GEOSTATISTICS AND MACHINE LEARNING

Introductory Lecture Outline

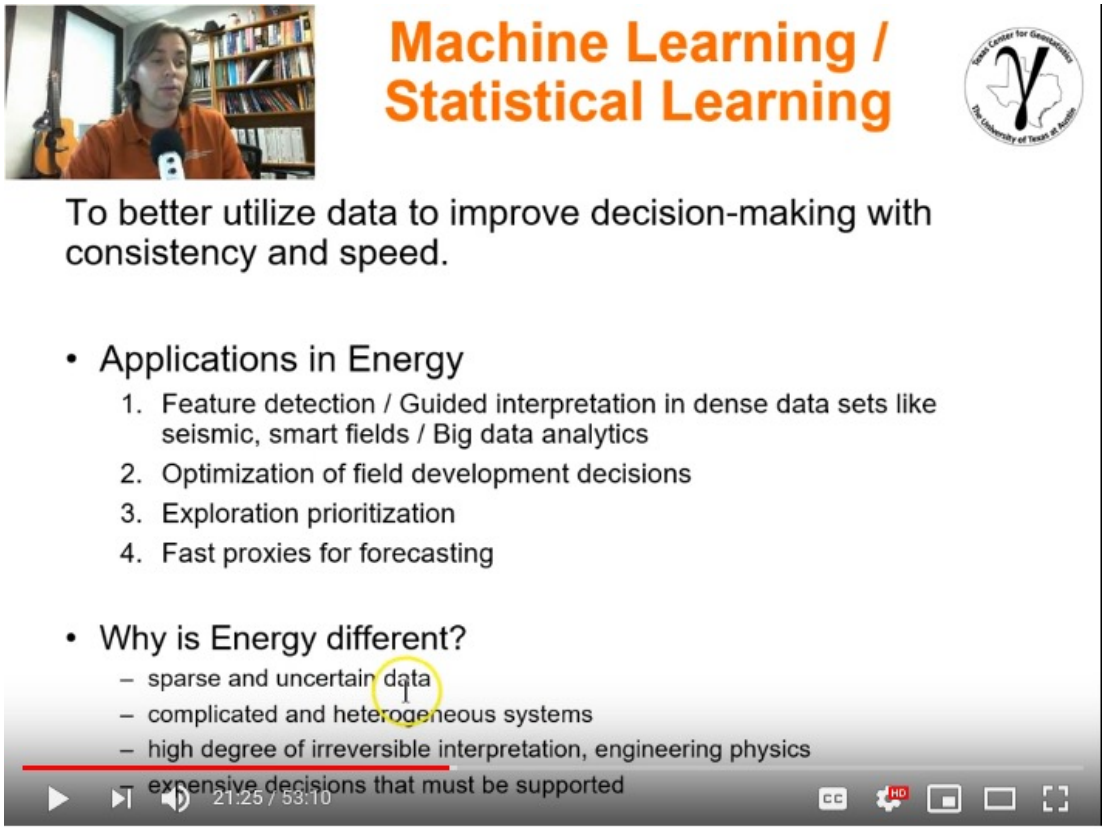
- General Comments
- Data Analytics / Geostatistics
- Machine / Statistical Learning
- Prediction and Inference

GEOSTATISTICS AND MACHINE LEARNING

Introduction

Other Resources:

- Recorded Lecture Statistical / Machine Learning

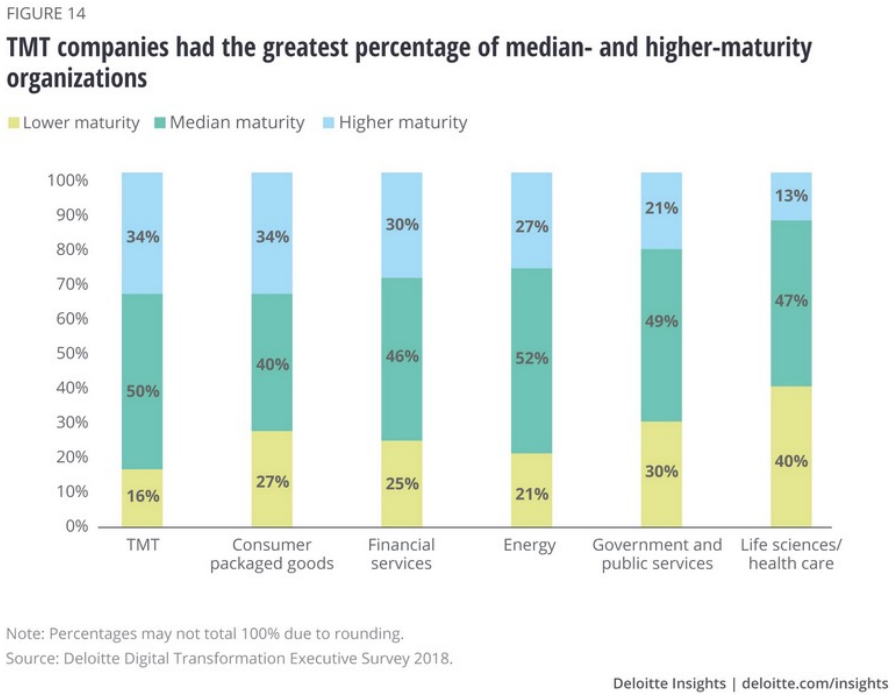


GOALS OF THIS LECTURE

- 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.

Source: 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 & expertise remains core to our business

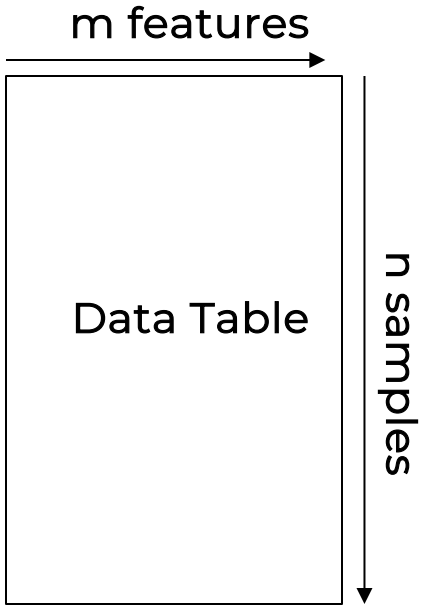
Digital transformation PricewaterhouseCoopers (PwC) panel April, 9th, 2019

