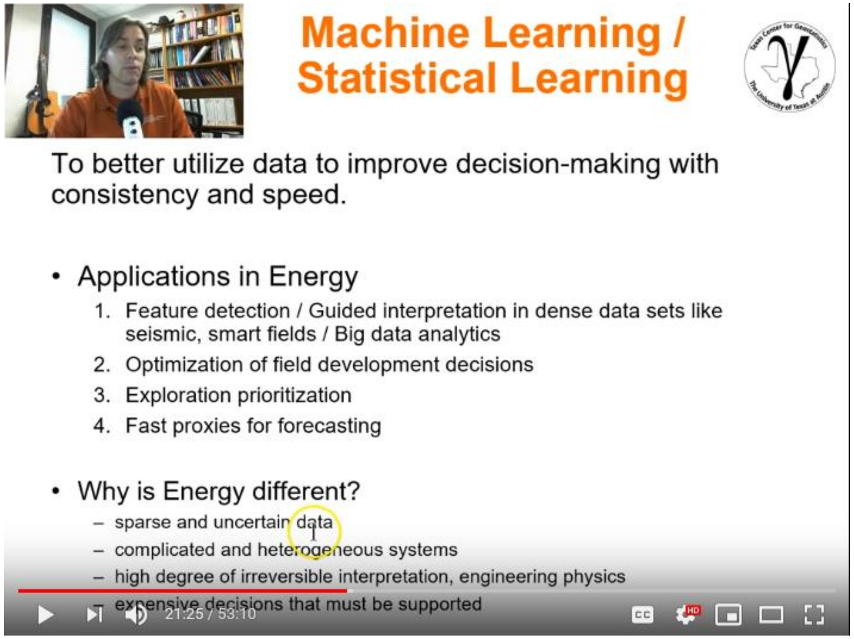
Multivariate Modeling: Statistical Learning

Lecture outline . . .

- Statistical Learning



Machine Learning / Statistical Learning

To better utilize data to improve decision-making with consistency and speed.

- Applications in Energy
  1. Feature detection / Guided interpretation in dense data sets like seismic, smart fields / Big data analytics
  2. Optimization of field development decisions
  3. Exploration prioritization
  4. Fast proxies for forecasting
- Why is Energy different?
  - sparse and uncertain data
  - complicated and heterogeneous systems
  - high degree of irreversible interpretation, engineering physics
  - expensive decisions that must be supported

