PGE 383 Subsurface Machine Learning

Lecture 18: Course Conclusion

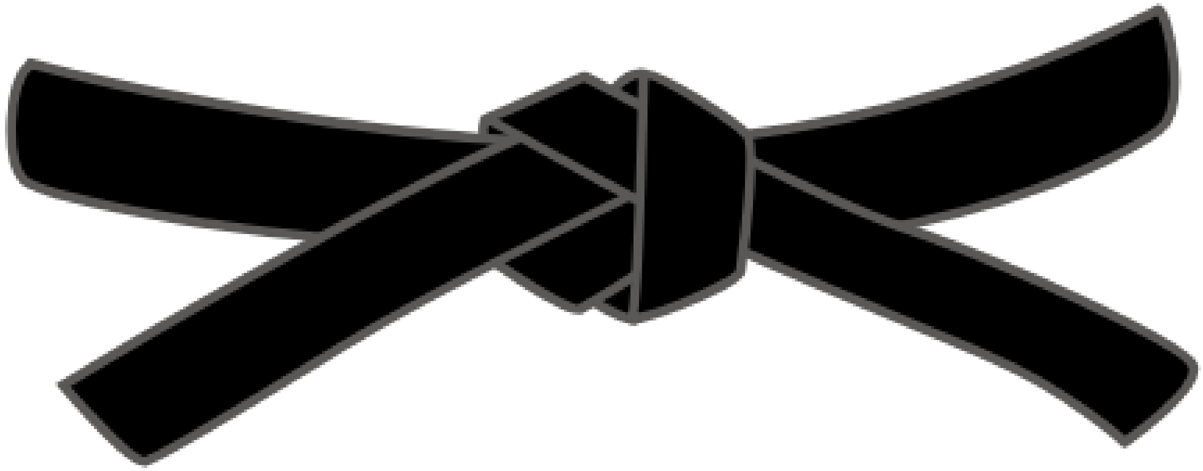
Lecture outline:

- **Lessons Learned**
- **Unsolved Problems**
- **Advanced Topics**

Motivation

Excellent performance, projects and results from this class!

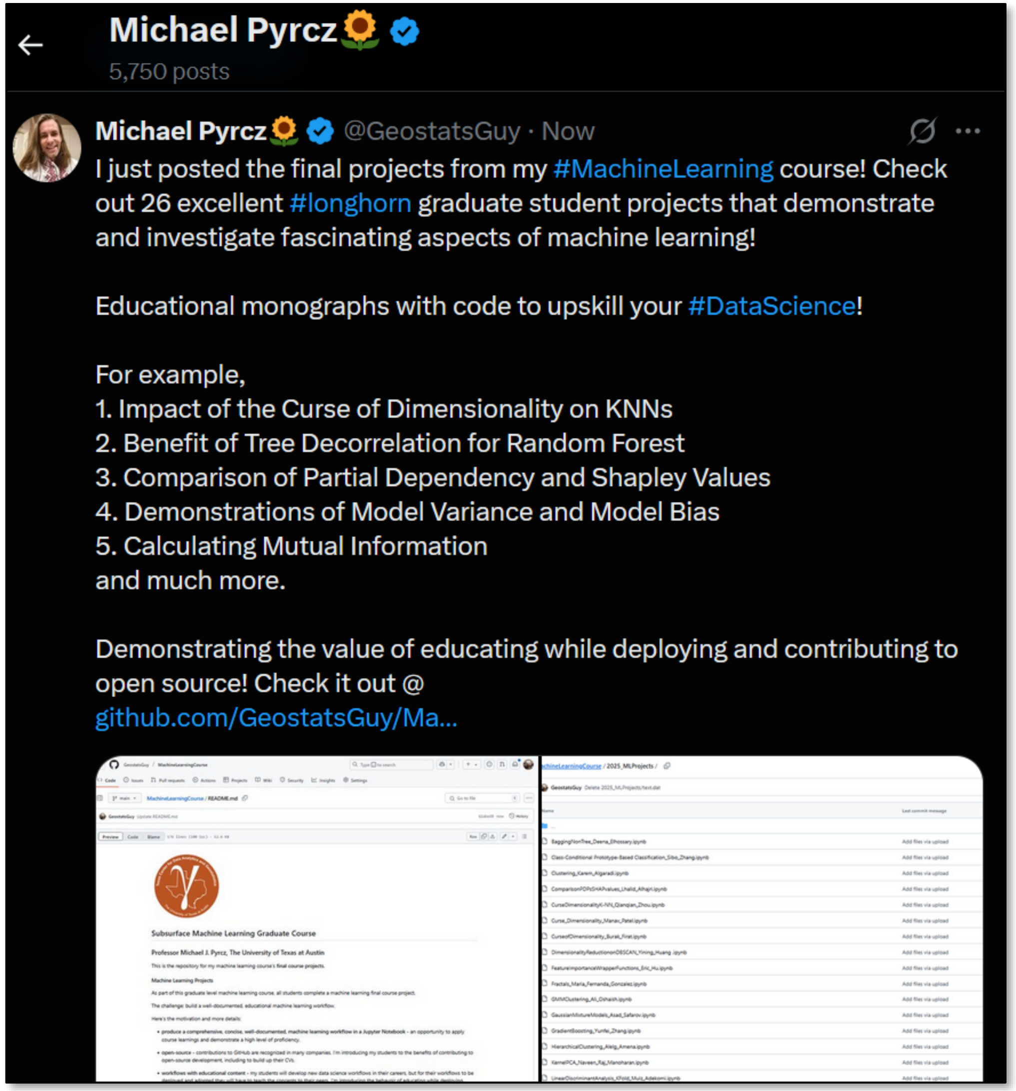
- You are ready, for deeper knowledge for more success with machine learning!