DATA ANALYTICS

BIG DATA

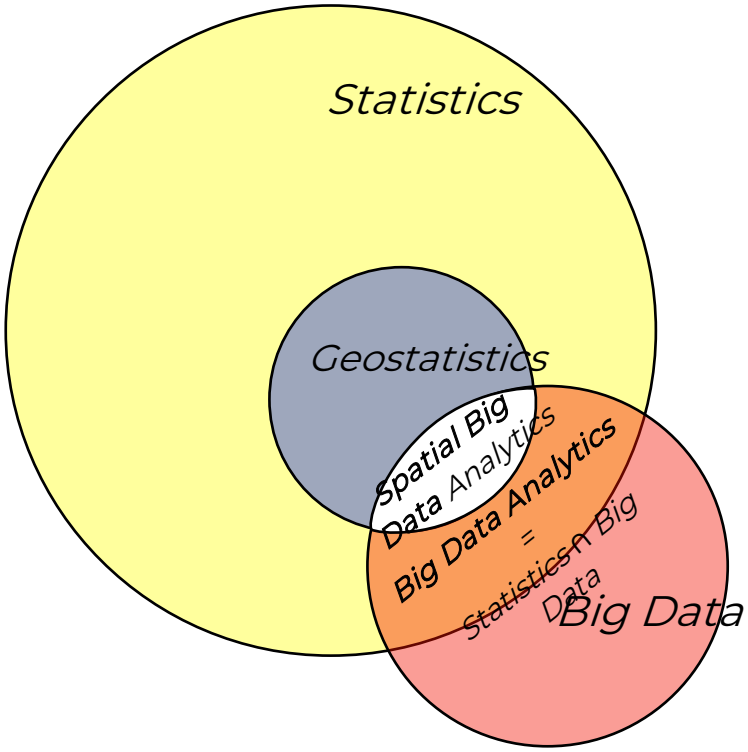
- **Big Data**: you have big data if your data has a combination of these:
- **Volume**: large number of data samples, large memory requirements and difficult to visualize
- **Velocity**: data is gathered at a high rate, continuously relative to decision making cycles
- **Variety**: data form various 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.
- **Spatial Big Data Analytics** = $Geostatistics \cap Big\ Data$
- Big data analytics is expert use of (geo)statistics on big data.

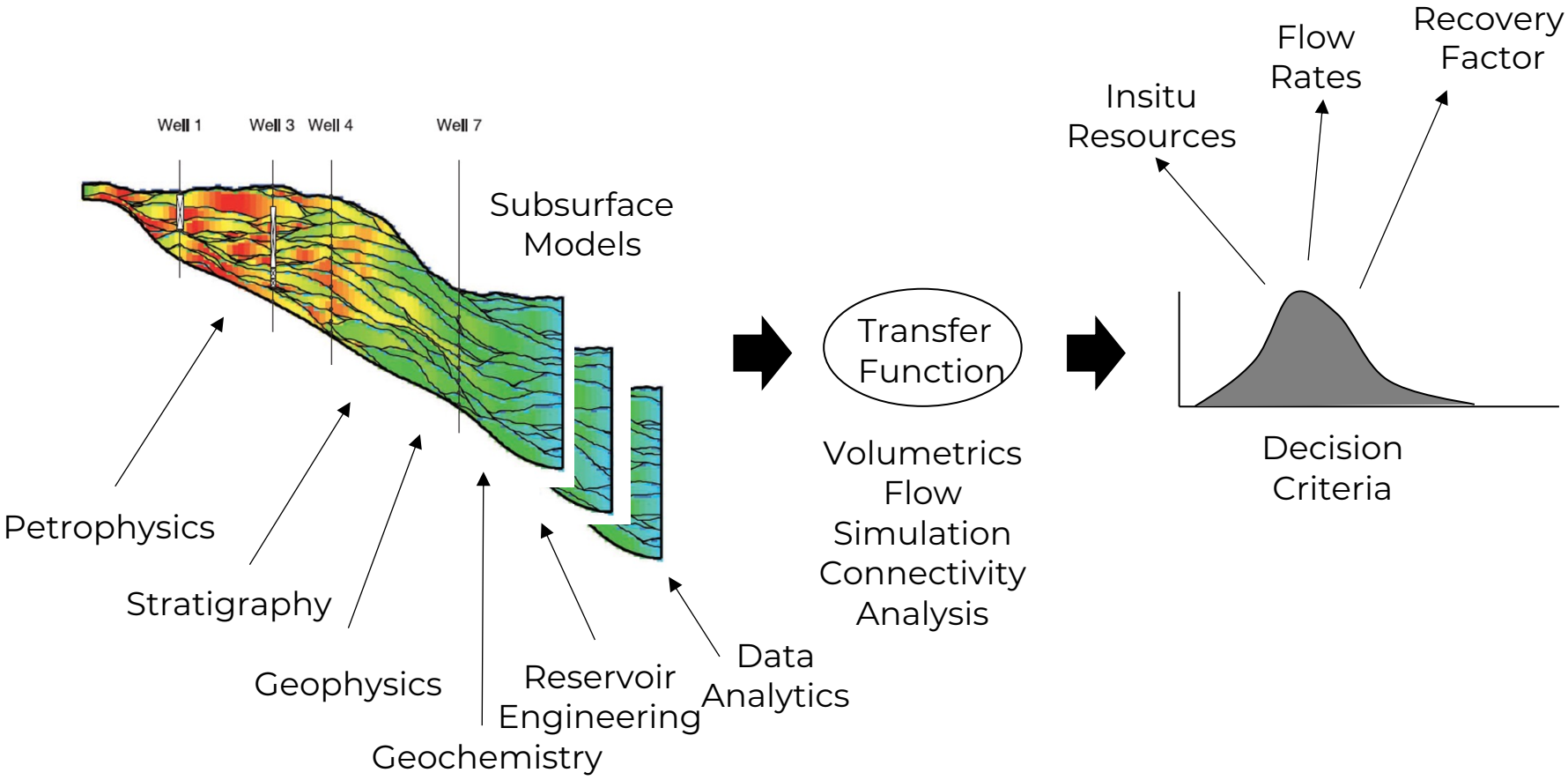


Proposed Venn diagram for spatial big data analytics.

