# A Statistical Perspective on Deep Learning

Chad M. Schafer
Department of Statistics & Data Science
Carnegie Mellon University
January 2019

# Main Point

Place deep learning into a statistical context, to help understand when this approach has the most potential to contribute

# Representations

What are the fundamental data science challenges in astronomy?

# Representations

What are the fundamental data science challenges in astronomy?

Recurring theme: Shift from variance-dominated to bias-dominated challenges

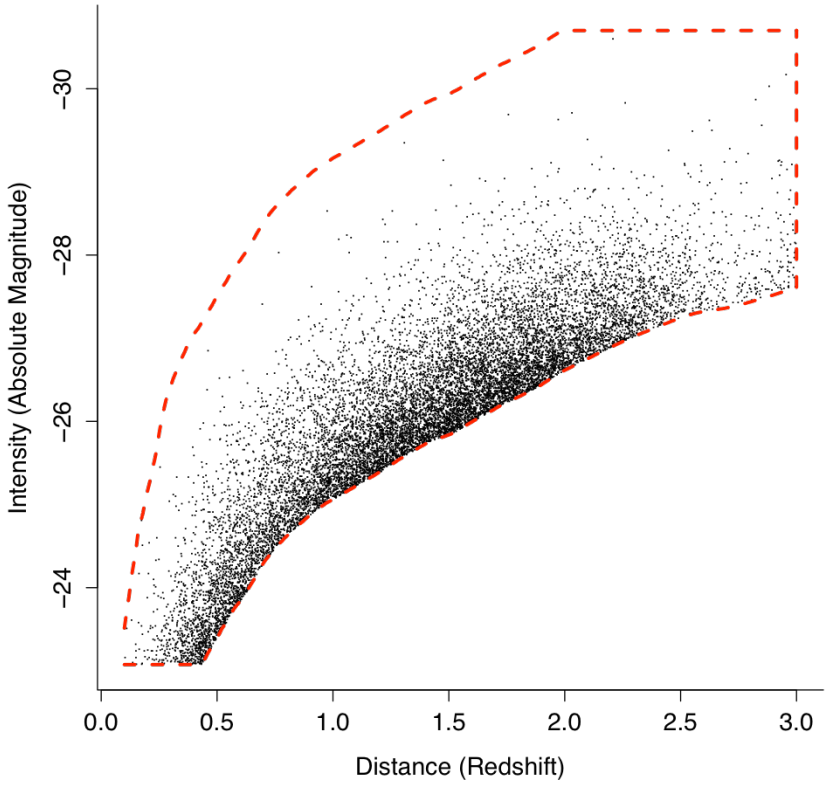
Data sufficient in quantity and quality that non-physical model assumptions limit progress

| Year | Count | Source |
|---|---|---|
| 1984 | 35 | Marshall, et al. |
| 1988 | 420 | Boyle, et al. |
| 1995 | 1,200 | Pei |
| 2002 | 3,814 | Schneider, et al. (SDSS Early Data Release) |
| 2003 | 16,713 | Schneider, et al. (SDSS DR1) |
| 2005 | 46,420 | Schneider, et al. (SDSS DR3) |
| 2007 | 77,429 | Schneider, et al. (SDSS DR5) |
| 2010 | 105,783 | Schneider, et al. (SDSS DR7) |
| 2018 | 526,356 | Paris, et al. (SDSS DR14) |
| 2019 | 2.0 million | Flesch (MILLIQUAS, Version 5.7) |
| 2030 | > 10 million | LSST |

## Progression of Quasar Sample Sizes

# Representations

How does one utilize this rich data source as part of a formal statistical analysis?




The luminosity function is a model for the distribution of magnitude (and distance) of quasars.

The estimated luminosity function can be thought of as a summary or representation of the full data

# Representations

Schechter (1976): Simple form for luminosity function:

$$n_\star 10^{0.4(\alpha+1)(M^\star-M)} \exp\left(-10^{0.4(M^\star-M)}\right)$$

"This formula was initially motivated by a simple model of galaxy formation (Press and Schechter, 1974), but it has proved to have a wider range of application than originally envisaged… With larger, deeper surveys, the limitations of the simple Schechter function start to become apparent."

Binney and Merrifield (1998), pages 163-164

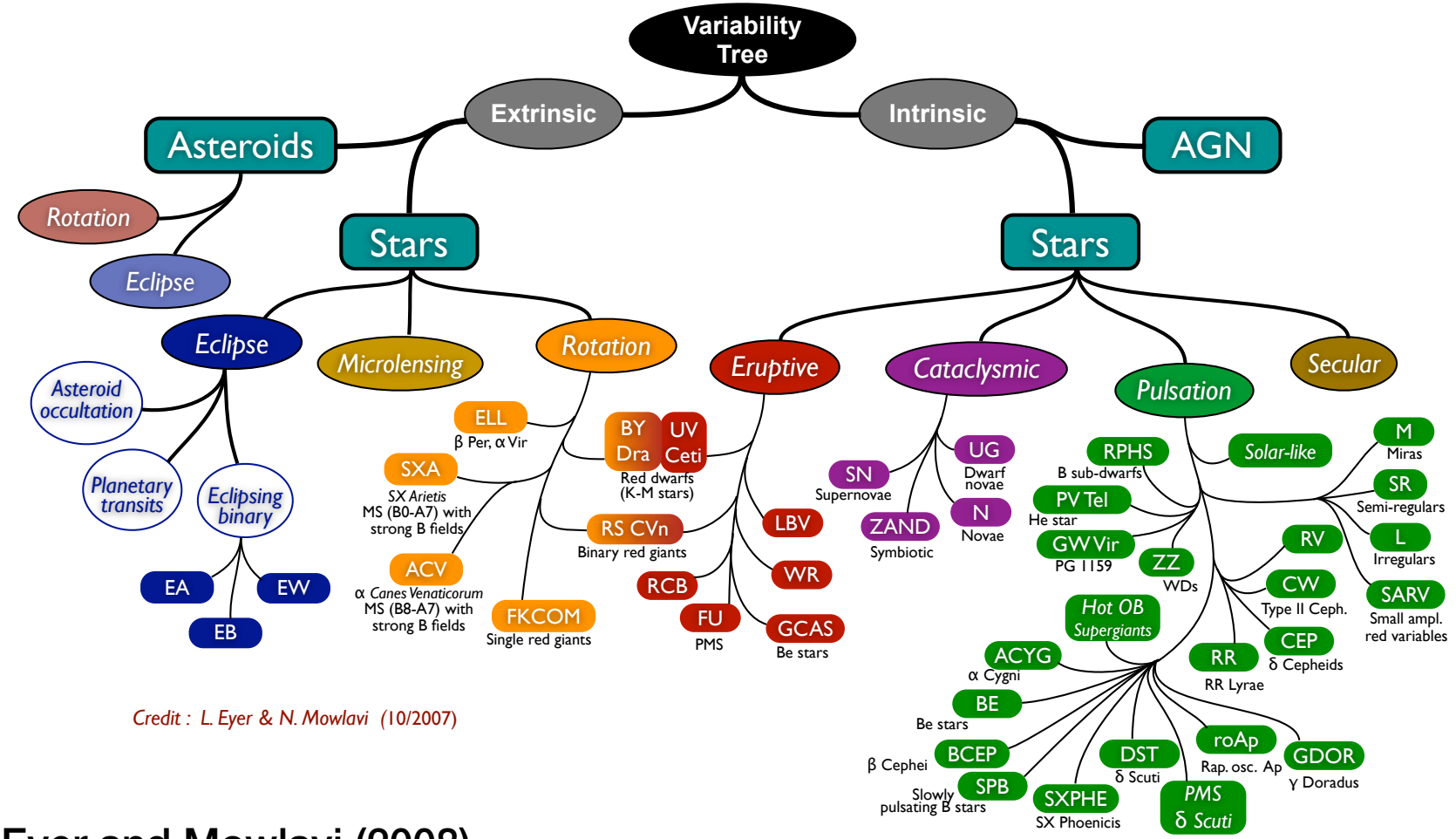
# Representations

Cataclysmic Variables (CV) – binary system in Milky Way with matter transfer from secondary (normal) star to primary white dwarf

