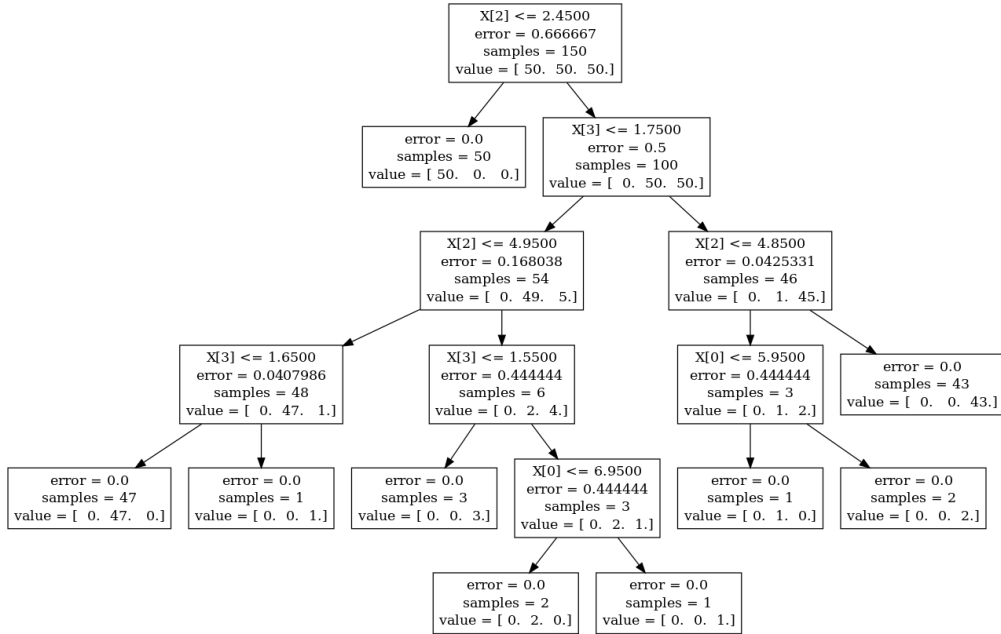


Figure 17: **Sample decision tree from scikit-learn [81]**. A single variable along with its decision rule is chosen at each level in the tree such that the cost function, such as gini or cross-entropy, is minimized.

box and applying a wrapper. Gradient Boosted Decision Trees (GBDTs) also attempt this and were used by the top physics team headed by Luboš Motl, via the XGboost package [32]. GBDTs employ a wrapper which is a linear combination of trees.

$$T_k = \sum_j^k b_j t_j \quad (22)$$

where T is the output of the forest of k trees and b_i are the coefficients giving the appropriate weight to each tree's output, t_j . Suppose you train K trees on the dataset with loss function $L(T_k)$. Then the GBDT pseudocode becomes: For k in range 1 through K

- Optimize $L(T_k)$ with respect to b_i
- Optimize $L(T_k)$ with respect to the decision coefficients of t_k

In this way, the decision rules for trees are partially finetuned during the training

of the wrapper. RGFs take this a step further. Instead of only optimizing over the coefficients of the last tree added, the RGF algorithm optimizes over the coefficients of all trees seen thus far: For k in range 1 through K

- Optimize $L(T_k)$ wrt b_i
- Optimize $L(T_k)$ wrt the decision coefficients of $T_k = \{t_1, \dots, t_k\}$

Additionally, an explicit regularization term is added to the cost function to combat overfitting. These Random Greedy Forests have been shown to perform better than GBDTs on a large number of tasks. In fact, the second place winner noted an advantage of RGFs over XGBoost on the ATLAS dataset in particular. In addition to the use of RGFs, the second place winner also used a slightly larger number of trees (48 vs. 44 for Motl's group) combined with 7-fold cross validation, two sets of training objectives, and feature engineering. There were two specific ways in which Salimans noted that he changed the features supplied. For one, he states that he can 'calculate the features with respect to new combinations of (pseudo)particles. For example, features like the transverse mass can be defined with respect to any combination of the particles in the detector, not just those for which it was calculated in the provided features'. Additionally, he transforms the eta and phi features to account for the symmetries outlined on the competition forum. This last portion is interesting because Melis also notes that he drops the phi features due to overfitting.

The dataset provided for the competition includes 4 components: primitives (PRI), derived (DER), weights and labels. Data is obtained from simulations with primitive quantities corresponding to those values which are measured by the ATLAS detector. Derived quantities are those which can be calculated from the primitives using physics concepts such as conservation of momentum. The weights in the dataset account for the mismatch between the probabilities of the event

occurring in the simulation and what would be observed in the selection window of the detector and are used to calculate the AMS. Finally, for the training data, labels are provided so that supervised learning can be employed. For more information regarding the dataset see Appendix F. In order to test the hypothesis that the derived quantities do give an edge, three sets of neural networks were trained. One on the full data set, one on the PRI quantities only and one on the DER quantities only.

This analysis was done as part of the Advancement to Candidacy Exam at Cornell. In the interest of time, it was necessary to be selective in the algorithms chosen. One such choice was to attempt to reproduce a version of Courtiol Perre's submission rather than that by first place competitor, Melis. This was due to the computational networks he used being much smaller than those of Melis and hence swifter to train. This choice also aided in debugging. Also, unlike Melis, the networks here make use of ReLU activations combined with dropout rather than channel out. Although this combination is not as effective, it is preferable to sigmoids. The reason for this choice was due to lack of support in Theano for the masking functions needed. To discard Theano for the training would have increased the time needed to both code and train the networks. Additionally, using channel out was not relevant to answering the question of how derived quantities affect the performance. Using Theano [6, 9] and a Theano implementation of dropout by Gabriel Synnaeve, 108 neural networks were trained on each of the 3 datasets. The networks were of sizes

1. 30x50x1
2. 30x50x25x1
3. 30x50x50x25x1

in equal proportion as described by Pierre. For each dataset, the last 20% were set aside for analysing the percentage of error of the trained network ensembles. The remaining 80% of events were randomly sorted to form 6 distinct sets of data for 2-fold cross validation. In keeping with the top performers, training was performed with backpropagation on a cross-entropy loss and normalized the data to have zero mean and standard deviation one. For each of the results presented, the cutoff between signal and noise was chosen such that the top 15-16% of events were labeled signal. This cutoff was found by Melis to maximize the AMS vs cutoff curve. To build on the tools of the top performers, a novel gradient method known as adadelta was used [114].

Adadelta is a technique which adaptively updates the step-size for gradient descent on a per dimension basis. Unlike contemporaries such as Newton’s Method, quasi-Newton methods, the work by Becker and Lecun [7] and that of Schaul et. al [95], adadelta does not require the Hessian or even an approximation to it. It uses only first order derivatives and has minimal additional computational cost above regular stochastic gradient descent. Like adagrad [45], adadelta has nice properties typically associated with second order methods or annealing. Small gradients correspond to large learning rates and vice versa so that progress over dimensions evens out over time. Unlike adagrad, adadelta’s learning rate does not continuously shrink over time resulting in slow optimization near the minima. Also, it is not sensitive to a initial global learning rate.

The results for each run are shown in Table 7. Each network was trained for 1000 epochs. These results indicate that the derived quantities perform better than their primitive counterparts. The full dataset performs the worst. Each ensemble performs substantially worse than the contest winners whose public AMS scores were 3.80581, 3.78913 and 3.78682 respectively. Motl’s group achieved an AMS score 3.76050. There were myriad reasons why the scores obtained here

were so much lower including the exclusion of the weight information from the training data. In the submission of Pierre, the weights were added during the final training epochs. The next step would be to run a finetuning phase on the network ensembles which includes the weights. Additionally, many more training epochs may be warranted.

One point to note is that the dimensionality of the full dataset was much larger than the sets of its PRI and DER constituents alone. This may increase the number of training epochs needed substantially. Another comparison which would be helpful would be one in which rather than fixing the number of training epochs, the training is stopped for PRI/DER when a score comparable to that of the full set is reached. Another way to perhaps answer this question is to, rather than retrain with stopping, simply reduce the number of networks in the PRI and DER sets until a comparable amount of expressiveness is reached. The results of this are also shown in Table 7 under PRI-1 and DER-1. The 1 denotes that the results quoted are the results for a single network. This result is very striking. One network achieved the same or better results in both cases while having lower accuracy on the holdout set. Why ensembling did not appear to achieve better results in this case is an open question. Although Pierre mentions that bagging, which was employed here, did not increase his AMS score over a simple ensemble, this does not explain how an ensemble in general does not outperform a single network. This surprising result, obtained using the shallowest network trained, may indicate that many more epochs of training are needed for the deeper networks. For example, the 4 layer Deep Belief Network of Hinton [55], requires 5000 training epochs which would have increased the training time of approximately four days five fold. Additionally, a combination of ReLU with dropout may be better suited to deeper networks with larger dimensionality such as those used by Melis.

The results presented highlight some of the issues with Machine Learning tech-

TEST	FULL	PRI	DER	PRI - 1	DER - 1
Dimension	30	17	13	17	13
AMS Public	0.78092	1.65326	2.68863	1.65836	2.79876
AMS Private	0.88286	1.61681	2.66465	1.62299	2.80616
Rank	#1678	#1591	#1287	#1591	#1206
Pct. Error	0.32775	0.27687	0.23356	0.30795	0.28869

Table 7: **Scores obtained after 1000 epochs of training with 108 neural networks on each dataset.** Dimension denotes the the number of entries for each event in the given dataset. AMS was evaluated by the Kaggle site on the unlabeled test.csv data provided. Percent Error was evaluated on the 20% of the training data held out.

niques. There were a huge number of choices to be made when setting up the training run. Although the choices made seem natural, the results clearly leave more questions than they resolve in terms of what choices result in superior algorithms. Although it is very interesting that more information seems to be encoded by applying physics knowledge, a good deal more exploration into how much can be gained by engineering features is warranted. It certainly seems that adding physics knowledge should increase the AMS score considering how much better the derived features perform. However, as many contestants pointed out, the CAKE features as well as the sophisticated features used by Motl’s group did not see the performance gains that a simple change in algorithm could produce. To quote Kaggle’s CEO:

Two pieces are required to be able to do a really good job in solving a machine-learning problem. The first is somebody who knows what problem to solve and can identify the data sets that might be useful in solving it. Once you get to that point, the best thing you can possibly do is to get rid of the domain expert who comes with preconceptions about what are the interesting correlations or relationships in the data and to bring in somebody who’s really good at drawing signals out of

data. -Anthony Goldbloom

There are several other directions that would be interesting to explore. It could be interesting, for example, to train the same networks or trees used in the challenge on different datasets. The challenge would then be to choose the appropriate datasets and training cutoff with which to make a fair assessment of how much better features can increase the score. Another interesting investigation could involve looking at the weights the networks themselves learn and interpreting what they can say about the relationships in the data.

6 JEFFREY'S PRIOR SAMPLING OF DEEP SIGMOIDAL NETWORKS

Neural networks such as those explored in the Kaggle challenge discussed above have been shown to have a remarkable ability to uncover low dimensional structure in data. To investigate this concept from the perspective of information geometry, we have examined the underlying model manifold of networks trained to reconstruct hand written digits. In particular, we analyze the manifold learned by Deep Belief Networks and Stacked Denoising Autoencoders using Monte Carlo sampling. What we find is that, contrary to what is found in other fields, the model manifold of these networks forms an only slightly elongated hyperball with actual reconstructed data appearing predominantly on the boundaries of the manifold. In connection with the results we present, we discuss problems of sampling high-dimensional manifolds as well as recent work by Transtrum et. al. [107] discussing the relation between high dimensional geometry and model reduction.

6.1 SLOPPY MODELS

Deep neural networks have proven to be state of the art on numerous machine learning benchmark tasks in recent years [96]. These networks are so called 'deep' due to their layered structure in which each subsequent layer appears to encode a different, more abstract representation of the data. Previously there has been much interest in the nature of the manifold learned by deep neural networks [1, 8, 90]. Through these studies, it has been conjectured that in the higher layers of the network, the representation of the data more uniformly fills the space of neural outputs and that high density regions of the manifold near which raw data concentrates tend to unfold. In other words, linear interpolation between data

points in higher layers of the network can be performed without leaving a region of high probability. In contrast, the same procedure in data space would produce a superposition of the two data points which typically results in a point having low probability of naturally occurring in the dataset. Recently, a method has been developed to approximately perform a Metropolis-Hastings sampling of the data distribution using a trained network [1]. The aim of this method is to generate samples which could be drawn with high probability from the distribution underlying the data. Here we ask a different question, that is, what is the geometry of the manifold the neural network has learned? We present a sampling of the model manifold implementing Metropolis-Hastings weighted by Jeffrey's Prior. Using this sampling method yields insight into the manifold learned by computational neural networks as well as into Jeffrey's Prior itself.

In general, models are created to form a low dimensional representation of the data space. This is true also for neural networks. In many fields — including systems biology, statistical physics and mathematical programming — it has been found that models often form a hierarchical structure in which certain combinations of parameters dominate the behavior while others barely contribute [31, 52, 112]. These directions in parameter space are referred to as 'stiff' and 'sloppy' respectively. Consider a fitting procedure where, given the model and data, the objective is to determine the parameter values. For a least squares cost function, the sloppiness of the model is reflected in the eigenvalues of the Hessian which span many orders of magnitude. Another indicator of sloppiness in a model is that the model manifold, which corresponds to the space of all possible predictions of the model, forms a hyper-ribbon [106]. The sloppiest combination of parameters results in the thinnest direction in data space while the stiffest affects the predictions of the model immensely, resulting in a very long direction.

Feed forward neural networks have been shown to display the signatures of

sloppiness [72]. In this work, however, we study deep networks whose aim is reconstruction. In particular, we examine Stacked Denoising Autoencoders and Deep Belief Networks trained to reconstruct images from the MNIST dataset, a dataset of handwritten digits in which each pixel takes a value between 0 and 1. In each case, the model manifold forms an only slightly elongated hyperball (Section 6.4) which, from the viewpoint of information geometry, indicates the network does well at weighting parameter combinations equally. In addition to this apparent lack of sloppiness, we also find that the actual data appears predominantly to lie near the boundaries of the manifold; a feature which may be due to the largely saturated pixels of the data set. That most of the images generated by the trained network do not correspond to actual images raises a very interesting question. Does this mean that the neural network is wasting a vast amount of expressive capability? Or does it point to a general feature in modeling that the interesting components of a manifold lie on its edges?

6.2 DEEP NETWORKS

The model manifolds we study belong to two types of prototypical deep networks; Deep Belief Networks (DBNs) [55] and Stacked Denoising Autoencoders (SdAs) [108]. At their heart, these networks rely on the following mapping between layers of 'neurons'

$$\vec{h} = \sigma(W\vec{d} + \vec{b}) \tag{23}$$

$$\vec{y} = \sigma(W'\vec{h} + \vec{b}') \tag{24}$$

where \vec{d} represents the input vector, \vec{h} the hidden activations or output, and \vec{y} the reconstructed input. W and W' are the weight matrices, \vec{b} and \vec{b}' are offset vectors, and σ denotes the sigmoid function. By stacking these learned encoding

and decoding maps, the network can be trained to compress and reconstruct data. The map learned by these networks is flexible in that the features in the top hidden layer can be used as input to a simple classifier such as a softmax layer or support vector machine.

DBNs and SdAs differ in the way they are trained. However, both of these types of networks take advantage of a layered structure in which each layer of neurons encodes a different representation of the data. For comparison, we train three of these deep networks on the MNIST dataset [67]. The first network we present is an SdA trained on a single digit, '1'. Since we are interested in networks that perform the reconstruction task rather than classification this is a natural choice to serve as a smaller testbed for our methods. The digit '1' was chosen due to the striking structure observed in PCA plots of the data. We refer to this network as the single-digit network. The other two networks we study are trained on the full dataset and are referred to as the DBN and SdA networks respectively. The particulars for each of the three networks are given in Table 8. Each SdA was trained using a cross-entropy loss and stochastic gradient descent via Theano [6, 9] and the classes given on deeplearning.net. The DBN was trained using MATLAB code provided by Hinton's group [55].

Each of these neural networks can be viewed as a fitting process in two distinct ways. For one, training the network is a fit. During training of the neural network, there are a large number of parameters (W , W' , b and b' for each layer) and the objective is to fit this multi-parameter model to a large number of images. The manifold associated with this fitting process has been shown to form a ribbon-like structure in high dimensional space in which the manifold becomes thinner and thinner in each successive dimension by a roughly constant factor [72]. This structure we refer to as a hyper-ribbon.

The second way to view the model as a fit occurs after the network has been

	SdA/Single-Digit	SdA	DBN
data	MNIST '1's	MNIST digits	MNIST digits
training set size	5678	50000	50000
testing set size	N/A	10000	10000
network size	784-100-100-10	784-1000-500-250-30	784-1000-500-250-30
class. error	N/A	1.28%	1.50%
corruption level	0.25	0.25	N/A
training method	Theano	Theano	Hinton

Table 8: **Training characteristics of the neural networks we study.** Classification accuracy on the MNIST dataset [67] was evaluated by training a support vector machine to classify the data given the top layer of features [81]. The network and support vector machine were then applied in tandem to the test set in order to calculate the error. Training of the networks were achieved using Theano [6, 9] and MATLAB code provided by the Hinton group [55]. The 4-layer SdA trained with theano was trained with a linear mapping at the top layer. This choice was made to ease comparison between the SdA and the Hinton group’s DBN which has this characteristic. For each of the 4-layer networks the layer sizes were chosen to correspond with Hinton’s original network. The top hidden layer in each case had a dimension of 30.

fully trained. Once the network is trained, we require that the weights and biases (W , W' , b and b' for each layer) remain fixed. The trained network provides a function f which maps a vector in the top hidden layer $\vec{\theta}$ to a distinct image in the reconstruction space \vec{y} such that $\vec{y} = f(\vec{\theta})$. A diagram is provided in Figure 18. The trained network provides us with a model in which a low dimensional parameter space $\vec{\theta}$ can be used to represent any image reconstructed from the dataset along with a host of others. The set of all images producible by this model forms the ‘model manifold’ of the trained network. As the parameters are varied, the images formed interpolate between the images the network reconstructs from the dataset. For example, examining our single-digit network, the ten neural outputs in the top hidden layer sweep out a 10 dimensional model manifold $\{\vec{y}\} = f(\{\vec{\theta}\})$ in the 784 dimensional space of possible MNIST images where the image space dimensionality is determined by the number of pixels in the images (28x28). The fit is then: given

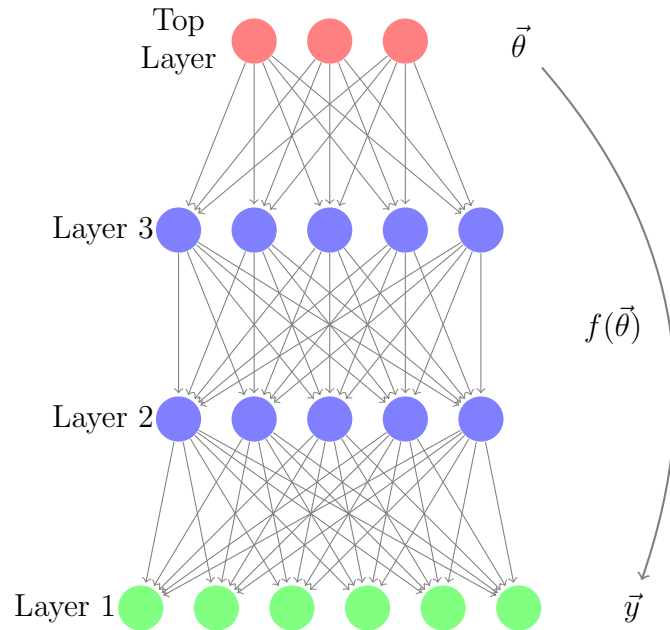


Figure 18: **Diagram of the decoding map learned by a generic 3-layer neural network.** Once trained, a DBN or SdA provides a function f such that, for any given neural activation in the top hidden layer $\vec{\theta}$, f provides a corresponding image in reconstruction space $\vec{y} = f(\vec{\theta})$.

an image \vec{d} , vary the parameters $\vec{\theta}$ such that the cost function $C(\vec{\theta}) = \|\vec{d} - f(\vec{\theta})\|^2$ is minimized.