GEOSTATISTICS

SUBSURFACE MODELS

- **Reservoir / Subsurface Modeling** is the integration of all subsurface information to build a suite of models representing uncertainty to support decision making



SUBSURFACE MODELS

- **Geostatistics** developed from practice of subsurface estimation and modeling in mining, theory added later
- Statistical, quantitative descriptions of concepts from geology!
- Geostatistics is the practical quantification of the subsurface to support decision making

Concept	Geological Expression	Geostatistical Expression
Major changes in relationships between reservoir bodies	Architectural complexes and complex sets	Regions—separate units and model with unique methods and input statistics
Changes in reservoir properties within reservoir bodies	Basinward and landward stepping Fining/Coarsening up	Nonstationary mean
Stacking patterns ~~if~~ reservoir bodies	Organization, disorganization, compartmentalization, compensation	Attraction, repulsion, minimum and maximum spacing distributions, interaction rules
Major direction of continuity	Paleo-flow direction	Major direction of continuity, locally variable azimuth model
Relationship between vertical and horizontal continuity	Walther's Law	Geometric and zonal anisotropy
Distinct reservoir property groups	Lithofacies, depositional facies, and architectural elements	Reservoir categories, stationary regions
Heterogeneity	Architecture	Spatial continuity model geometric parameters, training image patterns

SPATIAL STATISTICAL INFERENCE

- Any statistic requires replicates, repeated sampling (e.g. air or water samples from a monitoring station). In our geospatial problems repeated samples are not available at a location in the subsurface.

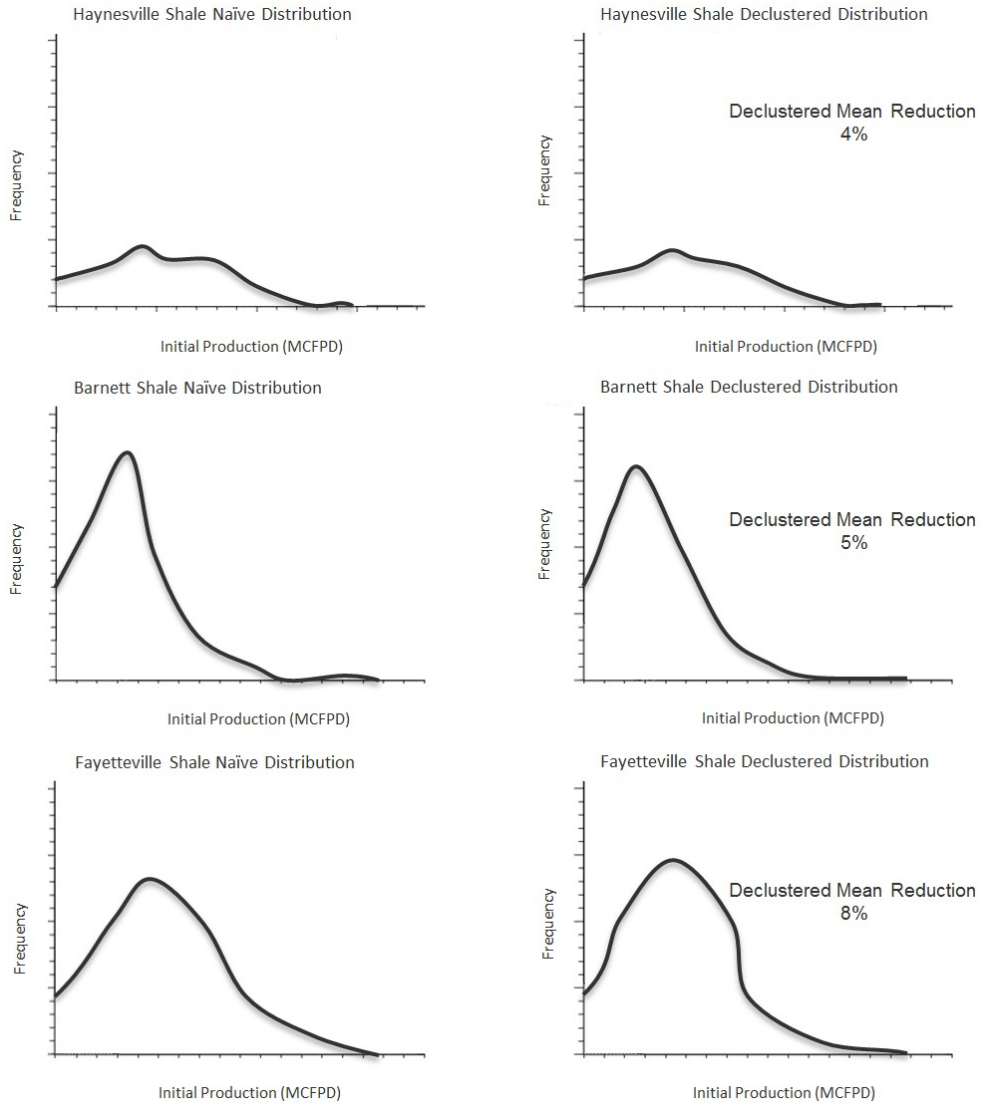
Extracted Sample at Location, $\mathbf{u}_1$

Hole Remains at Location, $\mathbf{u}_1$

- Instead of time, we must pool samples over space to calculate our statistics. This decision to pool is the decision of stationarity. It is the decision that the subset of the subsurface is all the "same stuff".