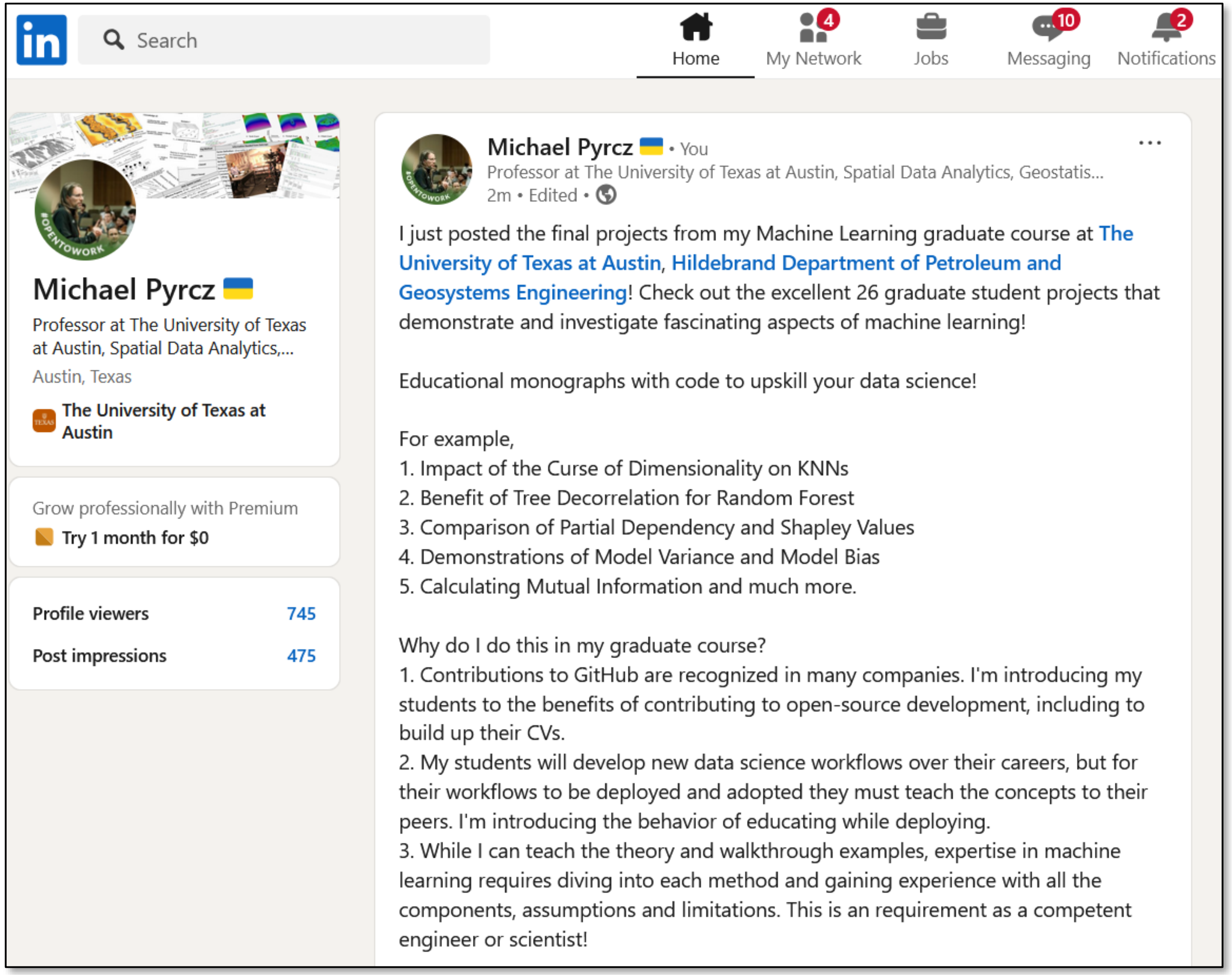
Black belt in machine learning.

Black belt image from https://upload.wikimedia.org/wikipedia/commons/0/01/Black_belt.svg

Motivation



Michael Pyrcz
5,750 posts
Michael Pyrcz @GeostatsGuy · Now
I just posted the final projects from my #MachineLearning course! Check out 26 excellent #longhorn graduate student projects that demonstrate and investigate fascinating aspects of machine learning!
Educational monographs with code to upskill your #DataScience!
For example,
1. Impact of the Curse of Dimensionality on KNNs
2. Benefit of Tree Decorrelation for Random Forest
3. Comparison of Partial Dependency and Shapley Values
4. Demonstrations of Model Variance and Model Bias
5. Calculating Mutual Information
and much more.
Demonstrating the value of educating while deploying and contributing to open source! Check it out @ github.com/GeostatsGuy/Ma...



in
Search
Home
My Network
4
Jobs
Messaging
10
Notifications
2
Michael Pyrcz
Professor at The University of Texas at Austin, Spatial Data Analytics,...
Austin, Texas
The University of Texas at Austin
Grow professionally with Premium
Try 1 month for $0
Profile viewers 745
Post impressions 475
Michael Pyrcz • You
Professor at The University of Texas at Austin, Spatial Data Analytics, Geostatis...
2m • Edited •
I just posted the final projects from my Machine Learning graduate course at The University of Texas at Austin, Hildebrand Department of Petroleum and Geosystems Engineering! Check out the excellent 26 graduate student projects that demonstrate and investigate fascinating aspects of machine learning!
Educational monographs with code to upskill your data science!
For example,
1. Impact of the Curse of Dimensionality on KNNs
2. Benefit of Tree Decorrelation for Random Forest
3. Comparison of Partial Dependency and Shapley Values
4. Demonstrations of Model Variance and Model Bias
5. Calculating Mutual Information and much more.
Why do I do this in my graduate course?
1. Contributions to GitHub are recognized in many companies. I'm introducing my students to the benefits of contributing to open-source development, including to build up their CVs.
2. My students will develop new data science workflows over their careers, but for their workflows to be deployed and adopted they must teach the concepts to their peers. I'm introducing the behavior of educating while deploying.
3. While I can teach the theory and walkthrough examples, expertise in machine learning requires diving into each method and gaining experience with all the components, assumptions and limitations. This is an requirement as a competent engineer or scientist!

PGE 383 Subsurface Machine Learning

Lecture 19: Dr. Pyrcz's Secrets for Success with Machine Learning

Lecture outline:

- **Set the Stage**
- **Use Our Geostatistics Domain Knowledge**
- **Use Our Domain Knowledge**
- **Uncertainty Always**
- **Make Powerful Plots**
- **Educate and Deploy**

PGE 383 Subsurface Machine Learning

Lecture 19: Dr. Pyrcz's Secrets for Success with Machine Learning

Lecture outline:

- **Set the Stage**

Motivation

Here's what you need to remember to succeed in the 4th paradigm of scientific discovery as an engineer or scientist.

1st Paradigm Empirical Science	2nd Paradigm Theoretical Science	3rd Paradigm Computational Science Simulation	4th Paradigm Data-driven Science
• Experiments	• Laws of classical mechanics, electrodynamics	• Continuum mechanics for heterogeneous systems • Computational fluid dynamics	• Detection of patterns and anomalies in big data • Artificial intelligence
<1600 BCE	1600s	1950s	2010s

Machine Learning's Role in Our Scientific Paradigms

Modified from https://www.nomad-coe.eu/news/147/39/NOMAD-establishes-new-fourth-paradigm-in-computational-materials-science

Machine Learning

Definition of Machine Learning

- define all terms
- state the prediction model
- list assumptions and limitations
- visualize the predictions

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

not a panacea

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

Demystify Machine Learning, Define and Describe

Quote from Machine Learning - Wikipedia

Machine Learning

Always When Applying Machine Learning,

- define all terms
- state the prediction model
- list assumptions and limitations
- visualize the predictions

The Model

Response **Predictors** **Random Error**

$$Y = f(X_1, \ldots, X_m) + \epsilon$$

Reminder that machine learning is a model, no matter how complicated an equation.