In addition to studying the model manifold corresponding to the reconstruction space, the layered structure allows for exploration of the model manifold the network has learned to represent the data in each layer of the network. Throughout, 'Layer 1' corresponds to the reconstruction space and 'Top Layer' to the top hidden layer of the network.

6.3 JEFFREY'S PRIOR

The first task in visualizing the model manifold is to generate images apart from those on which the neural network was trained. To get a picture of what the model manifold looks like, points in the model manifold are generated using Monte

Carlo sampling. To implement our sampling we use Metropolis Monte Carlo and an uninformative prior known as Jeffrey’s prior. This prior is the square root of the determinant of the Fisher Information Matrix. The choice of prior was made due to the fact that it weights volume in parameter space by volume in data space. In effect, it keeps the algorithm from getting stuck in small regions of data space which correspond to a wide range of parameter values. Additionally, unlike the uniform prior, Jeffrey’s prior is invariant to transformations of the parameters.

In order to implement Jeffrey’s Prior, the metric of the space must be defined. The mapping from the top hidden layer (parameter space) to the reconstruction space (data space) implies a natural fitting procedure for any given image. The cost for this fit is given by

$$C(\theta) = \frac{1}{2} \sum_k (y_k^\theta - d_k)^2 \quad (25)$$

where $\vec{y}^{\vec{\theta}}$ is the reconstruction with corresponding top hidden layer activations $\vec{\theta}$ and \vec{d} is the data point. The Hessian is then

$$H_{ij} = \frac{\partial^2 C}{\partial \theta_i \partial \theta_j} = \frac{\partial}{\partial \theta_i} \left(\frac{\partial}{\partial \theta_j} \sum_k \frac{(y_k^\theta - d_k)^2}{2} \right) \quad (26)$$

$$= \sum_k \left(\frac{\partial y_k^\theta}{\partial \theta_i} \frac{\partial y_k^\theta}{\partial \theta_j} + (y_k^\theta - d_k) \frac{\partial^2 y_k^\theta}{\partial \theta_i \partial \theta_j} \right) \quad (27)$$

The second term in this expression is computationally expensive and is exactly zero for data described by the model. Additionally, even for data points not lying on the model manifold, the values in this sum fluctuate between positive and negative, averaging to zero. These characteristics make an approximation which neglects

this term very natural:

$$H \approx \sum_k \left(\frac{\partial y_k^\theta}{\partial \theta_i} \frac{\partial y_k^\theta}{\partial \theta_j} \right) = J^T J \quad (28)$$

In addition to being less expensive to compute, the approximate Hessian, aka the Fisher information Matrix, is positive definite and data independent. Indeed it is the metric on the space of neural outputs (top hidden layer) induced by the least-squares metric in data (image) space. The distance between two nearby neural outputs is given by the squared difference of their corresponding images.

Now consider the singular value decomposition (SVD) of the Jacobian matrix, $J = U\Sigma V^T$, where V is a orthogonal matrix in parameter space, Σ is a diagonal matrix of the singular values and the columns of U form an orthonormal basis in data space which span the range of J . The metric can thus be written as

$$g = V\Sigma^2V^T \quad (29)$$

where the columns of V correspond to the eigenparameters and the eigenvalues are given by $\lambda_i = \Sigma_{ii}^2$. Geometrically, this states that the Jacobian maps metric eigenvectors into the data space vectors U_i stretched by a factor $\sqrt{\lambda_i}$. The mapping from hidden layers into data space expands volume (N-volume to N-dimensional surface area) by a factor $\prod_i \sqrt{\lambda_i} = \prod_i \Sigma_i$.

In order to sample according to Jeffrey's prior, the points are given a prior probability that is equal to the square root of the determinant of $g_{\alpha\beta}$. So we have

$$p(\theta) = \sqrt{|g_{\alpha\beta}|} = \sqrt{|J^T J|} = \prod_i \Sigma_i \quad (30)$$

where Σ_i are the singular values of the Jacobian. This probability density in the space of neural outputs thus samples surface volumes in the model manifold

equally. Performing Metropolis Monte Carlo on the model manifold using Jeffrey’s prior enables us to explore the model manifold of the single-digit network as well as the DBN and SdA.

6.4 RESULTS

Sampling of the model manifold of each trained neural network is performed using emcee - an MIT licensed pure-Python implementation of Goodman & Weare’s Affine Invariant Markov chain Monte Carlo (MCMC) Ensemble sampler [49, 51]. In particular, we found it useful to employ their parallel tempering package. For the single-digit network we used 6 temperatures with 20 walkers each chosen randomly from MNIST images of the digit ‘1’. The results presented are for a sampling of 500,000 steps for each walker at a temperature of 1.0. Similarly, for the DBN and SdA we used 6 temperatures each although the number of walkers was increased to 60 as the dimensionality of the parameter space increased three-fold. Walkers were again chosen randomly from MNIST digits. Each walker was run for 50,000 steps. Results shown are for a temperature of 1.0.

PCA images of the sampled points for the single-digit network are shown in Figure 20. This figure corresponds to PCA projections of the data where the $(i, j)^{th}$ figure in the grid is a plane spanned by singular vectors i and $j + 1$ of the centered data. Presenting PCA projections in this way reveals the striking structure which led us to consider the ones. The projections along the vector corresponding to the largest principal component are shown in the first four frames along the top row. In these images one can see that data (reconstuitions of the MNIST digit 1’s) lies in an arc. More discussion of this arc and how ones with different characteristic arrange themselves are presented in Figure 21. PCA projections of the sampling for the DBN are shown in 22 with sampled digits in Figure 23. Results for the

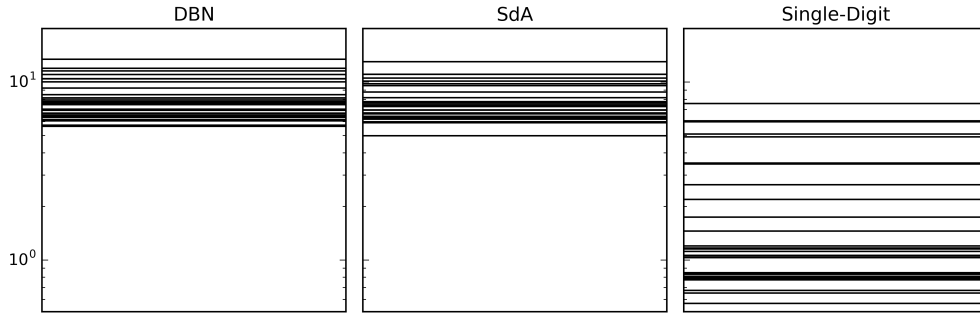


Figure 19: **Widths of the reconstructed manifold for each network along PCA directions.**

SdA displayed the same behavior and are collected in the Appendix G. In each of these images, the sampling forms a slightly elongated hyperball with corresponding hierarchy of PCA widths shown in Figure 19.

For the single-digit network, the top hidden layer has dimension 10 hence the reconstructions must be a 10 dimensional object embedded in the 784 dimensional pixel space. We find that in each successive PCA direction, the width of the manifold shrinks by a factor of 1.17 along the first 10 PCA vectors. For the DBN and SdA, the reconstructed manifold is 30 dimensional and the factor is a meager 1.03. In addition to PCA, there were other methods of determining widths we explored detailed in Section 6.5. These factors clearly contrast the structure of the neural network model manifold with the hyper-ribbon manifolds observed in sloppy models.

Another characteristic of each PCA projection plot is that, in each, points representing MNIST data are located along the edges of the model manifold while the bulk of the sampling remains deep in the interior. This is made more evident in 'Transparent' column of Figures 24 and 36 by increasing the transparency of the points corresponding to MNIST digits. Examining higher layers of each network reveals that the behavior of the single-digit network remains unchanged. For the

DBN and SdA, however, as we progress to higher layers, the behavior becomes reversed. Instead of the sampling lying deep in the interior, the sampling spans a much larger distance in parameter space. We believe this to be due to a linear top layer in the SdA and DBN in contrast to the single-digit network in which all layers were sigmoidal.

In addition to sampling, we also explore the location of the 'corners' in relation to the images and sampling. Corners correspond to datapoints for which the top hidden layer representation is a vector with each element at an extremum allowed by the network. More specifically, in the single-digit network, each parameter in the top hidden layer is constrained to lie between 0 and 1. Using this we can plot the representation for these 2^{10} corner points. For the DBN and SdA, the top layer activation is linear hence we set 'corners' to be those hidden layer activations for which $\vec{\theta} \in \{-10^6, 10^6\}$. The top layer of each of these networks has a dimension of 30, hence there are 2^{30} such corners. For this reason, only a random subset of 10,000 are plotted. The PCA projections for these are in the 'Corners' columns of Figures 24, 36 and 34. By definition the corners are extremely far apart in the top layer. For this reason each plot of the top layer is presented with a sigmoid applied. For each of the 4 layer networks the nonlinearity of the network results in the interior of the space bulging out such that the corners lie within the volume of the manifold in data space (Layer 1) rather than on the boundaries.

The 'digits' sampled by each of these networks do not correspond to actual images. Several of these sampled images for the DBN are shown along with the original images and 'eigen-images' of the dataset in Figure 23. The eigen-images are formed by taking the singular value decomposition of the MNIST dataset such that

$$X = U\Sigma V^T \tag{31}$$

where X is the data arranged as (number of samples) x (image size). The images in the matrix V which correspond to the eigenbasis of X compose the eigen-images. For the single-digit network and SdA network, the corresponding plots are found in the Appendix G. For each network, although the eigen-images do not form actual images, they show more identifiable structure than the sampled images. Additionally, they are located on the boundary of the manifold in reconstruction space just as are the actual images. This characteristic location provides further indication that the boundaries play an important role in the model.

6.5 DISCUSSION

A deep neural network with N hidden outputs provides a N dimensional representation of a high dimensional data set. As we vary N , how does the resulting hierarchy of descriptions reflect itself in the geometry of these models? Our hypothesis was that the neural manifolds would form a hyper-ribbon, with incremental expressiveness as $N \rightarrow N - 1$ reflected in geometrically thinner geodesic widths. This hypothesis, verified for multiparameter models in other disciplines [106] appears not to be the case for the neural manifolds we study. Instead, their model manifold forms an only slightly elongated hyperball. This is striking in that, for a typical model, such behavior would imply that the model tends to weight parameter combinations more or less equally. In other words, for the neural network, it would imply that there are no neurons which, in tandem, control the majority of behavior. Conversely, there are no sets whose values can change dramatically without affecting the output.

There are several techniques that have been found to be useful for training of neural networks; purportedly due to keeping the network from relying too heavily on any one neuron or set of neurons. For example, the L1 and L2 norm terms pe-

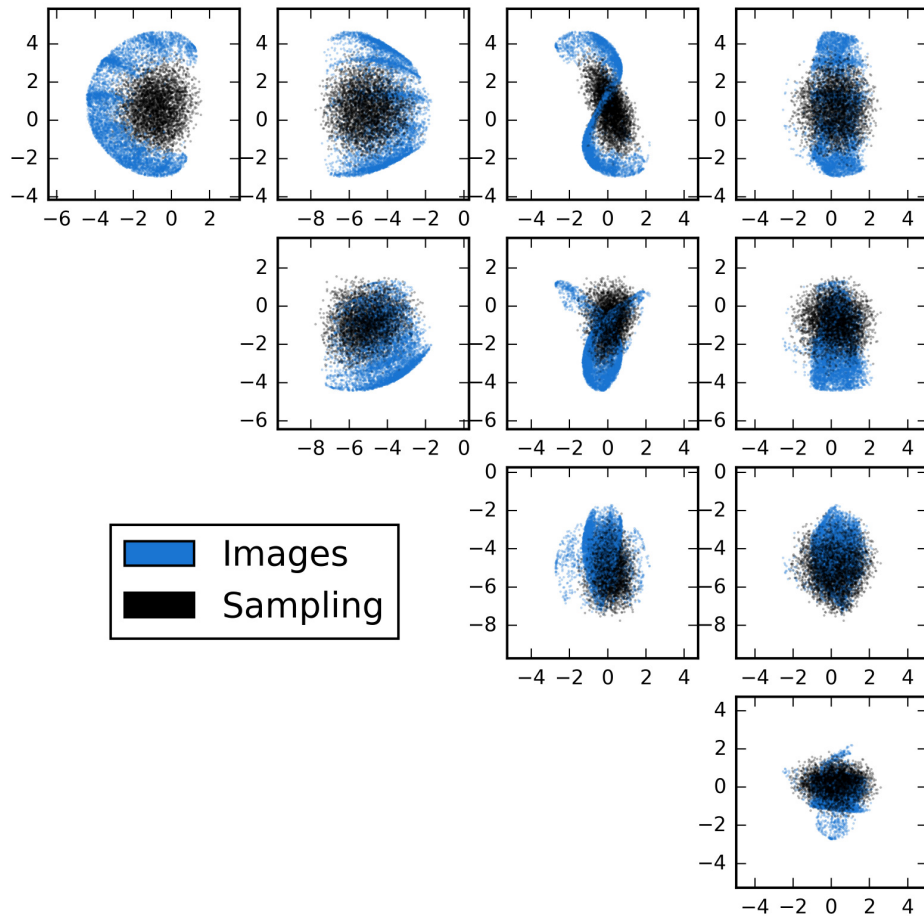


Figure 20: **Two dimensional PCA projections of the MNIST digit data for ones and the Jeffrey’s Prior sample of the model manifold.** The $(i, j)^{th}$ figure in the grid is a plane spanned by singular vectors i and $j + 1$ of the centered data. Image reconstructions lie on the edges of the manifold. Width of the network decreases slightly along each principal vector; however the aspect ratio is much closer to unity than that observed in other models [106].

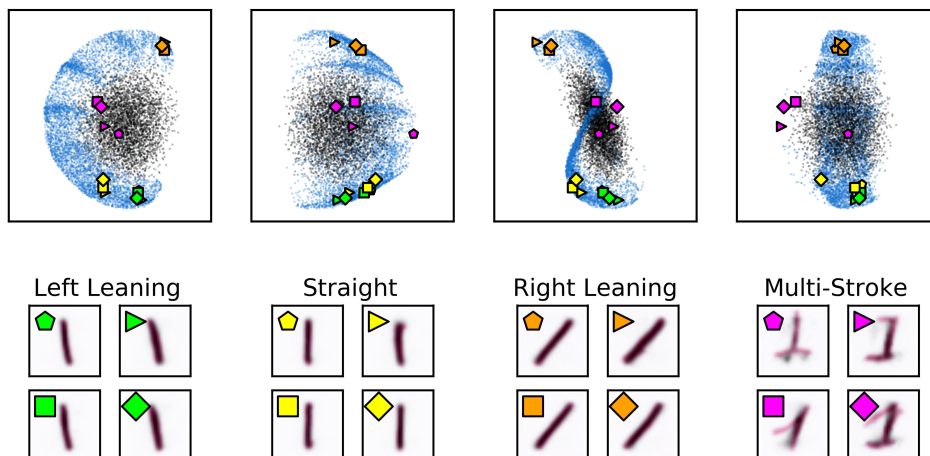


Figure 21: **Two dimensional PCA projections of the MNIST digit data for ones, the Jeffrey’s Prior sample of the model manifold and a few selected images from the dataset.** The top row corresponds to the top row in Figure 20. Ones with differing distinct characteristics such as tilt or bases have been labeled to display how the manifold is roughly arranged according to digit behavior. The images corresponding to these digits are shown on the bottom row. For each, the original digit has been plotted in pink with the reconstruction overlaid in black.

nalize large weights which would allow certain neurons to swamp the signal. Other techniques such as dropout [56] explicitly drop neurons at random during training forcing all neurons to ‘pull their own weight’. Another successful technique, channel out [102, 111], allows sets of neurons to be activated only for certain tasks. The goal, however, is still to make use of the whole network and avoid computational waste. Yet other techniques employ initializations designed to prevent the network from collapsing onto a few modes [94]. In the case presented here, only simple whitening of the data was used.