Blazars – Quasars with “jet” of energy pointed at Earth

Both produce light curves with irregular variability, lacking periodic structure

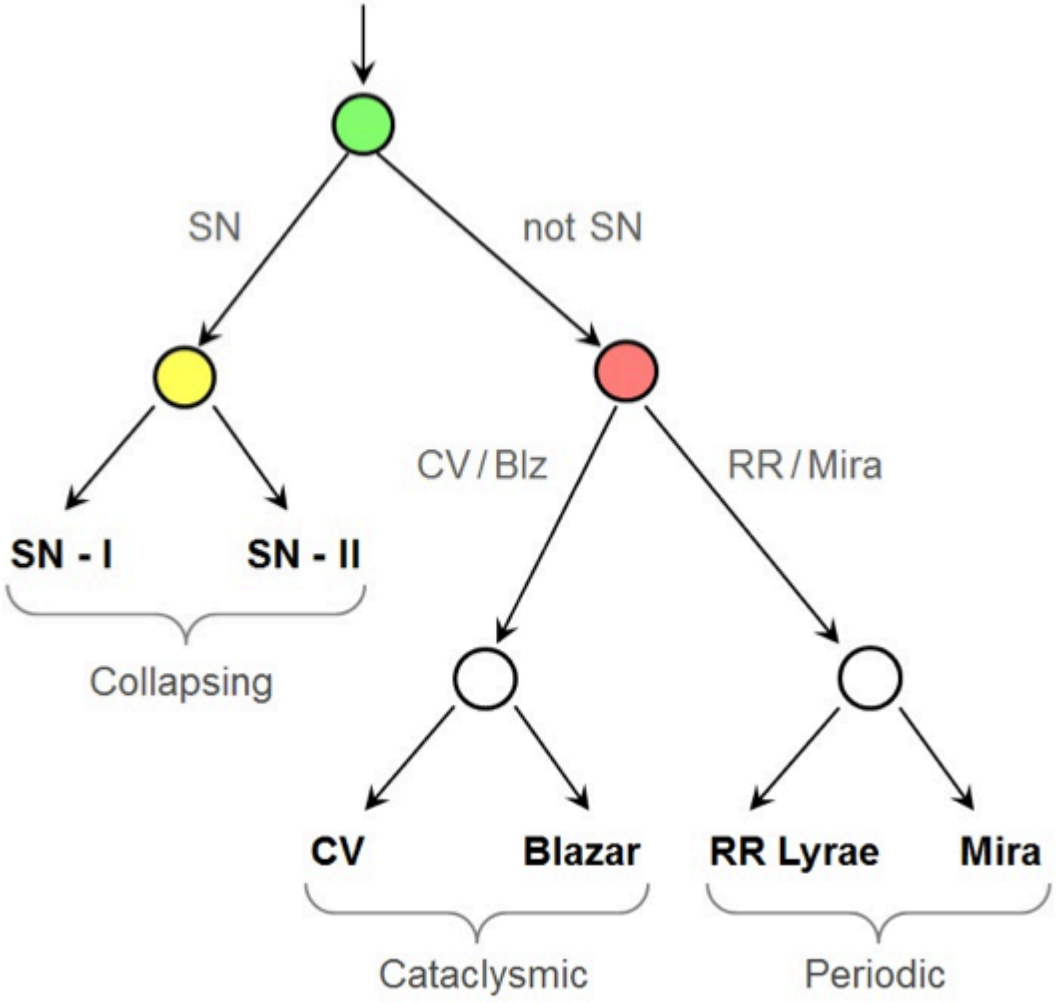
# Representations



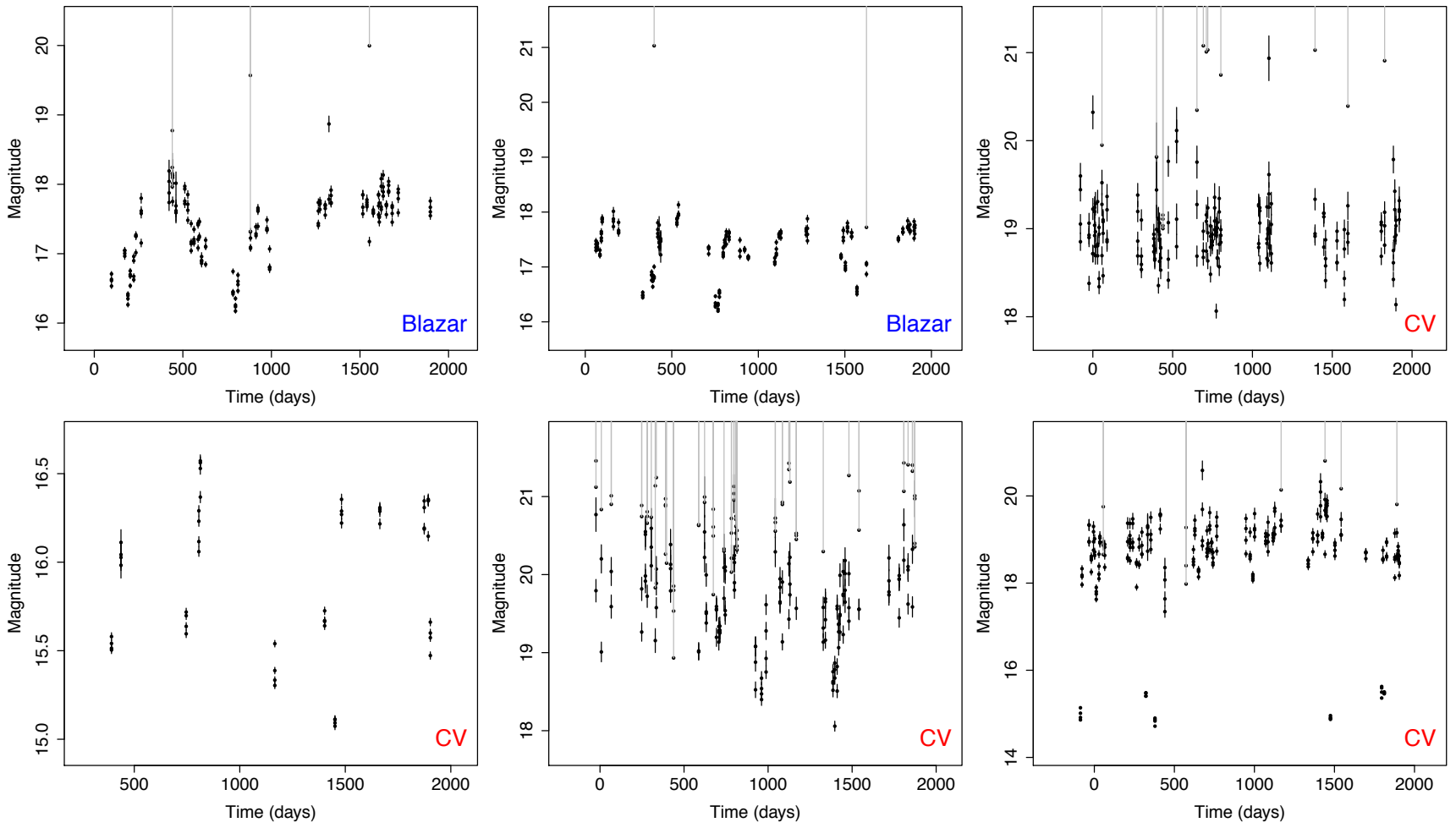

*Credit : L. Eyer & N. Mowlavi (10/2007)*

**Eyer and Mowlavi (2008)**

# Representations

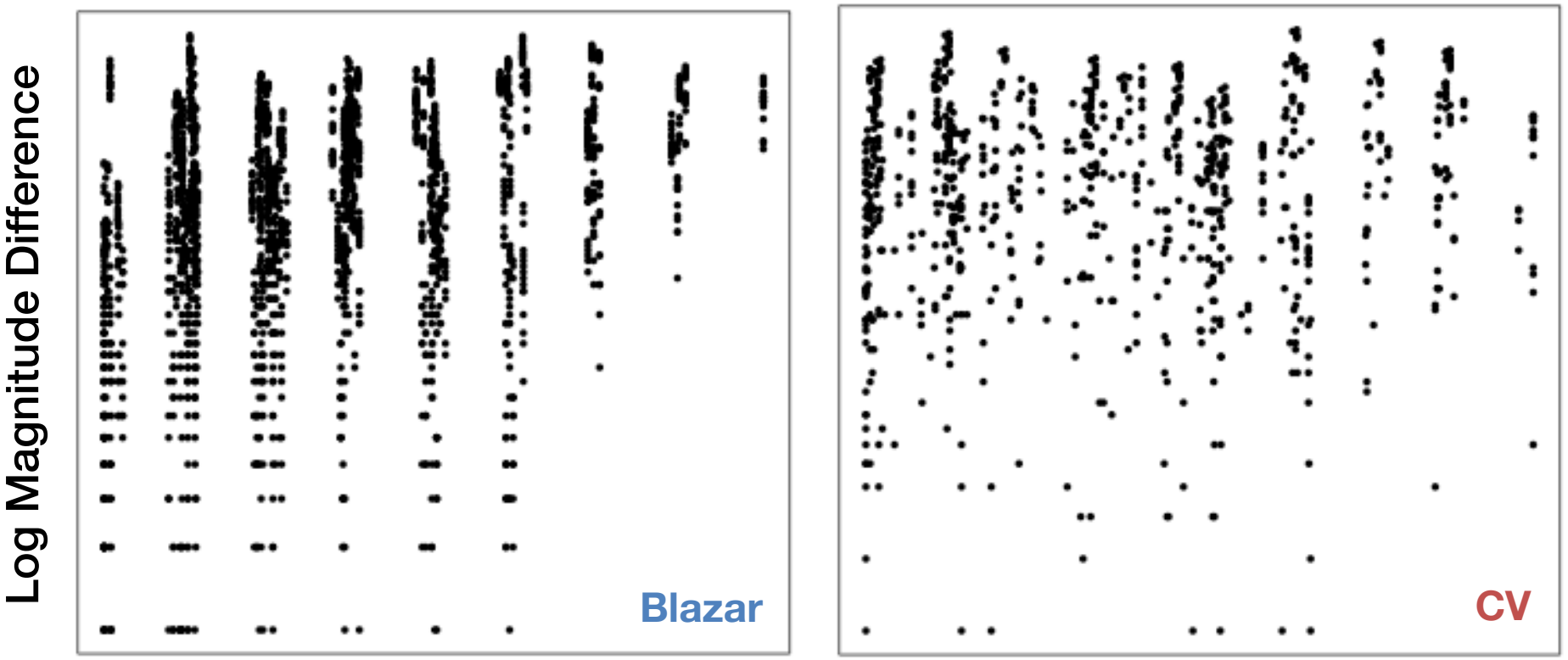


# Representations




Light Curves from Catalina Real-Time Transient Survey (Drake 2009)

# Representations




Comparison of Structure Functions

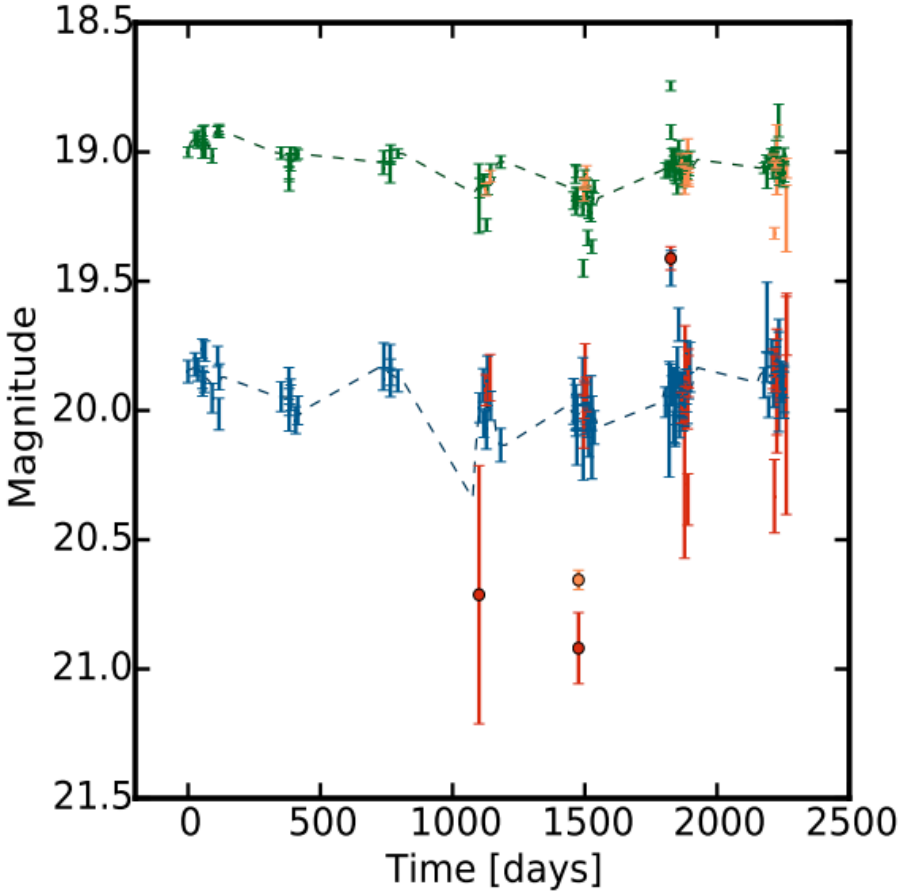
# Representations

Typical to fit model to structure function

- Power Law Form (Schmidt et al. 2010)
- Damped Random Walk (Kelly et al. 2009)

Effort to find a low-dimensional representation, avoiding the curse of dimensionality