Demystify Machine Learning, Define and Describe

Quote from Machine Learning - Wikipedia

Interpretability is Critical

Develop Methods and Workflows that Provide Useful Diagnostics

- Interpretability may be low
- Application may become routine and trusted
- The machine is trusted, becomes an ‘unquestioned authority’

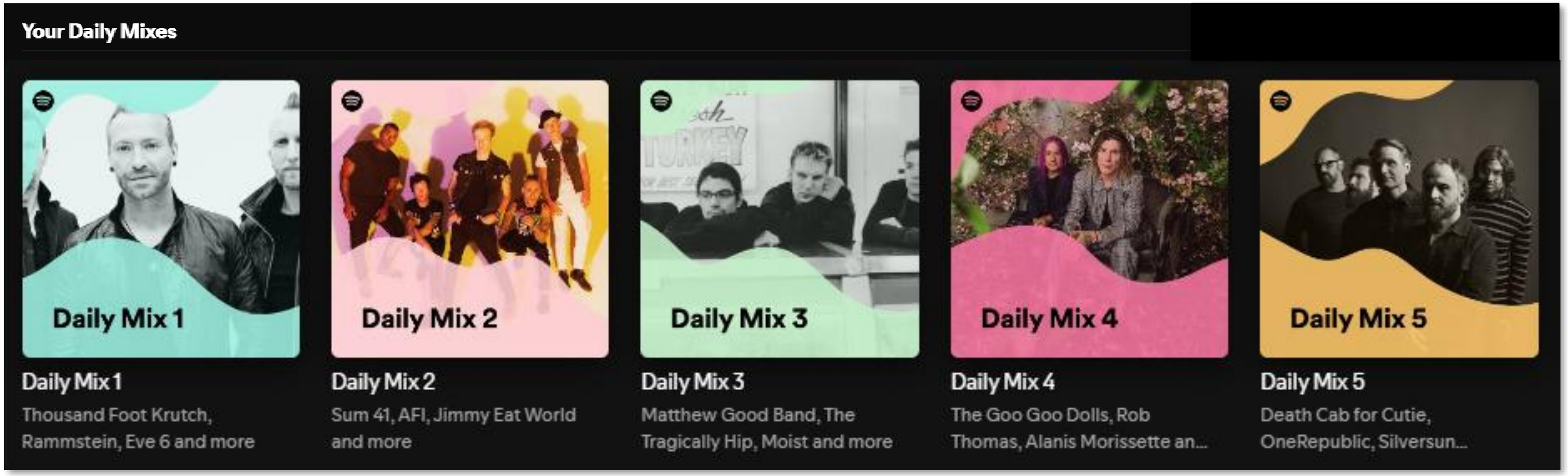
Machine learning-based logistic classifier to identify wolf or dog, image and example from Ribeiro et al. (2016) https://arxiv.org/pdf/1602.04938.pdf.

‘Even the developers that work on this stuff have no idea what it is doing’ ‘These systems do not fail gracefully!’
– Peter Haas TED Talk.

Fit-for-Purpose Application

Subsurface is Different and Needs New Solutions:

- Sparse, 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



My Spotify recommender system from my account summer, 2024.

We need to develop novel subsurface data analytics and machine learning solutions that integrate of geoscience and engineering → most data science tools is not ready off the shelf!

Keep Coding

Reasons All Geoscientists and Engineers Should Learn to Code

Transparency – *no compiler accepts hand waiving!* Coding forces your logic to be uncovered for any other scientist or engineer to review.

Reproducibility – *run it, get an answer, hand it over, run it, get the same answer.* This is a main principle of the scientific method.

Quantification – *programs need numbers.* Feed the program and discover new ways to look at the world.

Open-source – *leverage a world of brilliance.* Check out packages, snippets and be amazed with what great minds have freely shared.

Break Down Barriers – *don't throw it over the fence*. Sit at the table with the developers and share more of your subject matter expertise for a better product.

Deployment – *share it with others and multiply the impact*. Performance metrics or altruism, your good work benefits many others.

Efficiency – *minimize the boring parts of the job*. Build a suite of scripts for automation of common tasks and spend more time doing science and engineering!

Always Time to Do it Again! – *how many times did you only do it once?* It probably takes 2-4 times as long to script and automate a workflow. Usually worth it.

Be Like Us – *it will change you*. Users feel limited, programmers truly harness the power of their applications and hardware.

Lecture 19: Dr. Pyrcz’s Secrets for Success with Machine Learning

Lecture outline:

- **Use Our Geostatistics Domain Knowledge**

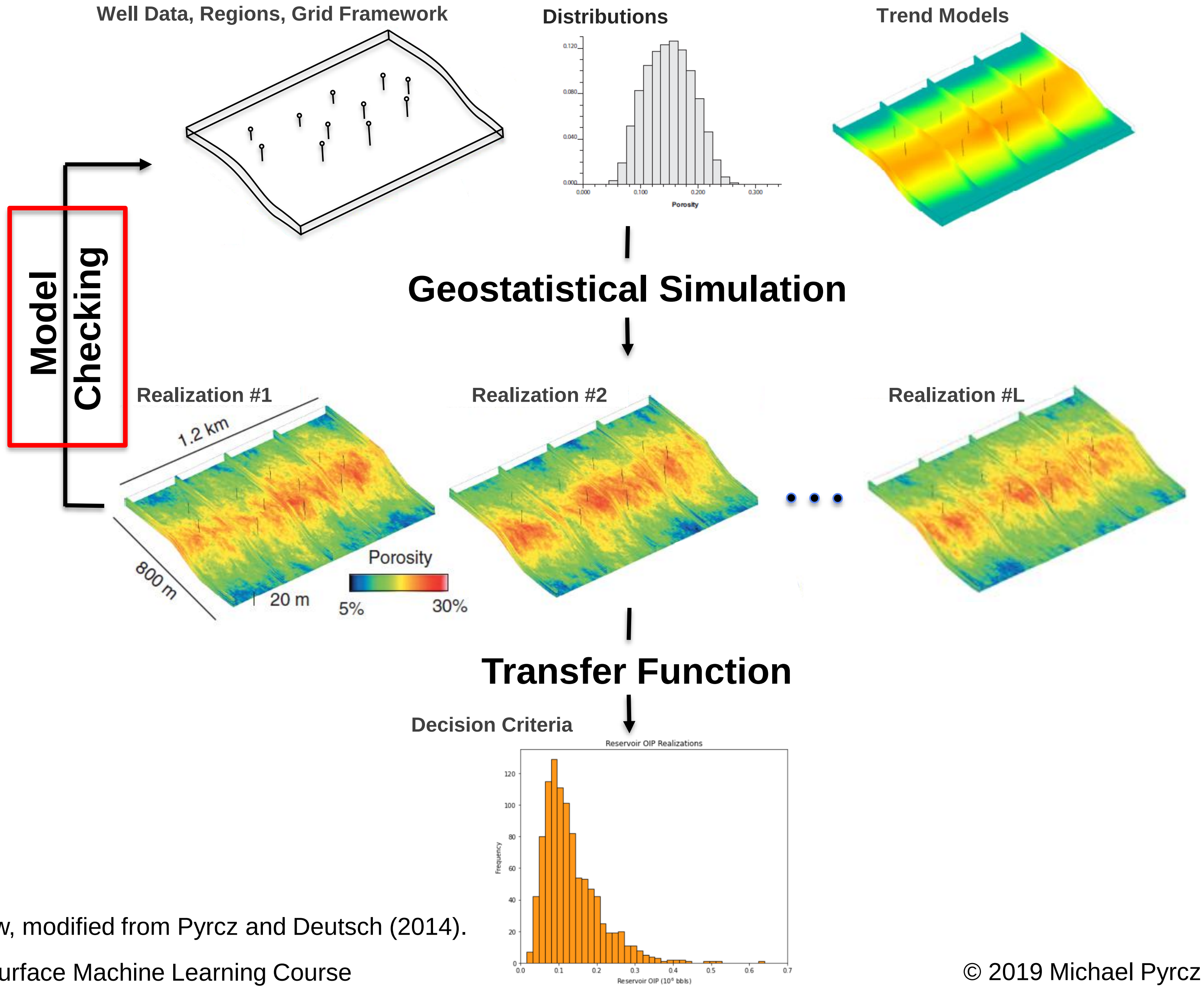
Model Checking

Established Best Practice

- over decades of supporting billion-dollar development decision making with data science!