In addition to the PCA plots, we employed a number of other tactics to search for a thin direction. These include sampling of slices of the manifold and a geodesic analysis. For the former, we found the two furthest points in the manifold and defined a slice at the midpoint of this vector. Sampling was then performed in the slice and the procedure repeated. This procedure yields a sequence of slices

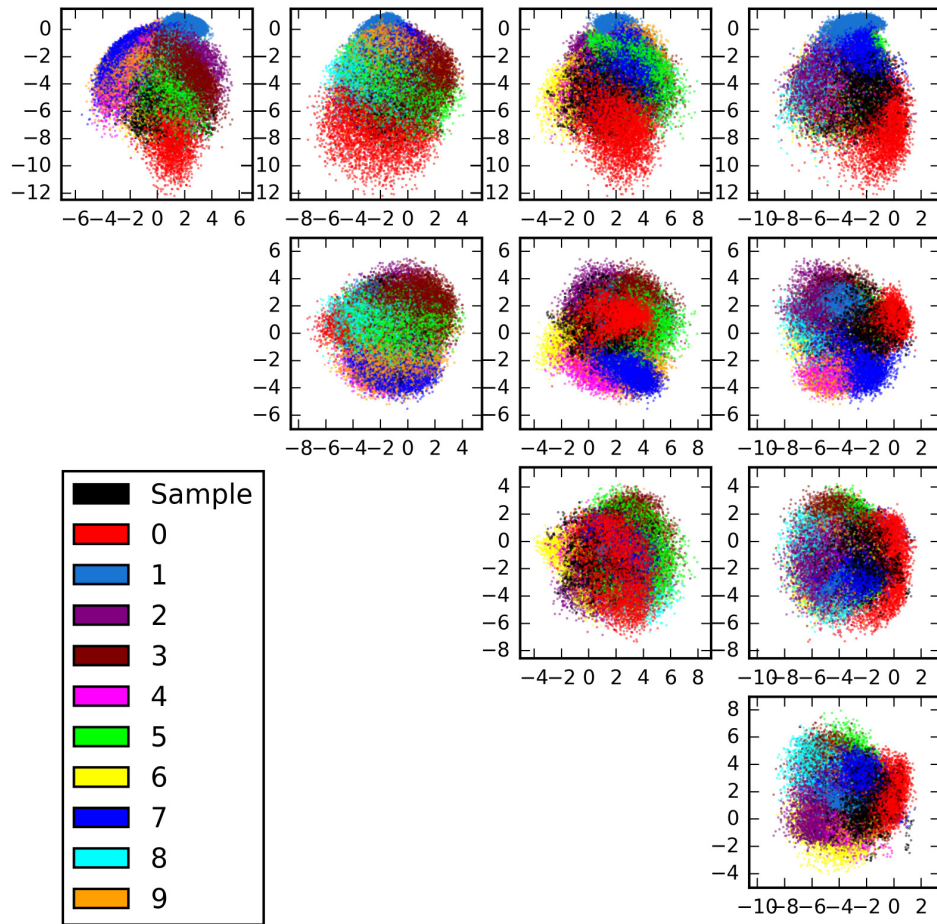


Figure 22: **Two dimensional PCA projections of the MNIST digit data and the Jeffrey's Prior sample of the model manifold for a 4 layer DBN.** The $(i, j)^{th}$ figure in the grid is a plane spanned by singular vectors i and $j + 1$ of the centered data.

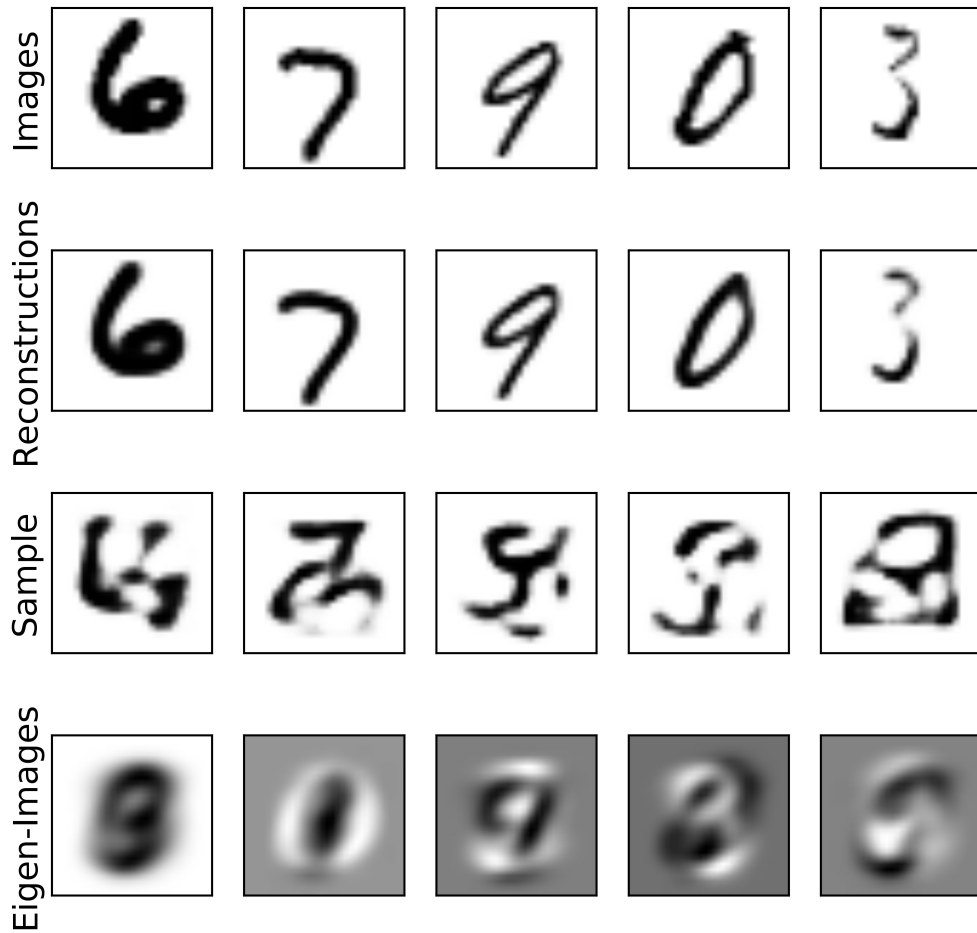


Figure 23: **Examples of MNIST images, their reconstructions, and images sampled using Jeffrey's Prior for the DBN.** For the sampled 'digits', each snapshot corresponds to the same walker. The final row corresponds to the top five 'eigen-digits' of the dataset.

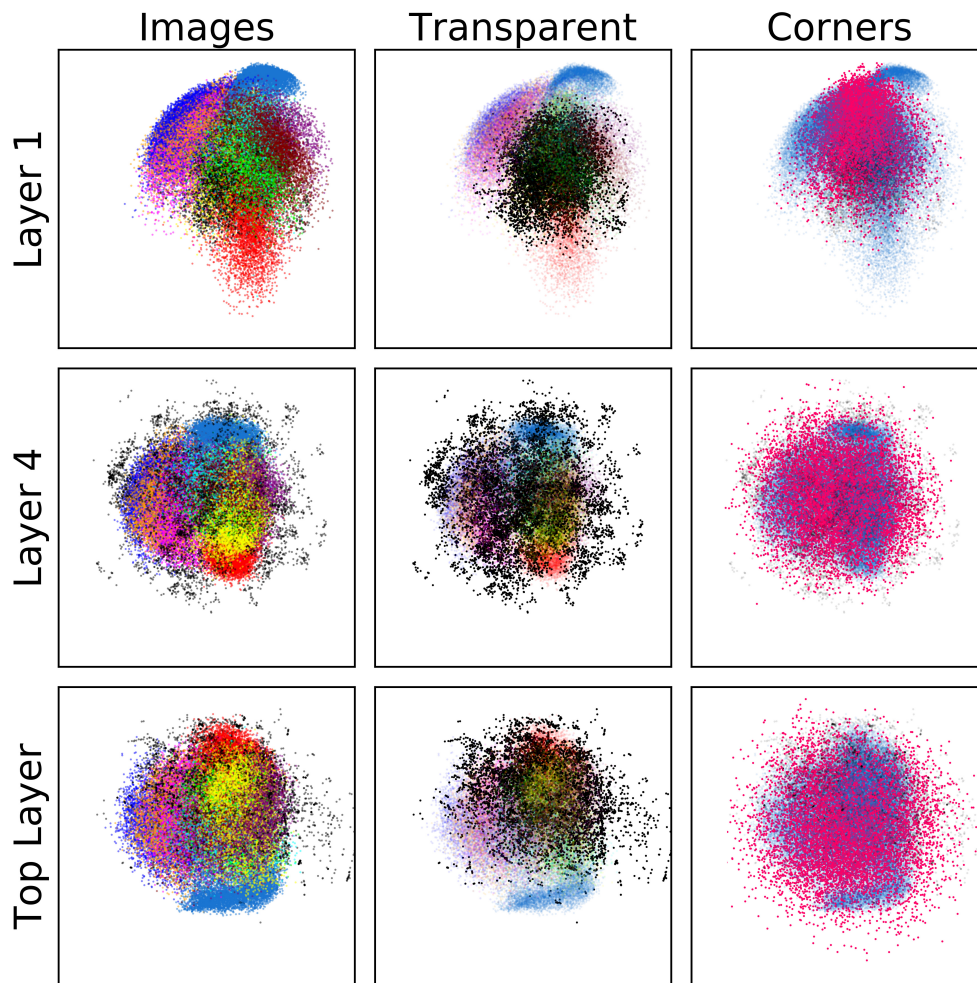


Figure 24: **PCA projection of the Jeffrey's Prior sampling with MNIST digit data for the DBN along the two largest principal component vectors.** In the middle column 'Transparent' the transparency of the MNIST points have been enhanced to show the position of the sampling. On the right under 'Corners' the digits and sampling have again been plotted in blue and black respectively with the corners added in pink. Corners correspond to activations in the top hidden layer ($\vec{\theta}$) for which $\theta_i \in \{-\infty, \infty\} \sim \{-10^6, 10^6\}$. The axes are shared along each row. In order to deal with the large values of the corners and sampling, the top layer is shown with a sigmoid applied. Although the sampling spans the parameter space (Top Layer), it is subsequently mapped to the interior of the manifold in reconstruction space ('Layer 1'). Note that the radius of the sampling is roughly $\frac{2}{3}$ that of the digits - resulting in a 30-D volume $(\frac{2}{3})^{30} = 5.2 \times 10^{-6}$ of the total volume. Notice also that the digits lie outside the corners.

which we constrained to be perpendicular. The hierarchy of widths uncovered by this procedure did not yield a thin direction. In addition, several variants were implemented in which we sliced along the densest section rather than midpoint and varied the width of the slice. We also explored the widths using geodesic analysis. Calculating geodesics from a central point radiating in a plane can be used to map the boundaries. We found a suggestion of thin directions. Under further scrutiny, these appeared to be due to curvature which caused the geodesic to strike a boundary prematurely. In all of our investigations, the manifold truly seems to form an only slightly elongated hyperball rather than a hyper-ribbon.

There is a characteristic of the dataset and networks we employ, however, which obscure the clear geometrical arguments made in other fields for the relative importance of parameter combinations. Namely, the data itself forms a hyperball. All images of digits have many saturated (white/black) pixels corresponding to the boundary of the manifold as dictated by the sigmoidal structure of the SdA and DBN. For this reason, it is difficult to state with certainty that the structure observed can be interpreted in the standard way.

Instead, we find evidence motivating a hypothesis that the $N-1$ and N -dimensional descriptions have a boundary relationship: $\mathcal{M}_{N-1} \approx \partial\mathcal{M}_N$. First, we find that the reconstructed data lies on the 'outside' of the neural manifold (Figures 24, 36 and 34), forming a shell around the L^2 Jeffrey's prior volume of the possible image reconstructions. This naturally suggests that lower dimensional representations will also hug the outside. Second, we have explicitly constructed lower dimensional neural manifolds by training an additional autoencoder layer on the 30 dimensional top hidden layer of the DBN and SdA. In each case, we first applied a sigmoid and the hidden layer consisted of 3 neurons. PCA projections of the corresponding 3 dimensional model manifold along with the original 30 dimensional version embedded in data space are shown in Figures 38 and 39. For the DBN

the lower dimensional representation is consistently on the boundary indicating that $\mathcal{M}_3 \lesssim \dots \lesssim \partial\mathcal{M}_{29} \lesssim \partial\mathcal{M}_{30}$. For the SdA, PCA along the first two principal components displays the lower dimensional manifold on the boundary. In the third, it appears to pierce the space. The emphasis on the boundaries for deep networks is in parallel to behavior observed in other models. Recently, work by Transtrum et. al. [107] has explored the relationship between manifold boundaries and emergent model classes. Their work in information topology has uncovered that boundaries of the model manifolds they explore each correspond to different classes of reduced models. The topology of the space in turn governs the best low dimensional description.

Geometry is strange in high dimensions. This can be illustrated by the following apparent contradiction: the majority of the volume of a hyperball is located near the boundary but it takes a huge sampling to obtain even one point in a cap of the ball. Given our observation that \mathcal{M}_{30} is roughly a slightly elongated hyperball, 95.8 % of the the volume is within 10% of the surface. This appears to imply that Jeffrey’s prior should be appropriate to sample the region near the boundary on which the actual reconstructed data lies. Conversely, consider a N dimensional hyperball of radius, r, with a cap of height, h. The volume of the cap is given by

$$V_n^{cap}(r) = \frac{1}{2}V_n(r)I_{\sin^2\phi}\left(\frac{n+1}{2}, \frac{1}{2}\right) \quad (32)$$

where V_N is the volume of the hyperball, I corresponds to the regularized incomplete beta function and $\phi = \sin^{-1}\frac{r-h}{r}$ is the colatitude angle such that $0 \leq \phi \leq \frac{\pi}{2}$ [68]. For a 10-D sphere with radius 1 and cap height 0.1, the volume of the cap corresponds to a tiny 0.0014% of the volume of the ball. In terms of sampling, this states that one would need to sample approximately 10^5 points with Jeffrey’s Prior to obtain even a single point in this cap. In this light, it seems that instead,

data located on the boundary in the way we observe suggests that a prior based on volume actually biases against the interesting regions.

The model manifolds formed by the neural networks we study have novel properties. In short, they do not appear to form the hyper-ribbons seen in other fields. Additionally, for Deep Belief Networks and Stacked Denoising Autoencoders trained on the MNIST digit dataset, the vast majority of the data arranges itself of the boundaries of the manifold. Sampling uncovers that the interior, which represents the bulk of the images the network can describe, does not correspond to actual images. This implies that either the network is in some sense wasteful or that Jeffrey's Prior is naturally biased to explore uninteresting regions of the model.

7 SEQUENTIAL IMPORTANCE SAMPLING

Sampling is a very important tool which is often employed in instances where it is impossible to analytically solve a problem. As discussed in the previous section, however, there are ways in which it can be problematic. Choice of algorithm, prior and a host of hyperparameters can greatly affect the performance and the results obtained. This section discusses another sampling algorithm in order to illustrate this truth. This sampling algorithm, now known as sequential importance sampling was used originally by Knuth to explore self avoiding random walks on a lattice. In this problem, however, the random variable of interest has a long tailed distribution. In addition, the variance grows exponentially compared to the mean as system size is increased. This results in an exponentially more difficult problem for which the sampling algorithm is not equipped.

In 1976, Donald E. Knuth created an algorithm to sample the self-avoiding paths traversing a 10 by 10 grid of squares from the bottom left corner to upper right [61]. This type of algorithm is now referred to as sequential importance sampling. An image of one of these original self-avoiding walks is shown in Fig 25. To generate these walks, each step is chosen with equal probability from the available paths. For example, starting from the bottom left, one can move up or to the right. Each path is chosen with equal probability. Since in this case there are two paths, the probability of this step is 1/2. Choices which result in a trapped walk are not considered. Continuing in this way, a path is generated which has probability

$$p(w) = \prod_i^L p(s_i) \quad (33)$$

where $p(s_i)$ is the probability of step i and L is the number of steps the walk takes to reach the top right corner.

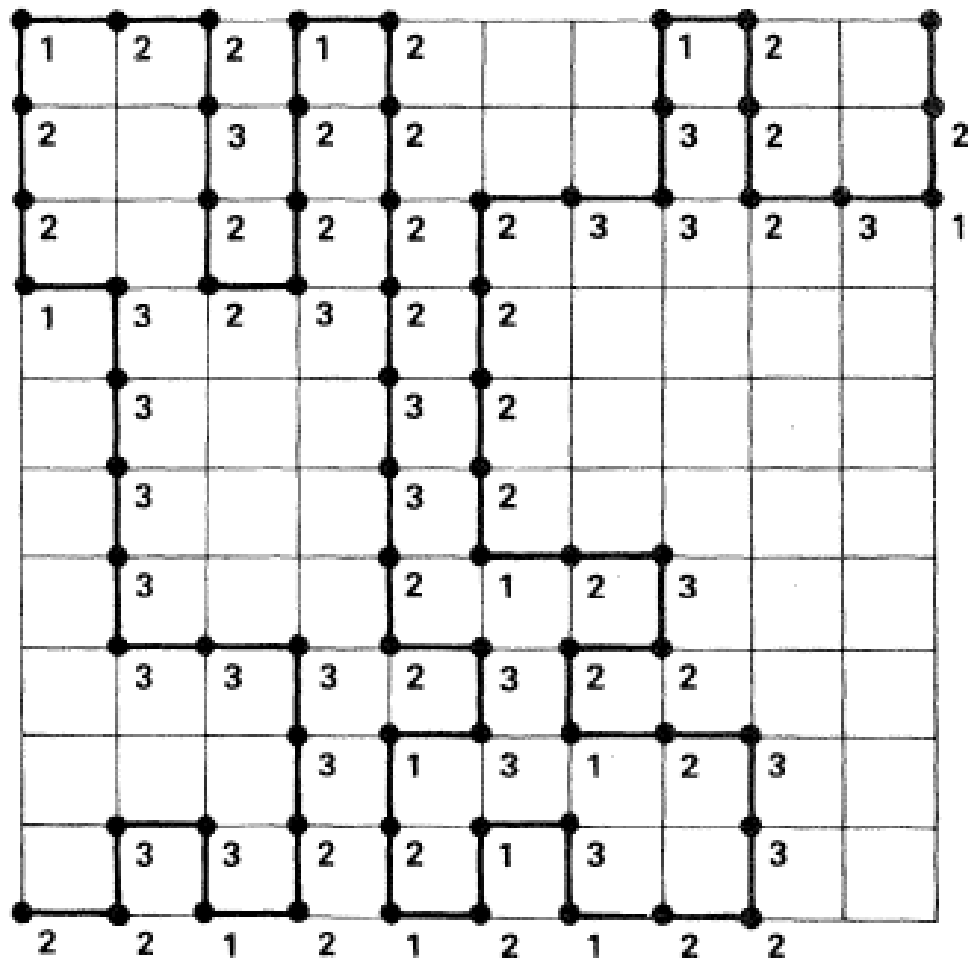


Figure 25: **One of Knuth's original samples on a 10x10 grid [61].** Numbers refer to the number of choices available at each vertex.