https://www.youtube.com/watch?v=9P4S0Wh4cho

Introduction

Prerequisites

Probability

Multivariate Analysis

Spatial Estimation

Statistical Learning

Feature Selection

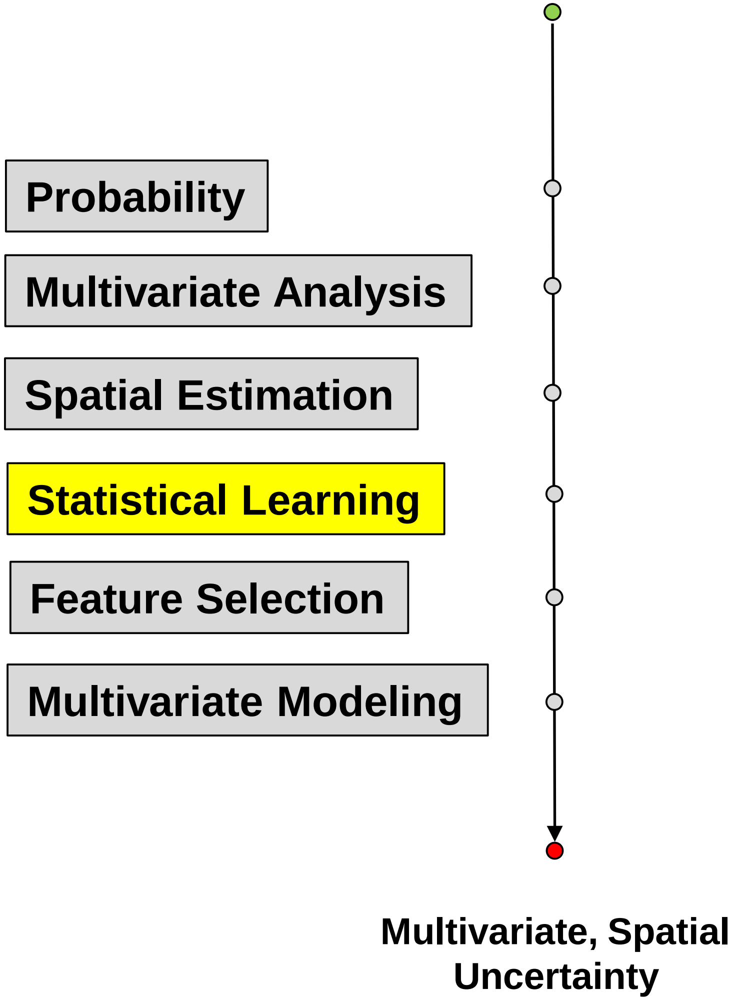
Multivariate Modeling

Conclusions

What Will You Learn?

Why Cover Statistical Learning?

- Consider inference and prediction
- Consider model training vs. testing
- Consider complexity vs. itnerpretability



Machine Learning / Statistical Learning

- I’m using this book for this section of the class: An Introduction to Statistical Learning with Applications in R, 2013, James et al., Springer. (http://www-bcf.usc.edu/~gareth/ISL/ISLR%20First%20Printing.pdf)

- Statistical Learning
  - vast set of tools for learning from data
  - based on initial assumptions and hypothesis

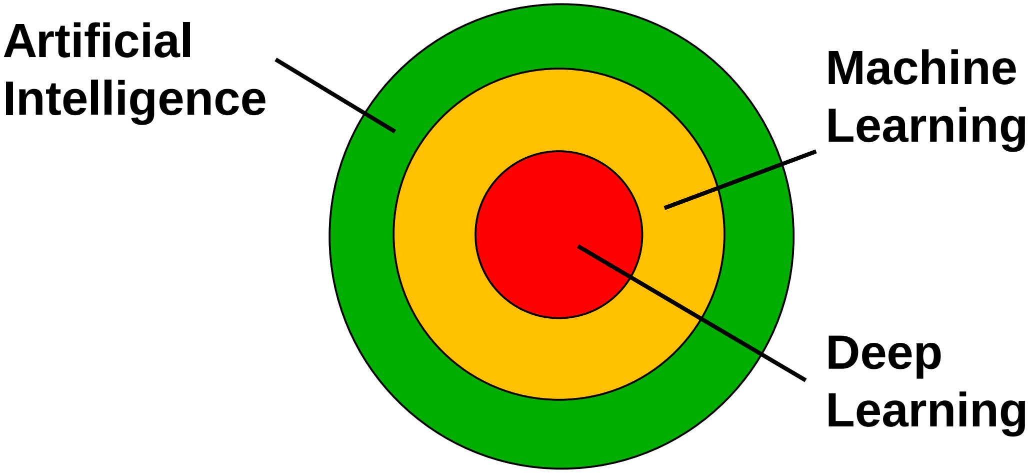
- Machine Learning vs. Statistical Learning
  - vast set of tools for learning patterns
  - very little if any prior assumptions

Note: Some consider statistical learning and machine learning to be the same
I’ll use them interchangeably

- Supervised Learning
  - building a predictive model for estimating an output given one or more inputs

- Unsupervised Learning
  - all inputs, no output
  - learn from the structures of the data alone

Machine Learning / Statistical Learning



Artificial Intelligence: the theory and development of computer systems able to perform tasks that normally require human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages (Google Dictionary)

Machine Learning: is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed (Google Dictionary). Access data and learn for themselves.

Deep Learning: subset of machine learning for unsupervised learning from unstructured, unlabeled data.