# Representations



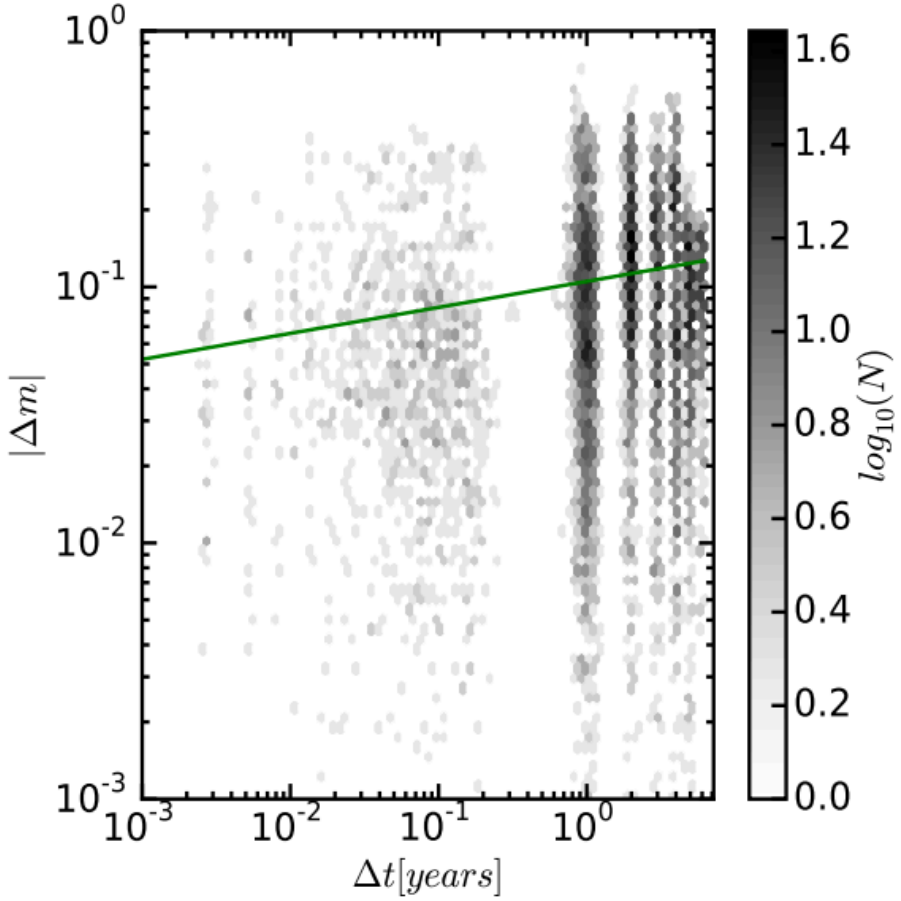




Figure 2 in Peters et al. (2015) Quasar light curve and SF

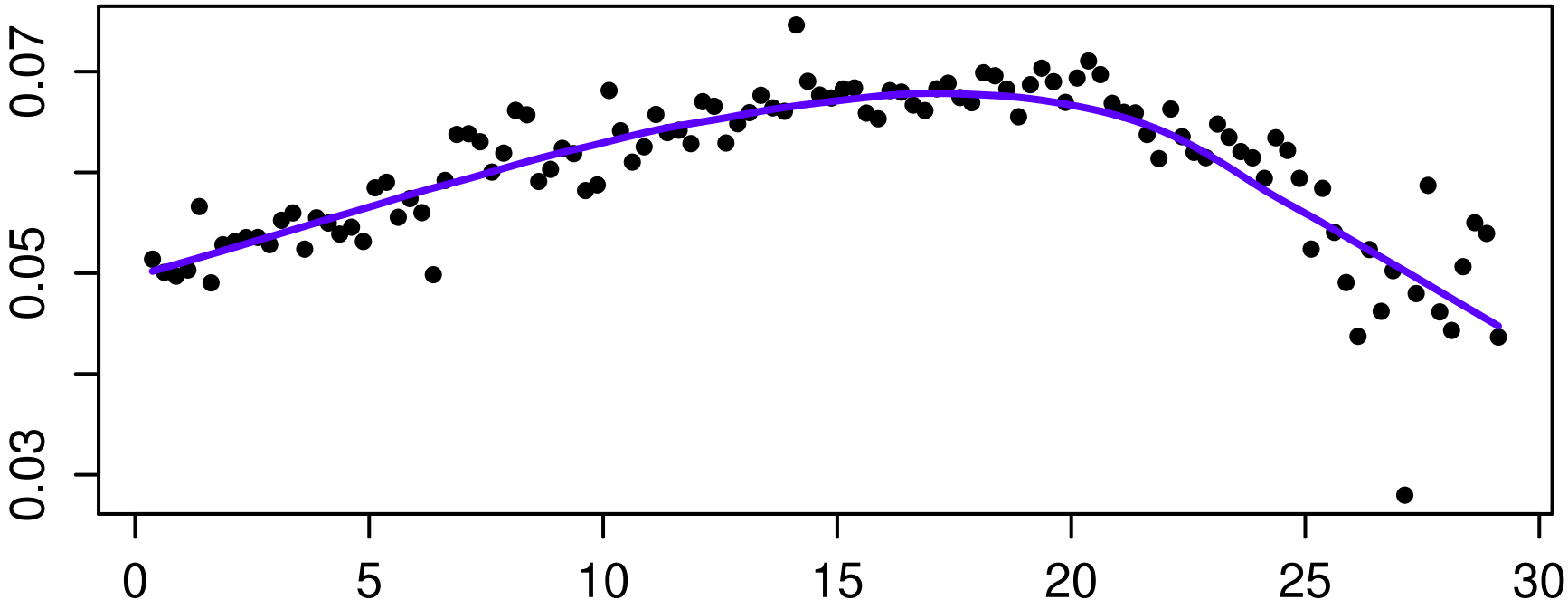
# Representations

A fundamental challenge of "Big Data": representing the raw data

What summary statistic retains the important information in the data?

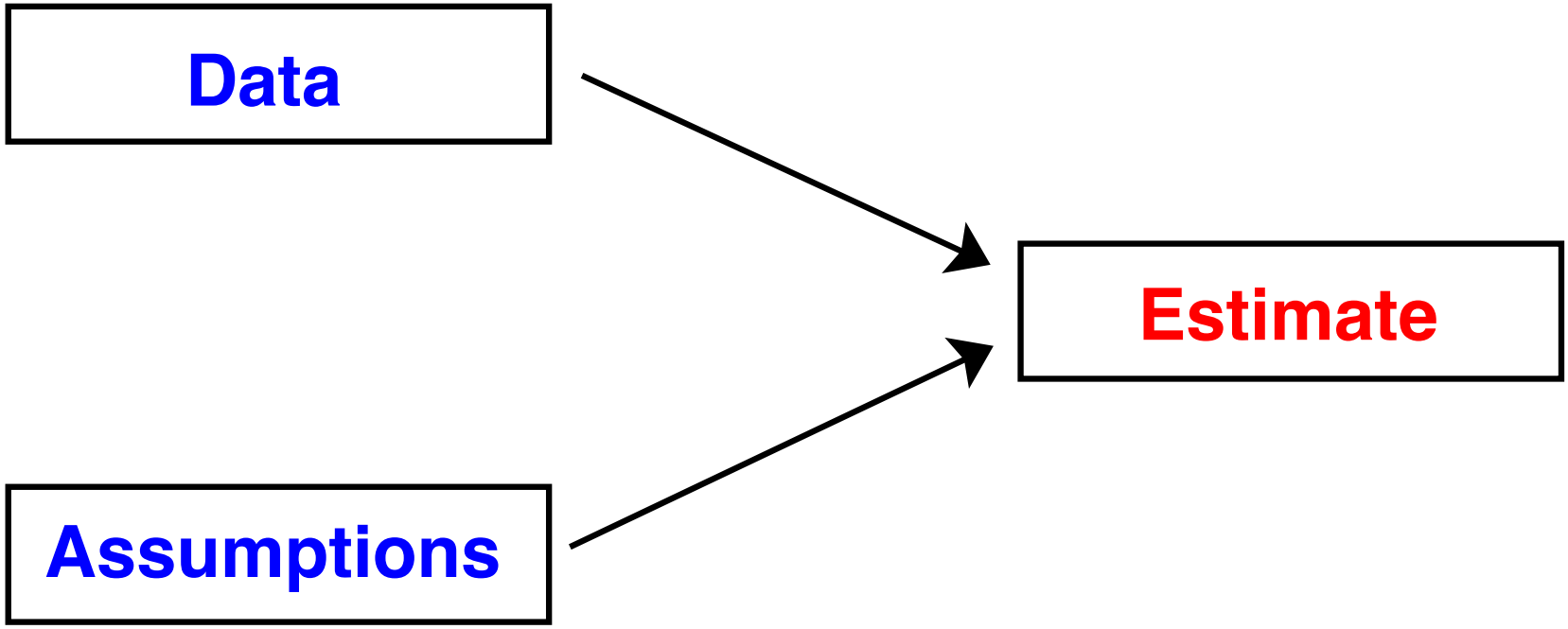
"Separating signal from noise" with minimal loss of information

# Nonparametrics



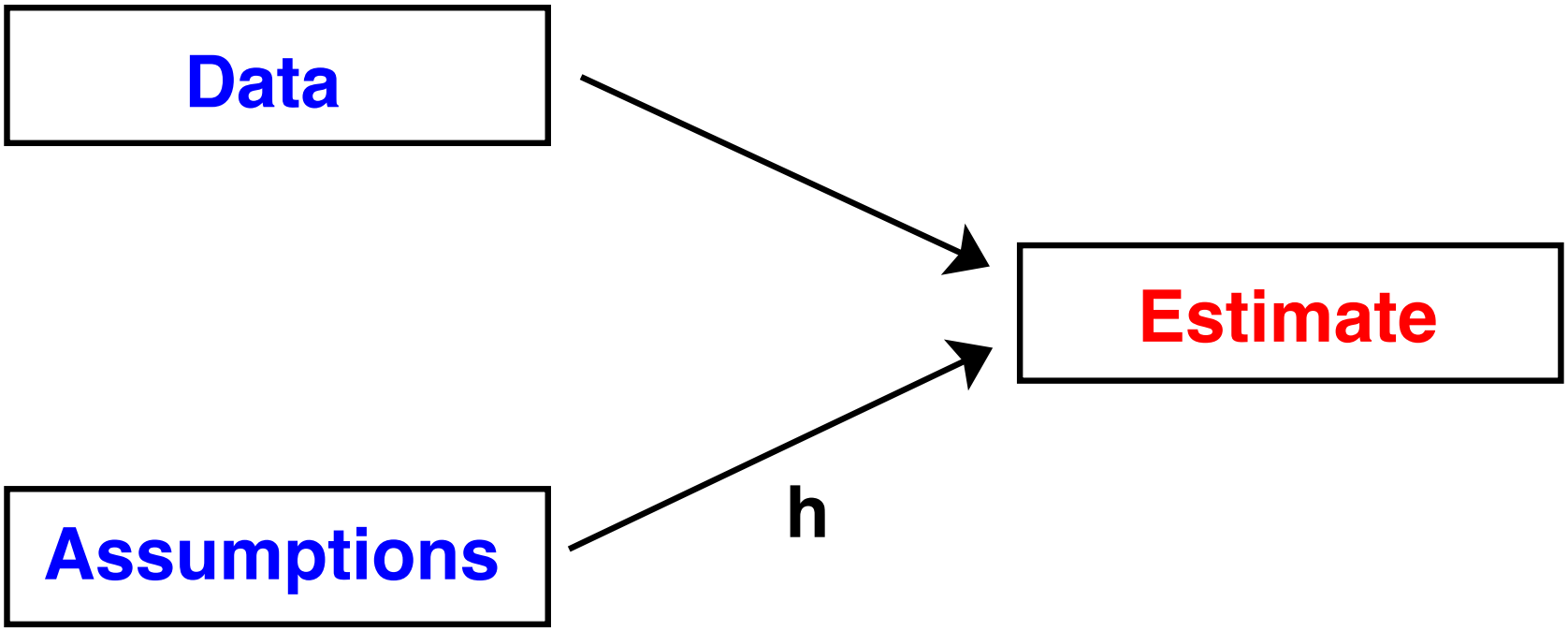

The curve is “separating signal from noise”

