# Subsurface Data Analytics and Machine Learning

## Analytics and Machine Learning

Lecture outline . . .

- General Comments
- Data Analytics
- Machine / Statistical Learning
- Prediction and Inference

**Introduction**

*Data Analytics*

*Inferential Methods*

*Predictive Methods*

*Advanced Methods*

Conclusions

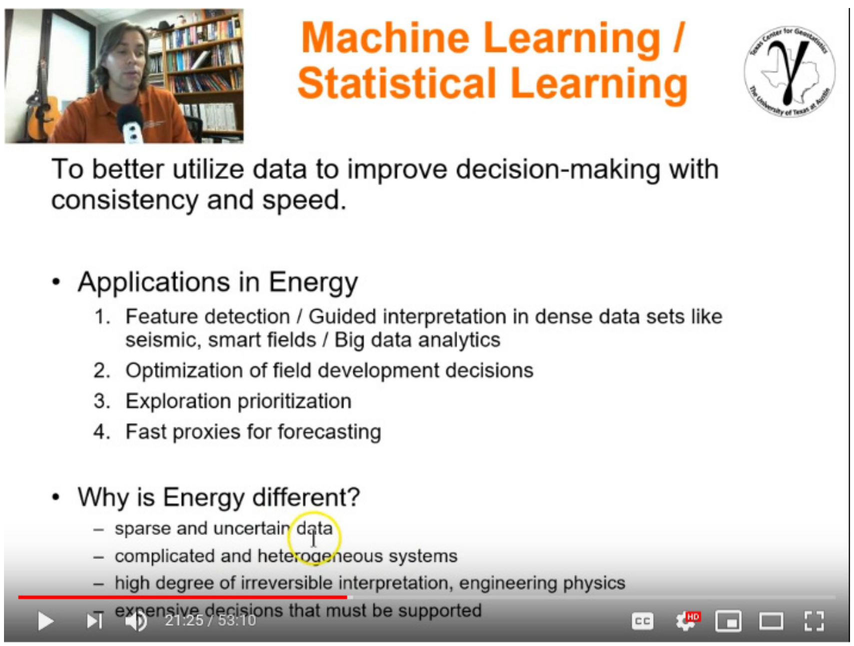
Instructor: Michael Pyrcz, the University of Texas at Austin

# Subsurface Data Analytics and Machine Learning

## Analytics and Machine Learning

### Other Resources:

- **Recorded Lecture Statistical / Machine Learning**



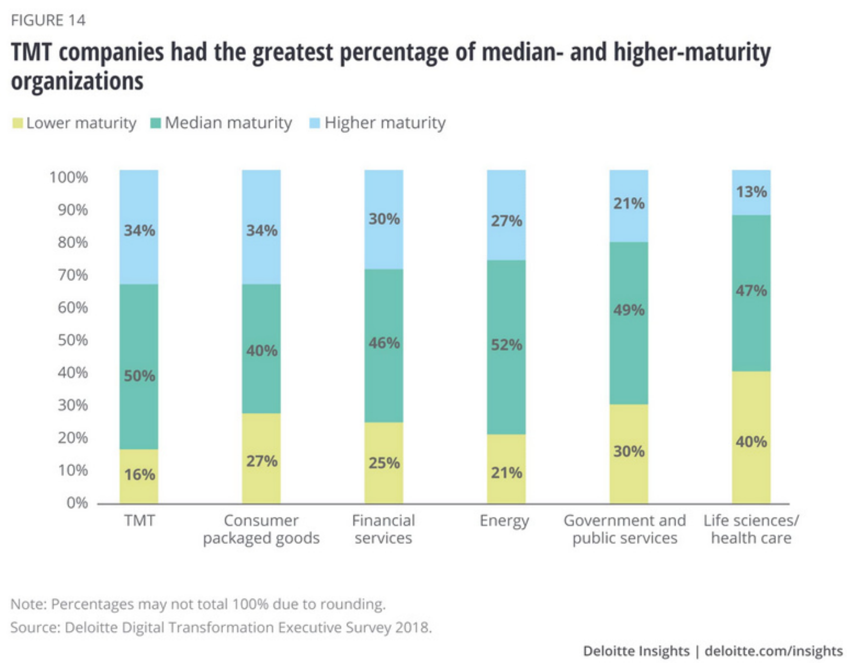

**Instructor: Michael Pyrcz, the University of Texas at Austin**

# Goals of This Lecture

- Motivation
- My biases
- Definition of terms and introduce concepts
- Then we will dive into data analytics, followed by machine learning.