Knuth's original algorithm does not describe how to avoid steps which lead to a trapped walk. One method, introduced by Mireille Bousquet-Mélou [18] is based on a winding number approach. The approach here uses a different method. Before each step, the points located one move away from the current location are listed. Then a cluster is grown from the top right to find the portion of the grid connected to the desired end point and avoiding the steps already taken. The intersection of the points connected to the current vertex and the cluster connected to the end point compose the allowed steps. The probability of making each choice (as given

by Knuth) is simply one over the number of allowed steps.

Using the combination of Knuth's algorithm and this cluster method, 200 walks were generated on a 100x100 grid. Some samples obtained are shown in Fig. 26. Generating samples allows for calculation of some statistics for the walks. One (seemingly straightforward) statistic is the number of possible walks. Consider a walk $w \in W_k$ where W_k is the set of all walks on a lattice composed of $k \times k$ squares. Let $X_k = 1/p(w)$ be the random variable of interest. We then have

$$E[X_k] = \sum_{w \in W_k} p(w) \frac{1}{p(w)} = |W_k| \quad (34)$$

So the number of walks can be approximated using

$$\frac{1}{N} \sum_i^N \frac{1}{p(w^{(i)})} \quad (35)$$

where N is the size of the sample. For the first set of 100 walks, $|W_k| = 3.00 \cdot 10^{1581}$. For the second, $|W_k| = 1.88 \cdot 10^{1544}$. These vary widely!! To make matters worse, the value has been shown to scale as λ^{k^2} where $1.628 < \lambda < 1.782$. This sets the bounds on $|W_k|$ to be $3.50 \cdot 10^{2116} < |W_k| < 1.19 \cdot 10^{2509}$. To explore the reason behind this discrepancy, it is important to understand how the variance scales with k . Namely, if we were to look at the *relative variance*:

$$Var\left(\frac{X_k}{E[X_k]}\right) \quad (36)$$

one would find that it grows exponentially with k . This indicates that as k increases, pinning down an estimate with reasonable bounds on the error becomes exponentially more difficult.

Recently, Mireille Bousquet-Mélou published work studying how the variance

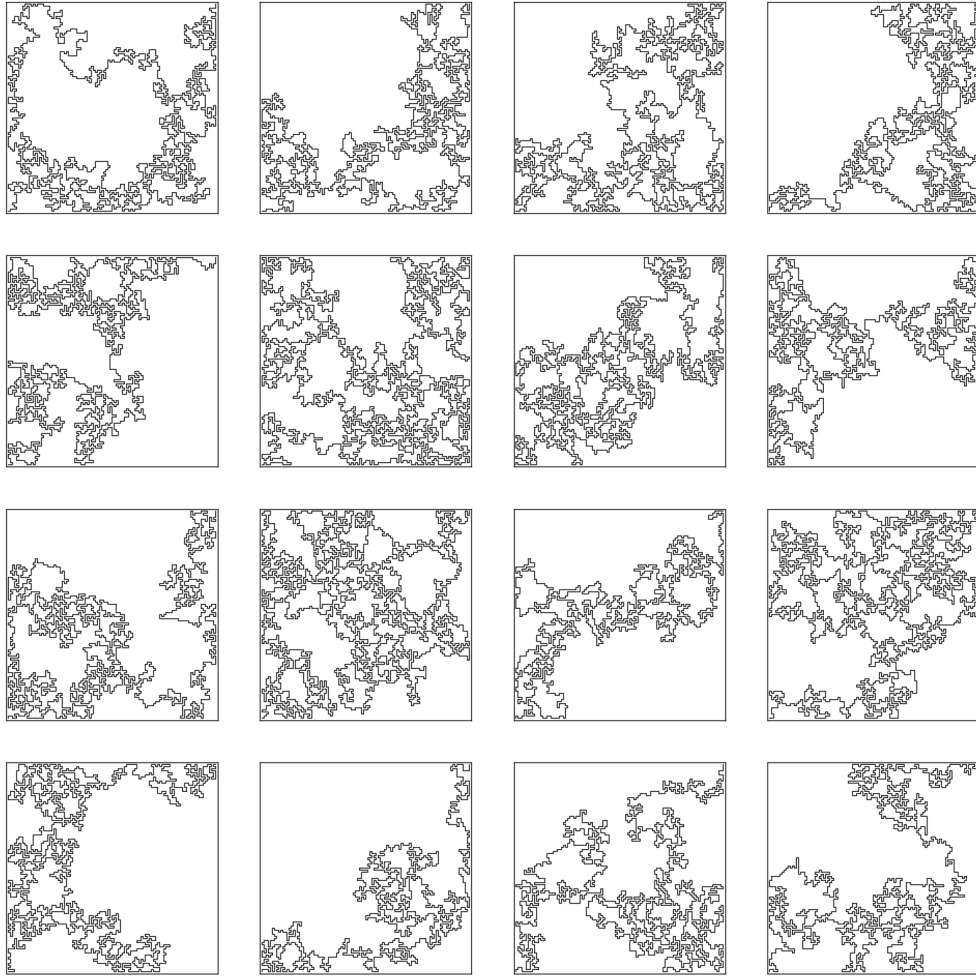


Figure 26: Images of the first few samples on a 100 by 100 grid.

scales in this problem [18]. There are three main quantities of interest: λ , β and κ . These are defined below.

$$E[X]^{1/k^2} \rightarrow \lambda \quad (37)$$

$$E[X^2]^{1/k^2} \rightarrow \beta \quad (38)$$

$$\frac{E[X^2]}{E[X]^2} \rightarrow \kappa^{k^2} \quad (39)$$

In a previous paper, λ was estimated to be 1.744550 ± 0.000005 [19]. To get an estimate for β and κ 50,000 walks each were sampled for $k = 4$ through $k = 10$.

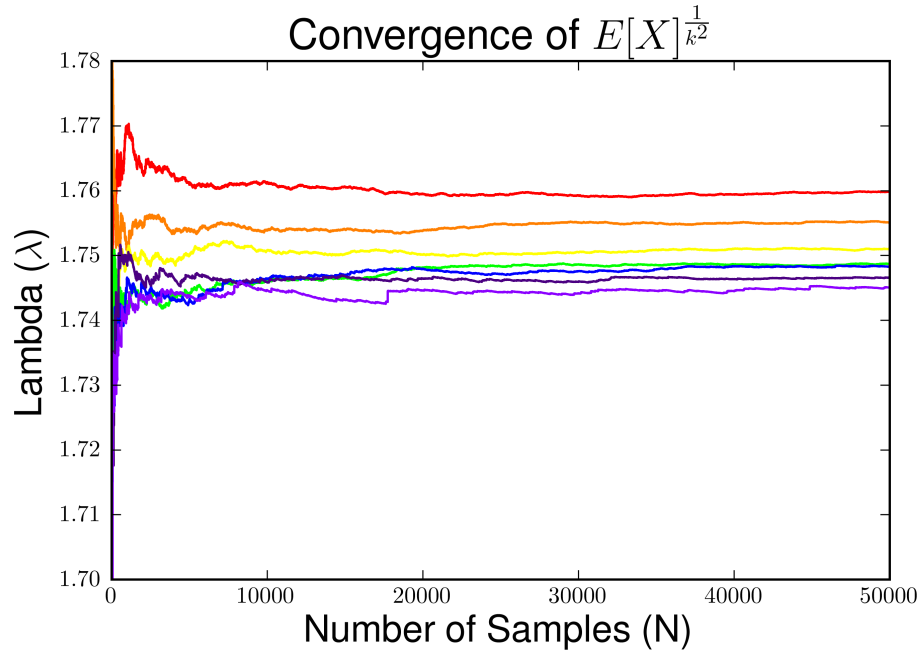


Figure 27: λ as a function of the number of walks sampled. (red: $k=4$, orange: $k=5$...purple: $k=10$)

Constant	Mean	Variance
λ	1.75063502174	2.30917727055e-05
β	3.22833544836	0.000670270353013
κ	1.05336386331	8.37442961221e-06
$\kappa = \beta/\lambda^2$	1.05338572854	8.37442961221e-06

Table 9: Numerical values of the constants obtained from sampling.

Convergence plots are shown in Figures 27 - 29. The average values obtained are given in Table 9. Comparison of the number of walks with the estimate given by sampling are given in Table 10.

The huge error in the estimate for the number of walks gives a sense of how difficult it would be to obtain an accurate value using this method. Combining this with knowledge of how the variance scales, however, we can get an even better feel

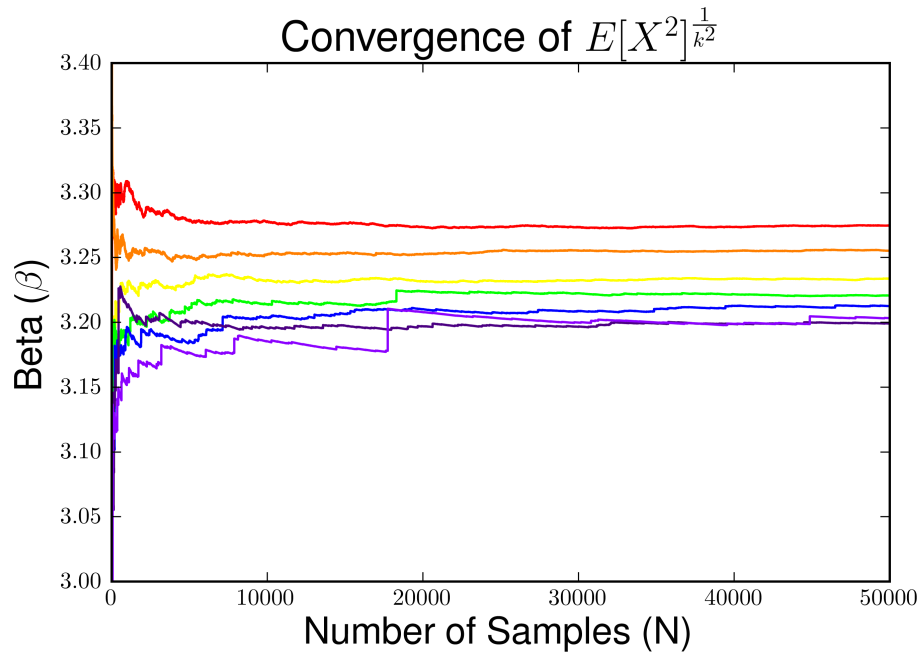


Figure 28: β as a function of the number of walks sampled. (red: $k=4$, orange: $k=5$...purple: $k=10$)

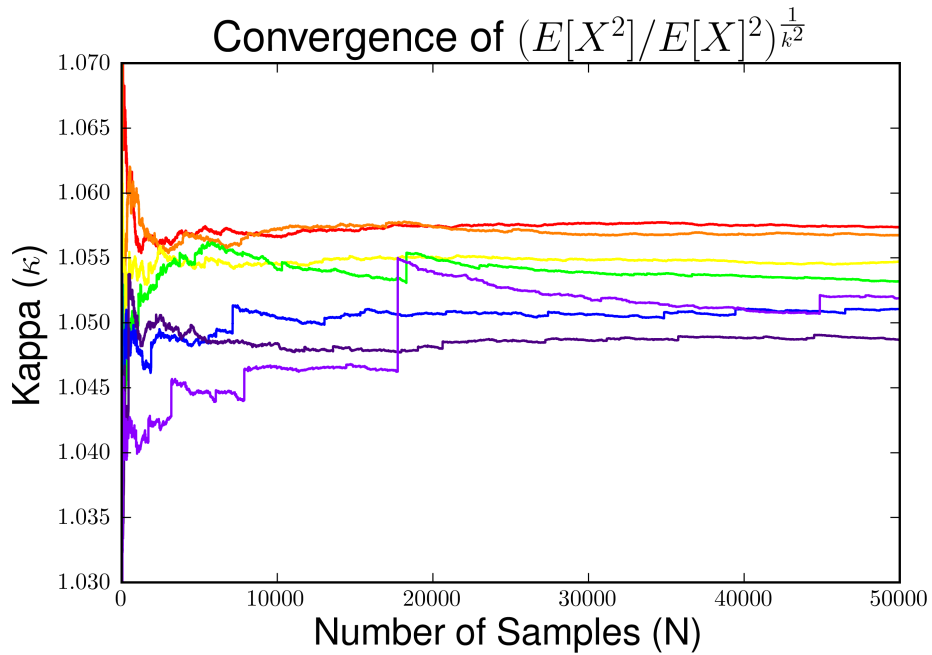


Figure 29: κ as a function of the number of walks sampled. (red: $k=4$, orange: $k=5$...purple: $k=10$)

True Value	Estimated
8512.0	8461.34936
1262816.0	1280892.66416
575780564.0	573422914.262
789360053252.0	781268497533.0
3.26659848698e+15	3.35978934841e+15
4.10442087026e+19	4.14704491265e+19
1.56875803046e+24	1.51295082827e+24

Table 10: **Comparison of the known number of walks with the estimates given by sampling.**

for how bad this problem really is. Consider Chebyshev’s Lemma which states:

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2} \quad (40)$$

for any random variable X with expected value μ , non-zero variance σ , and real number $k > 0$. As described in [18] the variance of our estimator is $Var(X_k)/N$. Letting $k = \epsilon/\sigma$ yields:

$$P(|X - \mu| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2} \quad (41)$$

Suppose we want the error in our estimate to be on the order of $|W_k| \approx \lambda^{k^2}$. Substituting for ϵ we have

$$P(|X - \mu| \geq \lambda^{k^2}) \leq \frac{1}{N} \frac{E[X^2] - E[X]^2}{E[X]^2} \rightarrow \frac{\kappa^{k^2}}{N} \quad (42)$$

Hence to get within an order of magnitude with probability greater than 90% one would need to sample approximately $6.08 \cdot 10^{226}$ walks. On 1 cpu, the algorithm presented here takes on the order of hours to generate 1 walk. Assume a scenario of 1 walk per hour. To generate enough walks for a semi-accurate estimate using this method would take about $5.07 * 10^{212}$ times the age of the universe. Putting it in this perspective makes it clear just how problematic sequential importance

sampling can be. Error estimates must be tied to an understanding of the relative variance in the problem or else estimates obtained are essentially meaningless.

8 DISCUSSION

Data analysis plays an important role in the Physics community and beyond. As such, new tools and techniques with differing ranges of applicability are continuously being developed. This thesis has discussed two philosophically different approaches, traditional methods involving the development of specialized mathematical theories that give insight into the underlying structure of a problem and new 'machine learning' approaches with vast applicability but limited interpretability.

The analysis of the two-dimensional random-field Ising model using a normal form theory of the Renormalization Group is a very nice example of the first type of approach. In this case, understanding the underlying mathematical structure of critical points and universality informs a broadly useful method to dealing with what are traditionally very difficult cases. Rather than needing to develop new renormalization group methods to extend its applicability on a case by case basis, all that is necessary is a basic understanding of the symmetries of the system or an afternoon to simply 'guess and check' candidate bifurcations. The straightforward, systematic approach provided enables an understanding of irregularities in the data, namely the inability to be collapsed using standard power law approaches, and presents a principled solution.

Through application of normal form theory we are able to elucidate that the distributions of the avalanche sizes and derivative of the magnetization with respect to the magnetic field are both well described by the presence of a transcritical bifurcation in the disorder flow equations. This finding suggests that the lower critical dimension of the non-equilibrium random-field Ising model is two. Our results are also consistent with a critical disorder, $r_c = 0$. In previous study of this model, collapses were only able to be performed over a small range in the disorder of $\sim 10\%$. In contrast, analyzing this data with an eye toward what physically

sets the two dimensional version apart allows for a collapse over a factor of 10 in the disorder.

Our novel approach to dealing with collapses in the lower and upper critical dimensions is informative and practical. In many cases, however, our ignorance of a governing theory precludes such a systematic approach. Machine learning is particularly well suited for dealing with these instances. In implementing these algorithms, however, it is very important to be aware of potential pitfalls. One such consideration is in the choice of algorithm. For example, preliminary data analysis of stock market returns reveals underlying tetrahedral structure. Without this initial exploration, it would have come down to sheer luck to have selected an algorithm so well suited to this particular problem. Choice of algorithm can have a huge impact on the effectiveness and interpretability of the results and as such should be made with care.

Another example of this can be seen in the results of the Higgs Boson Machine Learning Challenge. In this case, although it seems domain knowledge could have been leveraged to increase predictive performance, the choice in algorithm and training method far outweighed the incremental improvement hand engineered features were able to provide. In contrast to the unsupervised approach discussed in the context of stocks, however, the choice in algorithm here was made through 'domain knowledge' of the machine learning field itself rather than by any emergent structure in the data. By keeping abreast of current advances and adaptive adjustments to long-standing paradigms, data scientists in this instance were given a competitive edge.

The performance of machine learning algorithms can be difficult to conceptualize and reason about without extensive knowledge of the subfield. For example, one might expect that, like multi-parameter models found in other fields, these models should display a hierarchical structure in which certain combinations of parame-

ters dominate while others barely contribute. This type of underlying behavior is ubiquitous in physics and helps us to reason about why we can expect science to work in the first place and to make sense of concepts like emergence. This hierarchical structure is reflected in the presence of a hyper-ribbon model manifold which appears to be conspicuously absent in prototypical neural networks trained on a standard dataset. On the one hand, an absence of a hyper-ribbon could be in essence what defines them; if the underlying pieces and their interactions were able to be simply understood it would be unnecessary to attempt a machine learning approach to begin with. On the other hand, sampling and consequently visualizing high dimensional spaces is still a challenging problem. In recent years, variational autoencoders have become much more popular due to their superior performance [60]. These models are probabilistic in nature and could provide a much more natural way to sample the underlying space than Jeffrey's Prior. Despite this, preliminary results suggest that sampling these networks results in the same observed behavior; the sample is concentrated in the interior, while real data concentrates at the boundaries (Appendix H). Another potentially insightful approach would be to combine this sampling with high-dimensional visualization schemes such as InPCA [88] which could prove much more effective in analyzing the underlying manifold effectively through the lens of information geometry.

This thesis has explored the application of several machine learning approaches and also pointed out several potential hurdles. Although the philosophy of these approaches differs from canonical physics modelling, certain aspects remain the same. For example, both approaches provide robust methods to analyze data. Additionally, just as with physics models, learning more about the nature of these algorithms, why they work and in what cases they fail, can instruct the creation of new and better methods. It is my hope that each of these approaches may be leveraged to learn about the other and inform new paradigms. Indeed there has

already been interest in this direction [2]. Conversely, understanding at a deep level why machine learning performs so well could yield interesting insight into the way we think about physics modelling.