# What is Nonparametric?




In the parametric case, the influence of assumptions is fixed

# What is Nonparametric?




In the nonparametric case, the influence of assumptions is controlled by smoothing parameter h which shrinks with more data

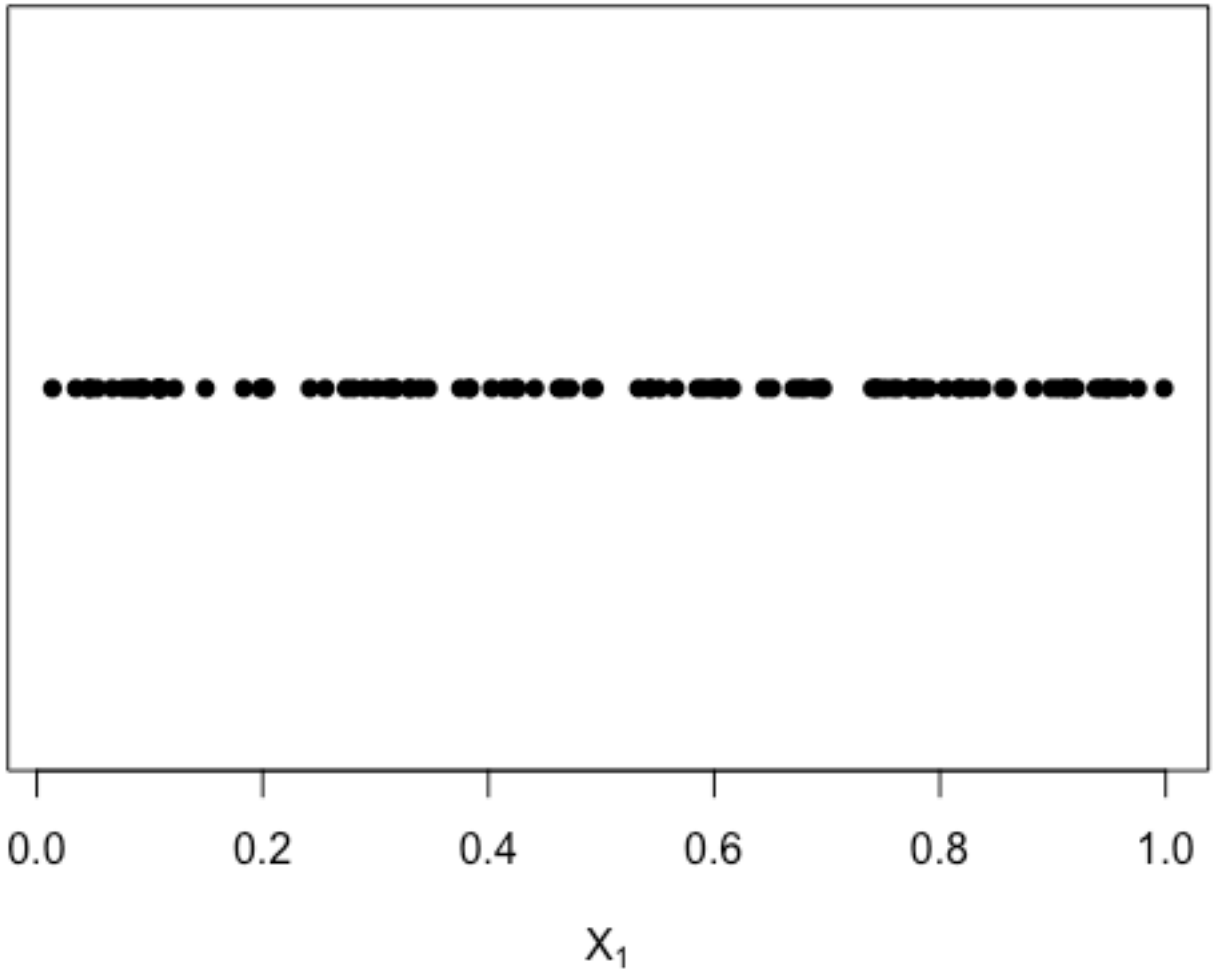
# Curse of Dimensionality

Despite the promise of nonparametrics, fitting models in high dimensions is a challenge

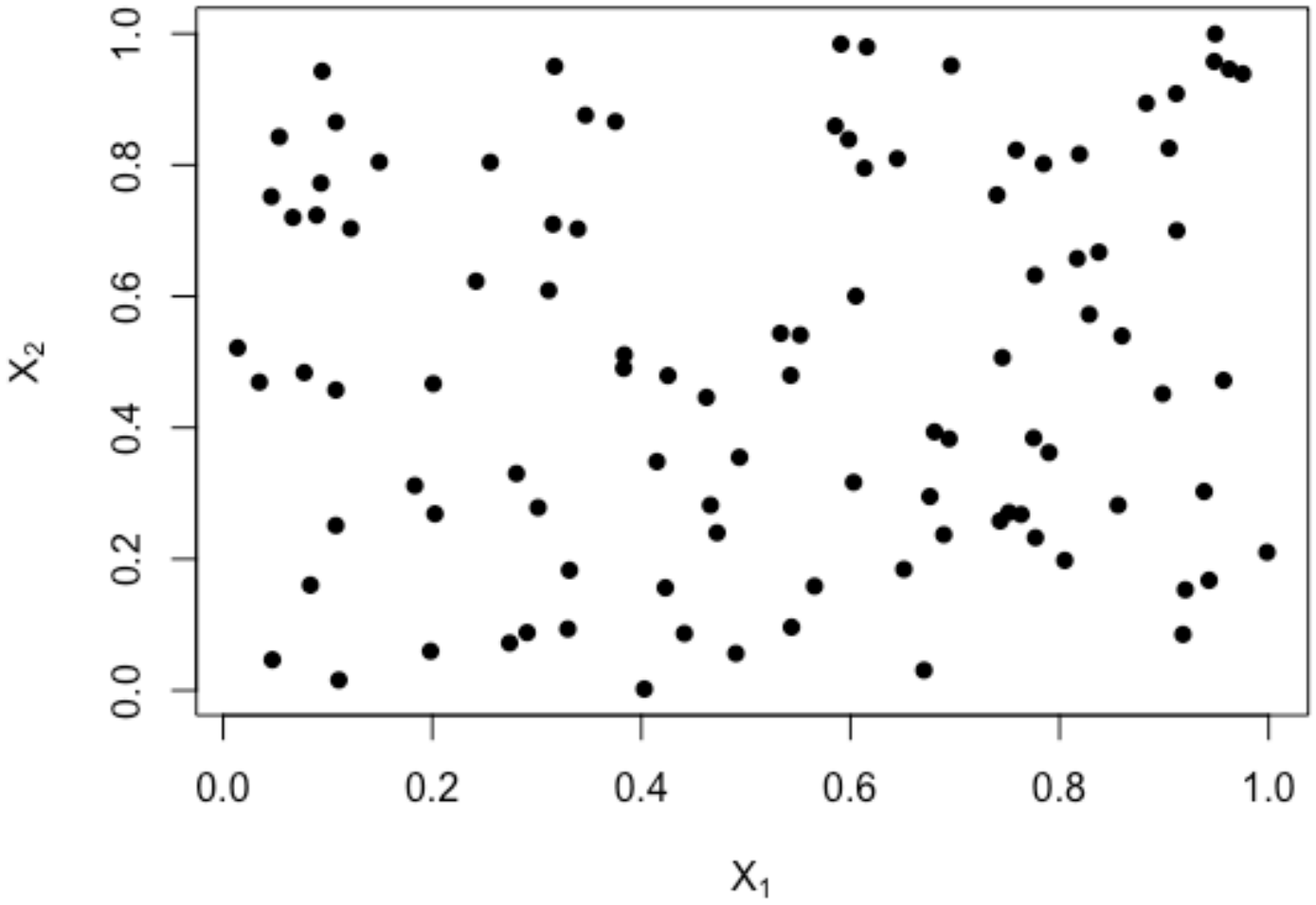
These fits require ample data in the "neighborhood" to be reliable, and data become sparse in high dimensions