# Digital Transformations

- We are not alone, digital transformations are underway in all sectors of our economy.
- Every energy company that I visit is working on this right now.




**Digital transformation study by Deloitte, 2019.**

https://www2.deloitte.com/insights/us/en/focus/digital-maturity/digital-maturity-pivot-model.html

# Digital Transformations

**My Biases:**

- There are opportunities to do more with our data
- There are opportunities to teach data analytics and statistical / machine learning methods to engineers and geoscientists to improve capability
- Geoscience and engineering knowledge and expertise remains core to our business


Digital transformation PricewaterhouseCoopers (PwC) panel April, 9th, 2019.

# Subsurface Data Analytics and Machine Learning

## Data Analytics and Machine Learning

Lecture outline . . .

- Data Analytics

**Introduction**

*Data Analytics*

*Inferential Methods*

*Predictive Methods*

*Advanced Methods*

Conclusions

# Big Data

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

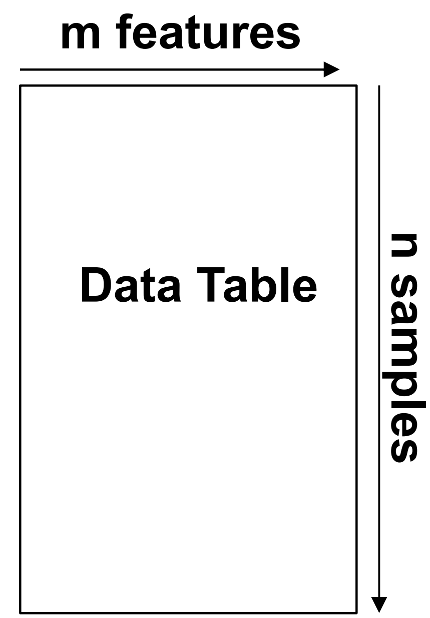
**Volume:** large number of data samples, large memory requirements and difficult to visualize

**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.

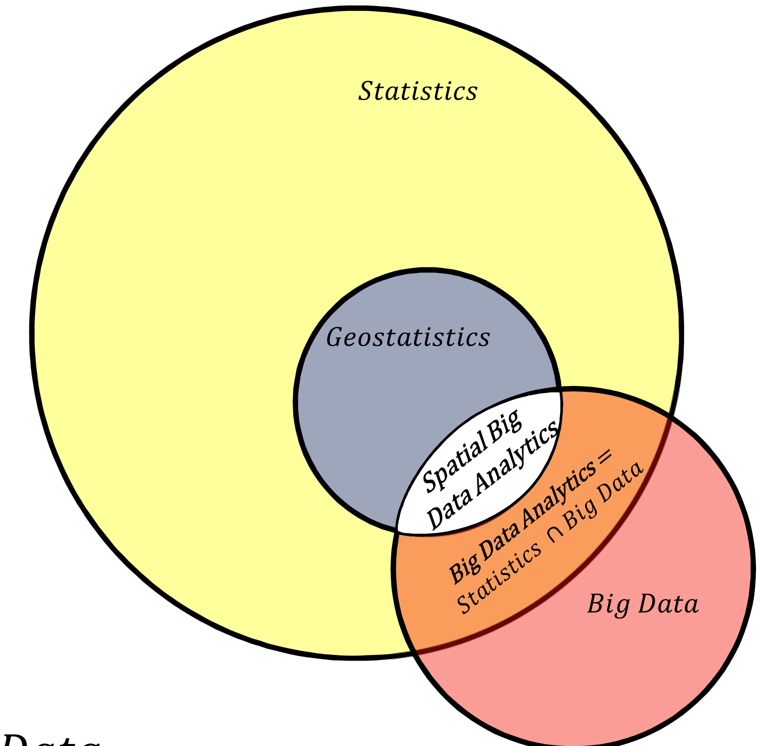
# Big Data Analytics

**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.

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




Proposed Venn diagram for spatial big data analytics.