Our models merge many information sources

- we must check that our model outputs match our model inputs.
- check every step, check every input
- check, check, check! Close the loop!



The standard geostatistical modeling workflow, modified from Pyrcz and Deutsch (2014).

Model Checking Demonstration

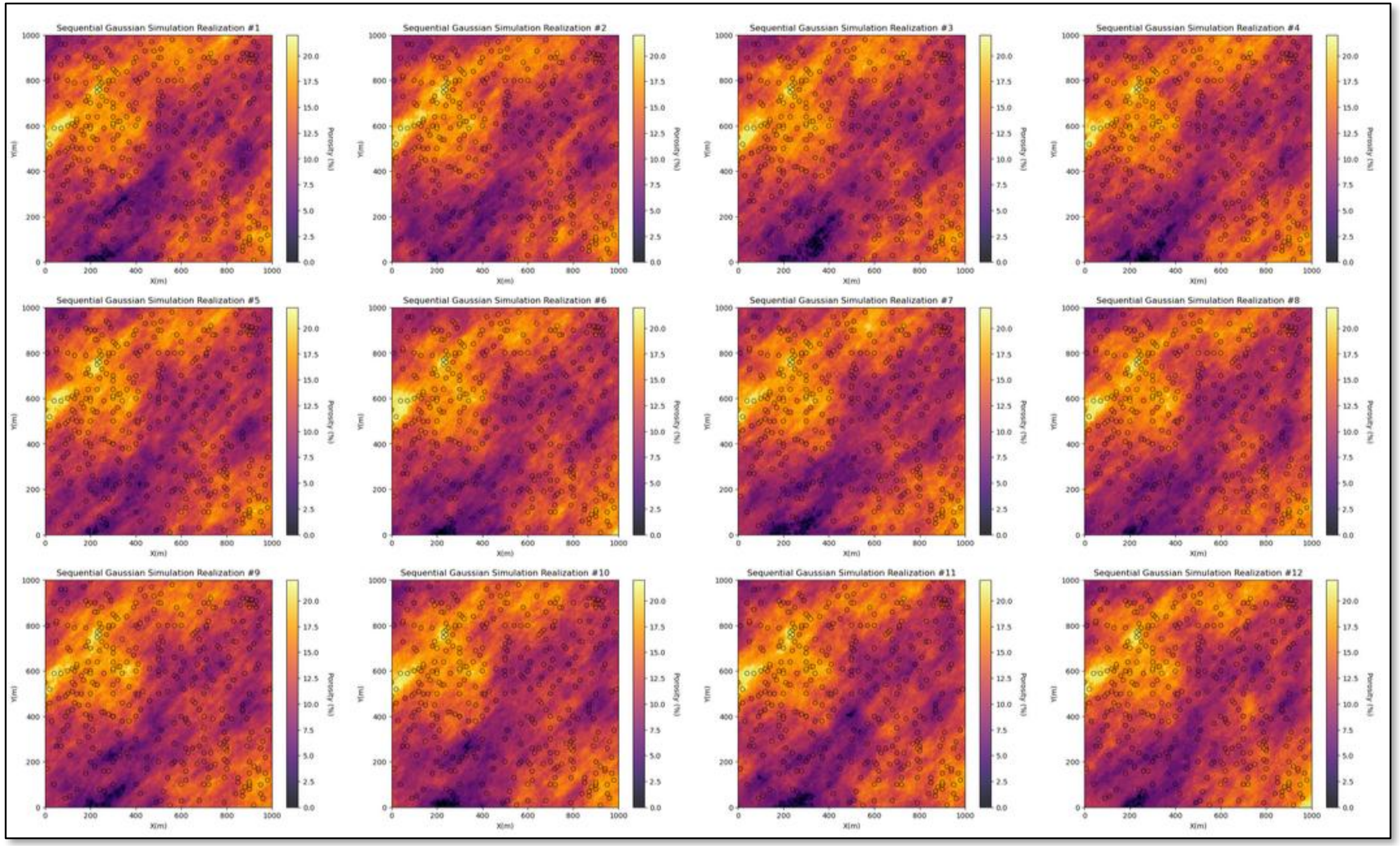
Use Visualization for Everything Possible

- Visualize your data and model!
- Look at the predictions! High dimensional, consider some sections!

Model Checking Demonstration from my New, Free e-book.

Chapter: Model Checking

Applied Geostatistics in Python: a Hands-on Guide with GeostatsPy.



Multiple geostatistical realizations of porosity (Pyrcz, 2014).

Try this out with my Applied Geostatistics in Python e-book: https://geostatsguy.github.io/GeostatsPyDemos_Book/GeostatsPy_model_checking.html

Model Checking Demonstration

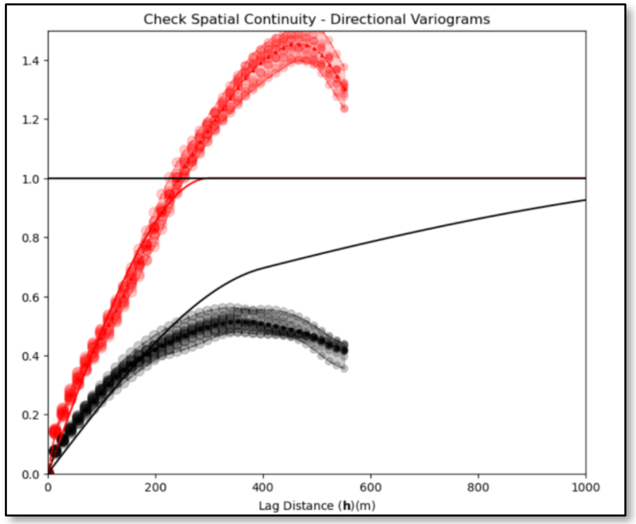
Go Beyond Visualization, Use a Suite of Statistics

- The basic checks for goodness of geostatistical realizations (Leuangthong et al., 2004).

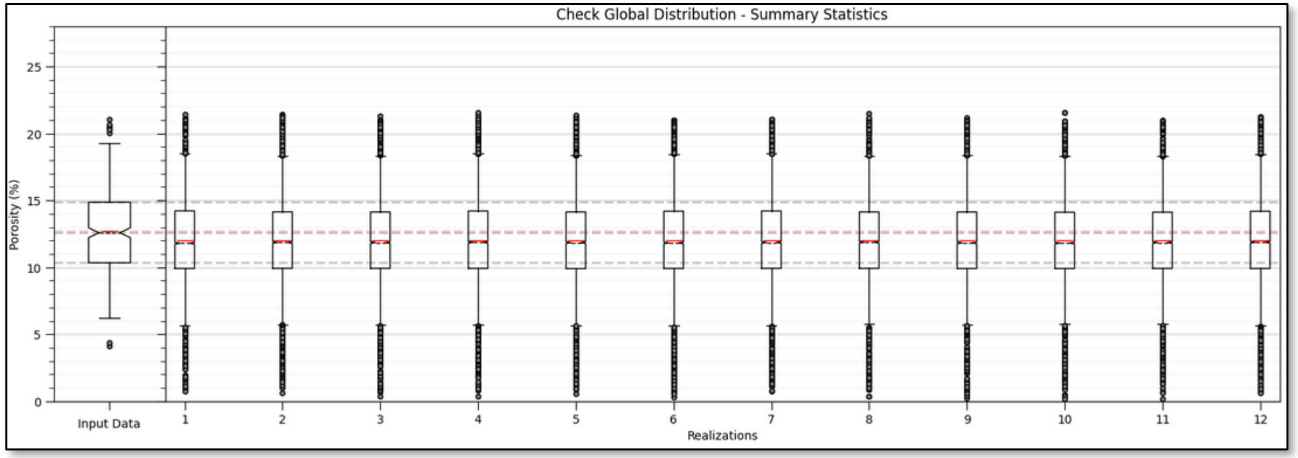
Statistical checks of all inputs are reproduced in the outputs

- data exactitude
- histogram
- variogram
- correlation

and check the uncertainty model goodness!