Choosing neighborhoods larger reduces the value of the approach

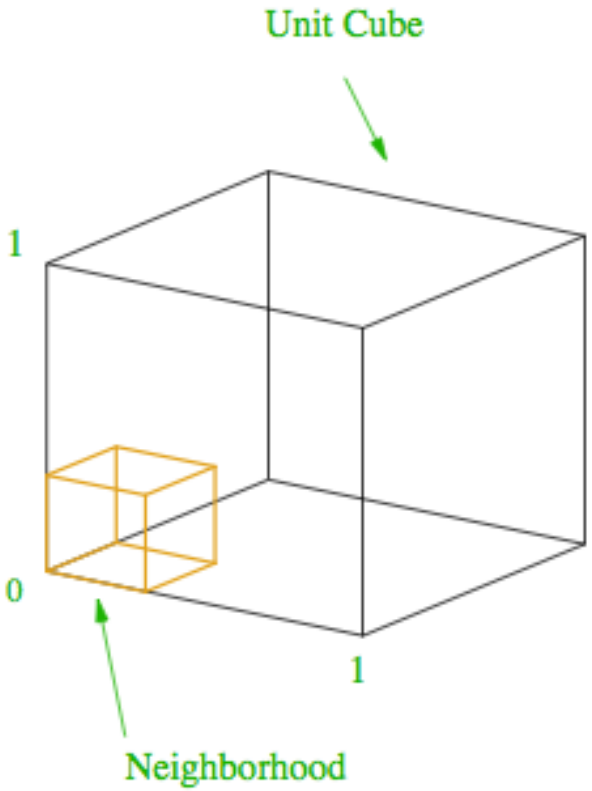
# Curse of Dimensionality



# Curse of Dimensionality



# Curse of Dimensionality



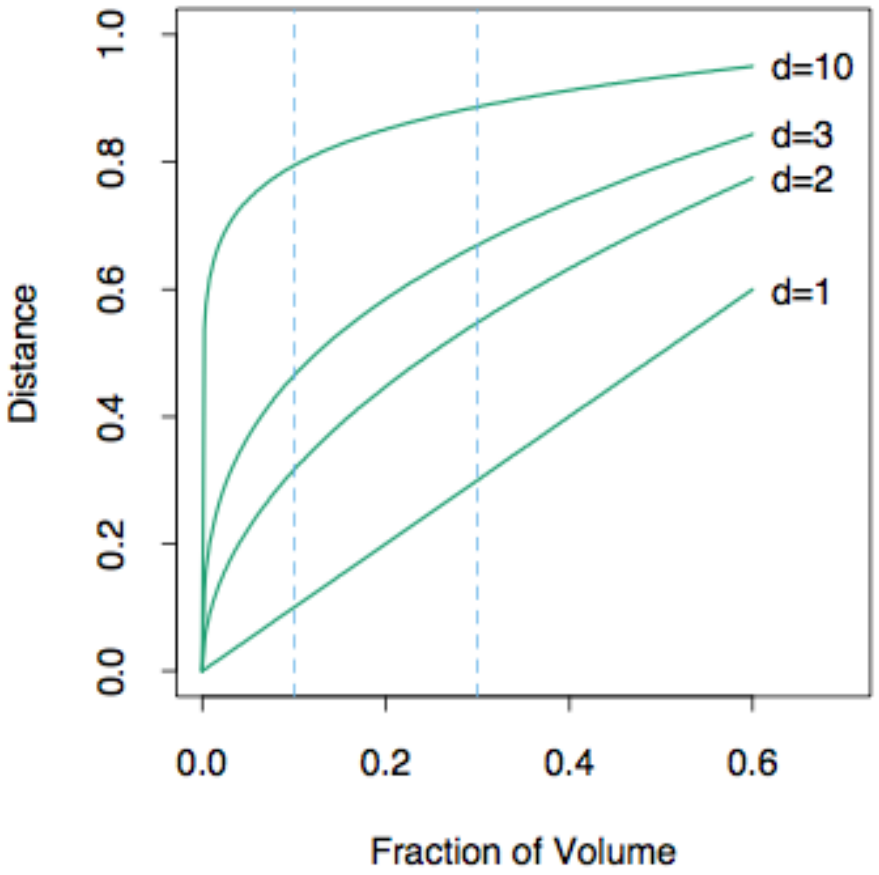
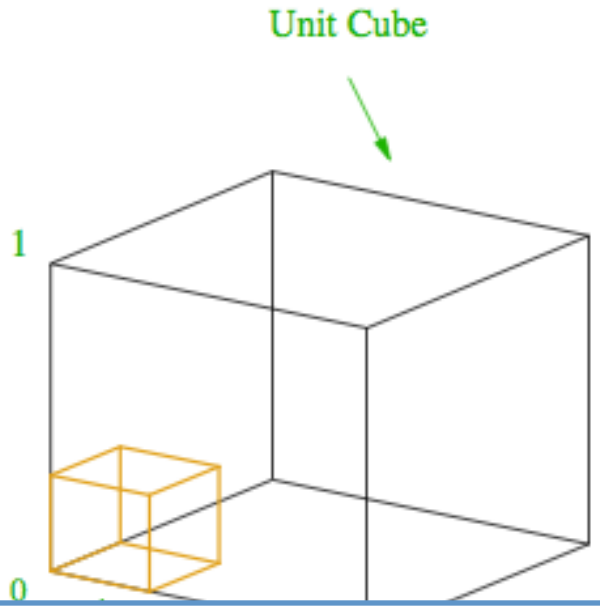




Figure 2.6 from Hastie, Tibshirani, and Friedman (2009)

# Curse of Dimensionality




In one dimension, the neighborhood has to cover 10% of the range in the one variable in order to capture 10% of the data points

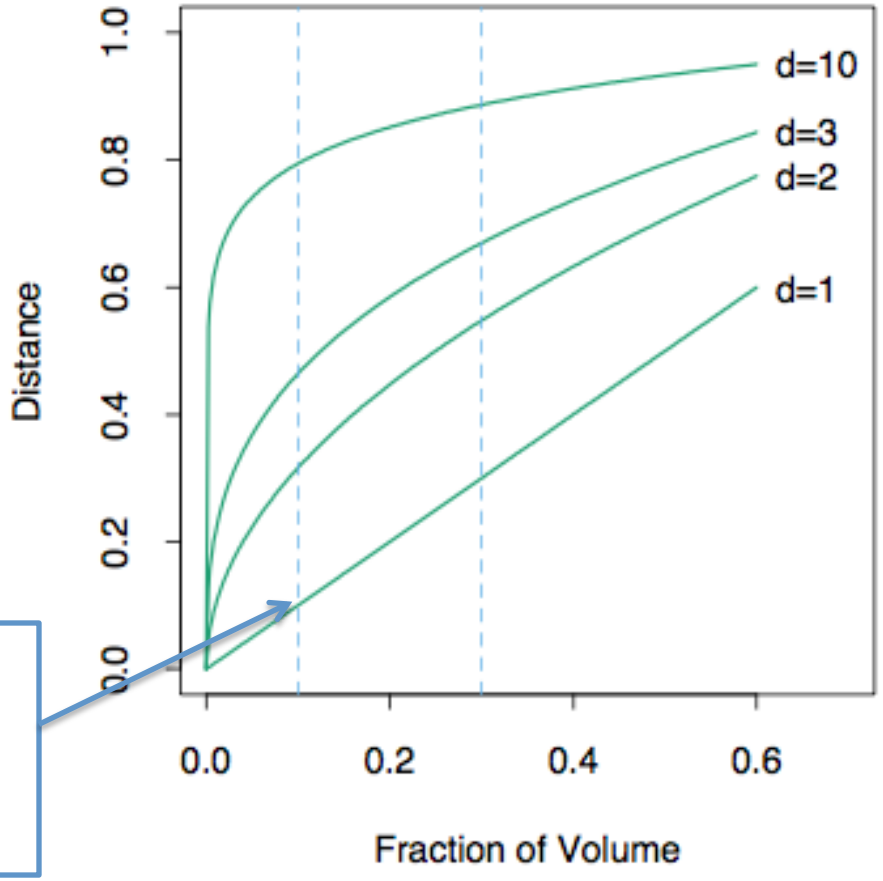
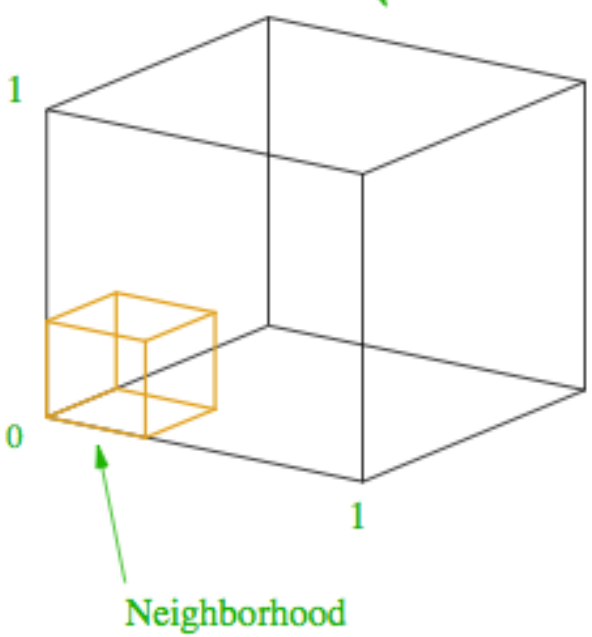



Figure 2.6 from Hastie, Tibshirani, and Friedman (2009)

# Curse of Dimensionality

In ten dimensions, the neighborhood has to cover 80% of the range in each variable in order to capture 10% of the data points



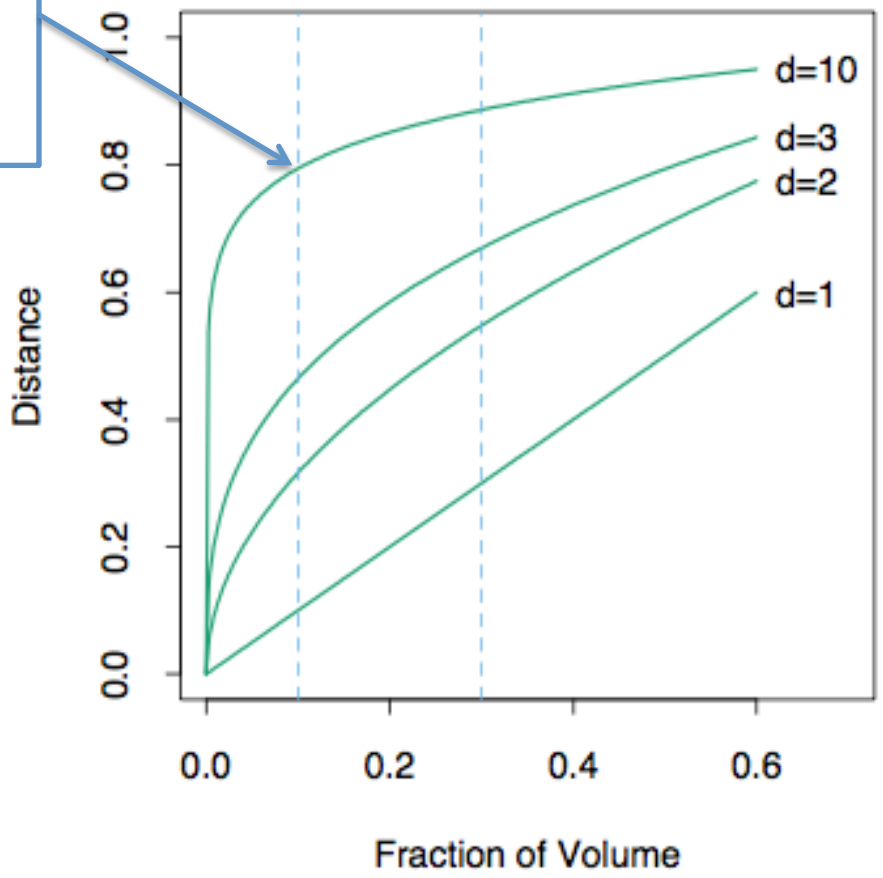




Figure 2.6 from Hastie, Tibshirani, and Friedman (2009)

# Additive Models

Additive models avoid the curse of dimensionality by making a strong, but not overly restrictive, assumption regarding the relationship between the response and predictors

# Additive Models

The fully nonparametric model:

$$f(x_1, x_2, \ldots, x_p)$$

with $f$ estimated from the data

The additive model:

$$f_1(x_1) + f_2(x_2) + \cdots + f_p(x_p)$$

with each $f_i$ estimated from the data

# Additive Models

The projection pursuit regression model:

$$\beta_0 + \sum_{k=1}^{M} f_k\left(\alpha_k^T \mathbf{x}\right)$$

with the $f_k$ and the $\alpha_k$ estimated from data

# Neural Networks

"The term neural network has evolved to encompass a large class of models and learning methods. Here we describe the most widely used "vanilla" neural net, sometimes called the single hidden layer back-propagation network, or single layer perceptron. There has been a great deal of *hype* surrounding neural networks, making them seem magical and mysterious. As we make clear in this section, they are just nonlinear statistical models, much like the projection pursuit regression model discussed above."