BIBLIOGRAPHY

- [1] Guillaume Alain and Yoshua Bengio. “What Regularized Auto-encoders Learn from the Data-generating Distribution”. In: *J. Mach. Learn. Res.* 15.1 (2014), pp. 3563–3593.
- [2] Alexander A Alemi and Ian Fischer. “TherML: Thermodynamics of machine learning”. In: *arXiv preprint arXiv:1807.04162* (2018).
- [3] Ivan Balog, Gilles Tarjus, and Matthieu Tissier. “Criticality of the random field Ising model in and out of equilibrium: A nonperturbative functional renormalization group description”. In: *Phys. Rev. B* 97 (9 2018), p. 094204.
- [4] Ivan Balog, Matthieu Tissier, and Gilles Tarjus. “Same universality class for the critical behavior in and out of equilibrium in a quenched random field”. In: *Phys. Rev. B* 89 (10 2014), p. 104201.
- [5] N. Basalto et al. “Clustering stock market companies via chaotic map synchronization”. In: *Physica A: Statistical Mechanics and its Applications* 345.1&A2 (2005), pp. 196–206.
- [6] Frédéric Bastien et al. *Theano: new features and speed improvements*. Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop. 2012.
- [7] S. Becker and Y. LeCun. “Improving the convergence of back-propagation learning with second order methods”. In: *Tech. Rep., Department of Computer Science, University of Toronto* (1988).
- [8] Yoshua Bengio et al. “Better Mixing via Deep Representations.” In: *ICML*. Vol. 28. 2013, pp. 552–560.
- [9] James Bergstra et al. “Theano: a CPU and GPU Math Expression Compiler”. In: *Proceedings of the Python for Scientific Computing Conference (SciPy)*. Oral Presentation. 2010.

- [10] Berry Petroleum Company History. 2013. URL: <http://www.bry.com/pages/history.html>.
- [11] G. Bertotti. *Hysteresis in Magnetism*. Academic, New York, 1998.
- [12] Michael Betancourt et al. “The Geometric Foundations of Hamiltonian Monte Carlo”. In: *Submitted to Statistical Science* (2014).
- [13] Giovanni Bonanno, Nicolas Vandewalle, and Rosario N. Mantegna. “Taxonomy of stock market indices”. In: *Phys. Rev. E* 62 (6 2000), R7615–R7618.
- [14] Giovanni Bonanno et al. “Topology of correlation-based minimal spanning trees in real and model markets”. In: *Phys. Rev. E* 68 (4 2003), p. 046130.
- [15] J. A. Bonetti et al. “Electronic Transport in Underdoped $\text{YBa}_2\text{Cu}_3\text{O}_{7-\delta}$ Nanowires: Evidence for Fluctuating Domain Structures”. In: *Phys. Rev. Lett.* 93 (8 2004), p. 087002.
- [16] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. “D3: Data-Driven Documents”. In: *IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis)* (2011).
- [17] Jean-Philippe Bouchaud. In: *J. Stat. Phys.* 151 (2013), p. 567.
- [18] M. Bousquet-Mélou. “On the Importance Sampling of Self-Avoiding Walks”. In: *Combinatorics, Probability & Computing* 23.5 (2014), pp. 725–748.
- [19] M. Bousquet-Mélou, A.J. Guttmann, and I. Jensen. “Self-Avoiding Walks Crossing a Square”. In: *Journal of Physics A: Mathematical and General* 38.42 (2005), pp. 9159–9181.
- [20] A J Bray and M A Moore. “Scaling theory of the random-field Ising model”. In: *Journal of Physics C: Solid State Physics* 18.28 (1985), p. L927.
- [21] Leo Breiman. “Bagging predictors”. In: *Machine Learning* 24.2 (1996), pp. 123–140.

- [22] Leo Breiman. “Random Forests”. In: *Machine Learning* 45.1 (2001), pp. 5–32.
- [23] L. Breiman and J. H. Friedman. “Learning classification trees”. In: *Statistics and Computing* 2 (1992), pp. 63–73.
- [24] Thomas Bury. “Market structure explained by pairwise interactions”. In: *Physica A: Statistical Mechanics and its Applications* 392.6 (2013), pp. 1375–1385.
- [25] E. W. Carlson et al. “Hysteresis and Noise from Electronic Nematicity in High-Temperature Superconductors”. In: *Phys. Rev. Lett.* 96 (9 2006), p. 097003.
- [26] CBOE[®] Oil Index. 2013. URL: <http://www.cboe.com/products/IndexComponentsAuto.aspx?PRODUCT=OIX>.
- [27] Djordje Spasojević, Sanja Janičević, and Milan Knežević. “Avalanche distributions in the two-dimensional nonequilibrium zero-temperature random field Ising model”. In: *Phys. Rev. E* 84 (5 2011), p. 051119.
- [28] Djordje Spasojević, Sanja Janičević, and Milan Knežević. “Numerical Evidence for Critical Behavior of the Two-Dimensional Nonequilibrium Zero-Temperature Random Field Ising Model”. In: *Phys. Rev. Lett.* 106 (17 2011), p. 175701.
- [29] Olga Perković, Karin Dahmen, and James P. Sethna. “Avalanches, Barkhausen Noise, and Plain Old Criticality”. In: *Phys. Rev. Lett.* 75 (24 1995), pp. 4528–4531.
- [30] Olga Perković, Karin Dahmen, and James P. Sethna. “Disorder-Induced Critical Phenomena in Hysteresis: A Numerical Scaling Analysis”. In: *arXiv:cond-mat/9609072* (1996).

- [31] Ricky Chachra, Mark K. Transtrum, and James P. Sethna. “Structural susceptibility and separation of time scales in the van der Pol Oscillator”. In: *Phys. Rev. E* 86.026712 (2012).
- [32] Tianqi Chen, Kailong Chen, and Tong He. *XGBoost*. <http://mloss.org/software/view/543/>. 2015.
- [33] Yan-Jiun Chen et al. “Avalanche Spatial Structure and Multivariable Scaling Functions; Sizes, Heights, Widths, and Views through Windows”. In: *Phys. Rev. E* 84 (Dec. 2011), p. 061103.
- [34] Francesca Colaiori et al. “Phase Transitions in a Disordered System in and out of Equilibrium”. In: *Phys. Rev. Lett.* 92 (25 2004), p. 257203.
- [35] The ATLAS Collaboration. “Evidence for higgs boson decays to tau+tau-final state with the atlas detector”. In: *Tech. Rep. ATLAS-CONF-2013-108* (2013).
- [36] T. Conlon, H.J. Ruskin, and M. Crane. “Cross-correlation dynamics in financial time series”. In: *Physica A: Statistical Mechanics and its Applications* 388.5 (2009), pp. 705–714.
- [37] C Coronello et al. “Sector identification in a set of stock return time series traded at the London Stock Exchange”. In: *Acta Physica Polonica B* 36 (2005), pp. 2653–2679.
- [38] G. Cowan et al. “Asymptotic formulae for likelihood-based tests of new physics”. In: *The European Physical Journal C* 71 (2011), 1554–1573.
- [39] Adele Cutler and Leo Breiman. “Archetypal Analysis”. In: *Technometrics* 36 (4 1994), pp. 338–347.

- [40] K. Dahmen and J. P. Sethna. “Hysteresis, avalanches, and disorder-induced critical scaling: A renormalization-group approach”. In: *Phys. Rev. B* 53 (June 1996), pp. 14872–14905.
- [41] F. Detcheverry et al. In: *Langmuir* 20 (2004), p. 8006.
- [42] Dow Jones[®] US Indices: Industry Indices. 2015. URL: www.djindexes.com/mdsidx/downloads/fact_info/Dow_Jones_US_Indices_Industry_Indices_Fact_Sheet.pdf.
- [43] Gabriel Doyle and Charles Elkan. “Financial Topic Models”. In: *NIPS Workshop on Applications for Topic Models: Text and Beyond*. Whistler, Canada, 2009.
- [44] B. Drossel and K. Dahmen. “Depinning of a domain wall in the 2d random-field Ising model”. In: *The European Physical Journal B - Condensed Matter and Complex Systems* 3.4 (1998), pp. 485–496.
- [45] John Duchi, Elad Hazan, and Yoram Singer. *Adaptive Subgradient Methods for Online Learning and Stochastic Optimization*. Tech. rep. UCB/EECS-2010-24. EECS Department, University of California, Berkeley, 2010. URL: <http://www.eecs.berkeley.edu/Pubs/TechRpts/2010/EECS-2010-24.html>.
- [46] Cheoljun Eom et al. *Topological Properties of Stock Networks Based on Random Matrix Theory in Financial Time Series*. Papers. arXiv.org, 2007.
- [47] Eugene F Fama and Kenneth R French. “Common risk factors in the returns on stocks and bonds”. In: *Journal of financial economics* 33.1 (1993), pp. 3–56.
- [48] Daniel J. Fenn et al. “Temporal evolution of financial-market correlations”. In: *Phys. Rev. E* 84 (2 2011), p. 026109.

- [49] D. Foreman-Mackey et al. “emcee: The MCMC Hammer”. In: *arXiv* 125 (2013), pp. 306–312.
- [50] I. J. Goodfellow et al. “Multi-digit number recognition from street view imagery using deep convolutional neural networks”. In: *In Proc. ICLR* (2014).
- [51] Jonathan Goodman and Jonathan Weare. “Ensemble Samplers with Affine Invariance”. In: *Communications in Applied Mathematics and Computational Science* 5.1 (2010).
- [52] Ryan N. Gutenkunst et al. “Universally Sloppy Parameter Sensitivities in Systems Biology”. In: *PLoS Comput Biol* 3(10).e189 (2007).
- [53] Lorien X. Hayden et al. “Canonical sectors and evolution of firms in the US stock markets”. In: *Quantitative Finance* 18.10 (2018), pp. 1619–1634.
- [54] Tapio Heimo, Kimmo Kaski, and Jari Saramäki. “Maximal spanning trees, asset graphs and random matrix denoising in the analysis of dynamics of financial networks”. In: *Physica A: Statistical Mechanics and its Applications* 388.2–3 (2009), pp. 145–156.
- [55] Geoffrey Hinton and Ruslan Salakhutdinov. “Reducing the Dimensionality of Data with Neural Networks”. In: *Science* 313.5786 (2006), pp. 504–507.
- [56] Geoffrey E. Hinton et al. “Improving neural networks by preventing co-adaptation of feature detectors”. In: *CoRR* abs/1207.0580 (2012).
- [57] Yoseph Imry and Shang-keng Ma. “Random-Field Instability of the Ordered State of Continuous Symmetry”. In: *Phys. Rev. Lett.* 35 (21 1975), pp. 1399–1401.
- [58] Rie Johnson and Tong Zhang. “Learning Nonlinear Functions Using Regularized Greedy Forest”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 36.5 (2014), pp. 942–954.

- [59] Dong-Hee Kim and Hawoong Jeong. “Systematic analysis of group identification in stock markets”. In: *Phys. Rev. E* 72 (4 2005), p. 046133.
- [60] Diederik P Kingma and Max Welling. “Auto-encoding variational bayes”. In: *arXiv preprint arXiv:1312.6114* (2013).
- [61] D.E. Knuth. “Mathematics and Computer Science: Coping with Finiteness”. In: *Science* 194.4271 (1976), pp. 1235–1242.
- [62] L Kullmann, J Kertész, and Rosario Nunzio Mantegna. “Identification of clusters of companies in stock indices via Potts super-paramagnetic transitions”. In: *Physica A: Statistical Mechanics and its Applications* 287.3 (2000), pp. 412–419.
- [63] M. C. Kuntz. “Barkhausen noise: simulations, experiments, power spectra, and two dimensional scaling”. PhD thesis. Cornell University, 1999.
- [64] Matthew C. Kuntz and James P. Sethna. In: *Phys. Rev. B* 62.11699 (2000).
- [65] Lobisor Kurbah, Diana Thongjaomayum, and Prabodh Shukla. “Nonequilibrium random-field Ising model on a diluted triangular lattice”. In: *Phys. Rev. E* 91 (1 2015), p. 012131.
- [66] Laurent Laloux et al. “Noise Dressing of Financial Correlation Matrices”. In: *Phys. Rev. Lett.* 83 (7 1999), pp. 1467–1470.
- [67] Yann LeCun and Corinna Cortes. *MNIST handwritten digit database*. <http://yann.lecun.com> 2010.
- [68] S. Li. “Concise Formulas for the Area and Volume of a Hyperspherical Cap”. In: *Asian Journal of Mathematics & Statistics* 4 (2011), pp. 66–70.
- [69] M. P. Lilly, A. H. Wootters, and R. B. Hallock. “Spatially Extended Avalanches in a Hysteretic Capillary Condensation System: Superfluid ^4He in Nucleopore”. In: *Phys. Rev. Lett.* 77 (20 1996), pp. 4222–4225.

- [70] Y. Liu and K. A. Dahmen. “Random-field Ising model in and out of equilibrium”. In: *Europhys. Lett.* 86.5 (2009), p. 56003.
- [71] Yang Liu and Karin A. Dahmen. “Unexpected universality in static and dynamic avalanches”. In: *Phys. Rev. E* 79 (6 2009), p. 061124.
- [72] Benjamin B. Machta et al. “Parameter Space Compression Underlies Emergent Theories and Predictive Models”. In: *Science* 342 (2013), pp. 604–607.
- [73] R. N. Mantegna. “Hierarchical structure in financial markets”. In: *The European Physical Journal B - Condensed Matter and Complex Systems* 11.1 (1999), pp. 193–197. ISSN: 1434-6028.
- [74] Amos Maritan et al. “Spin-flip avalanches and dynamics of first order phase transitions”. In: *Phys. Rev. Lett.* 72 (6 1994), pp. 946–946.
- [75] Andre C.R. Martins. “Random, but not so much a parameterization for the returns and correlation matrix of financial time series”. In: *Physica A: Statistical Mechanics and its Applications* 383.2 (2007), pp. 527–532.
- [76] N. Metropolis et al. “Equation of State Calculations by Fast Computing Machines”. In: *The Journal of Chemical Physics* 21.6 (1953), pp. 1087–1092.
- [77] Morgan Stanley[®] High-Tech 35 Index. 2005. URL: www.nasdaq.com/options/indexes/msh.aspx.
- [78] Morten Mörup and Lars Kai Hansen. “Archetypal analysis for machine learning and data mining”. In: *Neurocomputing* 80.0 (2012), pp. 54–63.
- [79] Nicolo Musmeci, Tomaso Aste, and Tiziana Di Matteo. “Relation between Financial Market Structure and the Real Economy: Comparison between Clustering Methods”. In: *SSRN* (2014).

- [80] Dave Nadig and Lara Crigger. “Signal From Noise”. In: *Journal of Indexes* 14 (2 2011), pp. 40–43, 50.
- [81] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.
- [82] Francisco J. Pérez-Reche and Eduard Vives. “Spanning avalanches in the three-dimensional Gaussian random-field Ising model with metastable dynamics: Field dependence and geometrical properties”. In: *Phys. Rev. B* 70 (21 2004), p. 214422.
- [83] E. W. Phillabaum B.and Carlson and K. A. Dahmen. “Spatial complexity due to bulk electronic nematicity in a superconducting underdoped cuprate”. In: *Nature Communications* 3 (2012), 915 EP.
- [84] Vasiliki Plerou et al. “Random matrix approach to cross correlations in financial data”. In: *Phys. Rev. E* 65 (6 2002), p. 066126.
- [85] Plum Creek[®] History. 2014. URL: <http://www.plumcreek.com/AboutPlumCreek/History/tabid/55/Default.aspx>.
- [86] Boris Podobnik and H. Eugene Stanley. “Detrended Cross-Correlation Analysis: A New Method for Analyzing Two Nonstationary Time Series”. In: *Phys. Rev. Lett.* 100 (8 2008), p. 084102.
- [87] William H. Press et al. *Numerical Recipes 3rd Edition: The Art of Scientific Computing*. 3rd ed. New York, NY, USA: Cambridge University Press, 2007. ISBN: 0521880688, 9780521880688.
- [88] Katherine N. Quinn et al. “Visualizing probabilistic models and data with Intensive Principal Component Analysis”. In: *PNAS* 201817218 (2019).
- [89] Archishman Raju et al. “Renormalization group and normal form theory”. In: *arXiv preprint arXiv:1706.00137* (2017).

- [90] Scott Reed et al. “Learning to Disentangle Factors of Variation with Manifold Interaction”. In: *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*. Ed. by Tony Jebara and Eric P. Xing. JMLR Workshop and Conference Proceedings, 2014, pp. 1431–1439.
- [91] F. Rosenblatt. “The Perceptron: A Probabilistic Model for Information Storage in the Brain”. In: *Psychological Review* 65.6 (1958).
- [92] Russell 3000[®] Index. 2015. URL: www.russell.com/indexes/data/fact_sheets/us/russell_3000_index.asp.
- [93] Chris H. Rycroft. “Voro++: A three-dimensional Voronoi cell library in C++”. In: *Chaos* 19 (2009), p. 041111.
- [94] Andrew M. Saxe, James L. McClelland, and Surya Ganguli. “Exact solutions to the nonlinear dynamics of learning in deep linear neural networks”. In: *CoRR* abs/1312.6120 (2013).
- [95] Tom Schaul, Sixin Zhang, and Yann LeCun. “No More Pesky Learning Rates”. In: *International Conference on Machine Learning (ICML)*. 2013.
- [96] Jürgen Schmidhuber. “Deep Learning in Neural Networks: An Overview”. In: *Neural Networks* 61 (2015), pp. 85–117.
- [97] Scottrade[®]. 2015. URL: www.scottrade.com.
- [98] J. P. Sethna, K. A. Dahmen, and O. Perkovic. “Random-Field Ising Models of Hysteresis”. In: *eprint arXiv:cond-mat/0406320* (2004). eprint: cond-mat/0406320.
- [99] Prabodh Shukla and Diana Thongjaomayum. “Criteria for infinite avalanches in the zero-temperature nonequilibrium random-field Ising model on a Bethe lattice”. In: *Phys. Rev. E* 95 (4 2017), p. 042109. DOI: 10.1103/PhysRevE.95.042109.