Machine Learning / Statistical Learning

Machine Learning:

toolkit

training with data

“is the study of algorithms and mathematical models that computer systems use to progressively improve their performance on a specific task. Machine learning algorithms build a mathematical model of sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task.”

learning

general

“where it is infeasible to develop an algorithm of specific instructions for performing the task.”

not a panacea

Machine Learning - Wikipedia

Machine Learning / Statistical Learning

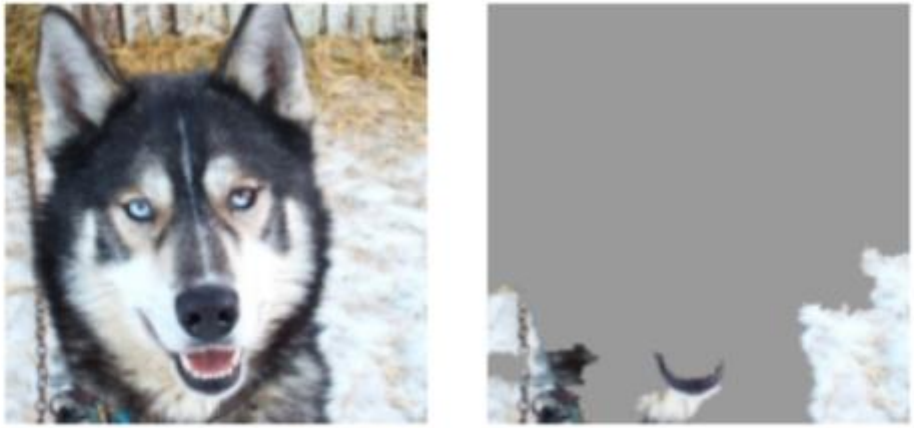
Concerns:

Biased training data

Rideiro et al. (2016) trained a logistic regression classifier with 20 wolves and dogs images to detect the difference between wolves and dogs.

The problem is:

- interpretability may be low
- application may become routine and trusted
- the machine is trusted, becomes an authority



(a) Husky classified as wolf

(b) Explanation

Figure 11: Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task.

Image and example from Ribeiro et al., (2016)
https://arxiv.org/pdf/1602.04938.pdf

Big Data

Big Data: you have big data if your data has a combination of these:

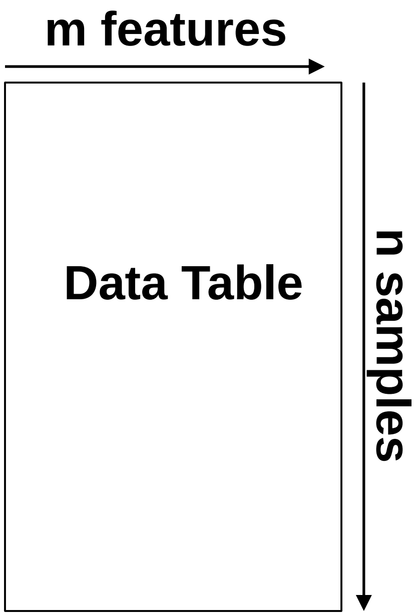
Volume: large number of data samples, large memory requirements and difficult to visualize

Velocity: data is gathered at a high rate, continuously relative to decision making cycles

Variety: data form various sources, with various types and scales

Variability: data acquisition changes during the project

Veracity: data has various levels of accuracy



"Energy has been big data before tech learned about big data."
– Michael Pyrcz

Big Data Analytics – methods to explore and detect patterns, trends and other useful information from big data to improve decision making.