Check reproduction of the anisotropic variogram model (Pyrcz, 2024).



Check reproduction of the input distribution (Pyrcz, 2024).

Try this out with my Applied Geostatistics in Python e-book: https://geostatsguy.github.io/GeostatsPyDemos_Book/GeostatsPy_model_checking.html

Model Checking Demonstration

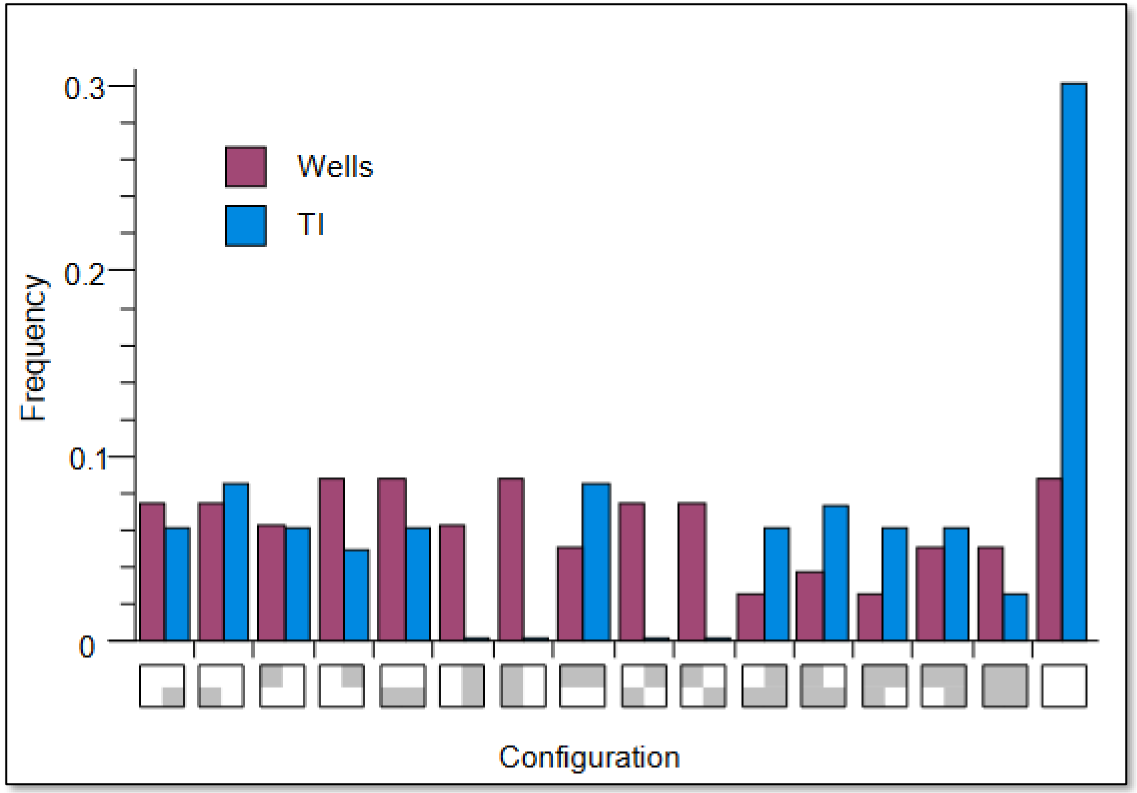
Always Quantify and New Quantification

Our subsurface workflows are complicated,

- every time we touch our data or models is an opportunity for a blunder.

When we developed new machine learning methods, develop new statistics to summarize data and models, for example,

- minimum acceptance criteria extended for multiple point simulation (MPS) realizations by Boisvert, Pyrcz and Deutsch (2006).



Checking reproduction of the multiple point spatial continuity model (Boisvert et al., 2006).

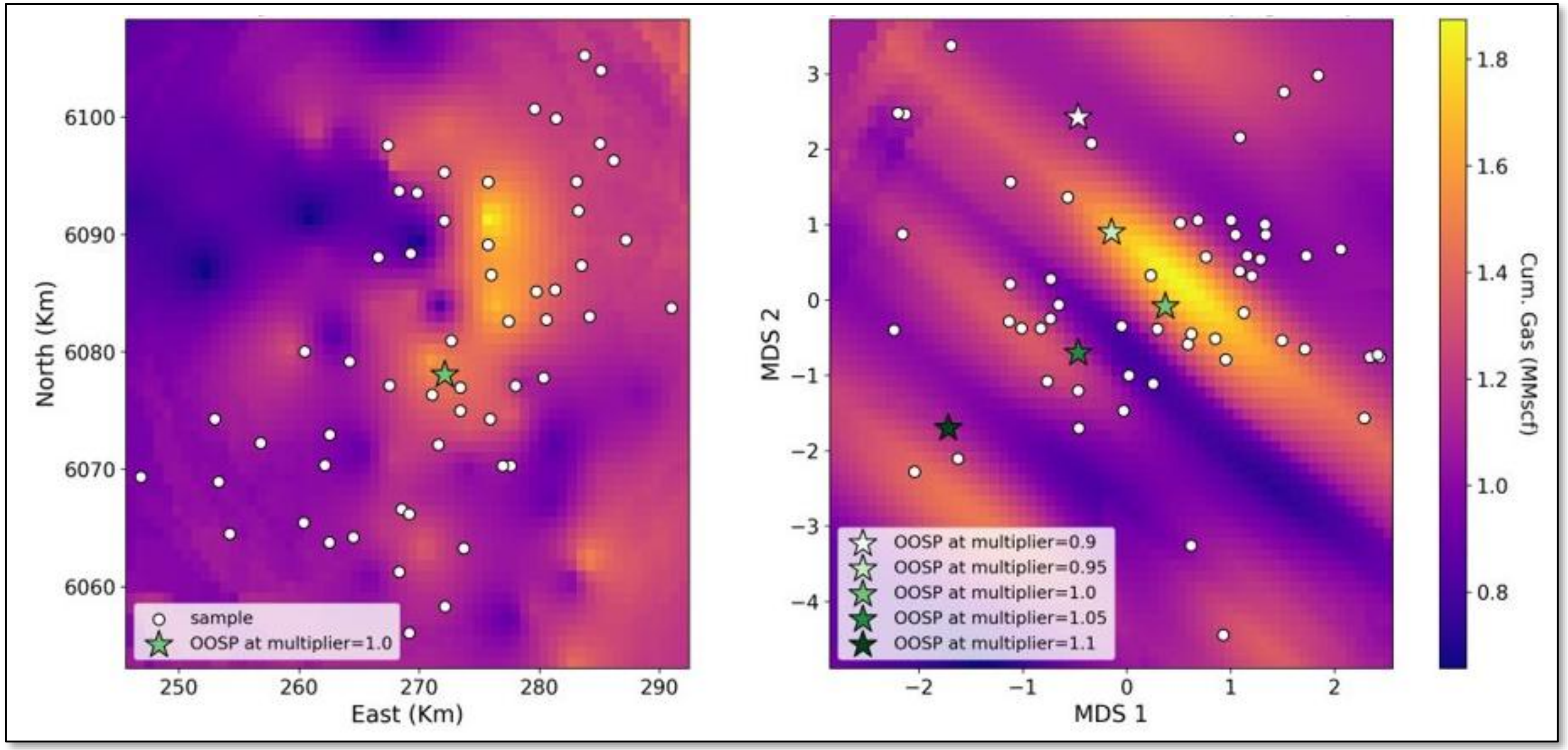
We continue to extend our checks as we add new machine learning modeling methods.

