Building Robust Data-Driven Machine Learning Models for Subsurface Energy Resource Applications: *Are We There Yet?*

Dr. Srikanta Mishra

BATTELLE

SPE Lima Section

August 26, 2021

The Attraction / Challenge

Big Data Analytics $\stackrel{?}{=}$ Game Changer

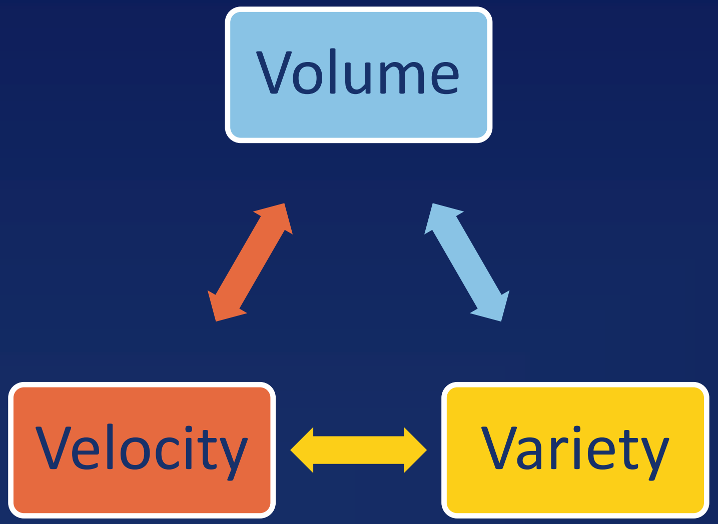
large volumes of data about subsurface, physical infrastructure and flows

New insights about reservoir from "data mining" can help increase operational efficiencies ????

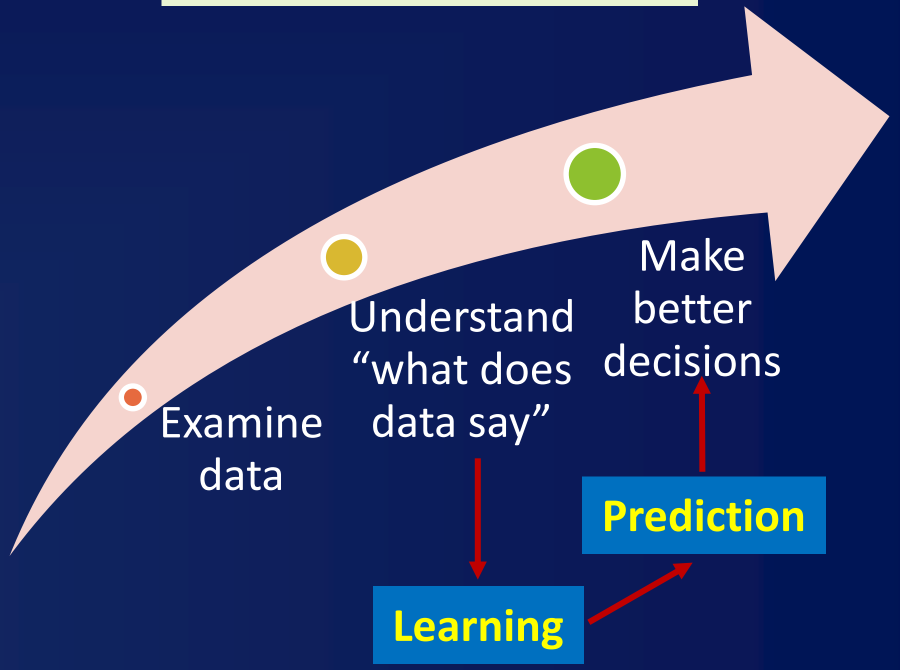
Actionable information

Big Data Analytics – What & Why?

Big Data



Data Analytics



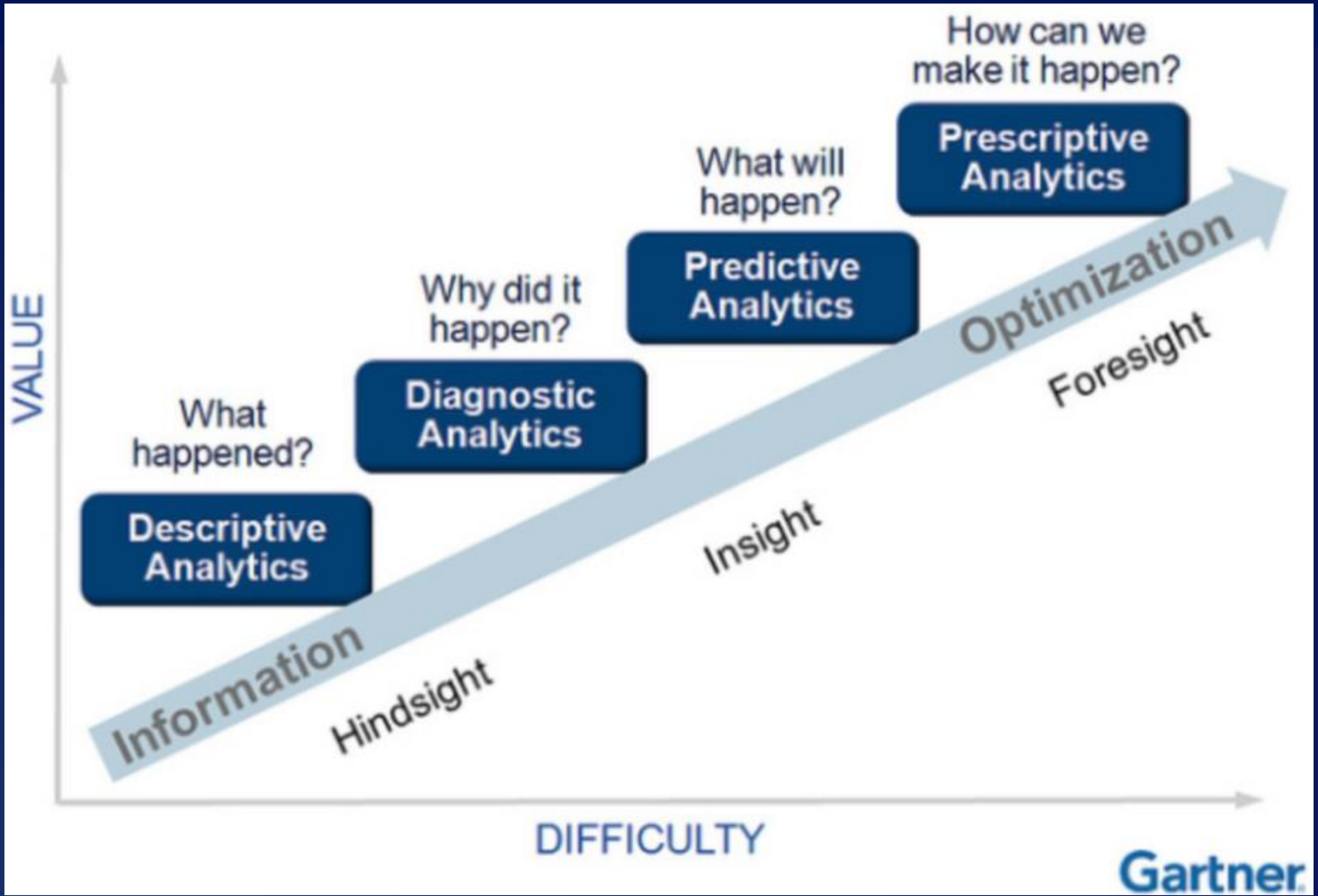
Data Analytics (*aka* Machine Learning, Data Mining)
helps understand <u>hidden patterns and relationships</u> in large, complex datasets