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

# Subsurface Data Analytics and Machine Learning

## Analytics and Machine Learning

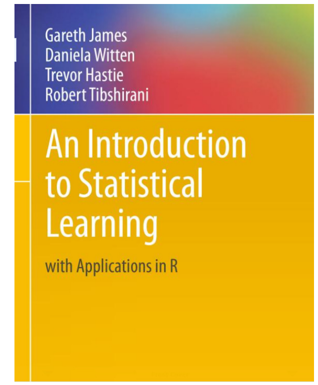
Lecture outline . . .

- Machine / Statistical Learning




Instructor: Michael Pyrcz, the University of Texas at Austin

# Machine Learning / Statistical Learning




- Excellent Reading on this Topic: 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

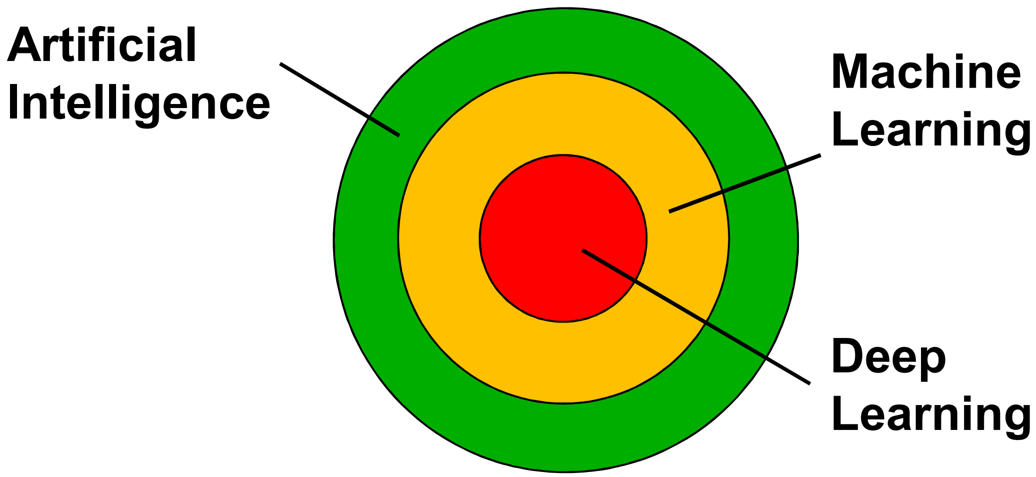
- **Machine Learning vs. Statistical Learning**
  - vast set of tools for learning patterns
  - very little if any prior assumptions

- **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

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

# 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 with complicated neural nets

# 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

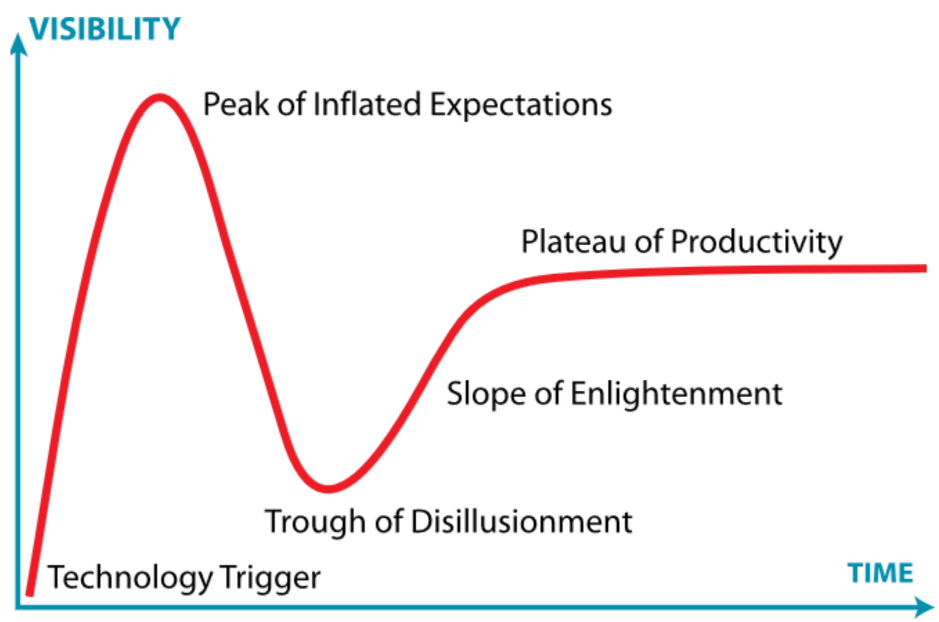
(a) Husky classified as wolf (b) Explanation

**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](https://arxiv.org/pdf/1602.04938.pdf)**

# Machine Learning / Statistical Learning

- Hype Cycle – from information technology firm, Gartner.