Model Checking Demonstration

Check the Uncertainty Model, Also

For example, checking our model realizations by visualizing a reduced dimensionality space based on pairwise dissimilarity (Caers, 2011).

- the range of realizations
- differences between model types and scenarios



Exploring the model uncertainty by visualizing production over location (left) and low dimensionality projection of the predictor features (right) (Mabadeje et al., 2024).

A generalized approach for any type of model.

Also, always check the uncertainty model.

What Did We Learn?

We must continue to check our models.

The old conditioning issues, implicit priority and artifacts remain! GenAI is often judged by ocular inspection.

New model checking methods must keep up with new modeling methods!

Always quantitatively check your models!

Logical and Model Consistency

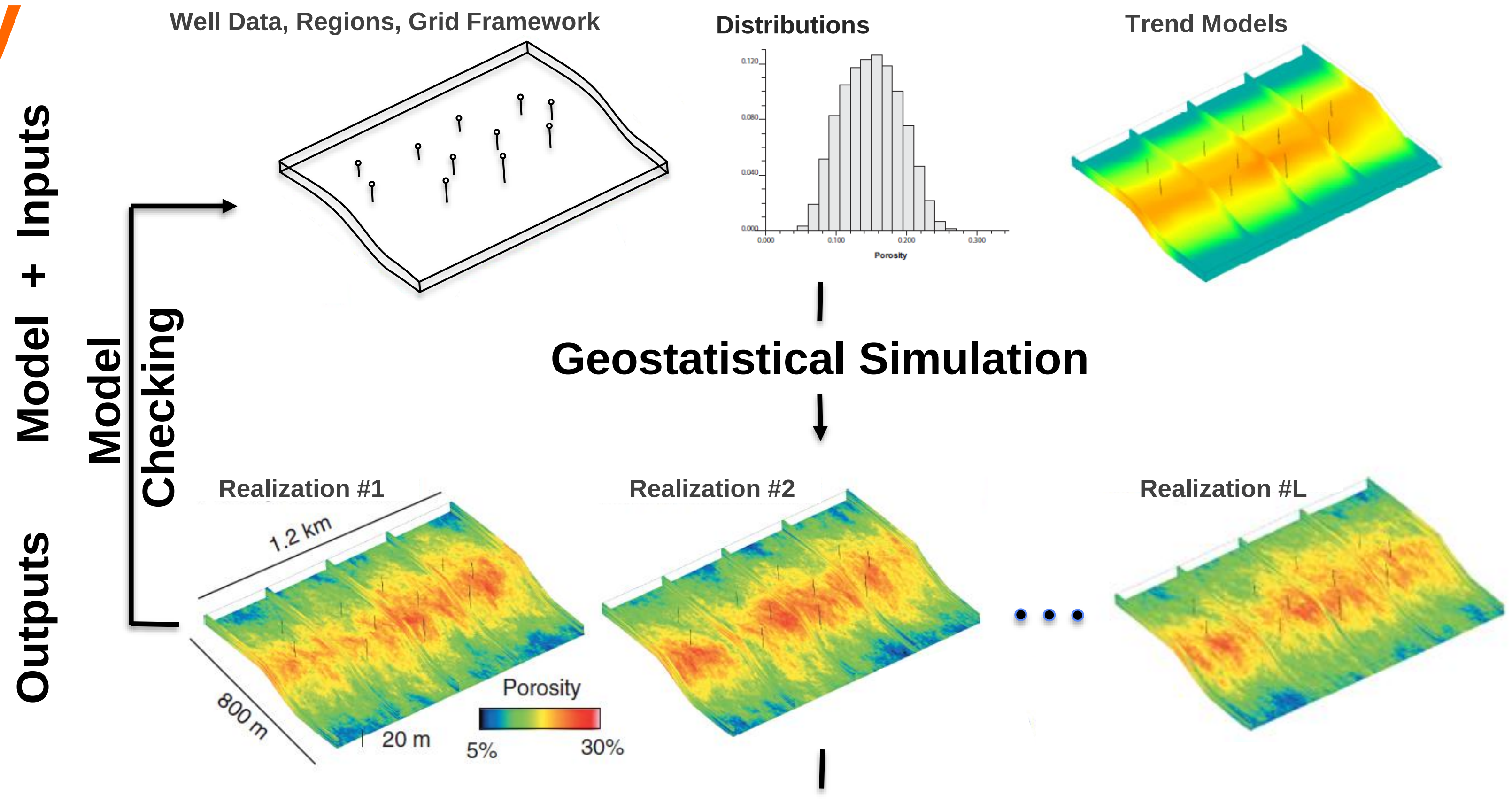
Logical Consistency

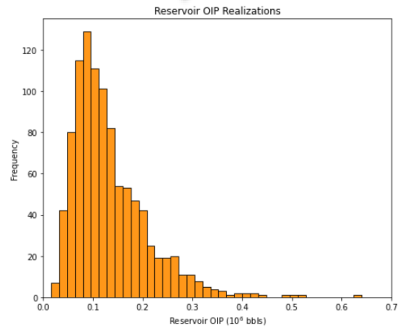
- flow of information, model choices all make sense

Model Consistency

- data and model
- prior and likelihood
- assumptions and interpretations
- applications and limitations

We battle Against Inconsistency!





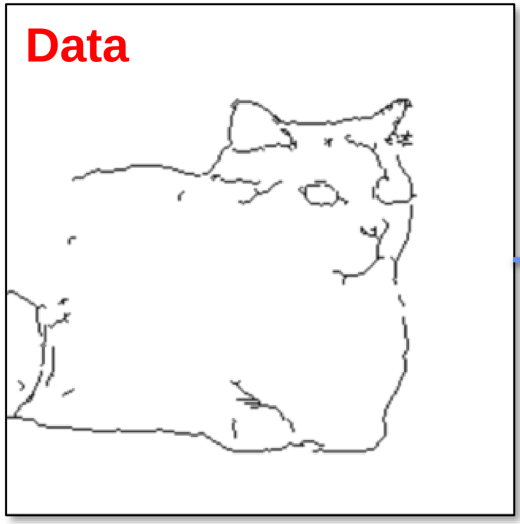
The standard geostatistical modeling workflow, modified from Pyrcz and Deutsch (2014).

Inconsistent Data and Model

Representative inputs and consistent data and model.

Now GenAI

Consistent Data and Model



Edges in a binary, hand drawn image.

pix2pix edges2cats

trained with 2k stock cat photos

Model

?

genAI Cat photo.

Calculated with the online edges2casts, a pix2pix deep learning method (https://affinelayer.com/pixsrv/).