A Few Definitions

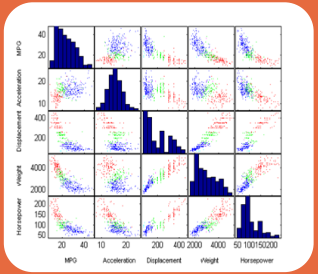
- ***Data analytics*** (DA) – sophisticated data collection + analysis
- ***Machine learning*** (ML) – building a model between predictors and response (often with a "black-box" algorithm)
- ***Artificial intelligence*** (AI) – applying predictive model with new data to make decisions

Mishra et al., 2021, JPT (March), 25-30.

Types of Analytics

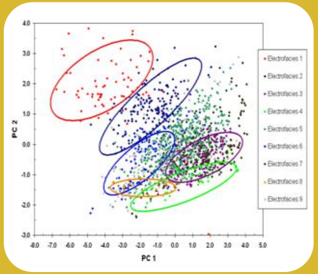


Data Analytics Process



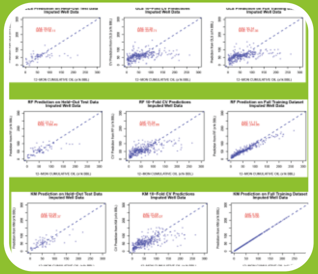
Exploratory Data Analysis

- Multi-dimensional data visualization
- Scatter-plot matrix, trellis plots



Unsupervised Learning

- Data reduction and clustering
- PCA, k-means, self-organizing maps



Supervised Learning

- Regression and classification
- Random forest, SVM, neural nets, kriging

Repertoire of Common ML Techniques

Regression & Classification Tree

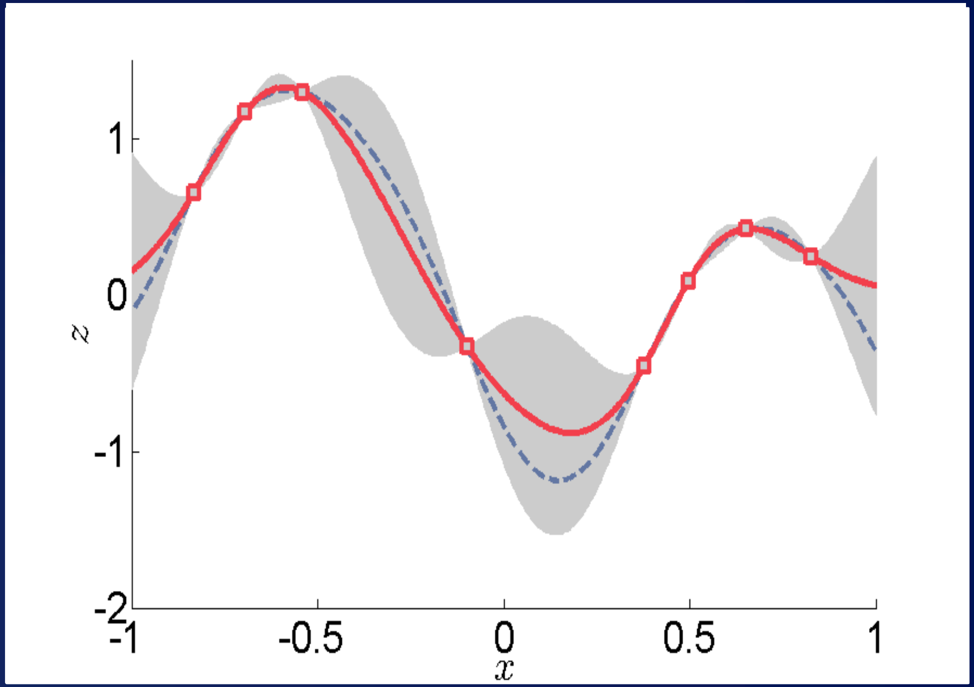
Random Forest

Gradient Boosting Machine

Support Vector Machine

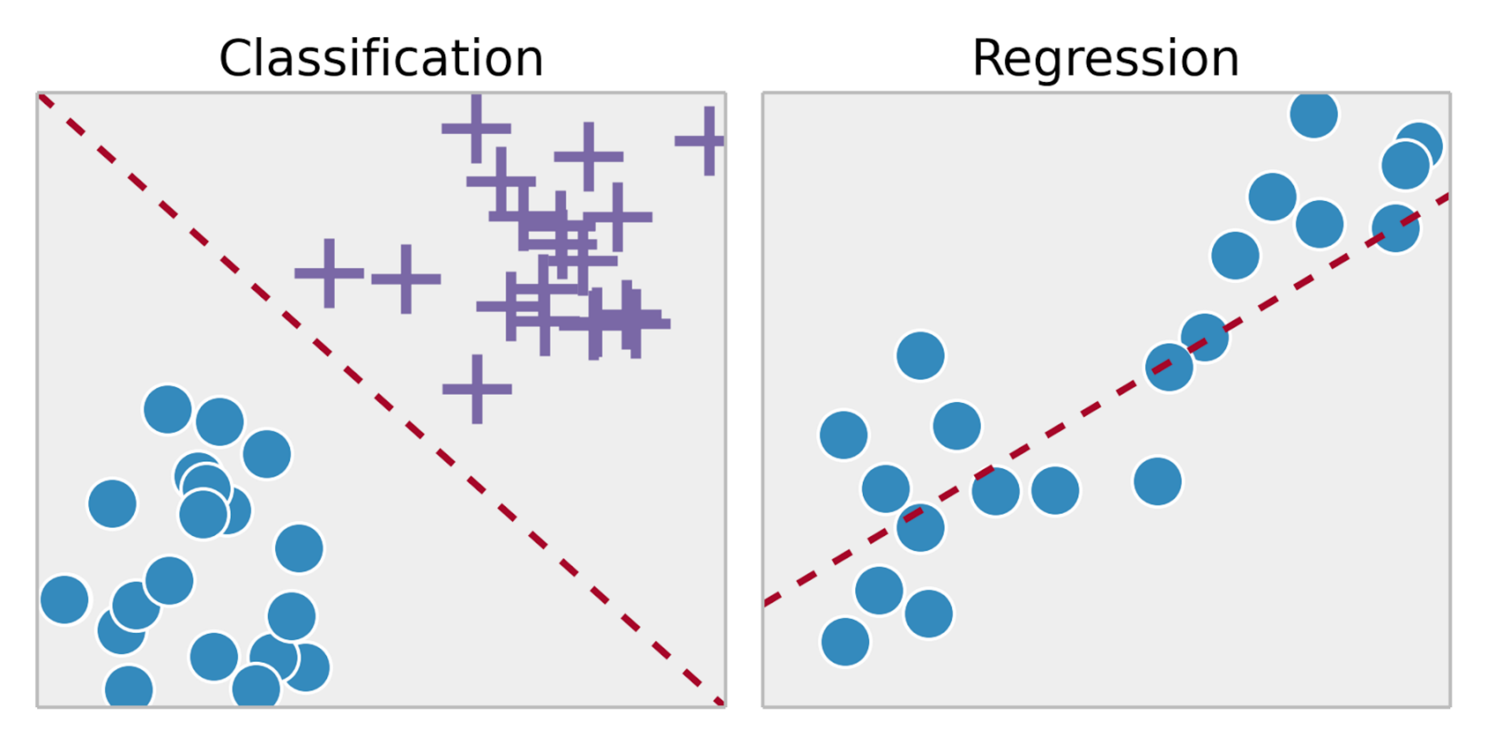
Artificial Neural Network

Gaussian Process Emulation



Multidimensional interpolation considering trend and autocorrelation structure of data