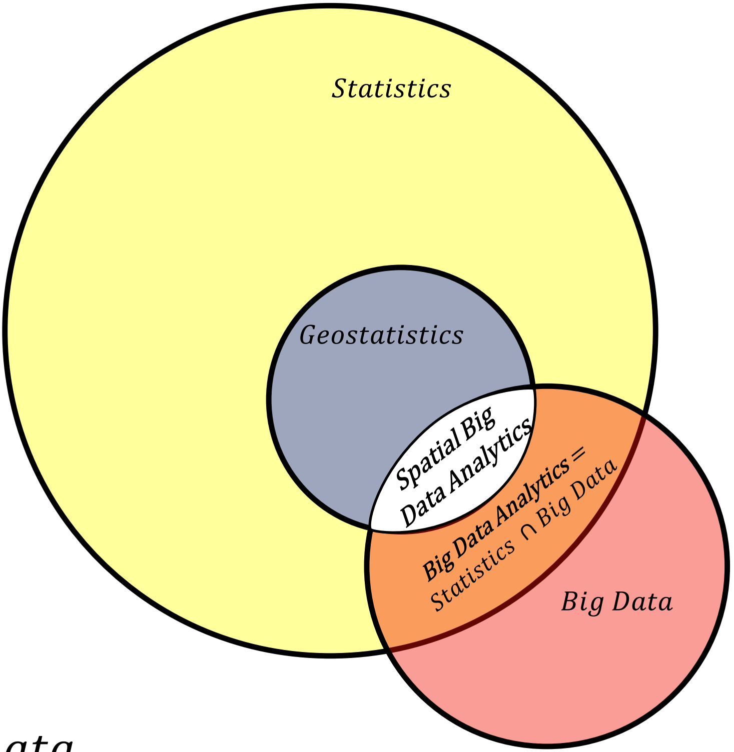
Big Data Analytics

Statistics is collecting, organizing, and interpreting data, as well as drawing conclusions and making decisions.

Geostatistics is a branch of applied statistics: (1) the spatial (geological) context, (2) the spatial relationships, (3) volumetric support, and (4) uncertainty.

Big Data Analytics is the process of examining large and varied data sets (big data) to discover patterns and make decisions.

$$\textit{Spatial Big Data Analytics} = \textit{Geostatistics} \cap \textit{Big Data}$$

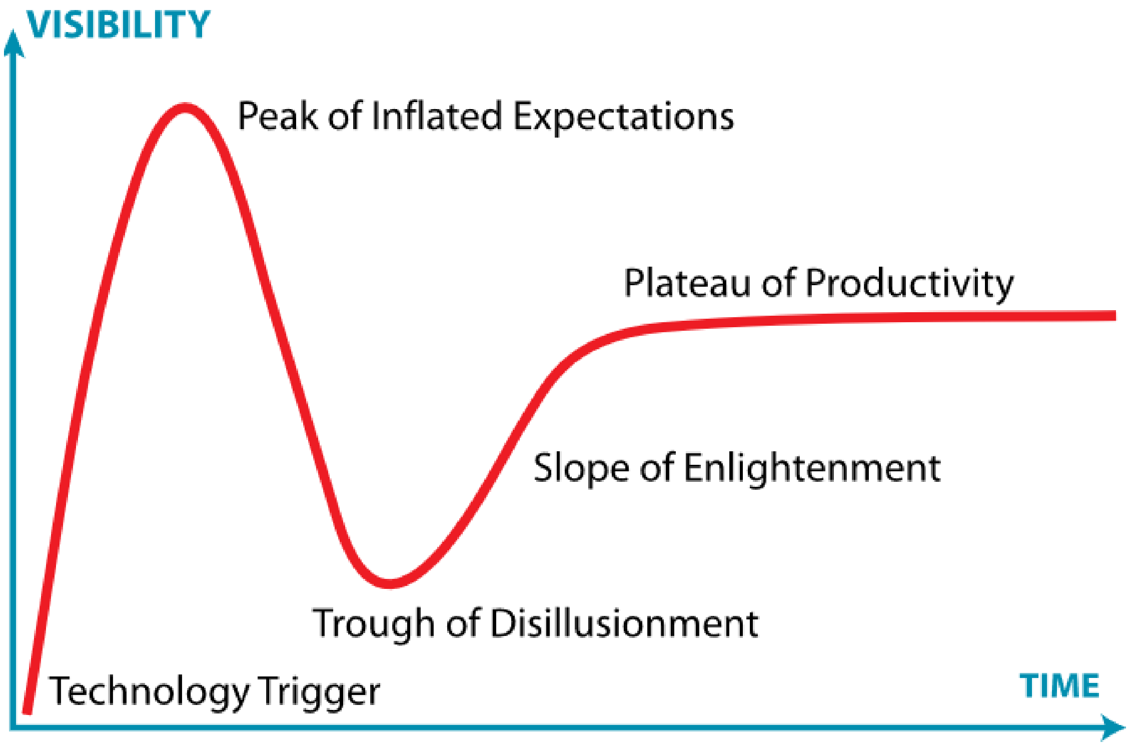

Statistics
Geostatistics
Spatial Big Data Analytics
Big Data Analytics = Statistics ∩ Big Data
Big Data

Proposed Venn diagram for spatial big data analytics.