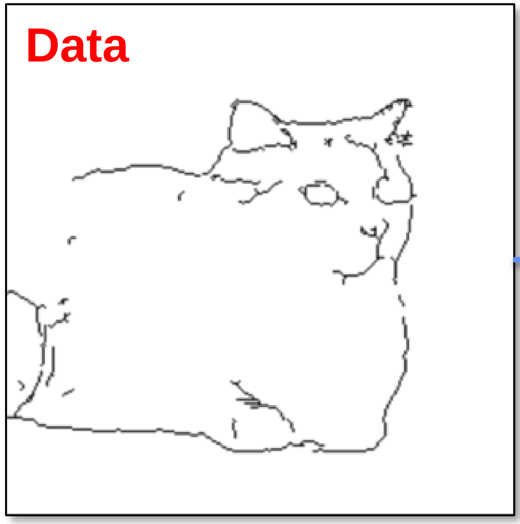
Inconsistent Data and Model

Representative inputs and consistent data and model.

Now GenAI

Consistent
Data and Model

Data



Edges in a binary, hand drawn image.

pix2pix edges2cats

trained with 2k stock cat photos

Model

genAI Cat photo.

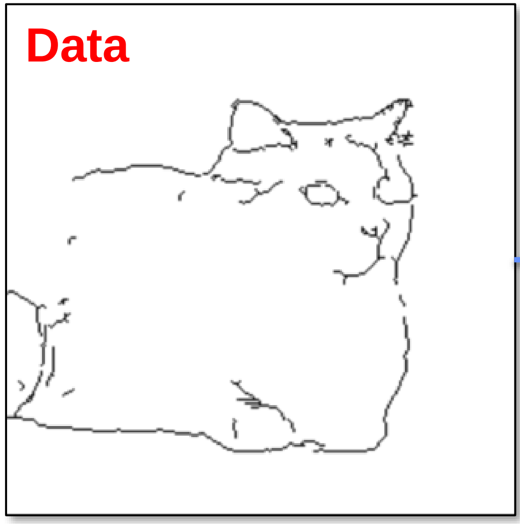
Calculated with the online edges2casts, a pix2pix deep learning method (https://affinelayer.com/pixsrv/).

Inconsistent Data and Model

Representative inputs and consistent data and model.

Now GenAI

Consistent Data and Model



Edges in a binary, hand drawn image.

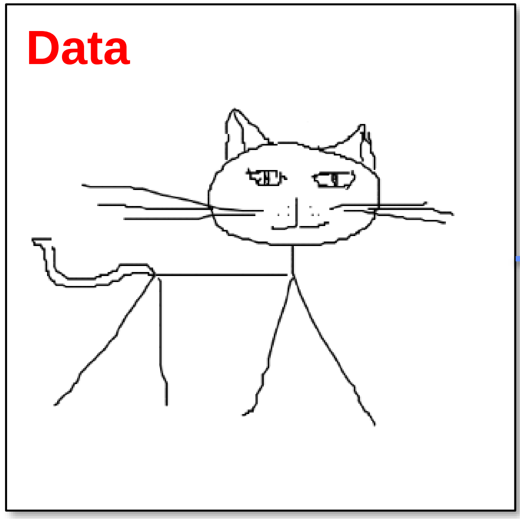
pix2pix edges2cats

trained with 2k stock cat photos

Model

genAI Cat photo.

Nonrepresentative Data



My hand drawn cat edges.

pix2pix

trained with 2k stock cat photos, no stick cats!

Model

?

genAI Cat photo.

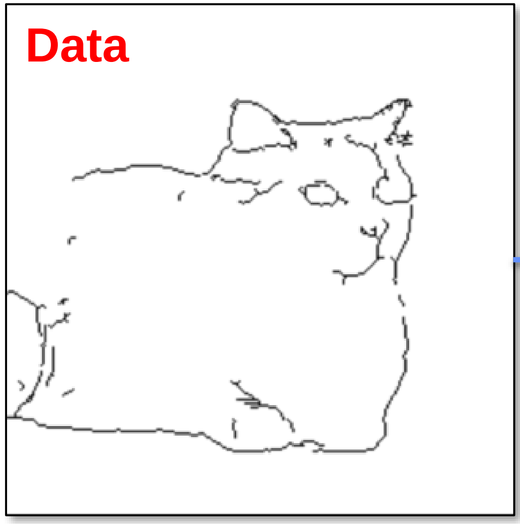
Calculated with the online edges2casts, a pix2pix deep learning method (https://affinelayer.com/pixsrv/).

Inconsistent Data and Model

Representative inputs and consistent data and model.

Now GenAI

Consistent Data and Model



Edges in a binary, hand drawn image.

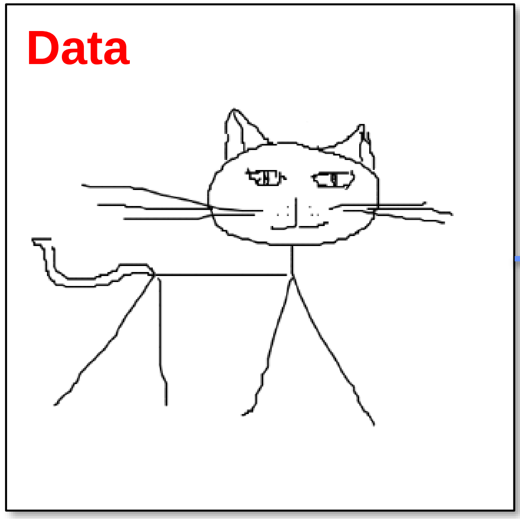
pix2pix edges2cats

trained with 2k stock cat photos

Model

genAI Cat photo.

Nonrepresentative Data



My hand drawn cat edges.

pix2pix

trained with 2k stock cat photos, no stick cats!

Model

genAI Cat photo.

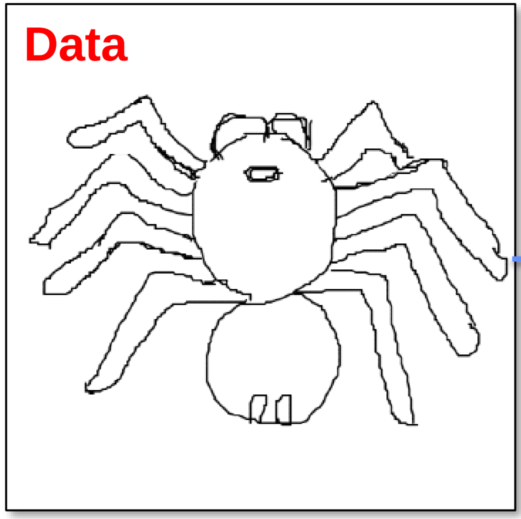
Calculated with the online edges2casts, a pix2pix deep learning method (https://affinelayer.com/pixsrv/).

Inconsistent Data and Model

Representative inputs and consistent data and model.