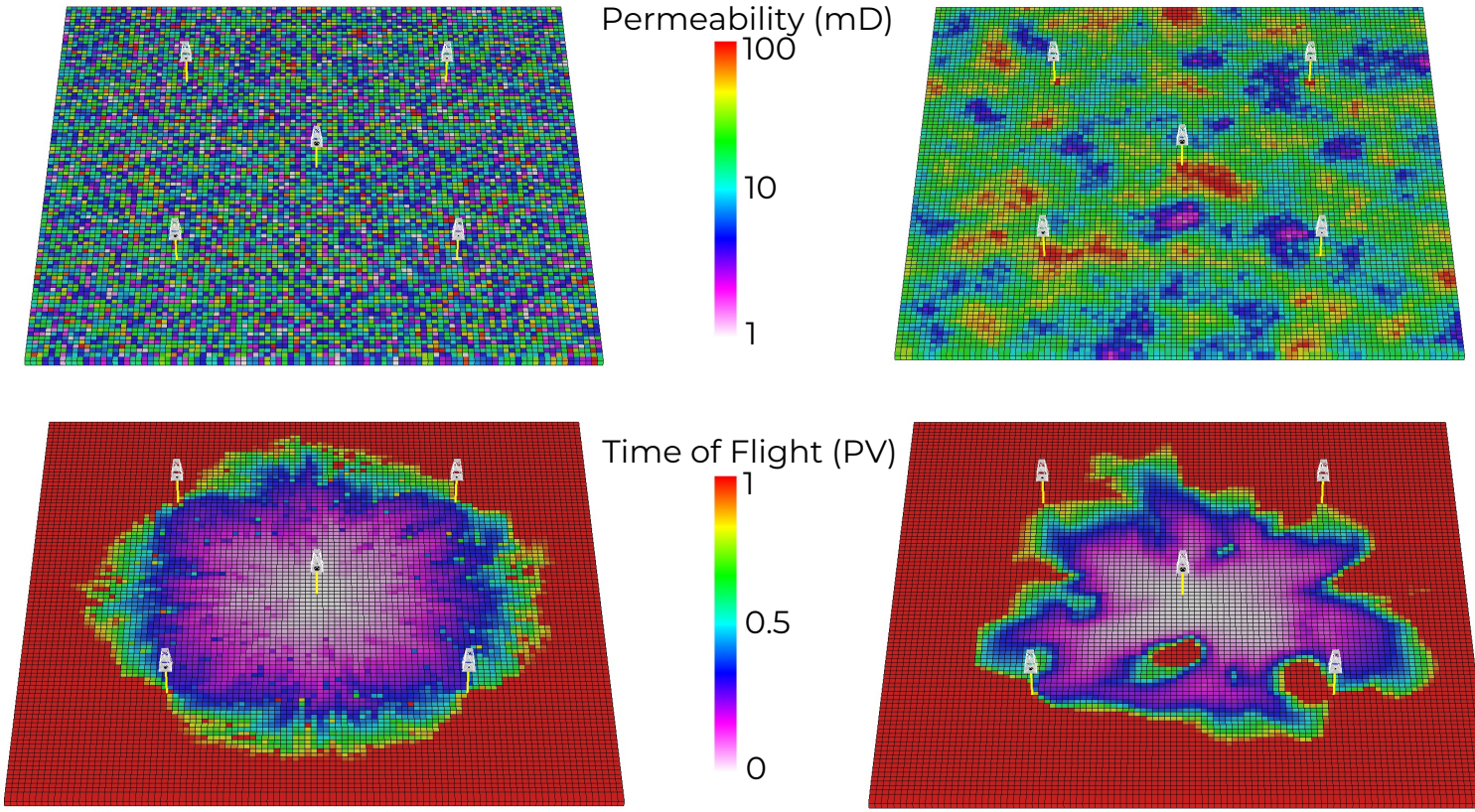
BIAS IN SUBSURFACE SAMPLING

- Virtually all subsurface datasets are biased
- Data is collected to:
  - Answer questions
  - Resolve risk
  - Maximize value
- These practices should not change.
- We must mitigate subsurface data bias before we build models



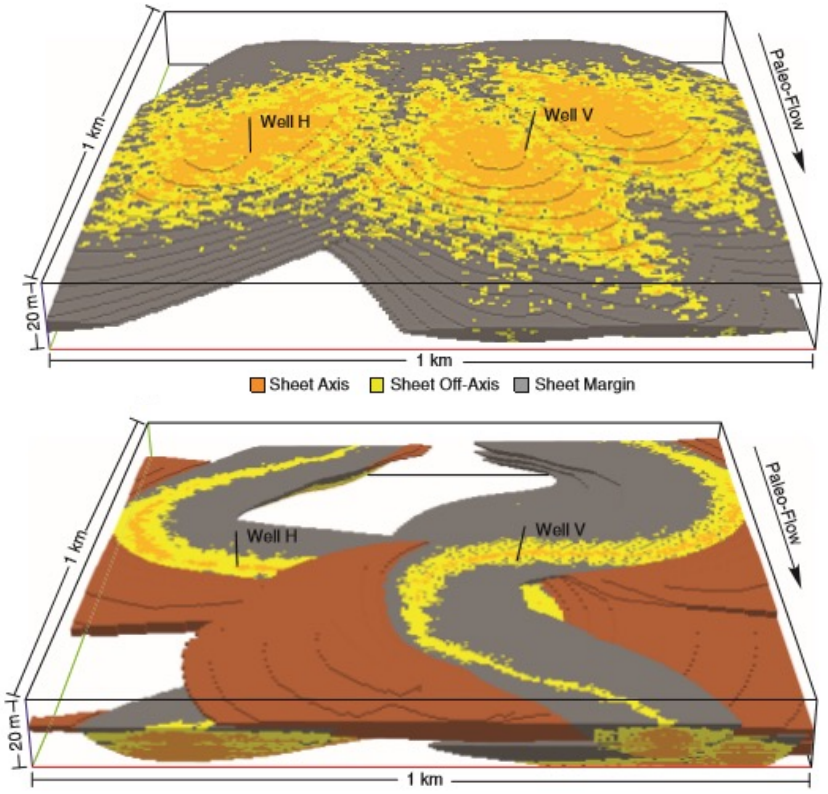
SPATIAL CONTEXT

- Spatial continuity varies significantly and impacts our analysis
- We must quantify & impose spatial continuity in our subsurface models



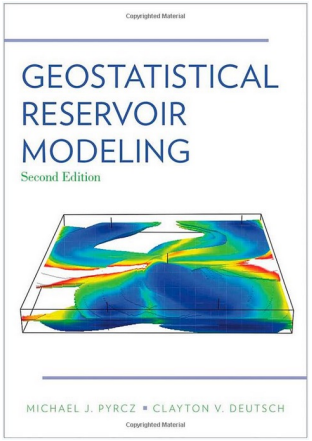
SUBSURFACE UNCERTAINTY

- Sources of uncertainty include:
  - Data measurement, calibration uncertainty
  - Decisions and parameters uncertainty
  - Spatial uncertainty in estimating away from data
- Uncertainty is due to our ignorance
- There is no objective uncertainty, it is a model
- Uncertainty in the uncertainty, don't go there!
- Ignoring uncertainty is assuming certainty

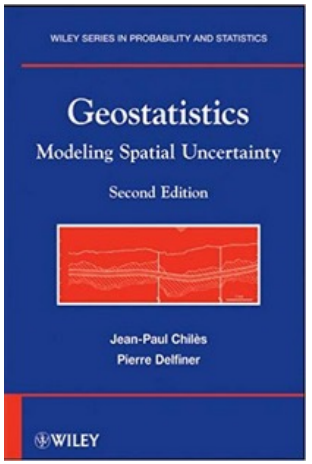


MACHINE LEARNING / STATISTICAL LEARNING

- Accessible treatment of subsurface data analytics and geostatistics
  Geostatistical Reservoir Modeling, 2014, Pyrcz, M.J., and Deutsch, C.V., Oxford University Press.