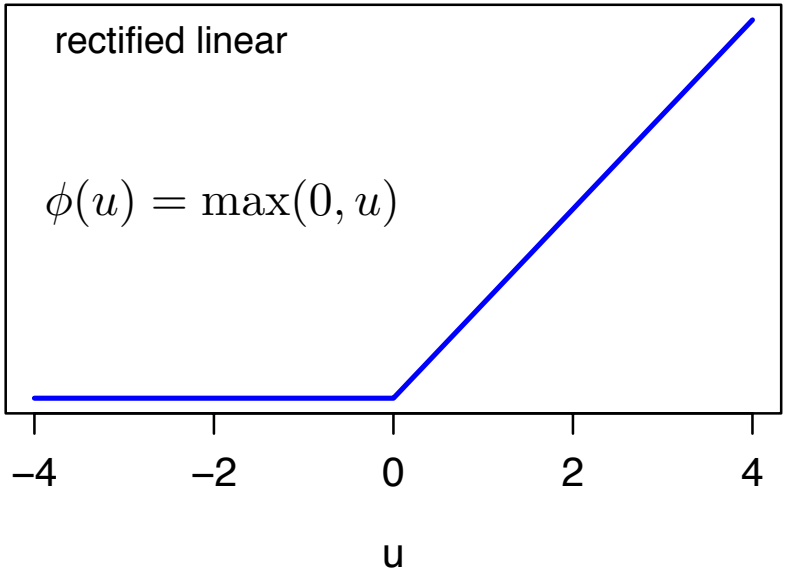
Quote from Hastie, Tibshirani, and Friedman (2009)

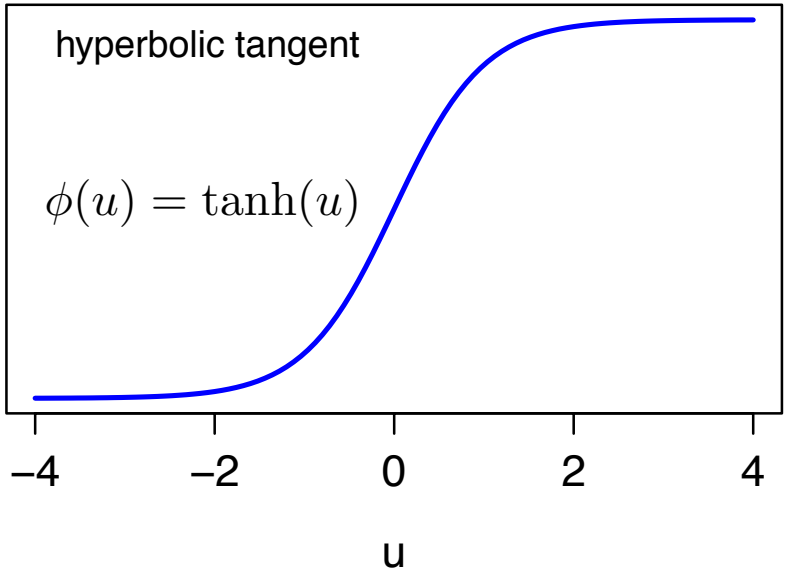
# Neural Networks

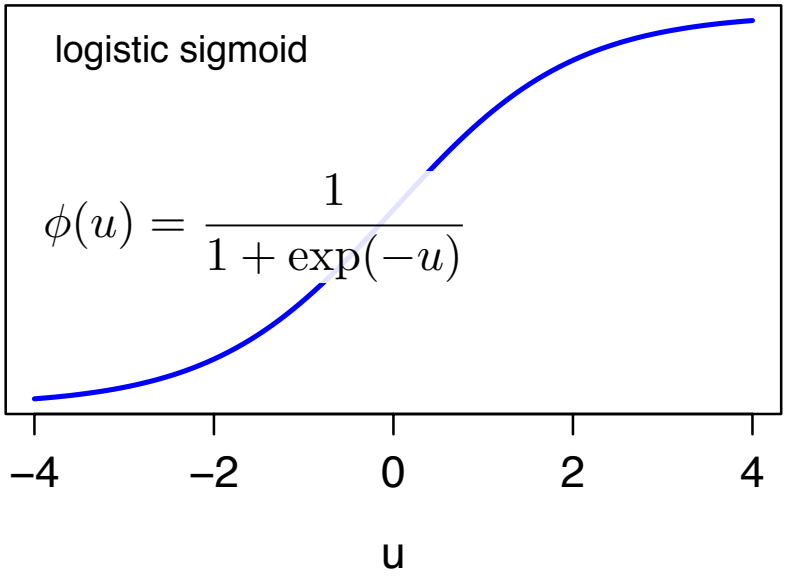
The single layer model:

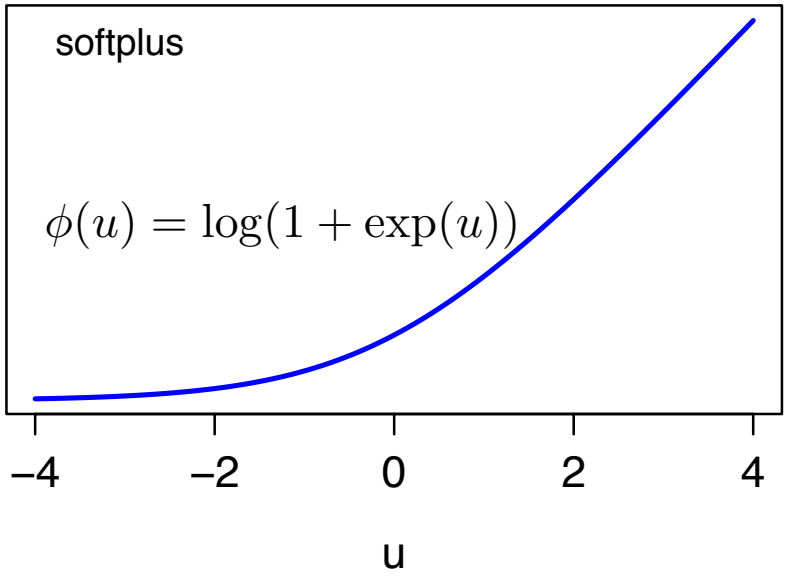
$$\beta_0 + \sum_{k=1}^{M} \beta_k \phi\left(\alpha_0 + \alpha_k^T \mathbf{x}\right)$$

with the $\beta_k$ and the $\alpha_k$ estimated from data, but $\phi$ is a user-chosen, simple nonlinear function


rectified linear
$\phi(u) = \max(0, u)$
−4
−2
0
2
4
u


hyperbolic tangent
$\phi(u) = \tanh(u)$
−4
−2
0
2
4
u


logistic sigmoid
$\phi(u) = \frac{1}{1 + \exp(-u)}$
−4
−2
0
2
4
u


softplus
$\phi(u) = \log(1 + \exp(u))$
−4
−2
0
2
4
u

# Neural Networks

The single layer model:

$$\beta_0 + \sum_{k=1}^{M} \beta_k \phi\left(\alpha_0 + \alpha_k^T \mathbf{x}\right)$$

Compare with projection pursuit model:

$$\beta_0 + \sum_{k=1}^{M} f_k\left(\alpha_k^T \mathbf{x}\right)$$

# Neural Networks

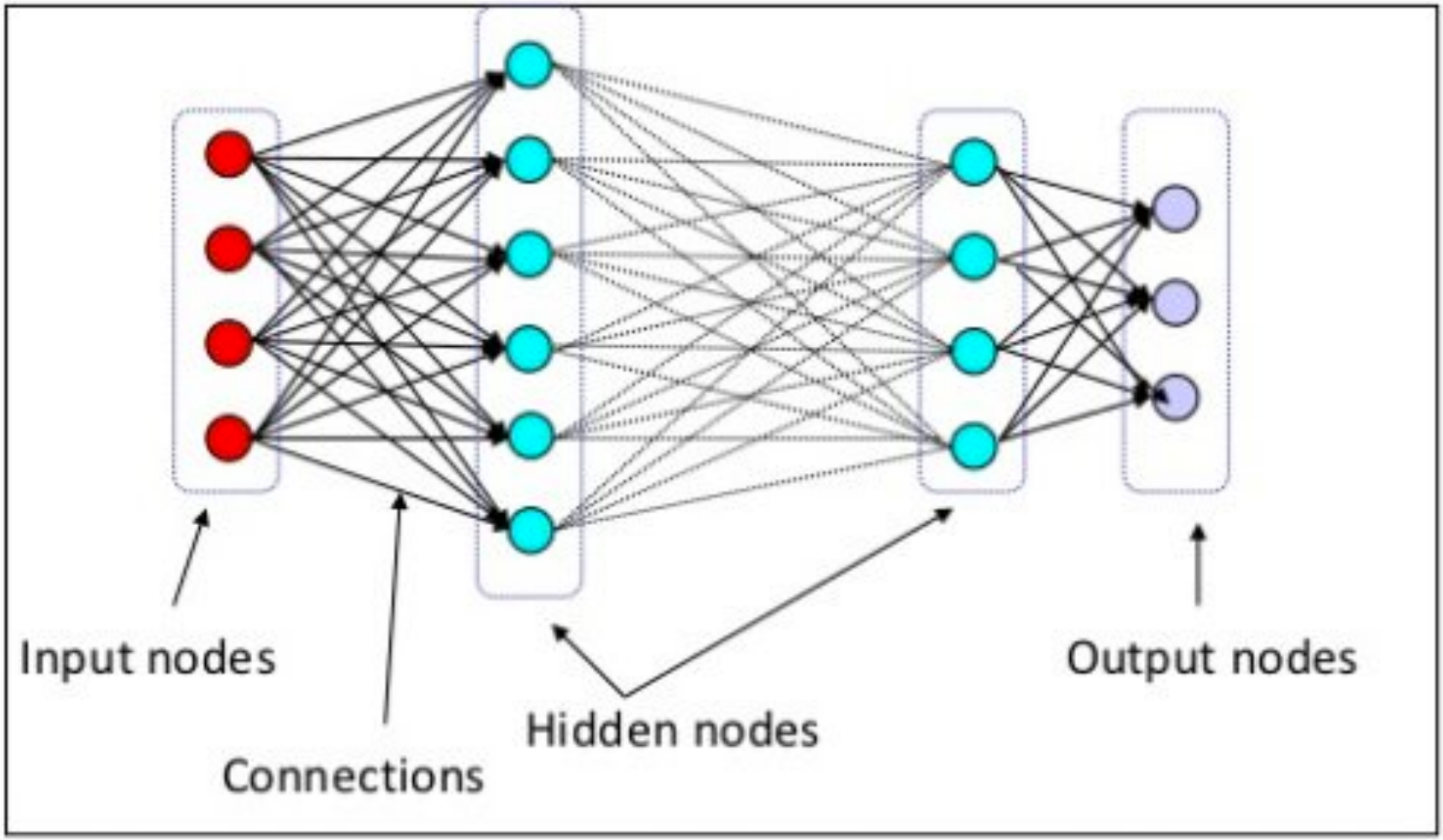
Sample size constrains the available model degrees of freedom (MDF)

Projection Pursuit: MDF "spent" on the $f_k$

Neural Network: MDF "spent" on increasing $M$

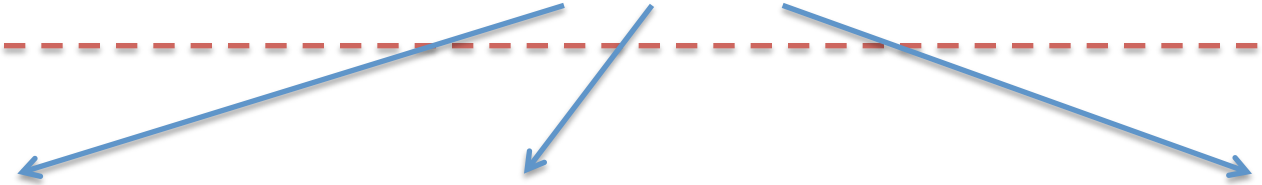
Larger $M$ means more linear compressions of data are considered