- [100] Prabodh Shukla and Diana Thongjaomayum. “Hysteresis in random-field Ising model on a Bethe lattice with a mixed coordination number”. In: *Journal of Physics A: Mathematical and Theoretical* 49.23 (2016), p. 235001.
- [101] S&P 500[®] Index. 2014. URL: us.spindices.com/indices/equity/sp-500.
- [102] Rupesh K Srivastava et al. “Compete to Compute”. In: *Advances in Neural Information Processing Systems 26*. Curran Associates, Inc., 2013, pp. 2310–2318.
- [103] S.H. Strogatz. *Nonlinear Dynamics And Chaos*. Studies in nonlinearity. Sarat Book House, 2007. ISBN: 9788187169857. URL: <https://books.google.com/books?id=PHmED2xxrE8C>.
- [104] Diana Thongjaomayum and Prabodh Shukla. “Effect of coordination number on the nonequilibrium critical point”. In: *Phys. Rev. E* 88 (4 2013), p. 042138.
- [105] C. Thureau, K. Kersting, and C. Bauckhage. “Convex Non-negative Matrix Factorization in the Wild”. In: *Data Mining, 2009. ICDM '09. Ninth IEEE International Conference on*. 2009, pp. 523–532.
- [106] Mark K. Transtrum, Benjamin B. Machta, and James P. Sethna. “Geometry of nonlinear least squares with applications to sloppy models and optimization”. In: *Phys. Rev. E* 83.036701 (2011).
- [107] M.K. Transtrum, G. Hart, and P. Qiu. “Information topology identifies emergent model classes”. In: *Submitted* (2014).
- [108] Pascal Vincent et al. “Extracting and Composing Robust Features with Denoising Autoencoders”. In: 2008, pp. 1096–1103.

- [109] Domenico Vistocco and Claudio Conversano. “Visualizing and clustering financial portfolios using internal compositions”. In: *SIS* (2009). Presented at Statistical Methods for the Analysis of Large Data-Sets Pescara, Italy, September 23-25.
- [110] Eduard Vives et al. “Universality in models for disorder-induced phase transitions”. In: *Phys. Rev. E* 52 (1 1995), R5–R8.
- [111] Qi Wang and Joseph JáJá. “From Maxout to Channel-Out: Encoding Information on Sparse Pathways”. In: *CoRR* abs/1312.1909 (2013).
- [112] Joshua J. Waterfall et al. “Sloppy-Model Universality Class and the Vandermonde Matrix”. In: *Phys. Rev. Lett.* 97.150601 (2006).
- [113] Website. “Project Website with additional figures and analyses”. 2013. URL: www.lassp.cornell.edu/sethna/Finance.
- [114] Matthew D. Zeiler. “ADADELTA: An Adaptive Learning Rate Method”. In: *CoRR* abs/1212.5701 (2012).

APPENDIX A CORRELATION LENGTH

The correlation length may be calculated directly from the flow equation of the disorder. For the transcritical form

$$\frac{dw}{d\ell} = w^2 + Bw^3 \quad (43)$$

we have

$$\int_{\ell_0}^{\ell^*} d\ell = \int_{w_0}^{w^*} \frac{dw}{w^2 + Bw^3} \quad (44)$$

where (w_0, ℓ_0) denotes an initial point and (w^*, ℓ^*) a fixed point of the RG, a constant. Performing the integration and letting $(w_0, \ell_0) \rightarrow (w, \ell)$ we obtain

$$\ell \sim B \log \left(B + \frac{1}{w} \right) - \frac{1}{w} \quad (45)$$

We have

$$\xi \sim \exp(-\ell) \quad (46)$$

hence

$$\xi \sim \left(\frac{1}{w} + B \right)^{-B} \exp \left(\frac{1}{w} \right) \quad (47)$$

APPENDIX B INVARIANT SCALING COMBINATIONS

B.1 POWER LAW FORM

As our invariant parameter combinations are unorthodox, we provide here a thorough derivation and a comparison to the usual power law ‘homogeneous’ variables seen at the usual hyperbolic fixed points. The invariant scaling combinations corresponding to traditional power law scaling may be simply derived from the

flow equations in 3 and higher dimensions. We have

$$\begin{aligned}\frac{dw}{d\ell} &= \frac{1}{\nu}w \\ \frac{ds}{d\ell} &= -\frac{1}{\sigma\nu}s \\ \frac{dh}{d\ell} &= \frac{\beta\delta}{\nu}h\end{aligned}\tag{48}$$

Taking $(dw/d\ell)/(ds/d\ell)$ and integrating gives

$$\int_{w_0}^{w^*} \frac{dw}{(1/\nu)w} = \int_{s_0}^{s^*} \frac{ds}{(-1/\sigma\nu)s}\tag{49}$$

Performing the integral and working through the algebra

$$\begin{aligned}\log w^* - \log w_0 &= -\sigma(\log s^* - \log s_0) \\ \Rightarrow \sigma \log(s_0) + \log w_0 &= \sigma \log s^* + \log w^* \\ \Rightarrow s_0^\sigma w_0 &= \text{constant}\end{aligned}\tag{50}$$

where (w^*, s^*) corresponds to the fixed point of the RG and is hence a constant.

The invariant scaling combination in this instance is thus

$$s^\sigma w\tag{51}$$

which agrees with the results in 3 and higher dimensions [30]. Similarly for h we have

$$\int_{w_0}^{w^*} \frac{dw}{(1/\nu)w} = \int_{h_0}^{h^*} \frac{dh}{(\beta\delta/\nu)h}\tag{52}$$

Performing the integral and working through the algebra

$$\begin{aligned}
\beta\delta(\log w^* - \log w_0) &= \log h^* - \log h_0 \\
\Rightarrow \log h_0 - \beta\delta \log w_0 &= \log h^* - \beta\delta \log w^* \\
\Rightarrow h_0 w_0^{-\beta\delta} &= \text{constant}
\end{aligned} \tag{53}$$

The invariant scaling combination is hence

$$h/w^{\beta\delta} \tag{54}$$

which again agrees with the literature [30].

B.2 TRANSCRITICAL FORM

The flow equations using the transcritical form for the disorder are as follows

$$\begin{aligned}
\frac{dw}{d\ell} &= w^2 + Bw^3 \\
\frac{ds}{d\ell} &= -d_f s - Csw \\
\frac{dh}{d\ell} &= \lambda_h h + Fhw
\end{aligned} \tag{55}$$

As before, we take the integral of $dw/d\ell$ over $ds/d\ell$ and obtain

$$\int_{s_0}^{s^*} (1/s) ds = \int_{w_0}^{w^*} \frac{-d_f - Cw}{w^2 + Bw^3} dw \tag{56}$$

Solving for s_0 we have

$$s_0 = \left(B + \frac{1}{w_0} \right)^{-Bd_f + C} \exp \left(\frac{d_f}{w_0} \right) f(w^*, s^*) \tag{57}$$

where $f(w^*, s^*)$ denotes a function of w^* and s^* and is therefore constant. The invariant scaling combination in this case is then

$$\frac{s}{\Sigma_{\text{th}}(w)} \tag{58}$$

where

$$\Sigma_{\text{th}}(w) = \left(B + \frac{1}{w}\right)^{-Bd_f + C} \exp\left(\frac{d_f}{w}\right) \tag{59}$$

Likewise for h we obtain an invariant scaling combination

$$\frac{h}{\eta_{\text{th}}(w)} \tag{60}$$

where

$$\eta_{\text{th}}(w) = \left(B + \frac{1}{w}\right)^{B\lambda_h - F} \exp\left(-\frac{\lambda_h}{w}\right) \tag{61}$$

B.3 ALTERNATIVE TRANSCRITICAL FORM

Applying our methods to the 2D equilibrium RFIM, we find that the fixed point is given by a pitchfork bifurcation corresponding to

$$\frac{dw}{d\ell} = w^3 - Dw^5 \tag{62}$$

In this instance, however, the behavior of the correlation length suggests an alternative choice for the normal form

$$\frac{dw}{d\ell} = \frac{w^3}{1 + Dw^2} \tag{63}$$

as discussed in [89]. This form, while retaining the pitchfork behavior, produces a well behaved correlation function that is also able to capture higher order cor-

rections to scaling which we expect to become important further from the critical point. We may apply the same procedure in the non-equilibrium case, although the function for the correlation length here appears well behaved. This yields an alternative form for the transcritical bifurcation given by

$$\begin{aligned}\frac{dw}{d\ell} &= \frac{w^2}{1-Bw} \\ \frac{ds}{d\ell} &= -d_f s - Csw \\ \frac{dh}{d\ell} &= \lambda_h h + Fhw\end{aligned}\tag{64}$$

As before, to determine $\Sigma(w)$, we take the integral of $dw/d\ell$ over $ds/d\ell$ and obtain

$$\int_{s_0}^{s^*} (1/s) ds = \int_{w_0}^{w^*} \frac{-d_f - Cw}{w^2/(1-Bw)} dw\tag{65}$$

Solving for s_0 we have

$$s_0 = w_0^{Bd_f - C} \exp\left(\frac{d_f}{w_0} + BCw_0\right) f(w^*, s^*)\tag{66}$$

where $f(w^*, s^*)$ denotes a function of w^* and s^* and is therefore constant. The invariant scaling combination in this case is then

$$\frac{s}{\Sigma_{\text{alt}}(w)}\tag{67}$$

where

$$\Sigma_{\text{alt}}(w) = w^{Bd_f - C} \exp\left(\frac{d_f}{w} + BCw\right)\tag{68}$$

Likewise for h we obtain an invariant scaling combination

$$\frac{h}{\eta_{\text{alt}}(w)}\tag{69}$$

where

$$\eta_{\text{alt}}(w) = w^{-B\lambda_h + F} \exp\left(-\frac{\lambda_h}{w} - BFw\right) \quad (70)$$

B.4 PITCHFORK FORM

The flow equations using a pitchfork form for the disorder are as follows

$$\begin{aligned} \frac{dw}{d\ell} &= w^3 + Bw^5 \\ \frac{ds}{d\ell} &= -d_f s - Csw \\ \frac{dh}{d\ell} &= \lambda_h h + Fhw \end{aligned} \quad (71)$$

As before, we take the integral of $dw/d\ell$ over $ds/d\ell$ and obtain

$$\int_{s_0}^{s^*} (1/s) ds = \int_{w_0}^{w^*} \frac{-d_f - Cw}{w^3 + Bw^5} dw \quad (72)$$

Solving for s_0 we have

$$\begin{aligned} s_0 &\sim w_0^{Bd_f} (1 + Bw_0^2)^{-\frac{Bd_f}{2}} \\ &\times \exp\left(\frac{d_f}{2w_0^2} + \frac{C}{w_0} + \sqrt{BC} \arctan(\sqrt{B}w_0)\right) \end{aligned} \quad (73)$$

The invariant scaling combination in this case is then

$$\frac{s}{\Sigma_{\text{pf}}(w)} \quad (74)$$

where

$$\begin{aligned} \Sigma_{\text{pf}}(w) &= w^{Bd_f} (1 + Bw^2)^{-\frac{Bd_f}{2}} \\ &\times \exp\left(\frac{d_f}{2w^2} + \frac{C}{w} + \sqrt{BC} \arctan(\sqrt{B}w)\right) \end{aligned} \quad (75)$$

Likewise for h we obtain an invariant scaling combination

$$\frac{h}{\eta_{\text{pf}}(w)} \quad (76)$$

where

$$\begin{aligned} \eta_{\text{pf}}(w) = & w^{-B\lambda_h} (1 + Bw^2)^{\frac{B\lambda_h}{2}} \\ & \times \exp\left(-\frac{\lambda_h}{2w^2} - \frac{F}{w} - \sqrt{BF} \arctan(\sqrt{B}w)\right) \end{aligned} \quad (77)$$

APPENDIX C UNIVERSAL SCALING FUNCTION FORMS

In order to perform our fits, we choose functional forms for the universal scaling functions. For the area weighted avalanche size distribution, we choose

$$\mathcal{A}(v_s) = \frac{1}{\mathcal{A}_N} v_s^a \exp(v_s^b) \quad (78)$$

where the leading power law v_s^x has been absorbed into v_s^a here and \mathcal{A}_N is the normalization factor $\mathcal{A}_N = [\Gamma(\frac{a}{b})\gamma(\frac{a}{b}, \Sigma(w)^{-2b})]/b$ where γ denotes the regularized upper incomplete gamma function.

For $d\mathcal{M}/dh$ we choose

$$\frac{d\mathcal{M}}{dh}(v_h) = \frac{1}{\frac{d\mathcal{M}}{dh}_N} \exp\left[\left(\frac{-v_h^2}{a + bv_h + cv_h^2}\right)^{d/2}\right] \quad (79)$$

where $\frac{d\mathcal{M}}{dh}_N$ is a normalization factor computed as a sum of $\frac{d\mathcal{M}}{dh}$ over the data range.

APPENDIX D FORM COMPARISON

In the lower critical dimension, we expect the fixed point to be governed by a transcritical bifurcation. Assuming compact avalanches, this yields directly

$$\Sigma_{\text{th}}(w) = \Sigma_s \left(B + \frac{1}{w} \right)^{-Bd_f} \exp \left(\frac{d_f}{w} \right) \quad (80)$$

where Σ_s is an unknown scale factor. We may compare this functional form for Σ with that derived assuming power law scaling and one assuming a pitchfork bifurcation as appears in the equilibrium model. For the power law case we have the invariant scaling combination

$$s^\sigma w \quad (81)$$

which is equivalent to

$$s/w^{-1/\sigma} \quad (82)$$

such that $\Sigma_{\text{pl}}(w)$ for the power law case would be given by:

$$\Sigma_{\text{pl}}(w) = \Sigma'_s w^{-1/\sigma} \quad (83)$$

where Σ'_s is an unknown scale factor determined by fitting to a power law form. For the pitchfork case we have from Section B.4

$$\begin{aligned} \Sigma_{\text{pf}}(w) = & \Sigma''_s w^{Bd_f} (1 + Bw^2)^{-\frac{Bd_f}{2}} \\ & \times \exp \left(\frac{d_f}{2w^2} + \frac{C}{w} + \sqrt{BC} \arctan(\sqrt{B}w) \right) \end{aligned} \quad (84)$$

where Σ''_s is the unknown scale factor. We have that $w = (r - r_c)/r_s$ for each of the functional forms considered. The comparison of the fits are shown in Figure 3.

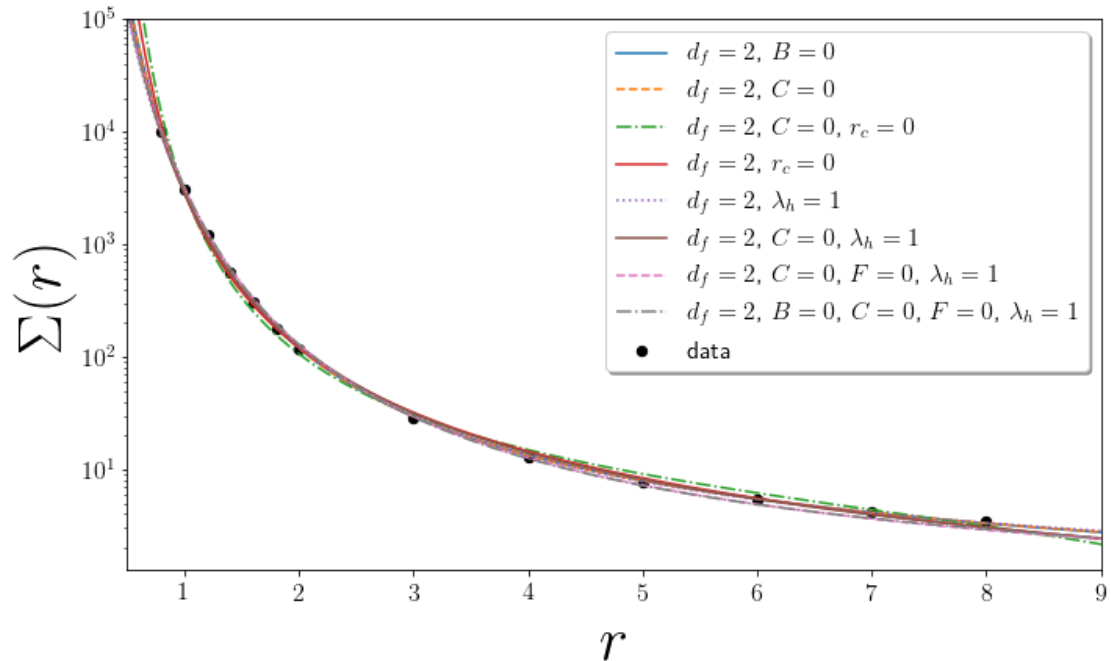


Figure 30: **Fit comparisons** $\Sigma_{\text{th}}(w)$, **transcritical form.**

APPENDIX E PARAMETER VALUES

The parameter values corresponding to reasonable fits are highly variable. For example, we may restrict $\lambda_h = 1$ corresponding to the Harris criteria for this model [30] and still obtain an acceptable fit. A wide range of fits with various restrictions are shown in Figures 30, 31, 32, and 33. Corresponding best fit parameter values are shown in Tables 11 and 12. As anticipated, the alternative form for the transcritical bifurcation is able to better capture the behavior far from the critical point.

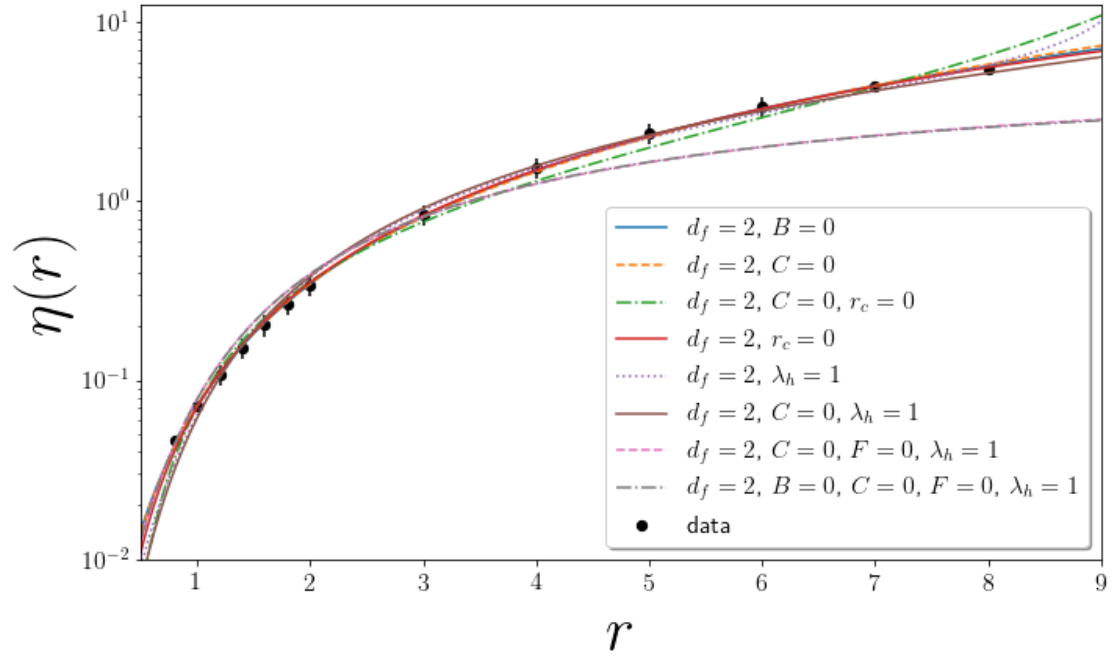


Figure 31: Fit comparisons $\eta_{\text{th}}(w)$, transcritical form.