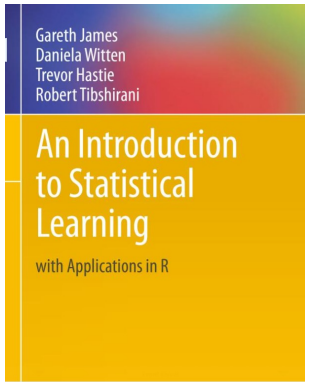
- A more theoretical, less accessible treatment *Geostatistics Modeling Spatial Uncertainty, 2012, Chilès and Delfiner, Wiley*



MACHINE LEARNING

MACHINE LEARNING / STATISTICAL LEARNING

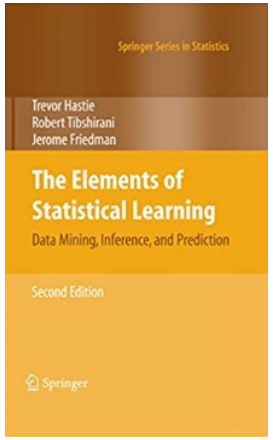
- Excellent Reading on this Topic - *An Introduction to Statistical Learning with Applications in R, 2013, James et al., Springer.*
  - (http://www-bcf.usc.edu/~gareth/ISL/ISLR%20First%20Printing.pdf)



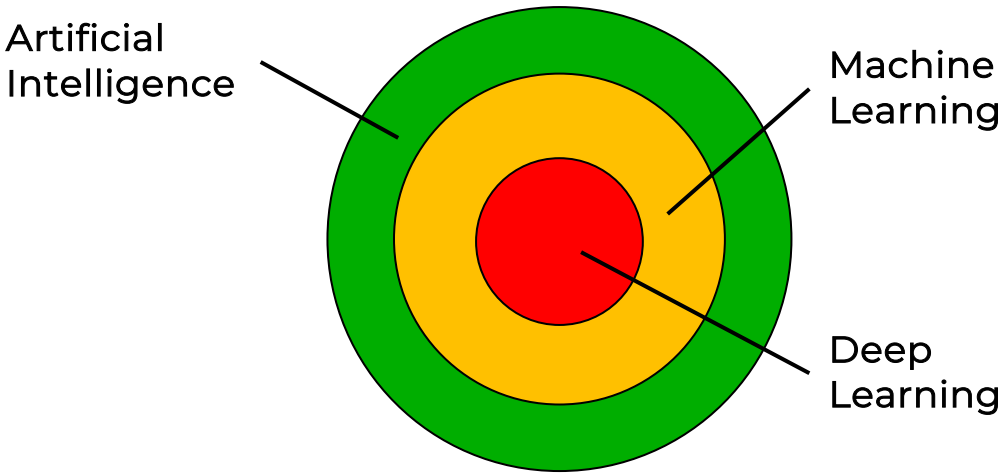
Statistical Learning

- I use the term statistical learning interchangeably with machine learning
- Remind us that these methods are statistical models, useful summarizations based on data to support decision making

- A more theoretical, less accessible, treatment *Hastie et al., 2009, The Elements of Statistical Learning: Data Mining, Inference and Prediction*



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

learning

- “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.”

training with data

general

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

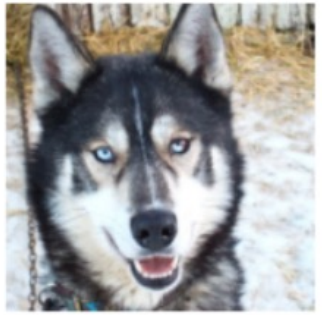
not a panacea

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



(a) Husky classified as wolf

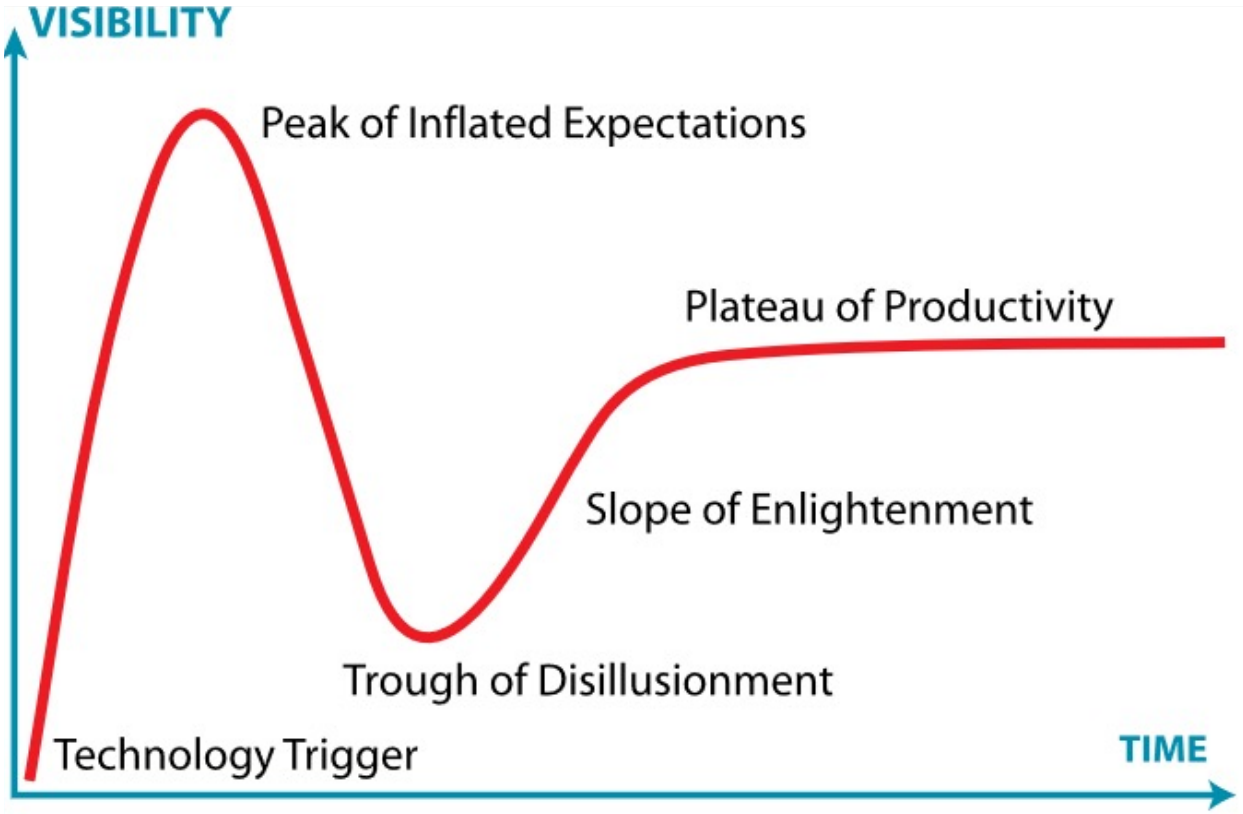
(b) Explanation

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

MACHINE LEARNING / STATISTICAL LEARNING

- Hype Cycle – from information technology firm, Gartner



Where are we currently for data analytics and machine learning?