Where are we currently for data analytics and machine learning?

Image from https://en.wikipedia.org/wiki/Hype_cycle.

# Machine Learning / Statistical Learning

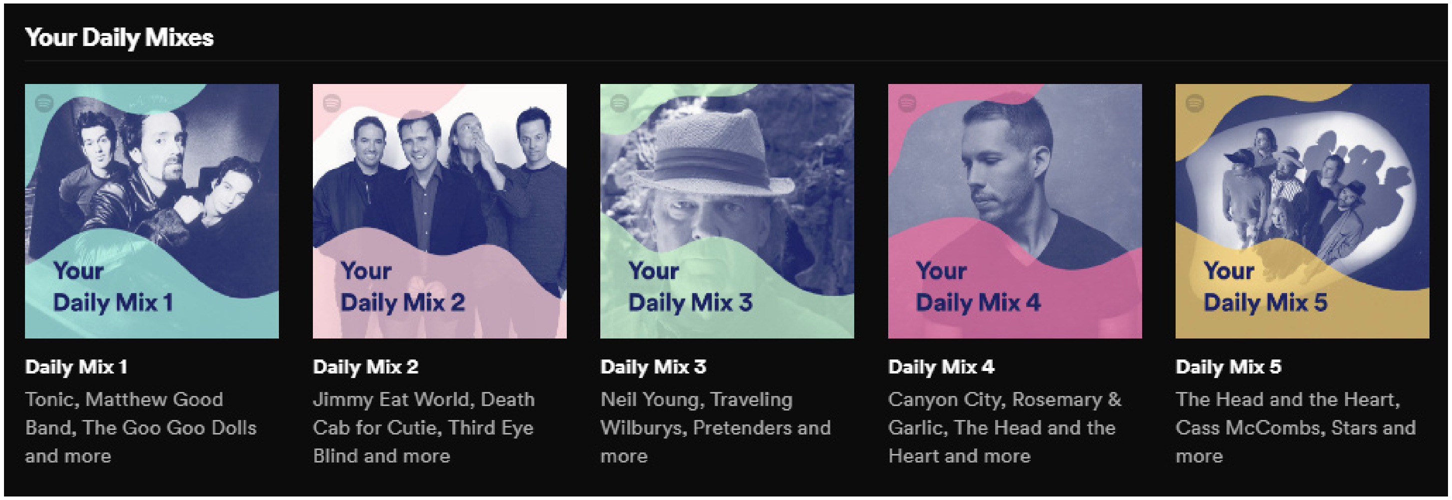
## Applications Around You / Societal Impacts

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

## But Energy is Very Different

- sparse and uncertain data
- complicated and heterogeneous, open earth systems
- high degree of **necessary geoscience and engineering interpretation and physics**
- **Expensive, high value decisions** that must be supported




Spotify recommendation engine, recommender system

# Machine Learning / Statistical Learning

**My recommendations for Machine Learning in Energy:**

**Support Subsurface Development when:**

- volume of data is too large to queried by hand
- the system is high dimensional and cannot be explained with geoscience and engineering
- the task is routine, highly repetitive and low value

**With systems that:**

- Streamline and automate
- Support expert and system interaction
- Interrogatable with excellent visualization and diagnostics

# Machine Learning / Statistical Learning

## Example Machine Learning Applications in Energy

1. Feature detection / guided interpretation in dense data sets like seismic and smart fields
2. Expert systems to detect anomalous operating conditions for safe drilling
3. Optimization of field development decisions with the integration of all relevant geoscience and engineering interpretations and physics
4. Model feedback with fast proxies for geologic and engineering processes to provide guidance for subsurface interpretation and modeling.

*"Significance, consistency, efficiency for more impact."*

# Machine Learning / Statistical Learning

## Data, Metadata and Databases

- 80% of any subsurface study is data preparation and interpretation
- We continue to face a challenge with data:
  - Data curation
  - Large volume
  - Large volumes of metadata
  - Variety of data, scale, collection, interpretation
  - Transmission, controls and security

- Databases are prerequisite to all data analytics and machine learning.