Two Example Applications

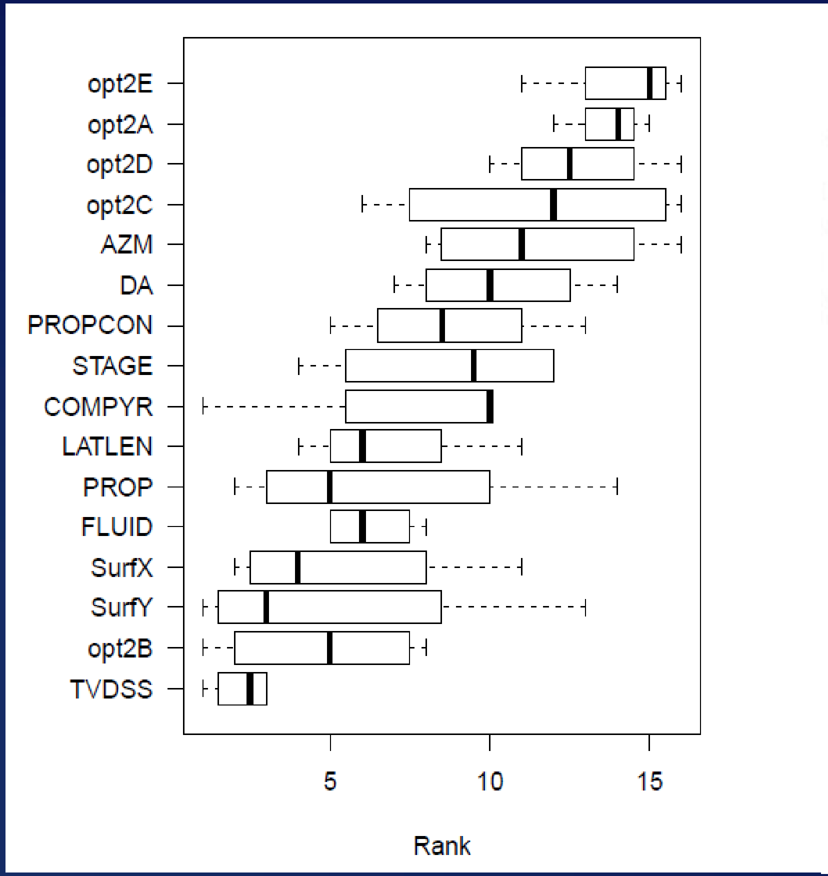


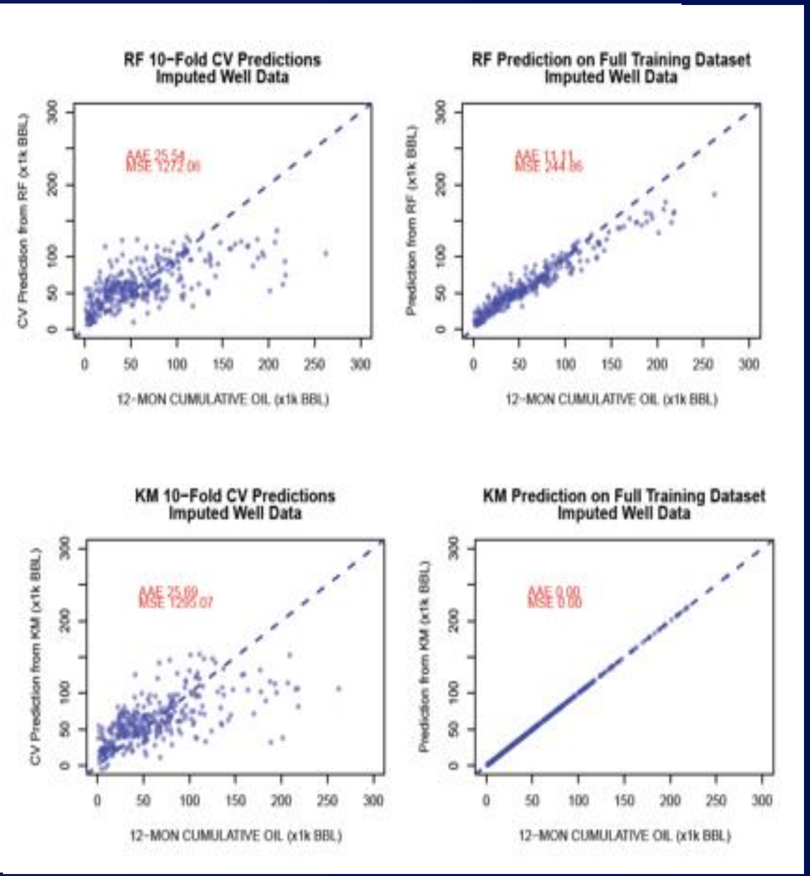
Identifying advanced log outputs (e.g., vug v/s no vug) using basic well log attributes

Explaining production from shale oil wells in terms of completion and well attributes

Example [1] – Key Factors Affecting Hydraulically Fractured Well Performance

- Wolfcamp Shale horizontal wells
  - Data from 476 Wells
  - **Goal** ⇨ Fit M12CO ~ *f* (12 predictors)
  - Multiple machine learning methods
  - Model validation + variable importance