Big data analytics is expert use of (geo)statistics on big data.

Machine Learning / Statistical Learning

- Hype Cycle – from information technology firm Gartner (https://en.wikipedia.org/wiki/Hype_cycle)



Where are we currently for machine learning?

Machine Learning / Statistical Learning

- Applications Around You / Societal Impacts
  1. Driving directions that crowd source and update improve traffic flow
  2. Air traffic routing
  3. Spam filters
  4. Plagiarism checkers
  5. Translation / computer reading
  6. Credit card fraud detection
  7. Face recognition (Facebook, Snapchat etc.)
  8. Recommendations (Amazon, Netflix, YouTube)
  9. Smart personal assistants

Machine Learning / Statistical Learning

To better utilize data to improve decision-making with consistency and speed.

- Applications in Energy
  1. Feature detection / Guided interpretation in dense data sets like seismic, smart fields / Big data analytics
  2. Optimization of field development decisions
  3. Exploration prioritization
  4. Fast proxies for forecasting

- Why is Energy different?
  - sparse and uncertain data
  - complicated and heterogeneous systems
  - high degree of irreversible interpretation, engineering physics
  - expensive decisions that must be supported

Machine Learning / Statistical Learning

- Just like spatial statistics / geostatistics, statistical learning is a set of tools to add to your tool box as an engineer

- Each is very dangerous to use as a black box. You will need to understand what's under the hood
  - methods, workflows, assumptions and limitations.
  - scope and trade offs between alternative methods

- Imagine you are a carpenter (all geostatistics workflows) (Pyrcz and Deutsch, 2014).
  - You would have a tool box
  - You would know each tool perfectly well
  - Understand performance over a variety of applications
  - You would understand the range of applications, weaknesses, strengths, limits.
  - Choice between tools would be based on expert judgement of circumstances and goals of a project
  - You would choose specific tools to have ready for use and other for more rare circumstances
  - Too few tools and a box overwhelmed with obscure tools are both issues.

The Model

- Predictors, Independent Variables, Features
  - input variables
  - for a model $Y = f(X_1, \dots, X_m) + \epsilon$, these are the $X_1, \dots, X_m$
  - note ϵ is a random error term

- Response, Dependent Variables
  - output variable
  - for a model $Y = f(X_1, \dots, X_m)$, this is Y

- Statistical / Machine Learning is All About
  - Estimating f for two purposes
    1. Prediction
    2. Inference

Prediction

- Estimating, $\hat{f}$, for the purpose of predicting $\hat{Y}$
  - We are focused on getting the most accurate estimates, $\hat{Y}$
  - We may not even understand what is happening between the X's!
  - We are concerned about the relationships between X and Y

- Accuracy of $\hat{Y}$ depends on reducible and irreducible error
  - $\hat{f}$ is not a perfect model. Error due to the estimate of f is reducible error
  - but even if we had f, $\hat{Y} = f(X)$, prediction would still have error
  - This is because Y is a function of ϵ, $Y = f(X) + \epsilon$
  - and ϵ is irreducible

Inference

- There is value in understanding the relationships
  - for $Y = f(X_1, \dots, X_m) + \epsilon$ we can understand the influence / interactions of each X_α on Y and eachother.

1. Which predictors are associated with the response?
   a) What data to collect? Value of information.
   b) What data to focus on? Simplification of the model. Communication. Big hitters.

2. What is the relationship between each response and each predictor?
   a) sense of the relationship (positive or negative)?
   b) shape of relationship (sweet spot)?
   c) relationships may depend on values of other predictors!

3. Can the relationship be modeled linearly?
   a) much simplified
   b) very low parametric representation
   c) use multiGaussian?