Source: https://en.wikipedia.org/wiki/Hype_cycle

MACHINE LEARNING / STATISTICAL LEARNING

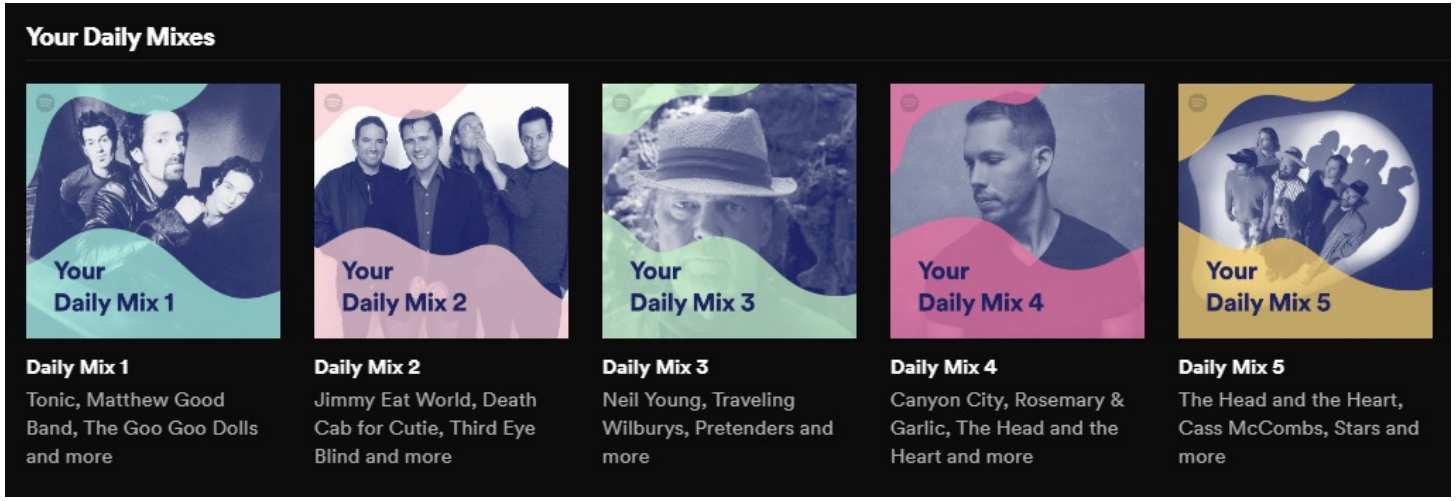
Applications Around You / Societal Impacts

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 for other 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

PREDICTION AND INFERENCE

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 ϵ 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 $Y = f(X_1, \ldots, X_m) + \epsilon$ we can understand the influence / interactions of each X_α 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 Xs!
- 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

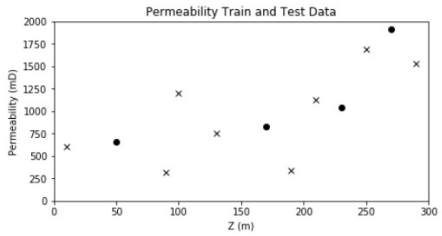
- Make no assumption about the functional form, shape
- More flexibility to fit a variety of shapes for f
- Less risk that $\hat{f}$ is a poor fit for f
- Typically need a lot more data for an accurate estimate of f

'Nonparametric is actually parametric rich!'

TRAINING AND TESTING

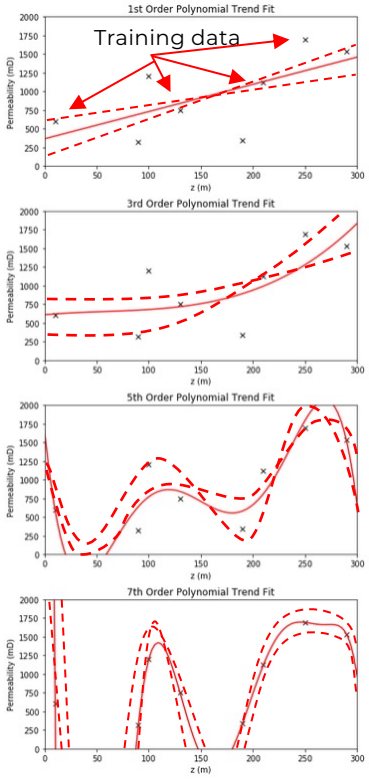
The Training and Testing Workflow

1. Split the Data into Train and Test Subsets



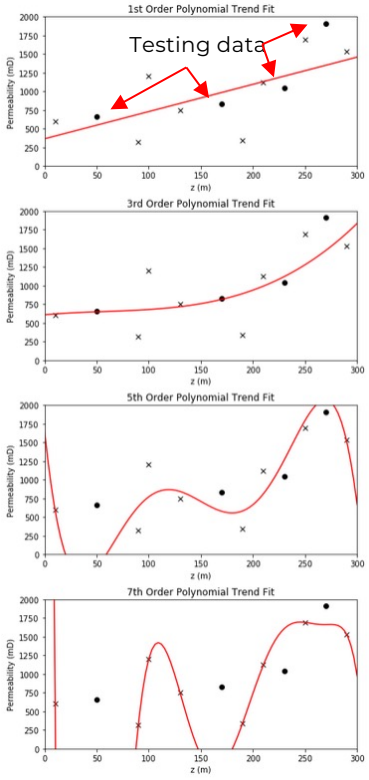
2. Build Models Over a Range of Hyperparameters Increasing Model Complexity

3. Train the Model Parameters with Training Data



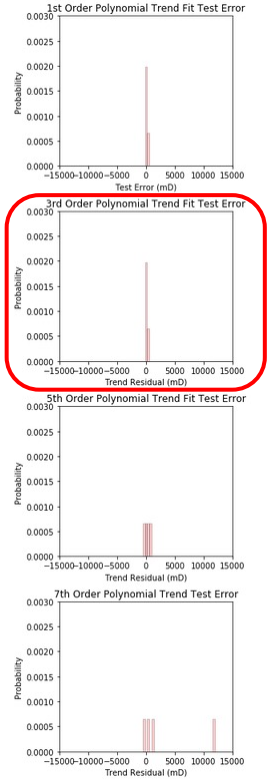
Best Fit Model to Training Data

4. Check Each Best Fit Model with Withheld Testing Data



Error Over the Testing Data

5. Tune the Model Hyperparameters with Testing Data



Machine learning model building workflow to avoid overfit.