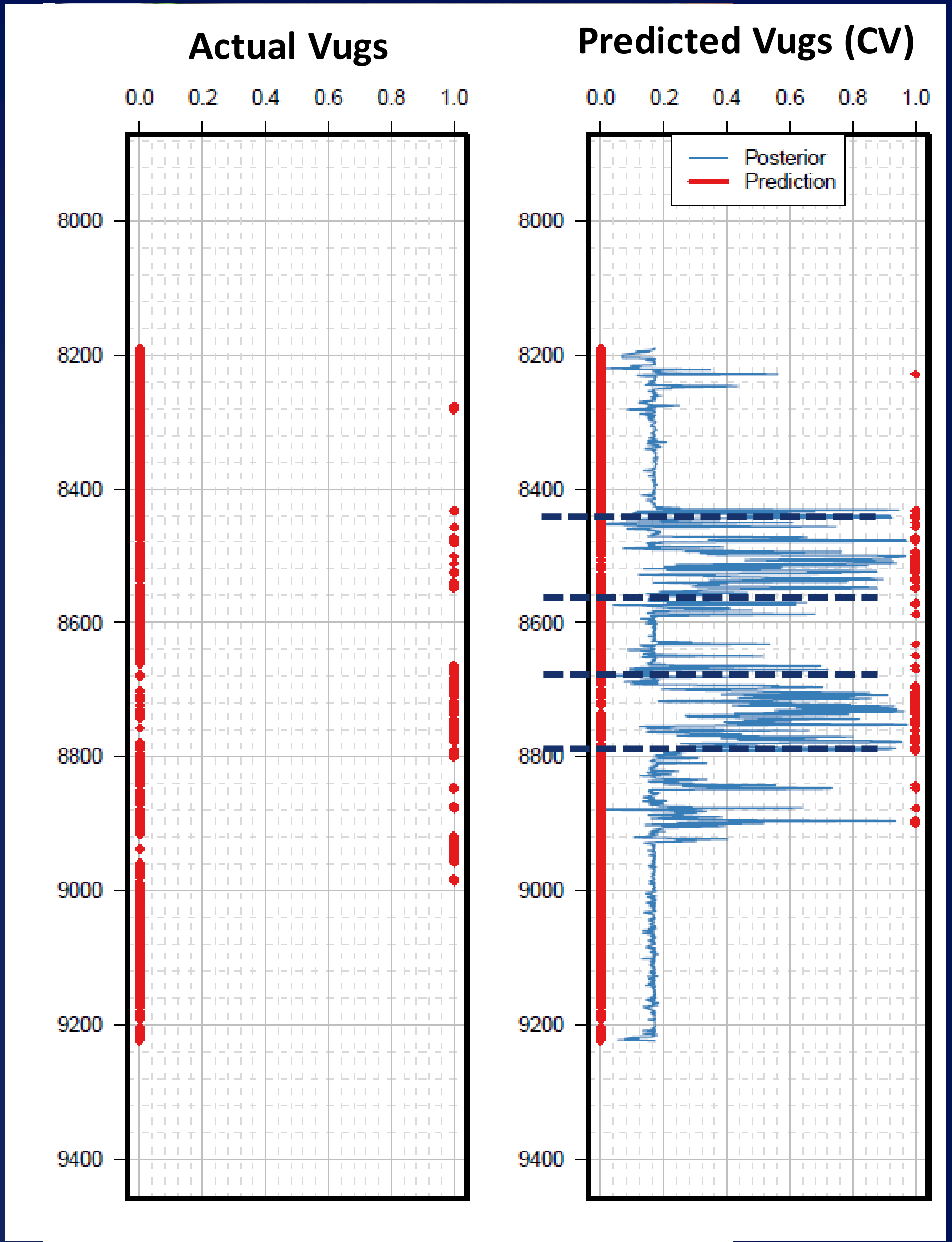


Schuetter, Mishra, Zhong, LaFollette, **2018, SPE Journal, SPE-189969-PA**

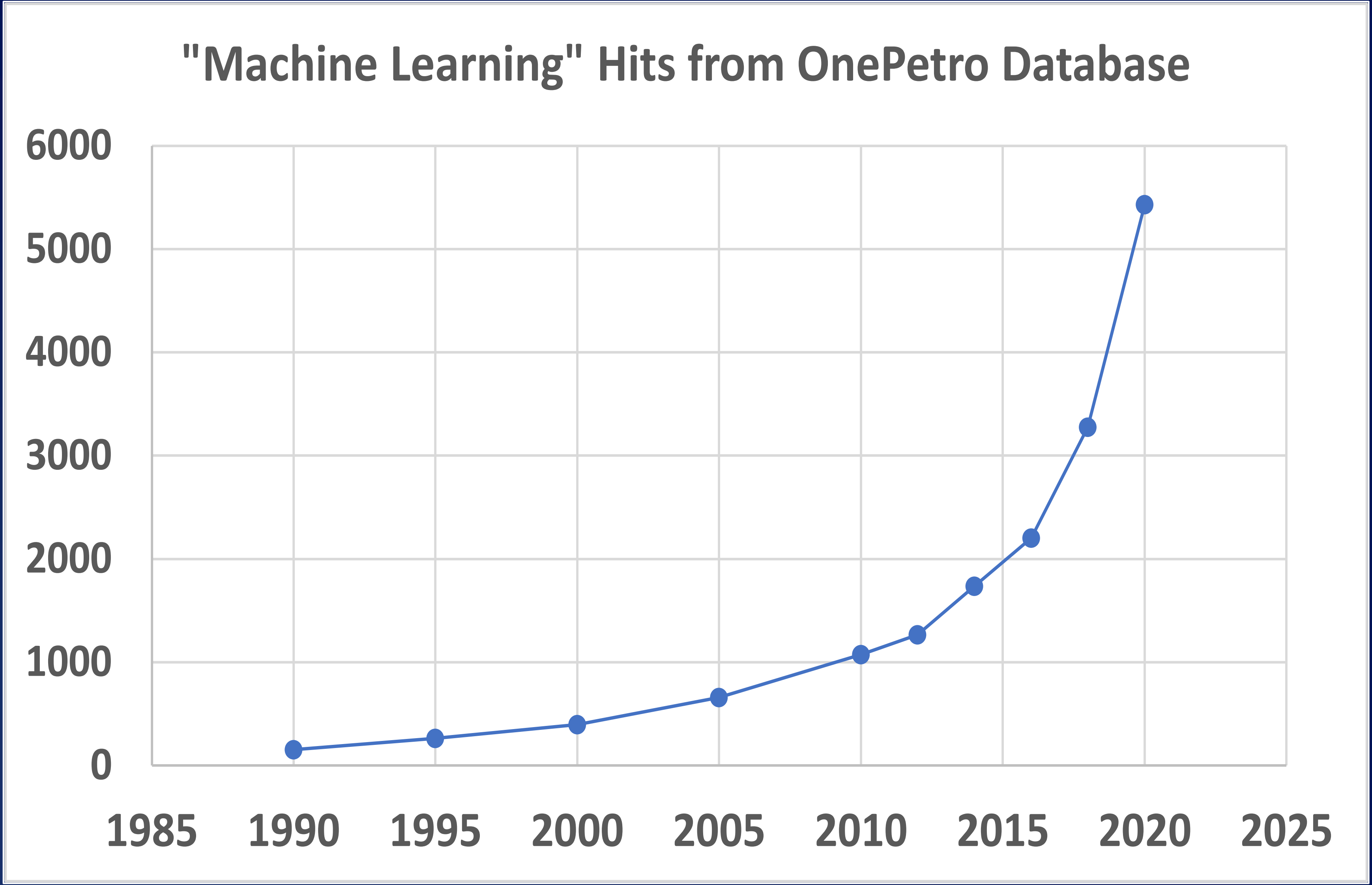
Example [2] – Vug Detection from Proxies

- Vuggy zones create high-permeability pathways in carbonate rocks
- Generally identified from cores and image logs
- **Challenge**: Identify vuggy zones from well-log response (PEF, GR, NPHI, RHOB)
- **Approach**: Use machine learning for classification

Haagsma et al, 2021, in "CO2 injection in network of fractures", de Dios et al. (Eds.,), Springer



Exponential Growth in ML Applications



Observations on Where Things Stand

- Two tracks (state of practice)
  - Some geoscientists and petroleum engineers may be applying these techniques in an ad-hoc manner
  - Others may be holding off on utilizing these methods because they do not have any formal ML training
- Some questions to ponder/discuss
  - Why ML models, and when?
  - One model or many?
  - Which predictors matter?
  - Can ML models become physics informed?
  - What are the challenges going forward?

***Mishra et al., 2021*, JPT (March), 25-30.**

Why ML Models and When?

- Historically, subsurface science and engineering analyses have relied on mechanistic (physics-based) models
- Incorporation of causal input-output relationship