# MetaData Definition

**'a set of data that describes and gives information about other data'** - Google dictionary

**'computing information that is held as a description of stored data'** – dictionary.com

- data collection, calibration, uncertainty, transformations, standardization, interpretation, correction, debiasing
- we have a massive amount of metadata

# Skilled Use

**Just like spatial statistics / geostatistics, statistical learning is a set of tools to add to your tool box as geoscientist or 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

# Skilled Use

**Imagine you are a carpenter (from 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.

# Skilled Use

Hadley Wickham, Chief Scientist at RStudio, known for development of open-source statistical packages for R to make statistics accessible and fun (http://hadley.nz/).

Read Hadley Wickham's, **Teaching Safe-Stats, Not Statistical Abstinence** (https://nhorton.people.amherst.edu/mererenovation/17_Wickham.PDF)

- **Teaching:** We need to rethink statistics curriculum – we risk becoming irrelevant!
- **Practice:** Stats tends to be taught as avoid, unless you are an "statistician" or with one
  - Otherwise you will cause great harm
  - But there are not enough professional statisticians
  - Rather than stigmatize amateur, new tools should be safer to use
- **Tools**: New tools should be easy and fun to use to encourage use
  - Flexible grammars, minimal set of independent components to build workflows

**Hadley Wickham photograph from https://en.wikipedia.org/wiki/Hadley_Wickham**

# Subsurface Data Analytics and Machine Learning

## Analytics and Machine Learning

**Lecture outline . . .**

- **Prediction and Inference**



**Instructor: Michael Pyrcz, the University of Texas at Austin**

# The Model

**Predictors, Independent Variables, Features**

- input variables
- for a model $Y = f(X_1, \ldots, X_m) + \epsilon$ , these are the $X_1, \ldots, X_m$
- note $\epsilon$ is a random error term

**Response, Dependent Variables**

- output variable
- for a model $Y = f(X_1, \ldots, X_m)$, this is $Y$

**Statistical / Machine Learning is All About**

- Estimating $f$ for two purposes
    1. Prediction
    2. Inference

# Inference

**There is value in understanding the relationships between predictor features**

- for $Y = f(X_1, \dots, X_m) + \epsilon$ we can understand the influence / interactions of each $X_\alpha$ on each other.

**What is the relationship between each predictor feature?**

- sense of the relationship (positive or negative)?
- shape of relationship (sweet spot)?
- relationships may depend on values of other predictors!

*'Inference is learning about the system.'*

# 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 $Y$

*'Prediction is modeling the system to make estimates, forecasts.'*

# Estimating $f$

## Parametric Methods

- make an assumption about the functional form, shape
- we gain simplicity and advantage of only a few parameters
- 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!

# Estimating $f$

## Nonparametric Methods

- make no assumption about the functional form, shape
- more flexibility to fit a variety of shapes for $f$
- less risk that $\hat{f}$ is a poor fit for $f$

- typically need a lot more data for an accurate estimate of $f$

*'Nonparametric is actually parametric rich!'*

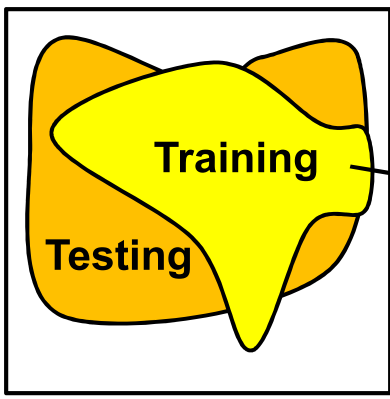
# Training and Testing

## The Training and Testing Workflow

**Train the Model**

Train the Model Parameters Maximize Accuracy with Training Data

**Separate The Data**




**Tune the Model**

Test the Model Tune the Hyperparameters to Optimize the Complexity

We avoid the overfit problem.