# Adding Layers



**Input Layer**

$$\mathbf{x} = (x_1, x_2, \ldots, x_d)$$

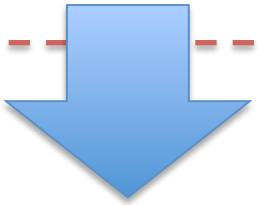


$$\phi(b_1 + \mathbf{w}_1^T\mathbf{x}) \quad \phi(b_2 + \mathbf{w}_2^T\mathbf{x}) \quad \cdots \quad \phi(b_m + \mathbf{w}_m^T\mathbf{x})$$

$\phi(\cdot)$ **is the activation function, a simple nonlinear mapping**

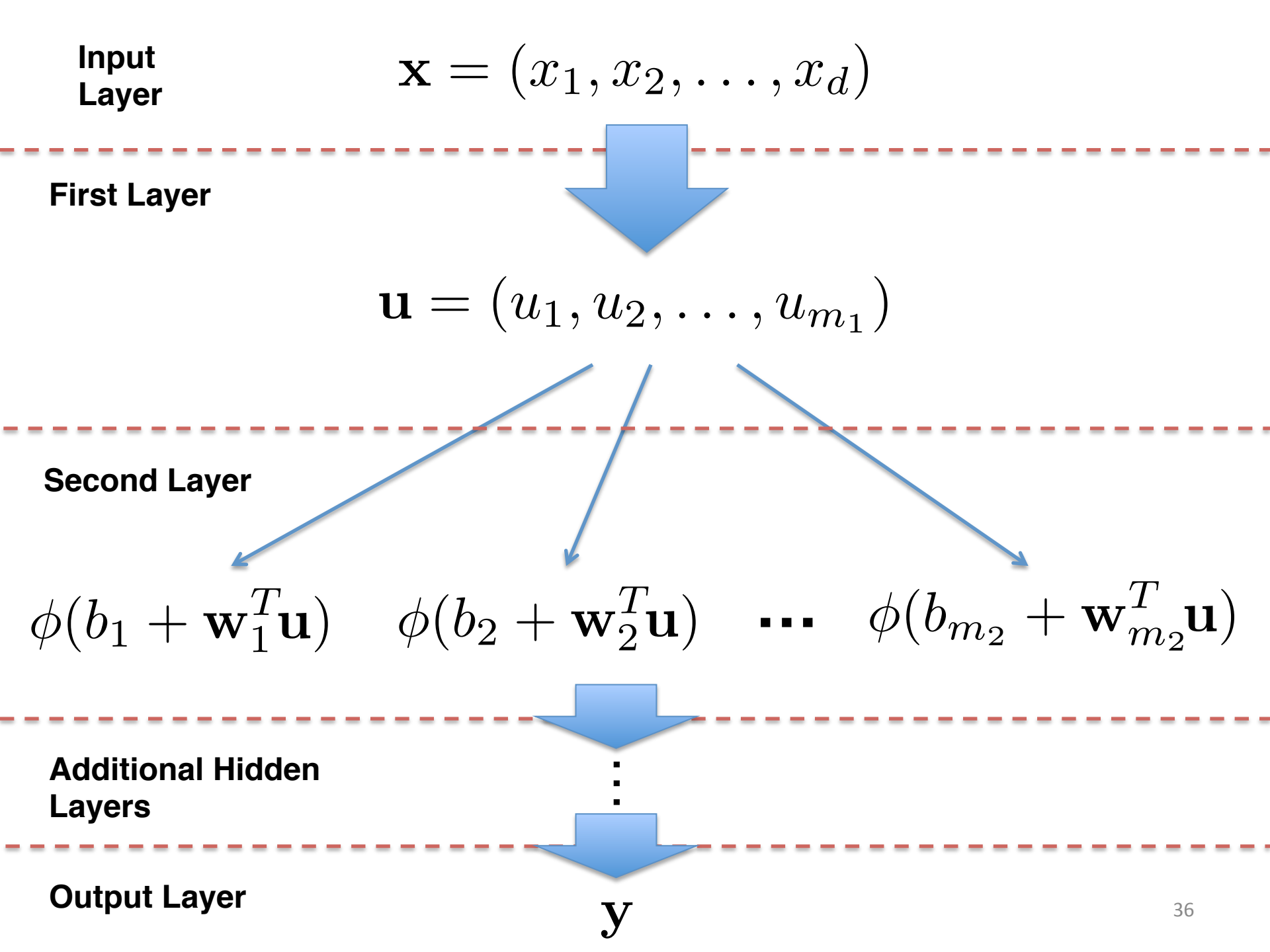
**Input Layer**

$$\mathbf{x} = (x_1, x_2, \ldots, x_d)$$



**First Layer**

$$\mathbf{u} = (u_1, u_2, \ldots, u_{m_1})$$


Input Layer
$\mathbf{x} = (x_1, x_2, \ldots, x_d)$
First Layer
$\mathbf{u} = (u_1, u_2, \ldots, u_{m_1})$
Second Layer
$\phi(b_1 + \mathbf{w}_1^T\mathbf{u})$
$\phi(b_2 + \mathbf{w}_2^T\mathbf{u})$
$\cdots$
$\phi(b_{m_2} + \mathbf{w}_{m_2}^T\mathbf{u})$
Additional Hidden Layers
Output Layer
$\mathbf{y}$

# Deep Learning

"Deep learning is a particular kind of machine learning that achieves great power and flexibility by representing the world as a nested hierarchy of concepts, with each concept defined in relation to simpler concepts, and more abstract representations computed in terms of less abstract ones."

--Page 8 in *Deep Learning,*
Goodfellow, Bengio, and Courville (2016)

# Resurgence of NN

Multiple factors contributed to growth of interest in Deep Learning:

- Increase in training set sizes
- Improved algorithms for training deeper networks (e.g., Hinton, et al. in 2006)
- Growth in computational resources
- Successes

# Flexibility

A primary appeal of the approach is the flexibility in constructing the layers

- How many units are there in each layer?
- What is the mapping from one layer to the next?
- How is the output constructed from the final hidden layer?

# The Contrast

Statistical: Carefully construct a model to relate the input to the output

Deep Learning: exploit a large collection of simple components to make a prediction

What is the role of expert knowledge?

# How Does it Work?

Universal Approximation Theorem (Hornik 1991): With enough units, a single hidden layer can approximate to arbitrary precision any “nice” function.

But: Deeper networks use units more efficiently, are easier to fit, and generalize better

# How Does it Work?

But: Deeper networks use units more efficiently, are easier to fit, and generalize better

Montufar, et al. (2014): "[f]or deep models, the maximal number of linear regions grows exponentially fast with the number of parameters, whereas, for shallow models, it grows polynomially fast with the number of parameters."

# Fitting the Model

A cost function is optimized to estimate the parameters (weights)

Choose cost function to maximize appropriate likelihood

Stochastic gradient descent with back propagation to estimate gradient

# Regularization

Overfitting is a huge concern

Approaches to regularization (smoothing) manage the bias/variance tradeoff

The model is "parametric," so $L^2$ (ridge) or $L^1$ (lasso) penalties on the cost function are commonly used

# Regularization

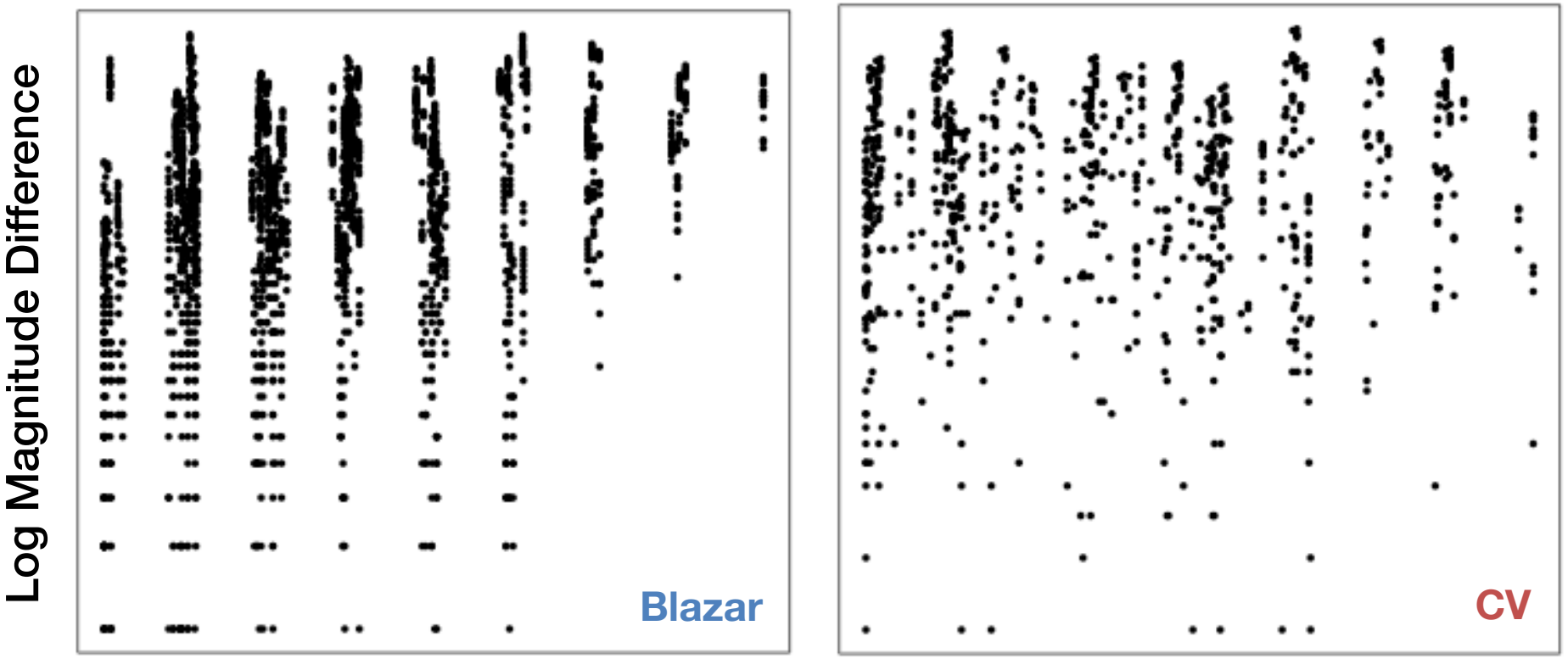
Dropout is a novel approach to regularization

Units are randomly included/excluded during training, approximating averaging over all possible submodels