- Experienced professionals are wary of purely data-driven “black-box” ML models that lack such understanding
- Nevertheless, the use of ML models is easy to justify - if
  - relevant physics-based model is computation intensive and/or immature
  - suitable mechanistic modeling paradigm does not exist

Three Cases for Black-Box Models (1)

- *When the cost of a wrong answer is low relative to the value of a correct answer, e.g.,*
  - using an ML-based proxy model to carry out initial explorations in the parameter space during history matching,
  - with further refinements in the vicinity of the optimal solution done using a full-physics model

Holm, 2019, Science, 367, 26-27.

Three Cases for Black-Box Models (2)

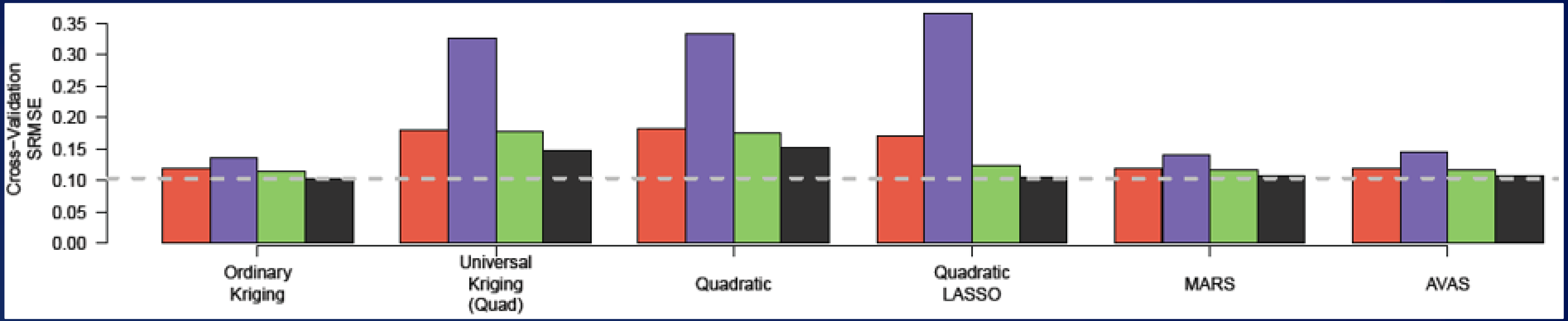
- *When they produce the best results, e.g.,*
  - using a large number of pre-generated images to seed a pattern recognition algorithm
  - Then matching the observed pressure derivative signature to an underlying conceptual model during well-test analysis

Three Cases for Black-Box Models (3)

- *As tools to inspire and guide human inquiry, e.g.,*
  - using operational and historical data for electrical submersible pumps in unconventional wells
  - understand the factors and conditions responsible for equipment failure or sub-optimal performance
  - perform preventative maintenance as needed

One Model or Many?