# Model Parameters Definition

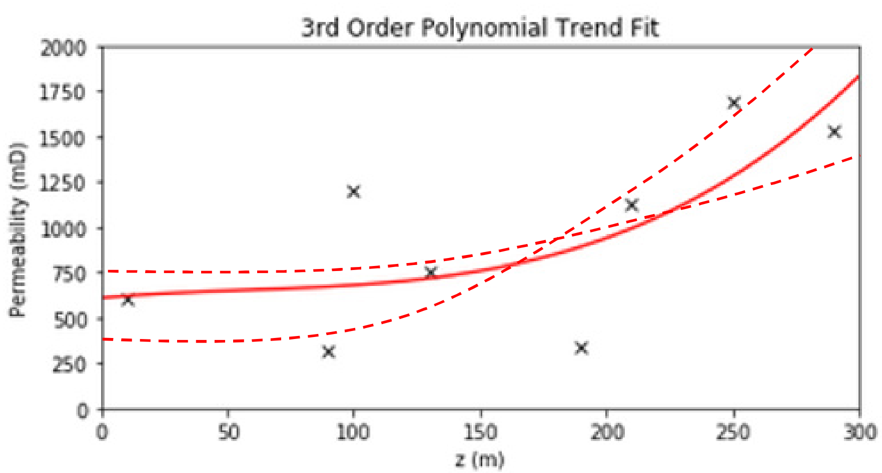
## Model Parameters

Derived during training phase to fit the model to the training data

**Parameters**

$$k = b_3 z^3 + b_2 z^2 + b_1 z + c$$

$b_3$, $b_2$, $b_1$ and $c$



# Model Hyperparameters Definition

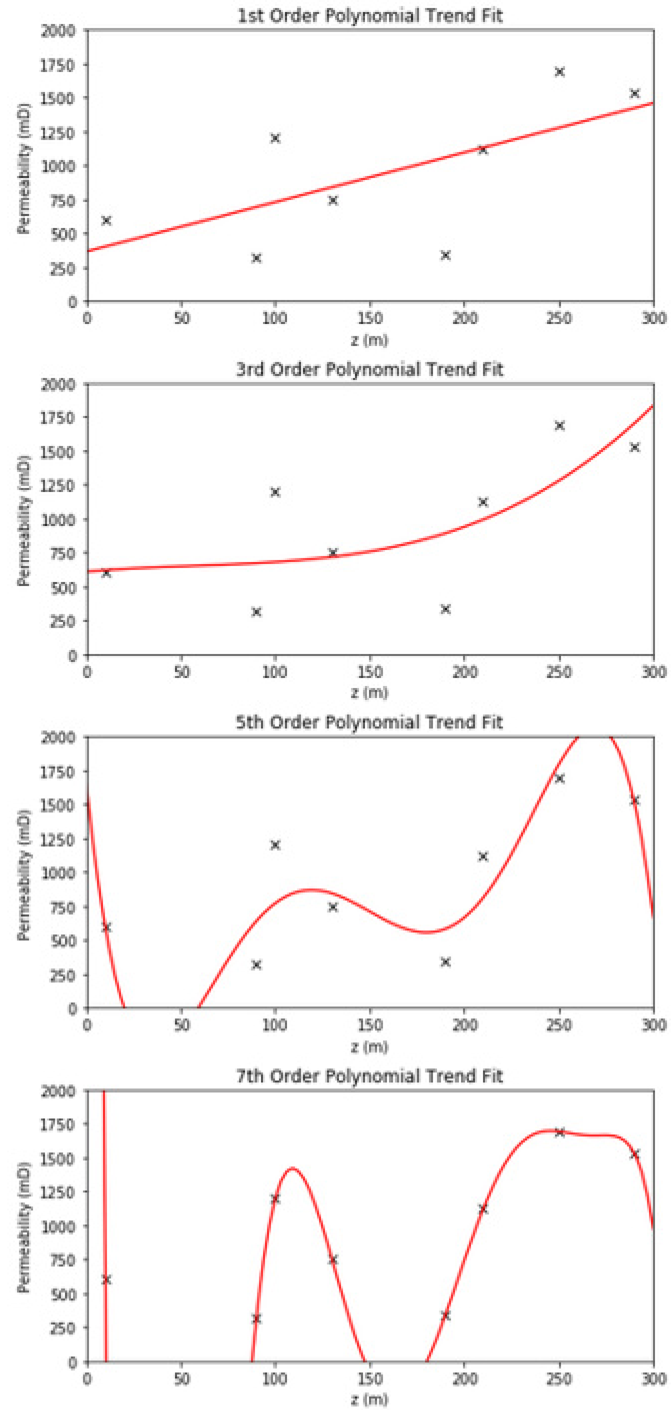
## Model Hyperparameters

Set prior to learning from the data. Impact the form of the model and often the complexity.