Estimating f

- Parametric Methods
  - make an assumption about the functional form, shape
  - we gain simplicity and advantage of only a few parameters
  - use training data to fit or train the model
  - test the model with withheld test data
  - for example, here is a linear model

$$Y = f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m$$

  - there is a risk that $\hat{f}$ is quite different than f, then we get a poor model!

- Model fitting
  - Apply training data
  - Solve for least squares solution for coefficients β_0, β_1,…, β_m

Estimating f

- Nonparametric Methods
  - make no assumption about the functional form, shape
  - estimate f that approaches the data without being too rough
  - more flexibility to fit a variety of shapes for f
  - less risk that $\hat{f}$ is a poor fit for f
  - does not reduce the problem to estimating a small set of parameters
    - » Typically need a lot more data for an accurate estimate of f
    - » Risk of overfitting is greater
    - » May also be parameter rich (e,g. a decision tree as a set of thresholds and averages)
    - » Lacks a compact expression for the model

Training and Testing

Training Phase

- The training subset of the data is applied to select the model parameters (fit the model) usually optimized to minimize the mean square error.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left[\left(y_i - \hat{f}(x_1^i, \dots, x_m^i)\right)^2\right]$$

Testing Phase

- Apply the model to the testing data (data withheld from training)
- Optimize the model hyper parameters (e.g. complexity) to minimize mean square error with the testing data

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left[\left(y_i - \hat{f}(x_1^i, \dots, x_m^i)\right)^2\right]$$

Prediction Accuracy vs. Model Interpretability / Explainability

- Interpretability / Explainability
  - is the ability to understand the model
  - how each predictor is associated with the response
  - for example, with a linear model is very easy to observe the influence of each predictor on the response
  - but for an artificial neural net it is very difficult

Complexity / Flexibility

Complexity / Flexibility

- Consider these potential polynomials $\hat{f}$ to predict $\hat{Y}$

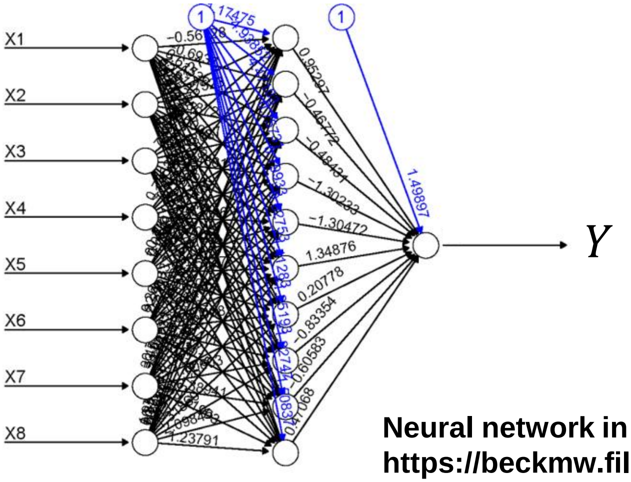
$$Y = \beta_0 + \beta_1 X$$

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4 + \beta_5 X^5 + \beta_6 X^6$$

- The 6th order polynomial is more complicated and more flexible to fit the relationship between feature, X, and response, Y
- Now, what if we use 8 bins on X and 10 nodes in a hidden layer of a neural net?:

Indicator Code X into Bins

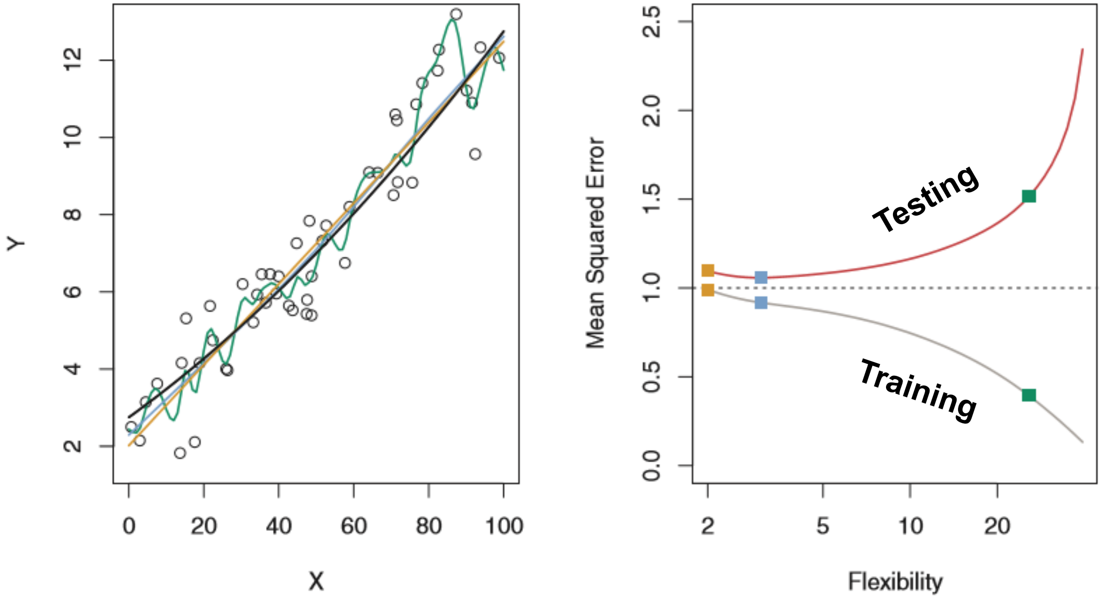
$$\boldsymbol{I(x; x_k)} = \begin{cases} \mathbf{1}, \boldsymbol{if\ x \in\ X_k} \\ \mathbf{0}, \boldsymbol{otherwise} \end{cases}$$



Neural network in R image from:
https://beckmw.files.wordpress.com/2013/11/neuralnet_plot.jpg

Assessing Model Accuracy

- Flexibility vs. Accuracy
  - Increased flexibility will generally decrease MSE on the **training dataset**
  - May result in increase MSE with **testing data**
  - Not generally a good idea to select method only to minimize training MSE



Data and model fits (left) and MSE for training and testing (right) from James et al. (2013).

  - High flexibility + minimize MSE = likely overfit.

Bias and Variance Trade-off

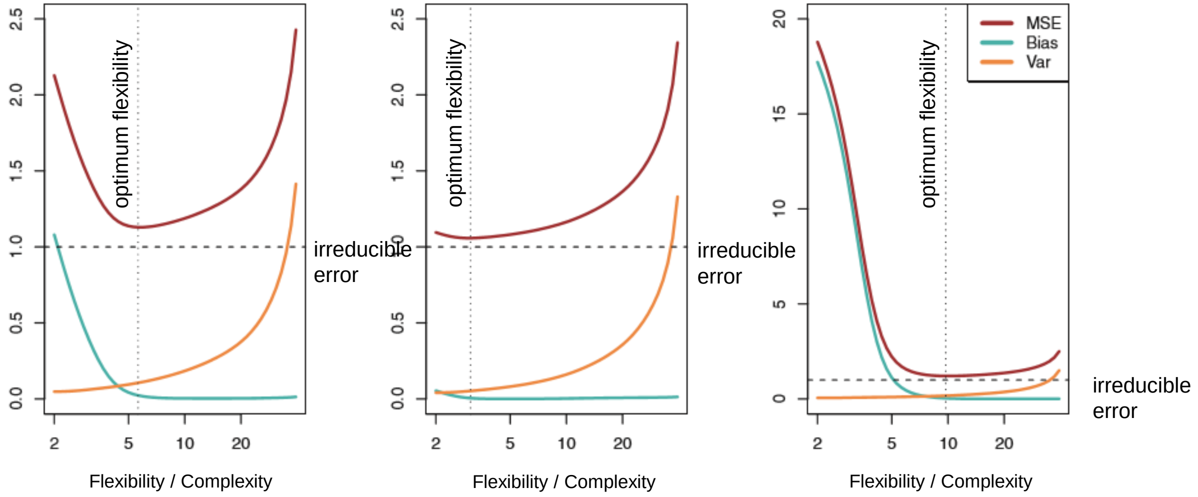
- The **Expected Test Mean Square Error** may be calculated as:

$$\mathrm{E}\left[\left(y_0 - \hat{f}(x_1^0,\dots,x_m^0)\right)^2\right] = \underbrace{Var\left(\hat{f}(x_1^0,\dots,x_m^0)\right)}_{\text{Model Variance}} + \underbrace{\left[Bias\left(\hat{f}(x_1^0,\dots,x_m^0)\right)\right]^2}_{\text{Model Bias}} + \underbrace{Var(\epsilon)}_{\text{Irreducible}}$$