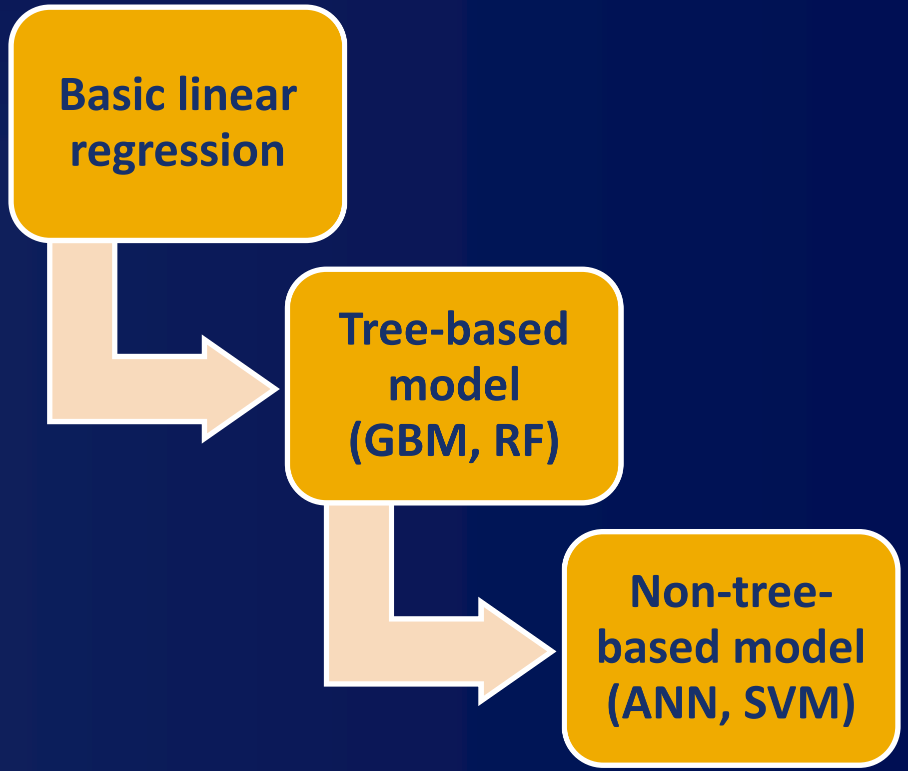
- Model fits measured in terms of training or test error – multiple competing models may arise!



- Aggregating over a large set of acceptable models can provide more robust understanding and predictions
- Ensemble models (with predictions aggregated) have been top performers in data science competitions

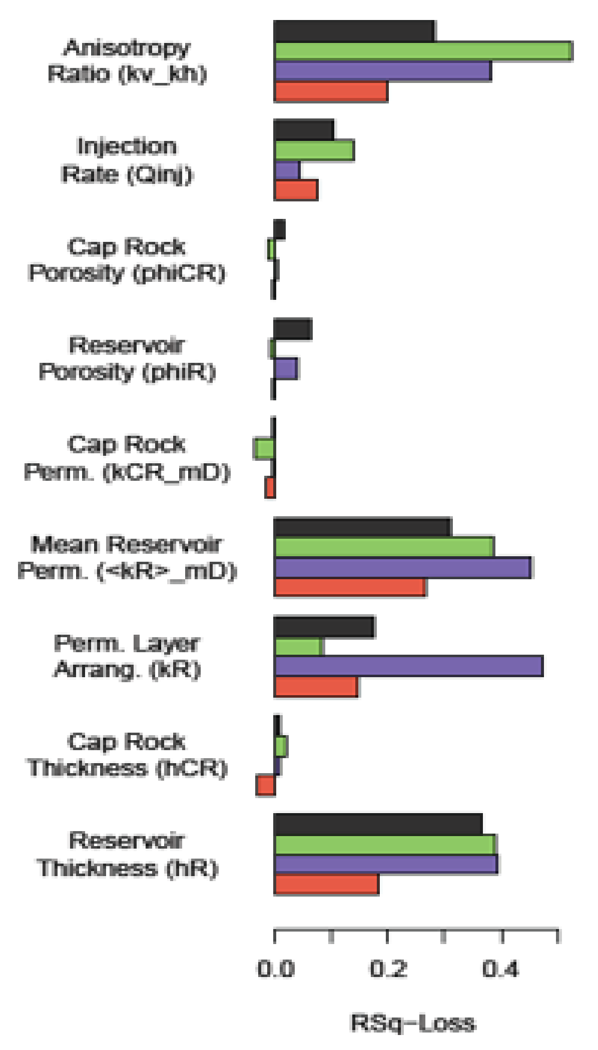
Ensemble Modeling Methods

- Model aggregation strategies
  - Simple averaging (direct average of constituent model predictions, e.g., using arithmetic average)
  - Weighted averaging (weighted averaging of constituent model predictions, e.g., using inverse of RMSE)
  - Stacking (predictions from the constituent models are used as predictors in an aggregate model, e.g., NN training)



Which Predictors Matter?