**3rd Order:** $$k = b_3 z^3 + b_2 z^2 + b_1 z + c$$

**2nd Order:** $$k = b_2 z^2 + b_1 z + c$$

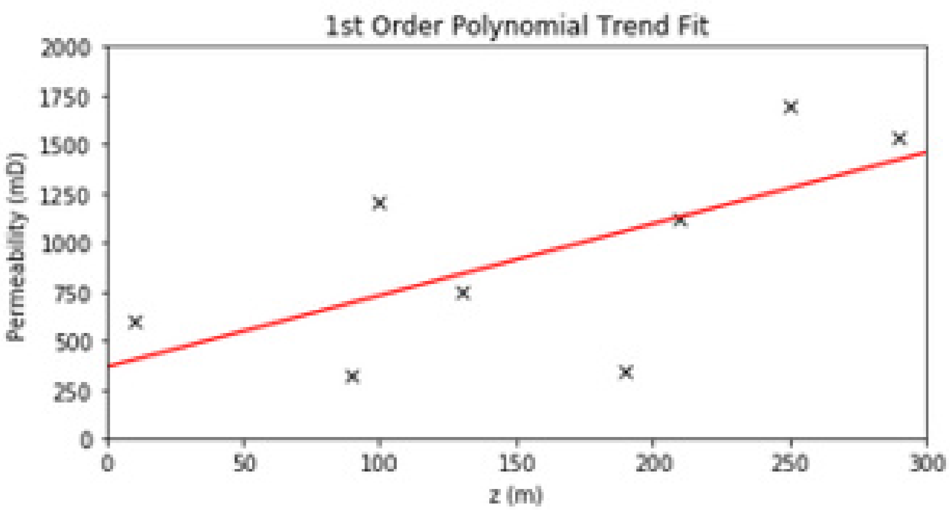
**1st Order:** $$k = b_1 z + c$$



# Prediction Accuracy vs. Model Interpretability / Explainability

## Interpretability / Explain-ability

- 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 $\hat{Y}$

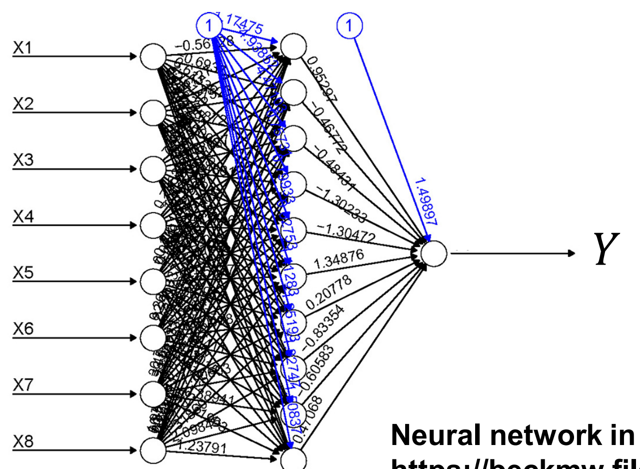
$$Y = \beta_0 + \beta_1 X$$

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

- The 6$^{th}$ 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**

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



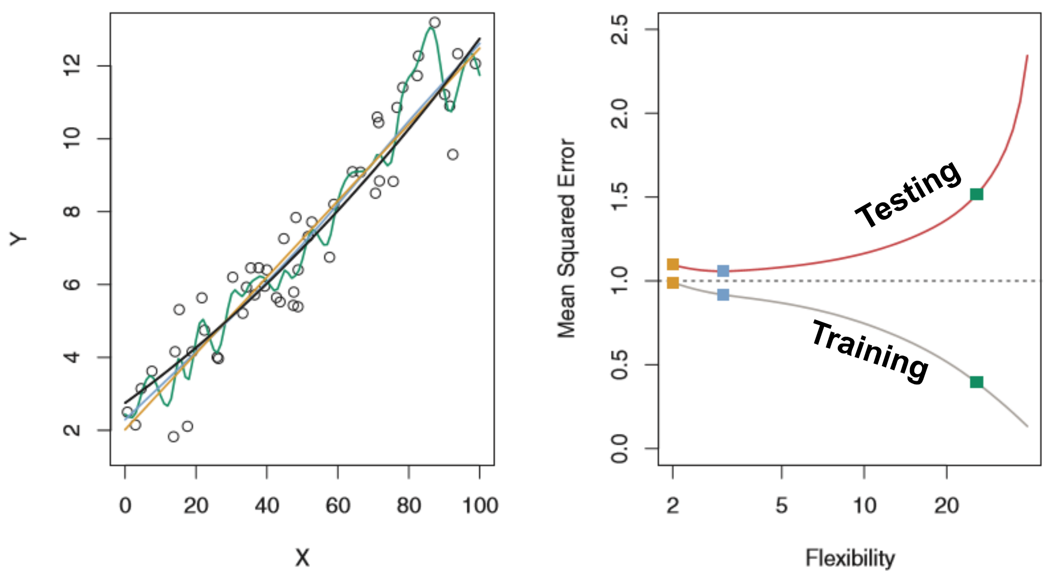

We will discuss neural nets later.

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

# Assessing Model Accuracy

## Flexibility / Complexity 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).

# Bias and Variance Trade-off

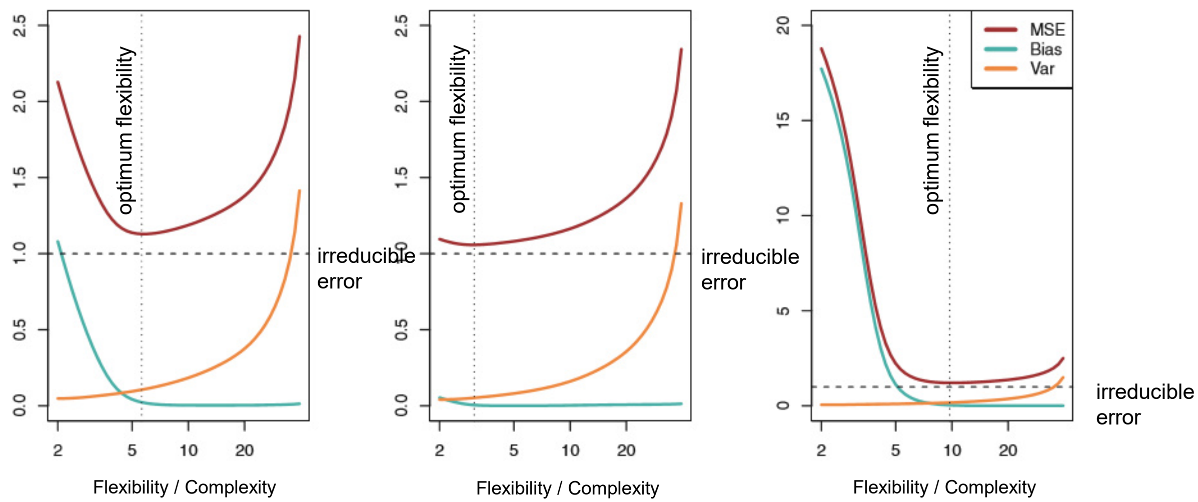
- The **Expected Test Mean Square Error** may be calculated as:

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

**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.

# Statistical Learning
# New Tools

| Topic | Application to Subsurface Modeling |
| --- | --- |
| Data Analytics is the use of statistics, geoscience and engineering with data. | Learn applied statistics and workflows to support your work with data.<br><br>*Growing new competencies to augment geoscience and engineering expertise is a great solution, consider open source packages in Python.* |
| Parametric and Nonparametric | Parametric models need less data to train but may have model bias, nonparametric models often are parametric rich and may be overfit.<br><br>*Be aware of the performance of your selected modeling methods.* |
| Model bias, Model variance and Irreducible Error | There is an error trade-off for accuracy with testing data.<br><br>*Low complexity models may outperform high complexity models.* |

# Subsurface Data Analytics and Machine Learning

## Analytics and Machine Learning

Lecture outline . . .

- General Comments
- Data Analytics
- Machine / Statistical Learning
- Prediction and Inference

**Introduction**

*Data Analytics*

*Inferential Methods*

*Predictive Methods*

*Advanced Methods*

Conclusions

Instructor: Michael Pyrcz, the University of Texas at Austin