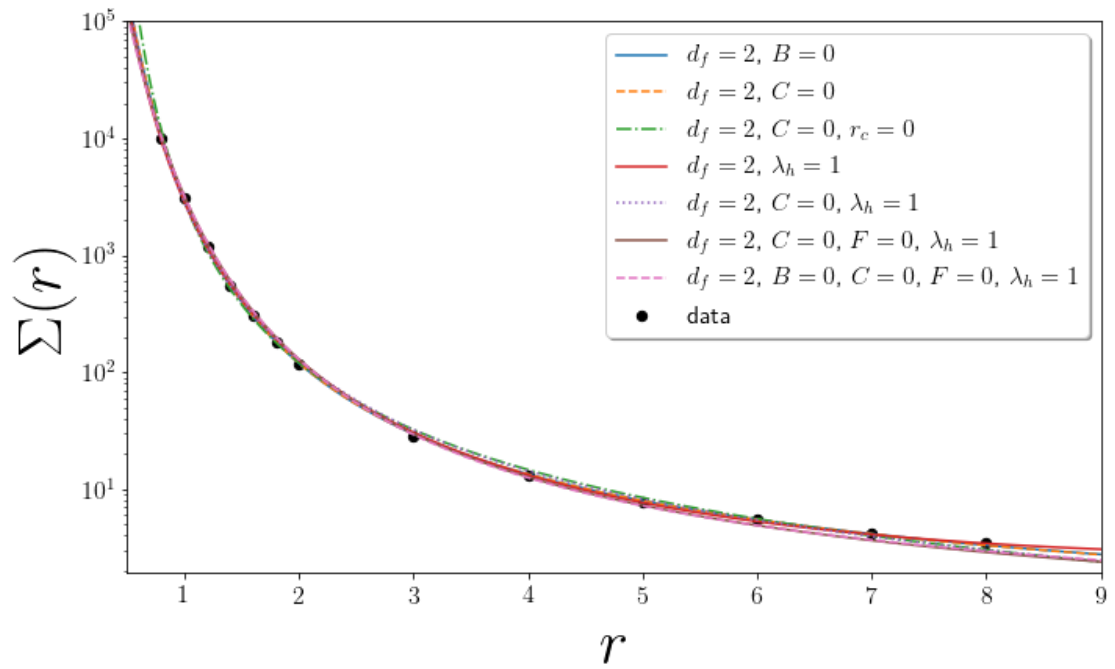


Figure 32: Fit comparisons $\Sigma_{\text{alt}}(w)$, alternative transcritical form.

r_s	5.11 ± 0.54	5.49 ± 0.18	2.91 ± 0.27	2.04 ± 0.34	6.89 ± 0.63	4.93 ± 0.15	6.78 ± 0.20	7.40 ± 0.12
r_c	-0.42 ± 0.11	-0.46 ± 0.06	0	0	-0.65 ± 0.10	4.93 ± 0.15	-0.64 ± 0.06	-0.70 ± 0.03
Σ_s	1.24 ± 0.66	1.11 ± 0.05	5.27 ± 1.96	14.40 ± 8.84	0.68 ± 0.22	1.64 ± 0.04	0.64 ± 0.04	0.54 ± 0.005
η_s	3.16 ± 0.11	3.11 ± 0.44	1.02 ± 0.52	0.50 ± 0.41	6.26 ± 0.25	4.33 ± 0.42	5.55 ± 0.52	6.08 ± 0.82
d_f	2	2	2	2	2	2	2	2
λ_h	0.44 ± 0.04	0.52 ± 0.07	0.61 ± 0.05	0.24 ± 0.08	1	1	1	1
B	0	-0.15 ± 0.01	-0.27 ± 0.03	0.039 ± 0.007	-0.69 ± 0.21	-0.25 ± 0.03	-0.09 ± 0.01	0
C	0.46 ± 0.10	0	0	1.76 ± 0.28	-1.38 ± 0.52	0	0	0
F	1.72 ± 0.08	1.33 ± 0.12	0.73 ± 0.02	2.02 ± 0.13	-0.37 ± 0.35	0.45 ± 0.06	0	0

Table 11: Table of the best fit values corresponding to Figures 30 and 31. Values in bold correspond to values fixed in the fit.

r_s	5.10 ± 0.54	5.05 ± 0.36	2.12 ± 0.33	1.81 ± 0.08	3.62 ± 0.18	6.57 ± 0.39	7.40 ± 0.12
r_c	-0.42 ± 0.11	-0.42 ± 0.09	0	-0.15 ± 0.07	-0.29 ± 0.09	-0.62 ± 0.10	-0.70 ± 0.03
Σ_s	1.24 ± 0.66	-1.27 ± 0.38	13.16 ± 5.72	21.44 ± 1.31	3.19 ± 0.45	-0.67 ± 0.32	0.54 ± 0.005
η_s	3.16 ± 0.11	2.49 ± 0.20	0.56 ± 0.40	0.69 ± 0.04	2.48 ± 0.09	5.42 ± 0.17	6.08 ± 0.82
d_f	2	2	2	2	2	2	2
λ_h	0.44 ± 0.04	0.54 ± 0.08	0.70 ± 0.05	1	1	1	1
B	0	-0.24 ± 0.01	-0.76 ± 0.14	-1.70 ± 0.16	-0.56 ± 0.05	-0.13 ± 0.01	0
C	0.46 ± 0.10	0	0	-0.31 ± 0.04	0	0	0
F	1.72 ± 0.08	1.22 ± 0.16	0.45 ± 0.04	-0.031 ± 0.038	0.35 ± 0.03	0	0

Table 12: Table of the best fit values corresponding to Figures 32 and 33. Values in bold correspond to values fixed in the fit.

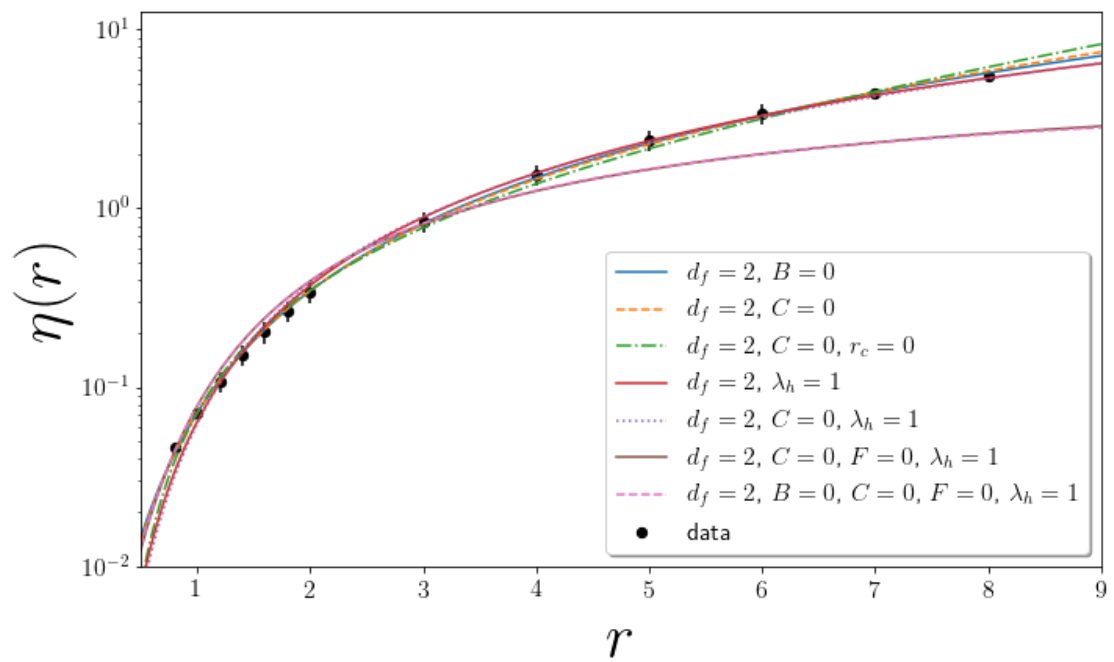


Figure 33: **Fit comparisons $\eta_{\text{alt}}(w)$, alternative transcritical form.**

APPENDIX F KAGGLE DATASET DETAILS

The Higgs boson has many channels through which it can decay. In the Kaggle challenge, the task was to identify only one type of signal event; the Higgs to tau tau recently reported by ATLAS [35]. To produce enough signal events on which to train a classifier, the data provided is created through simulation. Hence, in order to calculate what the approximate mean significance (AMS) would be in a real event, the dataset must include a weight which corrects for the difference between the natural (prior) probability of the event occurring and that of its occurrence in the simulation. Additionally, for the challenge, the only event types included in the dataset contain either one electron or one muon along with one hadronic tau. Events with b-quark jets were also discarded.

In each proton collision, part of the kinetic energy is converted into new particles, most of which are highly unstable and decay into a cascade of lighter particles. The ATLAS detector itself measures 3 quantities of the particles or pseudoparticles which reach it: type of particle, energy and direction. Using these data the original heavier particles identities are inferred. There are two main features of the Higgs to tau tau decay that make it difficult to infer. One is that neutrinos produced cannot be measured which makes the mass of the Higgs difficult to infer on an event by event basis. The other reason the signal for the Higgs is difficult to disentangle from background is that other processes such as the decay of the Z boson produce similar signals. In the contest dataset Z boson decay, decays of two top quarks and W boson decay were included as each of these decays can result in a lepton and a hadronic tau.

The dataset provided for the challenge is composed of four parts. The weights, which were discussed above, event ids, primitives (PRI) and derived quantities (DER). The primitives are quantities measured by the detector. There are five

'particles' for which measurements are provided. These include the hadronic tau, the lepton, the missing transverse energy and the leading and subleading jets. For each of these particles the transverse momentum, the azimuthal angle and the pseudorapidity are provided. Additionally, the number of jets is provided as well as the scalar sum of the transverse momentum of all the jets. The transverse energy and azimuthal angle correspond to the plane perpendicular to the beamline. The pseudorapidity is defined as

$$\eta = -\ln(\tan(\theta/2)) \quad (85)$$

where θ is the polar angle. Particles in the range $\eta \in [-2.5, 2.5]$ can be identified. Those for $\eta \in \pm[2.5, 5]$ cannot be identified, however, their momentum can be measured. For $\eta > 5$ the particle escapes down the beam pipe. The DER quantities are those which can be derived from the raw output using physics knowledge. These include simple things such as pseudorapidity separation and invariant masses to more complicated calculations such as the approximate mass of the original particle. For those quantities which are easily calculated, formulas were provided by the challenge organizers.

Several people, including the top performers, made use of physics knowledge for their submissions. These included simply calculating quantities with the given formulas if they were not provided, making use of rotational invariance to remove features and more complicated calculations such as those used by the CAKE team and Motl's group. The top two performers noted an increase in score by engineering of features.

APPENDIX G JEFFREY'S PRIOR SAMPLING: ADDITIONAL RESULTS

This appendix contains the results for the single-digit and SdA networks which were left out of the main text: Figures 34- 39.

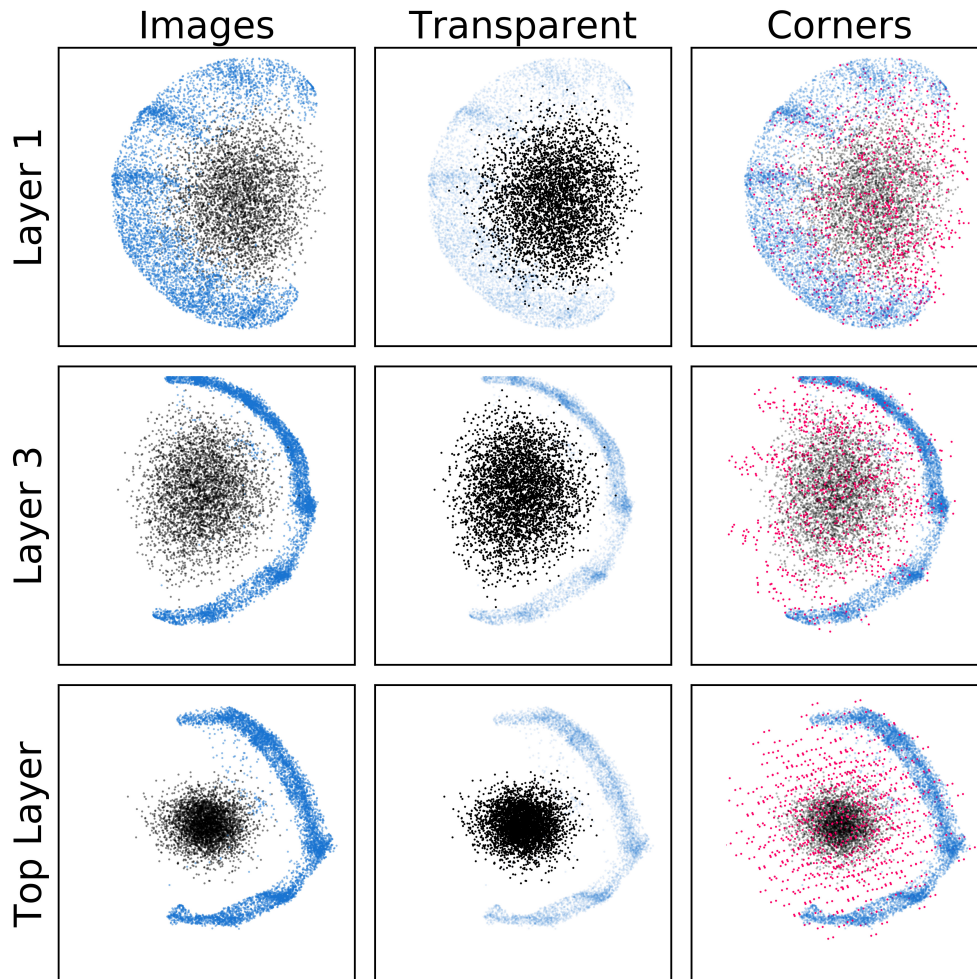


Figure 34: **PCA projection of the Jeffrey's Prior sampling with MNIST '1' digit data for the single-digit network along the two largest principal component vectors.** In the middle column 'Transparent' the transparency of the MNIST points have been enhanced to show the position of the sampling. On the right under 'Corners' the '1's and sampling have again been plotted in blue and black respectively with the corners added in pink. Corners correspond to activations in the top hidden layer ($\vec{\theta}$) for which $\theta_i \in \{0, 1\}$.

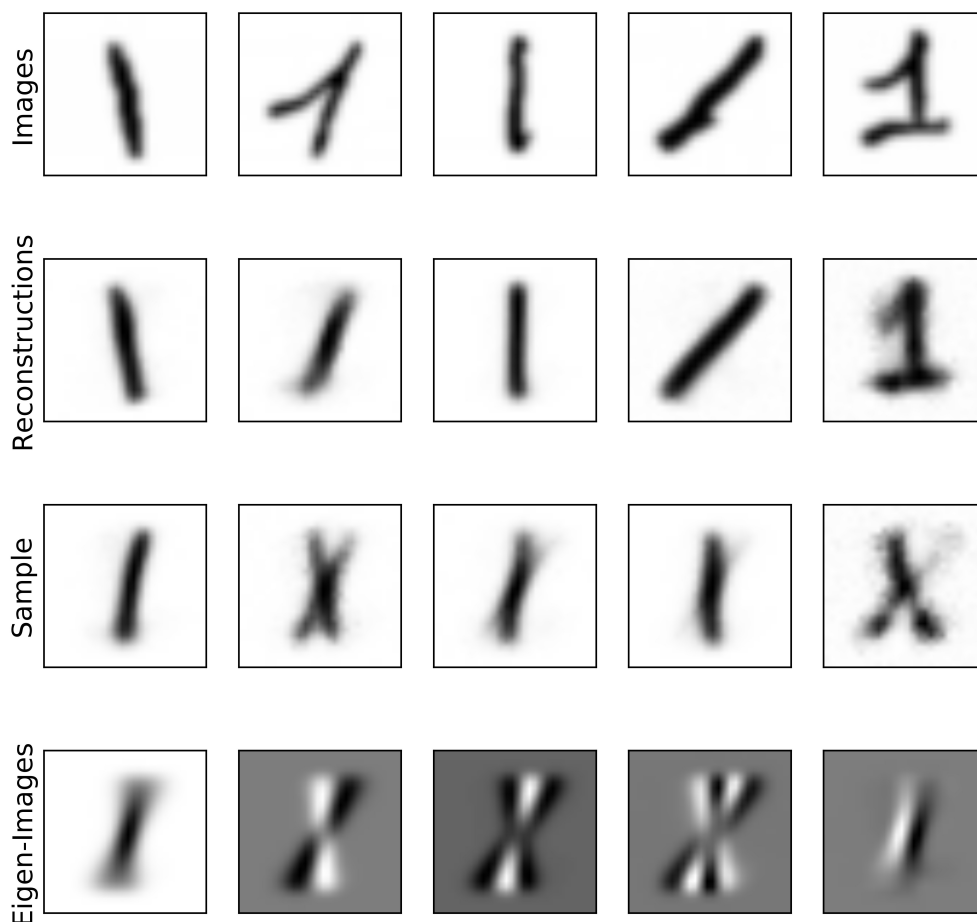


Figure 35: **Examples of MNIST '1' images, their reconstructions, and images sampled using Jeffrey's Prior for the single-digit network.** For the sampled '1's, each snapshot corresponds to the same walker. The final row corresponds to the top five 'eigen-digits' of the '1's dataset.