- Identification of variable importance can be model specific (e.g., for RF, GBM)
- Model independent metric based on R^2-loss
  - Extension of feature randomization used in RF for variable importance
  - [R^2 for full model] minus [R^2 for model without predictor of interest]
  - larger R^2–loss ⇨ greater influence



Variable Importance Strategies

Strategy	*Notation*	*Description*
Removing a variable	Remove	Remove a variable from the model, re-train the model and compare the reduction in pseudo-R^2, i.e. R^2 loss.
Permuting a variable	Permute	Permute a variable's values, which breaks the relationship between the variable and the true outcome, then compare the reduction in pseudo-R^2, i.e. R^2 loss, of the dataset with permuted values to that with true values.
Partial Dependent Plot	PDP	The partial dependence plot shows the marginal effect of different variables on the predicted outcome. PDPs are "flat" for less important variables while the variables whose PDP vary across a wider range of the response are more likely to be important.
Accumulated Local Effects Plot	ALE	Compare how the model predictions change in a small "window" of different variables. ALE plots are faster and unbiased alternative to partial dependence plots.
Local Interpretable Model-Agnostic Explanations	LIME	LIME attempts to understand the model by perturbing the input of data samples and interpreting how the predictions change. Variable weights can then be extracted from a simple local model on the permuted dataset to explain local behavior.
Shapley Addictive exPlantations	SHAP	SHAP is a method to explain individual predictions based on the game theoretically optimal Shapley values. A prediction can be explained by assuming that each feature value of the instance is a "player" in a game where the prediction is the payout. Shapley values – a method from coalitional game theory, tells us how to fairly distribute the "payout" among the features.

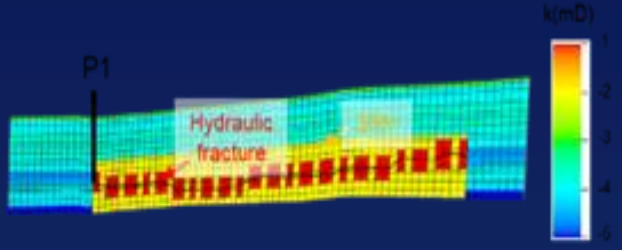
https://christophm.github.io/interpretable-ml-book/

Can ML Models Become Physics-Informed?

- Standard data-driven ML algorithms trained solely based on data
- No assurance that model predictions are physically consistent
  - Pressure versus viscosity
  - Relative permeability versus saturation
- Physics-informed ML approaches => more general "loss" function
  - Standard data misfit term (i.e., predicted v/s observed)
  - Additional residual term (i.e., based on governing equations)
  - In general, better fits to model results

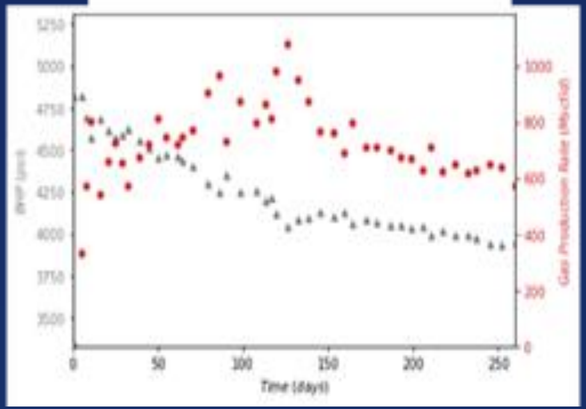
Example PINN Results

Reservoir model description

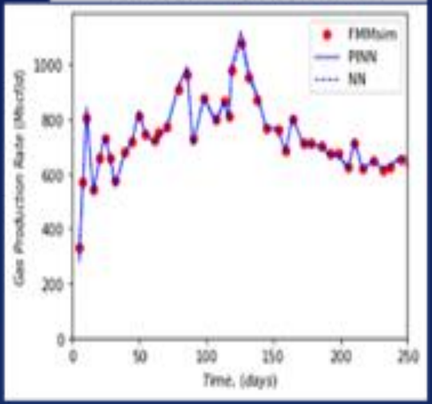


3-D unconventional reservoir, cross section view (based on Zhang et al., 2016)

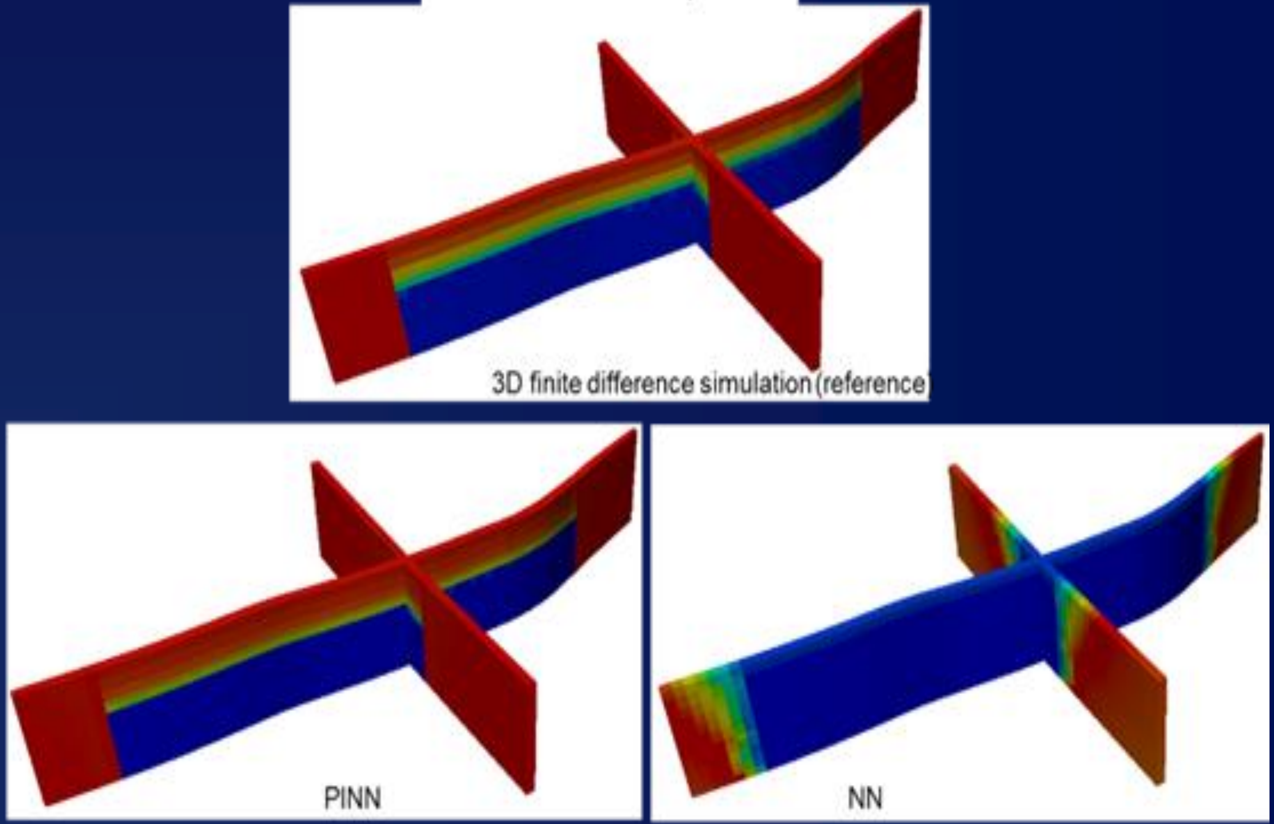
Gas production rate and BHP



Gas production rate



Pressure map



Challenges for Acceptance of ML

- Our ML models are not very good.
- If I don’t understand the model, how can I believe it?
- We are still waiting for the “Aha” moment!
- My staff need to learn data science, but how?