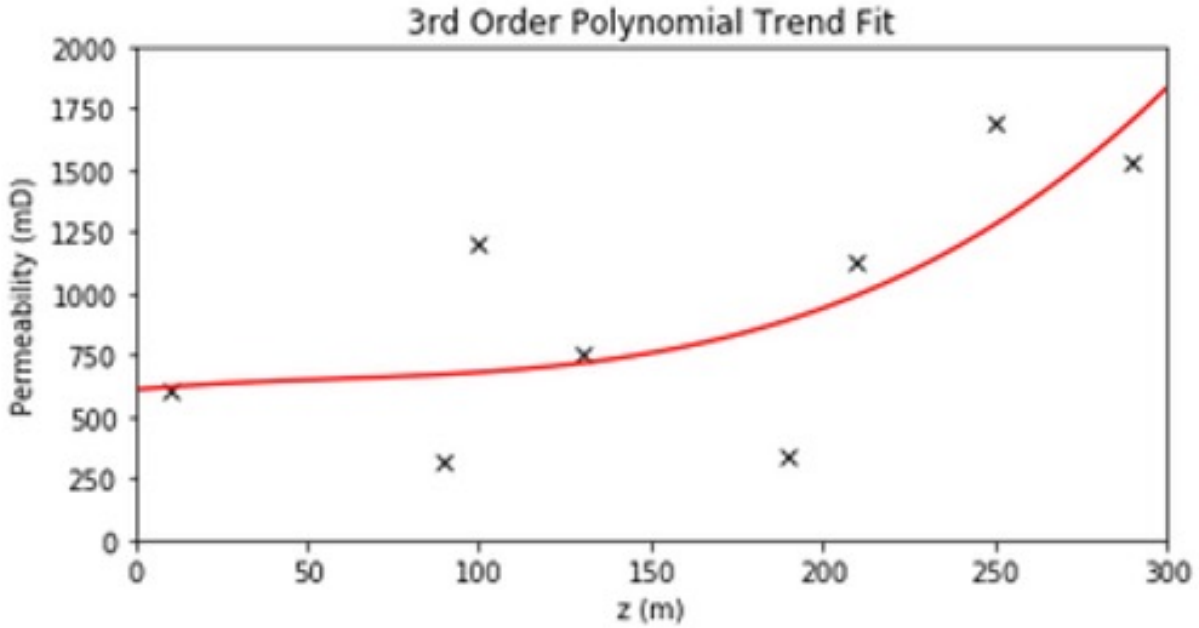
MODEL PARAMETERS

Definition

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

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

Parameters
b_3, b_2, b_1 and c



MODEL HYPERPARAMETERS

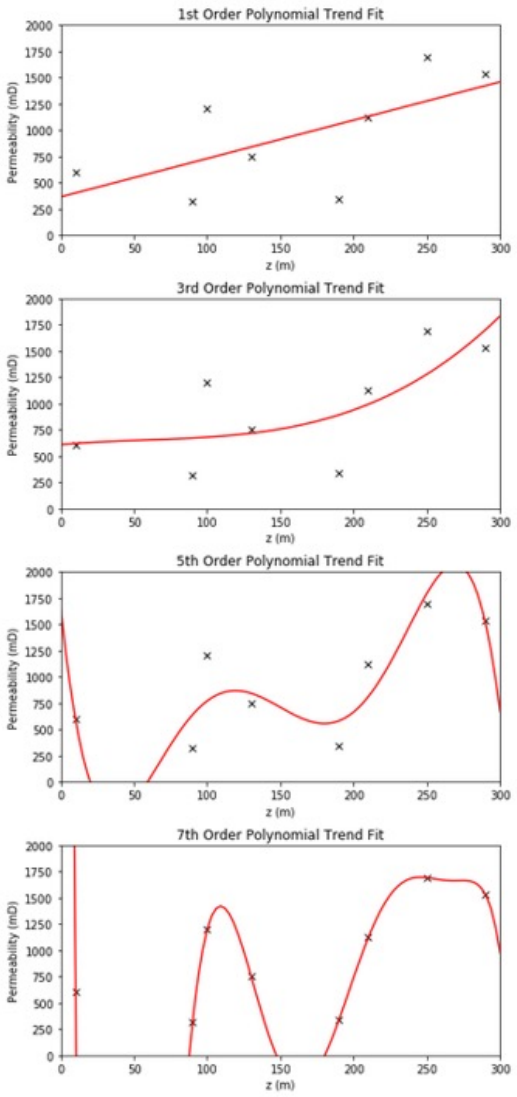
Definition

- 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: $k = b_2 z^2 + b_1 z + c$