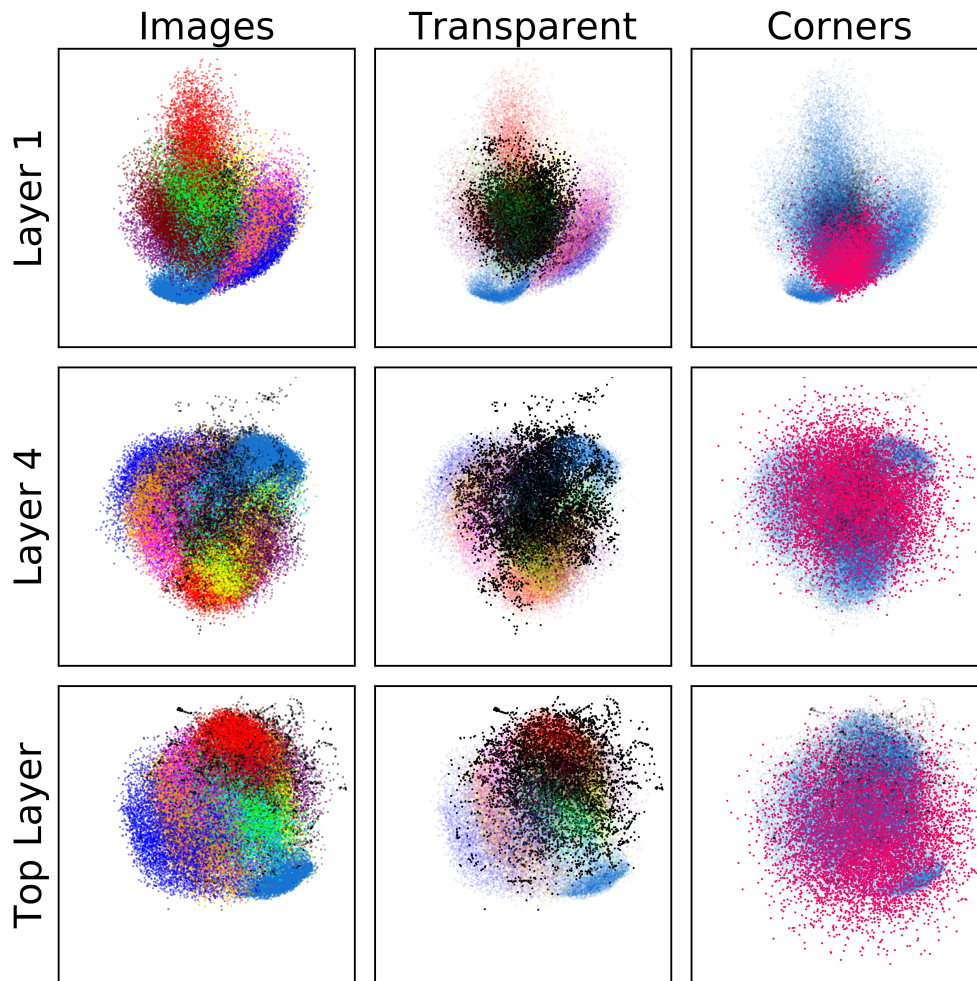


Figure 36: **PCA projection of the Jeffrey's Prior sampling with MNIST digit data for the SdA along the two largest principal component vectors.** In the middle column 'Transparent' the transparency of the MNIST points have been enhanced to show the position of the sampling. On the right under 'Corners' the digits and sampling have again been plotted in blue and black respectively with the corners added in pink. Corners correspond to activations in the top hidden layer ($\vec{\theta}$) for which $\theta_i \in \{-\infty, \infty\} \sim \{-10^6, 10^6\}$. The axes are shared along each row. In order to deal with the large values of the corners and sampling, the top layer is shown with a sigmoid applied. Although the sampling spans the parameter space (Top Layer), it is subsequently mapped to the interior of the manifold in reconstruction space ('Layer 1').

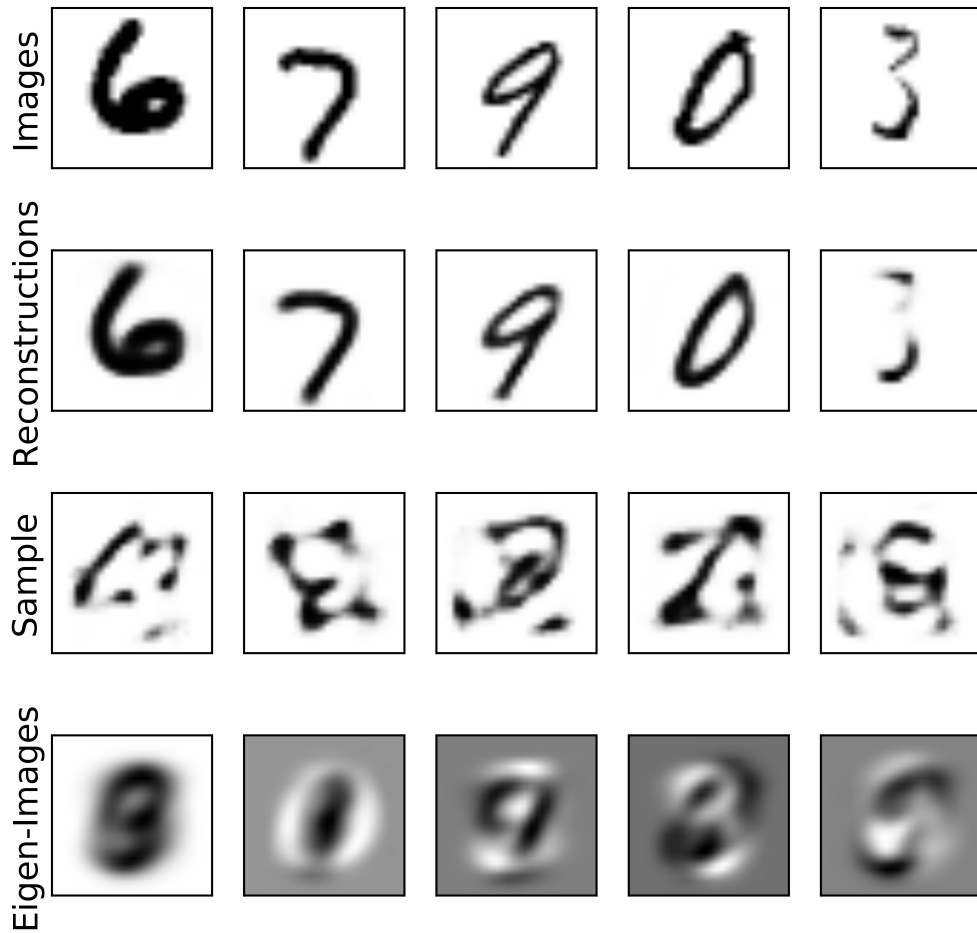


Figure 37: **Examples of MNIST images, their reconstructions, and images sampled using Jeffrey’s Prior for the SdA.** For the sampled ‘digits’, each snapshot corresponds to the same walker. The final row corresponds to the top five ‘eigen-digits’ of the dataset.

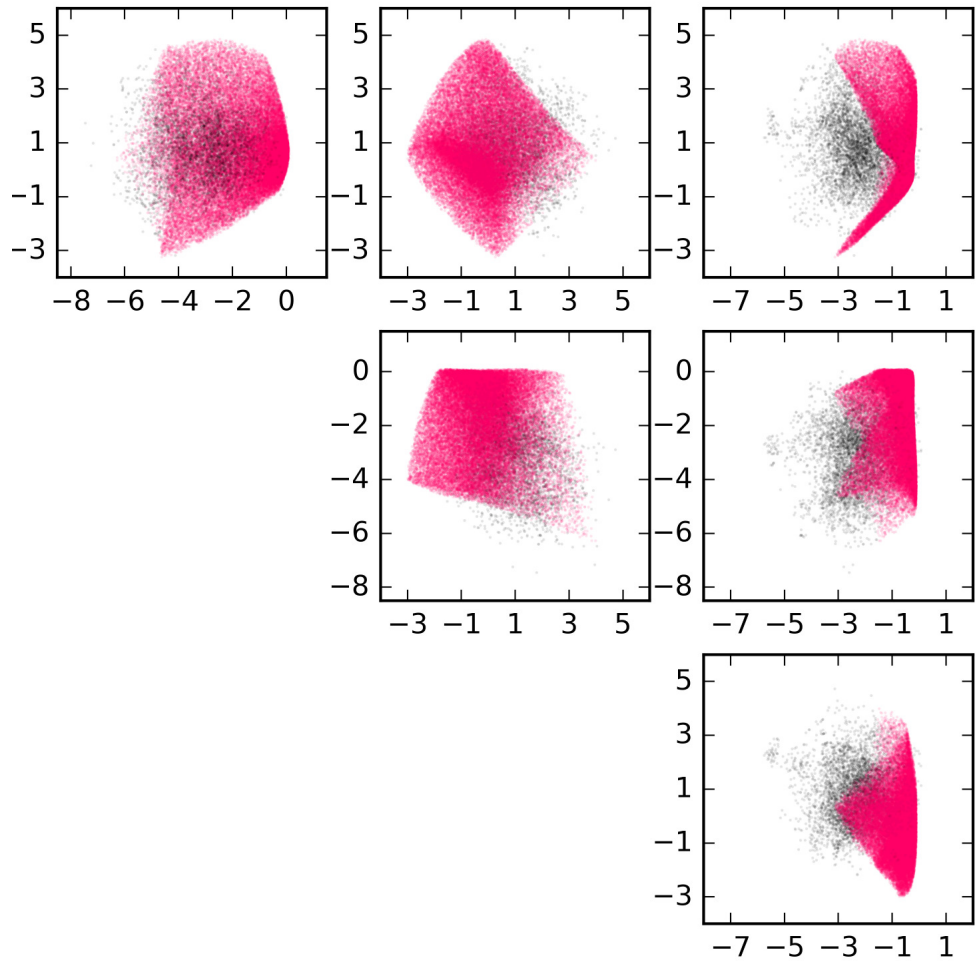


Figure 38: PCA projection of the Jeffrey's Prior sampling for the 3D manifold (Pink) and 30D manifold (Black) in reconstruction space for the DBN.

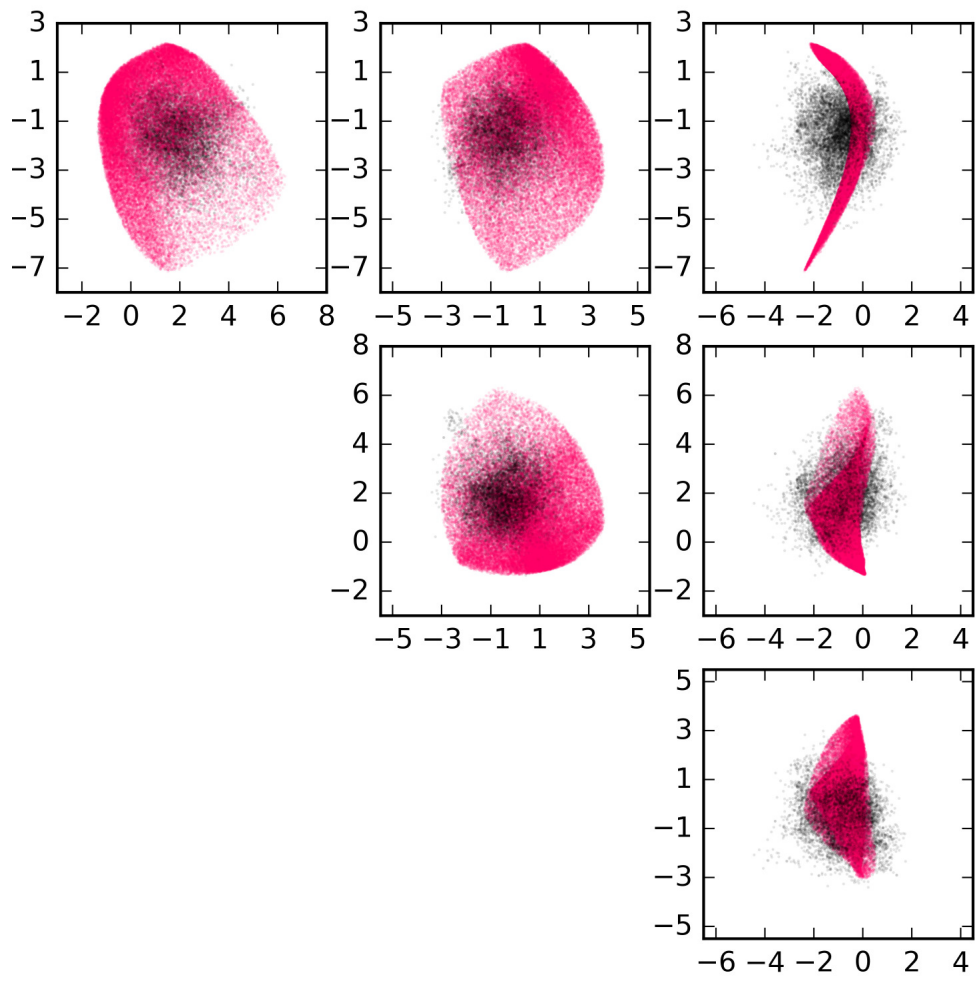


Figure 39: PCA projection of the Jeffrey's Prior sampling for the 3D manifold (Pink) and 30D manifold (Black) in reconstruction space for the SdA.

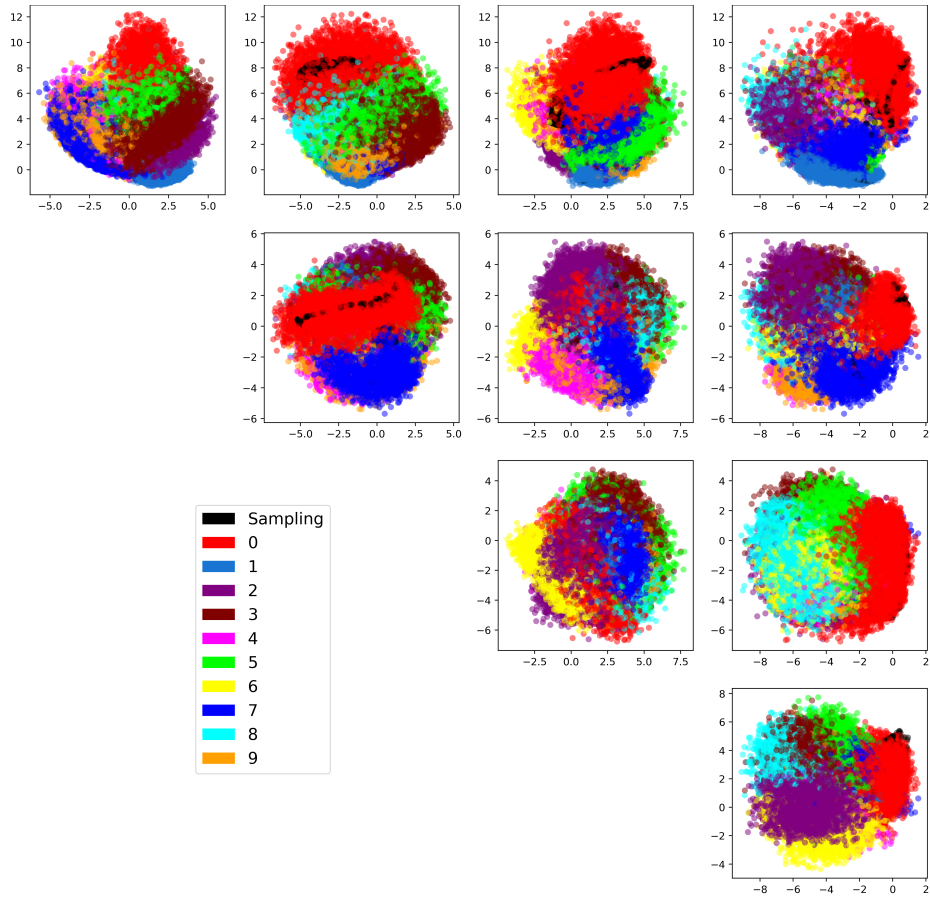


Figure 40: **PCA projection of the sampling of a variational autoencoder and the MNIST digits.**

APPENDIX H SAMPLING OF A VARIATIONAL AUTOENCODER: PRELIMINARY RESULTS

Variational autoencoders provide a natural method for sampling the space of digits learned by the neural network. Figures 40 and 41 show PCA projections of the sampling generated from a variational autoencoder trained on MNIST. The sample appears to lie within the interior of the space while MNIST digits populate the boundaries.

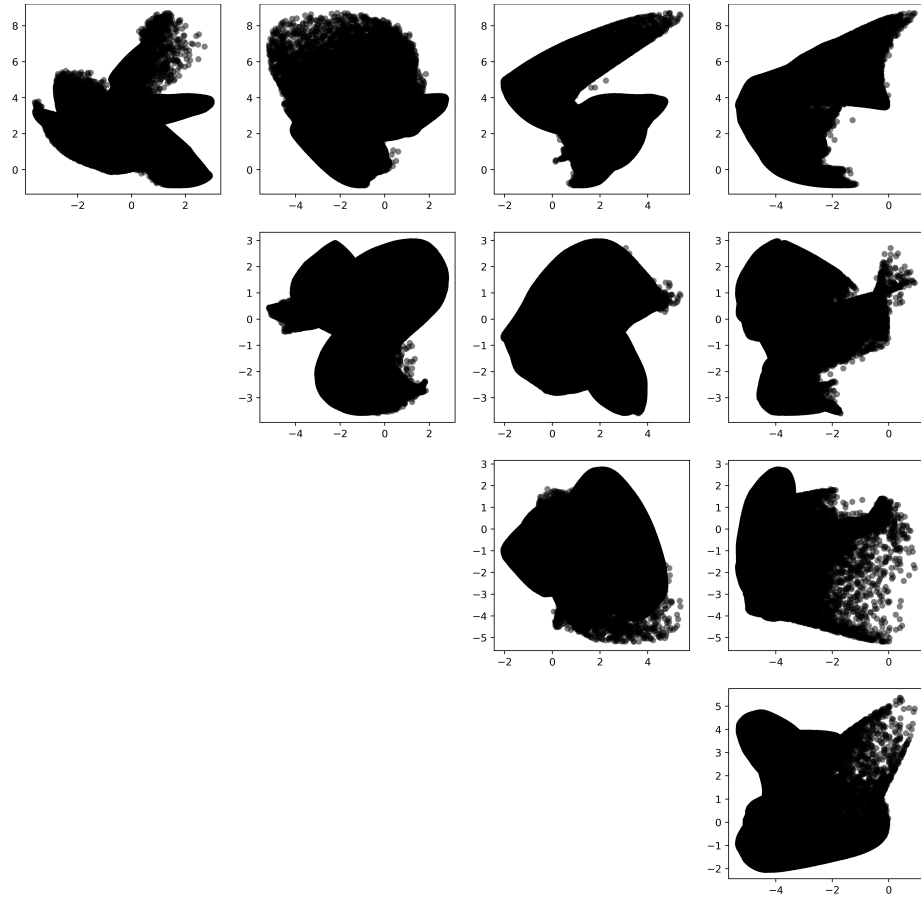


Figure 41: PCA projection of the sampling of a variational autoencoder.