Addressing These Challenges ....

Poor Model Quality

- Consumer marketing ML/AI models are not necessarily highly accurate!
- Need to manage expectations re. quality of fit for subsurface models
- Focus more on added value from ML models + complementary role

Lack of Understanding

- Articulate adequacy of predictors
- Demonstrate model robustness
- Explain inner workings (key variables)
- Use creative visualizations

"Aha" Moment?

- ML model may or may not produce new insights
- Provides an alternative quantitative input-output relationship
- Useful when physics-based model is slow, data-intensive or immature

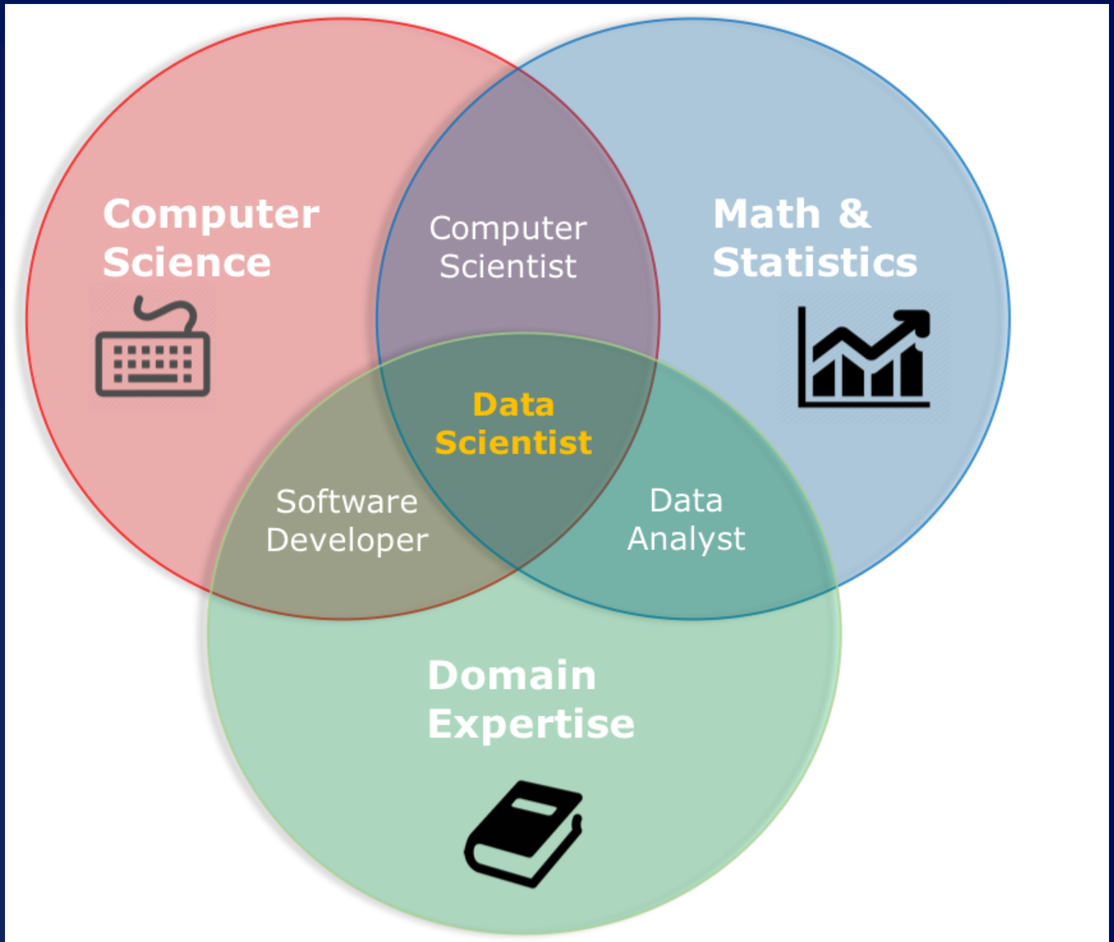
Learning Data Science

- Significant (informal) self-learning to become "citizen data scientists"
- Need formal knowledge of conventional data analysis, python/R programming, and machine learning

Learning Petroleum Data Science

- Petroleum data scientist/analyst (*one who learns from data*)
  - Better at statistics than programmer, better at programming than statistician, and better at petroleum engineering than both
- Core competencies
  - Data collection, preparation and exploration
  - Data storage and retrieval
  - Computing with data
  - Applied machine learning
  - Data visualization/communication

Donoho, J. Comp. Graphical Statistics, 2017



https://www.linkedin.com/pulse/new-venn-diagram-data-science-pierluigi-casale/

Closing Thoughts – Present

- Buzz about DA and ML/AI ⇨ growing O&G applications ⇨ misplaced expectations?
- Significant ongoing activity related to technology adaptation/development + formal/informal upskilling of geo-energy professionals in data science
- Current status of this field ⇨ somewhat immature
  - Similarity to Geostatistics in 1990s
  - Potential realized by industry
  - Not yet fully adopted for mainstream applications

Closing Thoughts – Future

- Focus on issues for making data-driven models more robust (i.e., accurate, efficient, understandable, and useful)
- Promote foundational understanding of ML-related technologies among petroleum engineers and geoscientists
- Appropriate mindset
  - NOT curve-fitting exercises using very flexible and powerful algorithms
  - BUT extraction of insights consistent with mechanistic understanding

ACKNOWLEDGMENTS

Battelle Memorial Institute

US DOE-NETL

Thank you for your attention

mishras@battelle.org