Model Variance is the variance if we had estimated the model with a different training set (simpler models ⇩ lower variance)

Model Bias is error due to using an approximate model (simpler models ⇧ higher bias)

Irreducible error is due to missing variables and limited samples ⇨ can't be fixed with modeling



Model variance, model bias and test MSE for 3 datasets with variable flexibility (Fig 2.12, James et al., 2013), labels added for clarification.

James, G, Witten, D., Hastie, T. and Tibshirani, R., 2013, An Introduction to Statistical Learning with Applications in R, Springer, New York

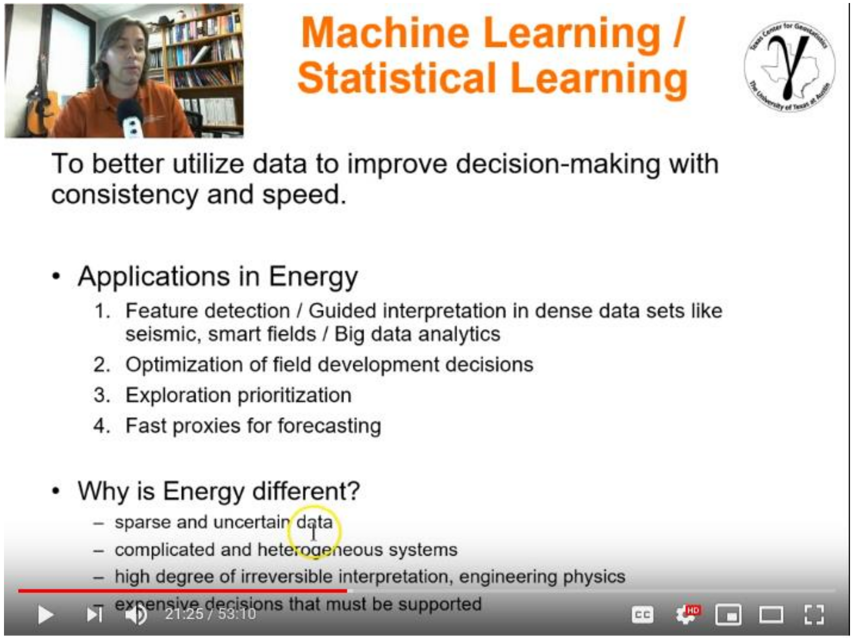
Statistical Learning New Tools

Topic	Application to Subsurface Modeling
Consider inference and prediction	Value of working with inference and prediction. *Permeability was deemed to be redundant with porosity and was removed from the prediction model.*
Consider model training vs. testing	Maximize model prediction accuracy with testing not training. *A reduced complexity model was adopted for predicting porosity from acoustic impedance due improved testing accuracy.*

Multivariate Modeling:
Statistical Learning

Lecture outline . . .

- **Statistical Learning**



Machine Learning / Statistical Learning

To better utilize data to improve decision-making with consistency and speed.

- Applications in Energy
  1. Feature detection / Guided interpretation in dense data sets like seismic, smart fields / Big data analytics
  2. Optimization of field development decisions
  3. Exploration prioritization
  4. Fast proxies for forecasting

- Why is Energy different?
  - sparse and uncertain data
  - complicated and heterogeneous systems
  - high degree of irreversible interpretation, engineering physics
  - expensive decisions that must be supported

https://www.youtube.com/watch?v=9P4S0Wh4cho

Introduction

Prerequisites

Probability

Multivariate Analysis

Spatial Estimation

Statistical Learning

Feature Selection

Multivariate Modeling

Conclusions