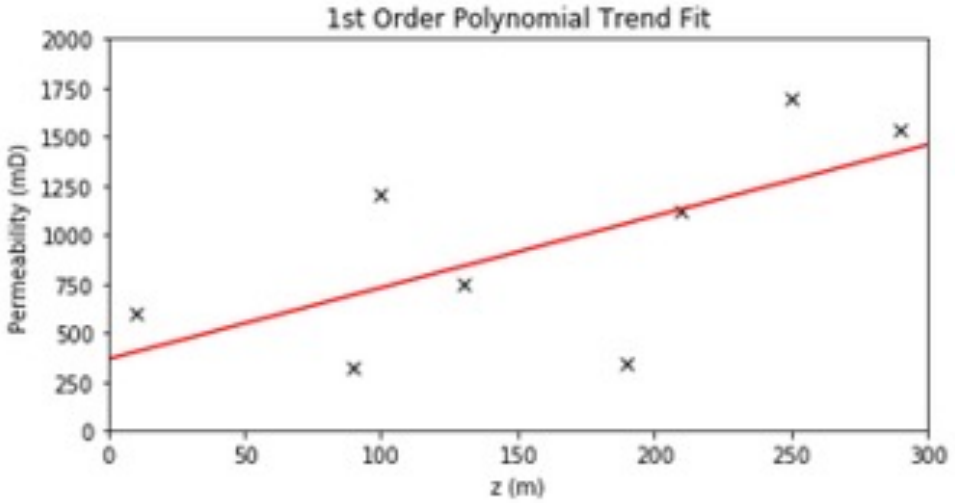
1st Order: $k = b_1 z + c$



PREDICTION ACCURACY VS. MODEL

Interpretability / 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

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

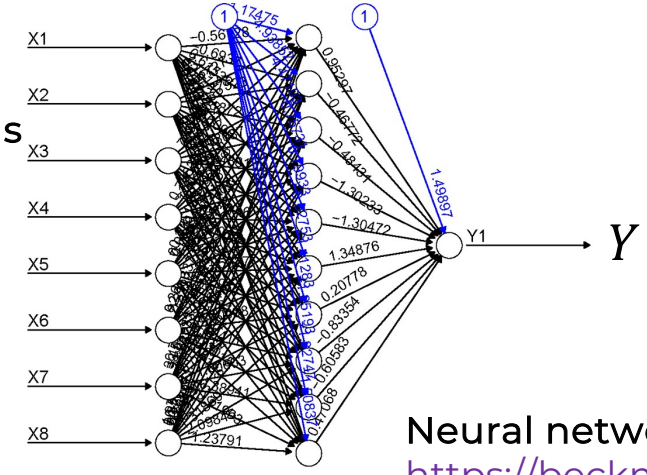
$$Y = \beta_0 + \beta_1 X$$

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

- 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

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



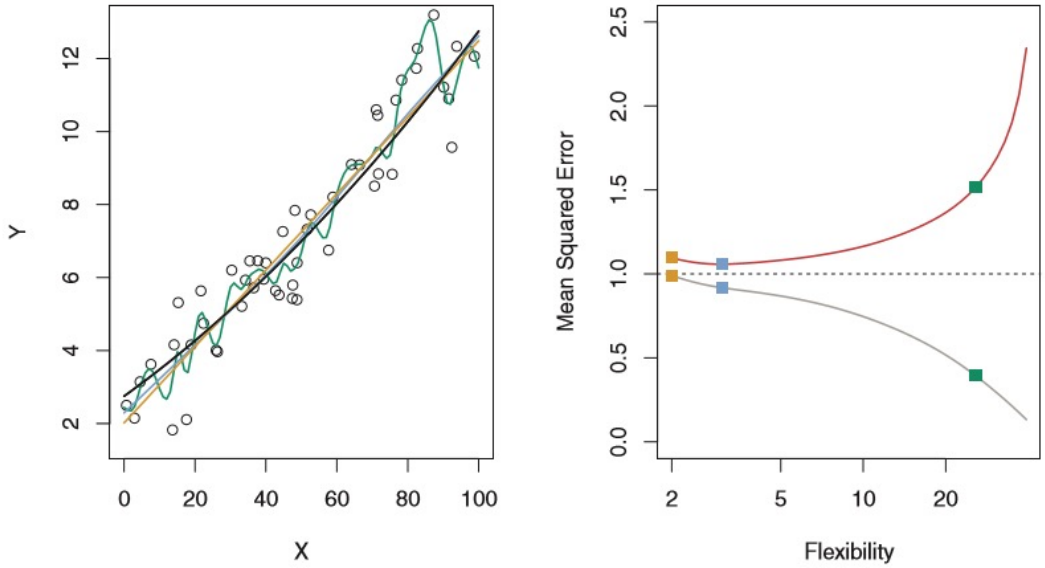
We will discuss neural nets later

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

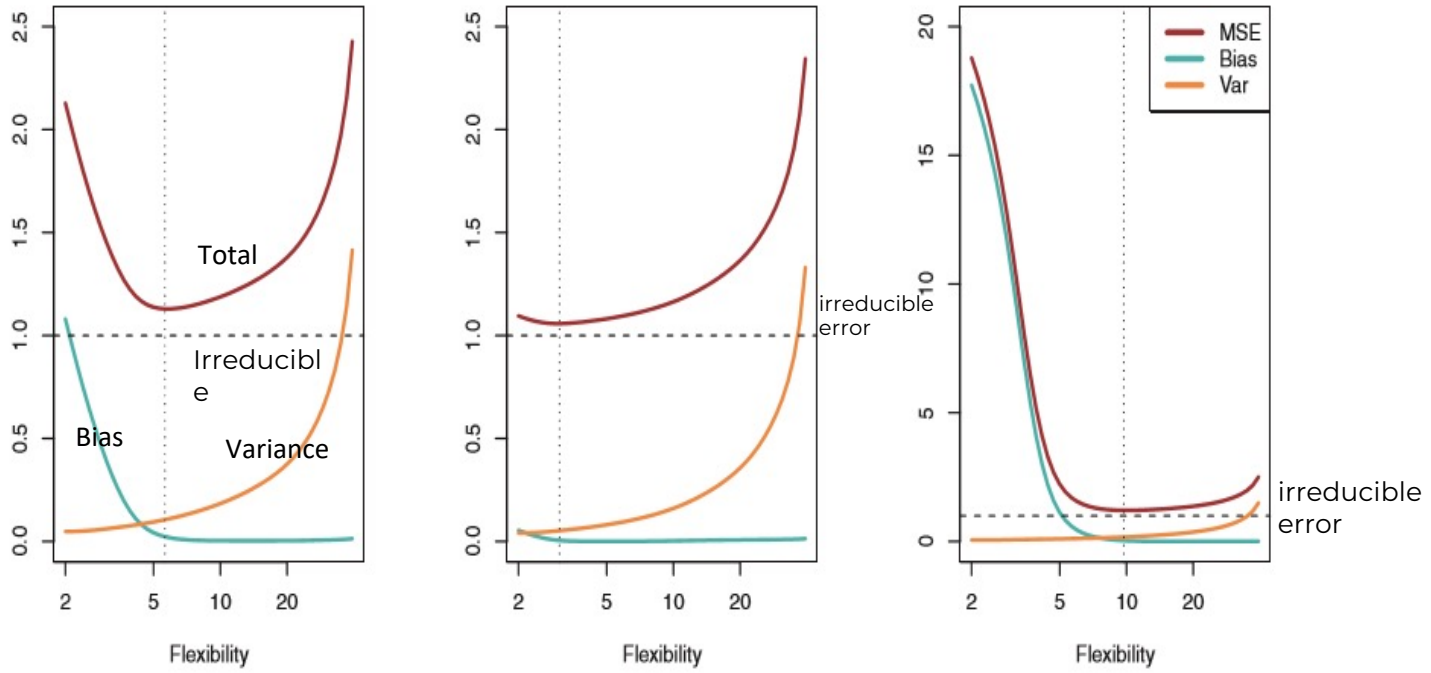
- The **Expected Test Mean Square Error** may be calculated as:

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

- **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

BIAS AND VARIANCE TRADE-OFF (CONT'D)

- **Model Variance** – due to sensitivity to specific training set
- **Model Bias** – due to model simplicity
- **Irreducible error -** is due to missing information



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. *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. *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. *Low complexity models may outperform high complexity models.*

GEOSTATISTICS AND MACHINE LEARNING

Lecture Outline Recap

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