Variant of bagging

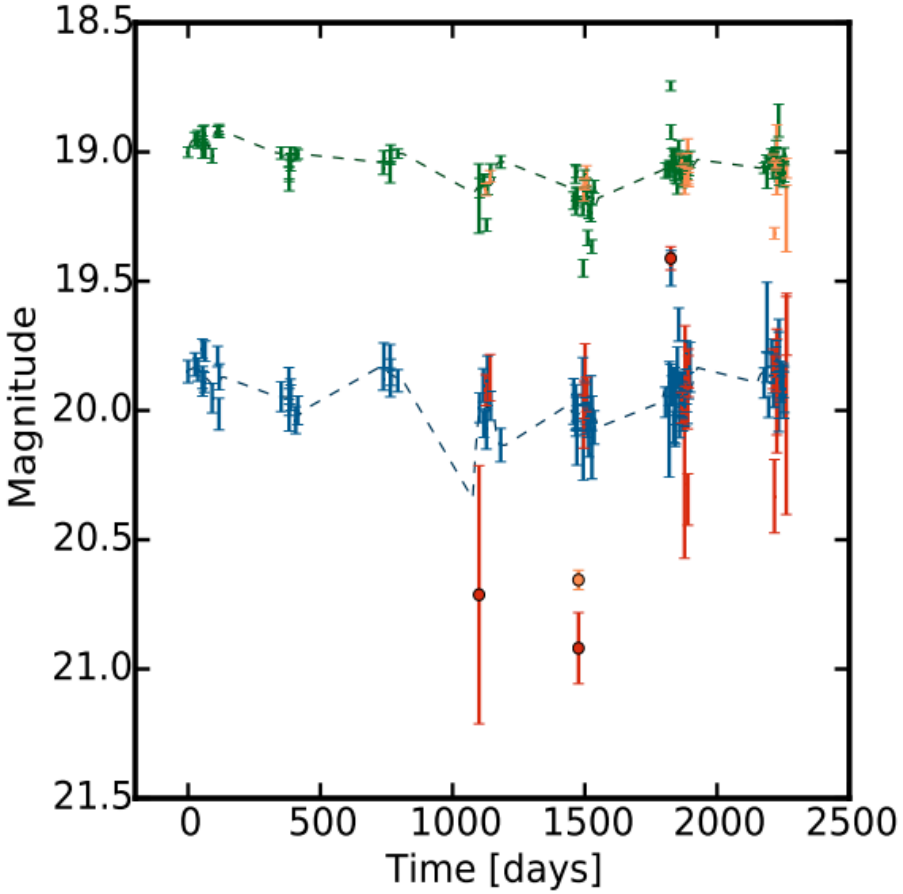
Reduces potential influence of any individual unit

# Blazars versus CVs




Comparison of Structure Functions

# Summarizing the SF



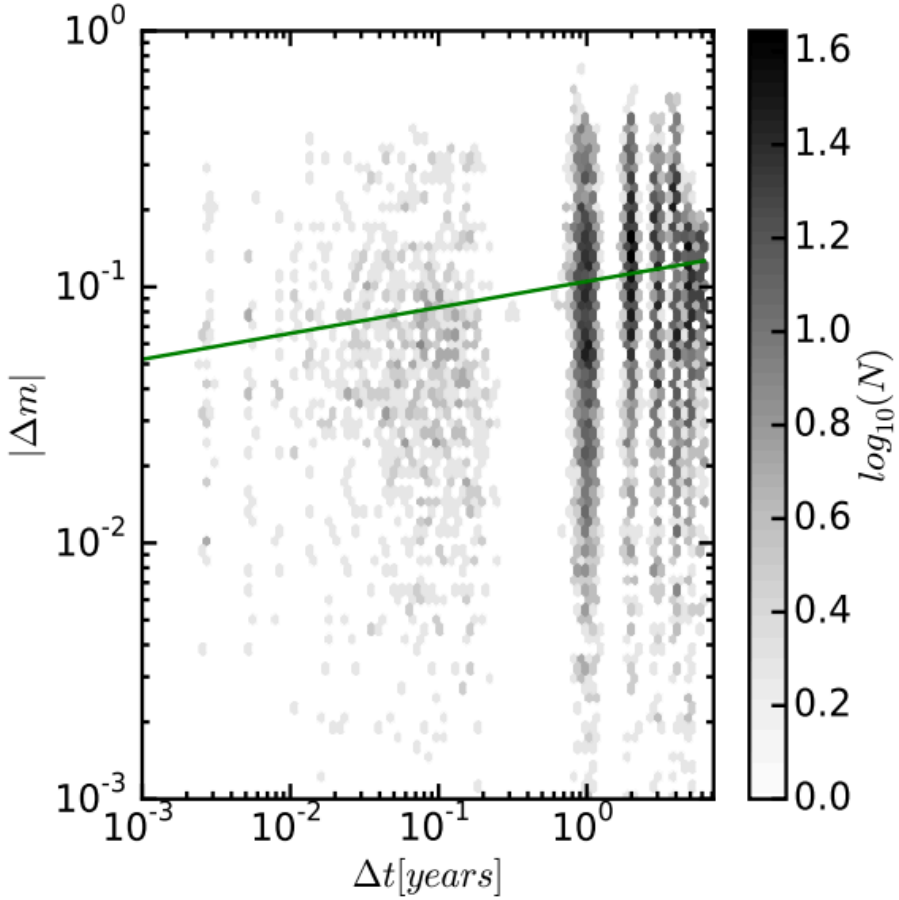




Figure 2 in Peters et al. (2015) Quasar light curve and SF

# Summarizing the SF

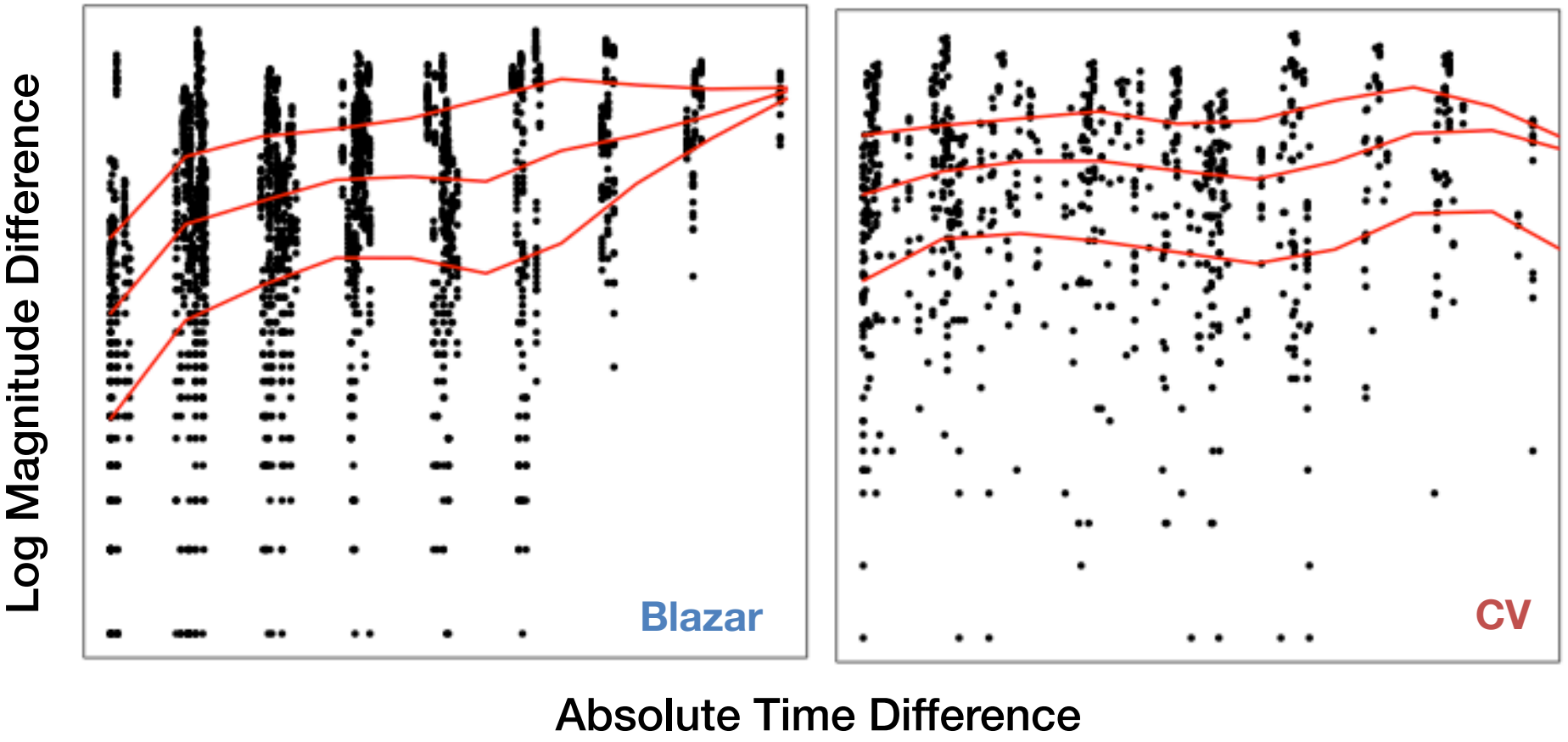
Typical to fit model to structure function

- Power Law Form (Schmidt et al. 2010)
- Damped Random Walk (Kelly et al. 2009)

Effort to find a low-dimensional representation, avoiding the curse of dimensionality

Ideally, could utilize higher-dimensional representation

# Blazars versus CVs




Quantile regression fits

# Blazar versus CV

Fit model with three hidden layers, using Dropout

128 nodes per layer

Rectified linear units as the activation functions

958 CVs, 318 Blazars from Catalina Real-Time Transient Survey (Drake 2019)

# Blazar versus CV

Performance on test set:

| Prediction \ Truth | Blazar | CV |
|---|---|---|
| Blazar | 18 | 10 |
| CV | 8 | 91 |

# Blazar versus CV

Performance on test set:

**Deep Learning**

| Prediction \ Truth | Blazar | CV |
|---|---|---|
| Blazar | 18 | 10 |
| CV | 8 | 91 |

**Random Forest**

| Prediction \ Truth | Blazar | CV |
|---|---|---|
| Blazar | 12 | 8 |
| CV | 14 | 93 |

# Potential of Deep Learning

Best suited to situations where high-dimensional input is required

Avoid the curse of dimensionality

Seems particularly relevant for classification challenges

Can be extended to unsupervised learning - autoencoders

# References


Binney and Merrifield 1998. *Galactic Astronomy*. Princeton, NJ: Princeton University Press
Boyle BJ, et al. 1988. MNRAS (235): 935
Drake, AJ, et al. 2009. ApJ (696): 870
Eyer and Mowlavi 2008. *Journal of Physics Conference Series*
Goodfellow, Bengio, and Courville 2016. *Deep Learning.* MIT Press
Hastie, Tibshirani, and Friedman 2009. *Elements of Statistical Learning*
Hinton, et al. 2006. *Neural Computation* (18): 1527
Hornik 1991. *Neural Networks* (4): 251
Kelly, et al. 2009. ApJ (698): 895
Marshall, et al. 1984. ApJ (283): 50
Flesch 2019. http://quasars.org/milliquas.htm
Montufar, et al. *NIPS* 2014
Paris, et al. 2014. *Astronomy and Astrophysics* (563): A54
Pei 1995. ApJ (438): 623
Peters, et al. 2015. ApJ (811): 95
Press and Schechter 1974. ApJ (187): 425
Schechter 1976. ApJ (203): 297
Schmidt et al. 2010. ApJ (714): 1194
Schneider, et al. 2002. AJ (123):567
Schneider, et al. 2003. AJ (126): 2579
Schneider, et al. 2005. AJ (130): 367
Schneider, et al. 2007. AJ (134): 102
Schneider, et al. 2010. AJ (139): 2360


AJ = *Astronomical Journal*. ApJ = *Astrophysical Journal*, MNRAS = *Monthly Notices of the Royal Astronomical Society*