Now GenAI

Inconsistent Data and Model



Edges in a binary, hand drawn image.

pix2pix

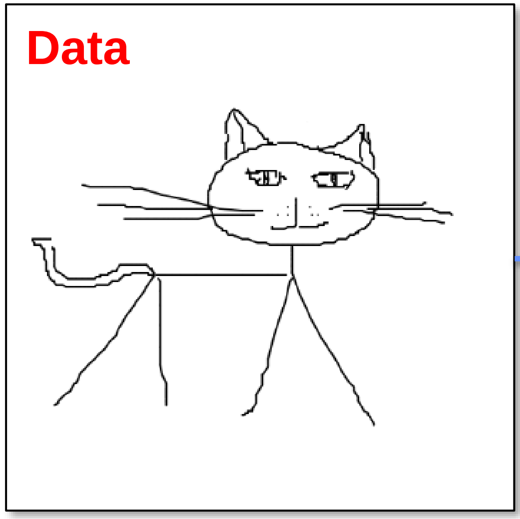
trained with 2k stock cat photos, no spiders!

Model

?

genAI ‘Cat’ photo.

Nonrepresentative Data



My hand drawn cat edges.

pix2pix

trained with 2k stock cat photos, no stick cats!

Model

genAI Cat photo.

Calculated with the online edges2casts, a pix2pix deep learning method (https://affinelayer.com/pixsrv/).

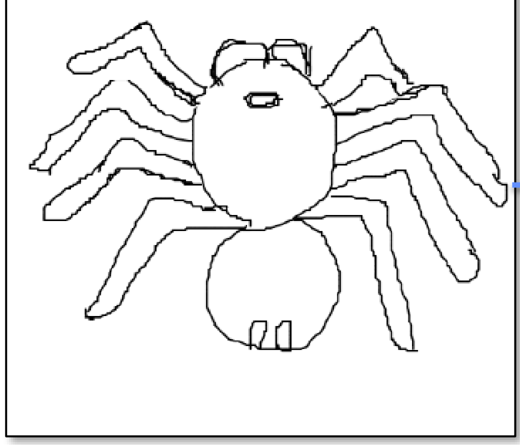
Inconsistent Data and Model

Representative inputs and consistent data and model.

Now GenAI

Inconsistent Data and Model

Data



Edges in a binary, hand drawn image.

pix2pix

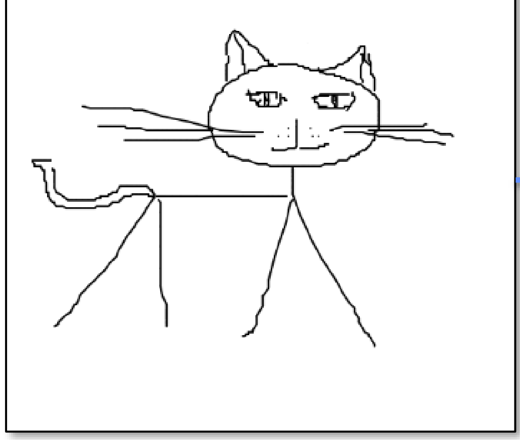
trained with 2k stock cat photos, no spiders!

Model

genAI 'Cat' photo.

Nonrepresentative Data

Data



My hand drawn cat edges.

pix2pix

trained with 2k stock cat photos, no stick cats!

Model

genAI Cat photo.

Calculated with the online edges2casts, a pix2pix deep learning method (https://affinelayer.com/pixsrv/).

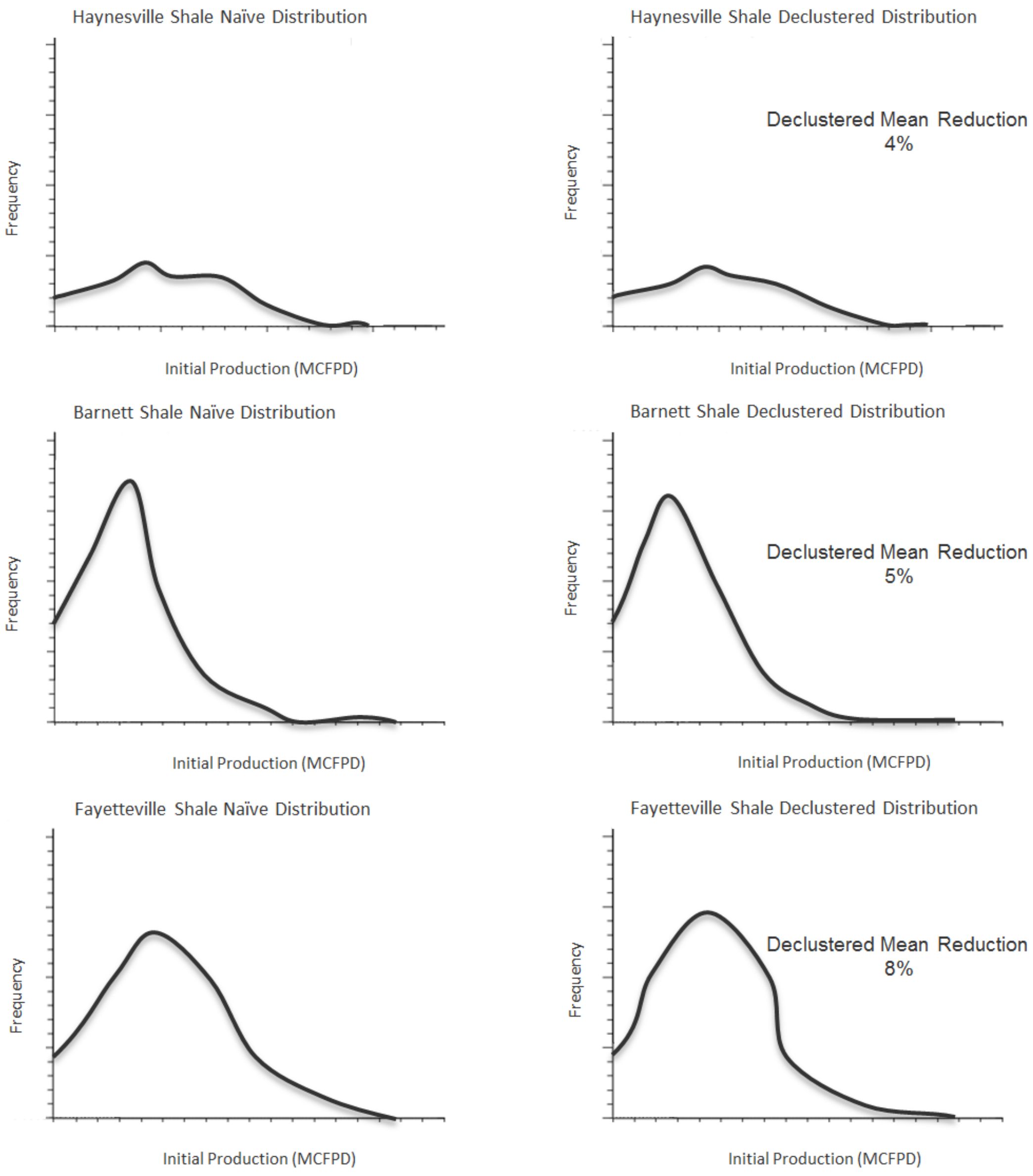
Is Our Data Biased?

Virtually all sparsely sampled, spatial datasets are biased.

Data is collected to maximize value and minimize uncertainty – we should not change this!

Variety of methods have been developed for debiasing.

- data weighting
- data distribution imputation



Naïve and declustered distributions from cell-based declustering (modified from Pyrcz et al. 2017).

Do Our Data-Driven Methods Debias?

Do Data Anchor the Model and debias the Predictions?

- here’s an example with 3 input distributions, and 50 wells.
- we honor the data each time!

Our data alone are rarely sufficient to impose the global distributions, spatial continuity and trends!

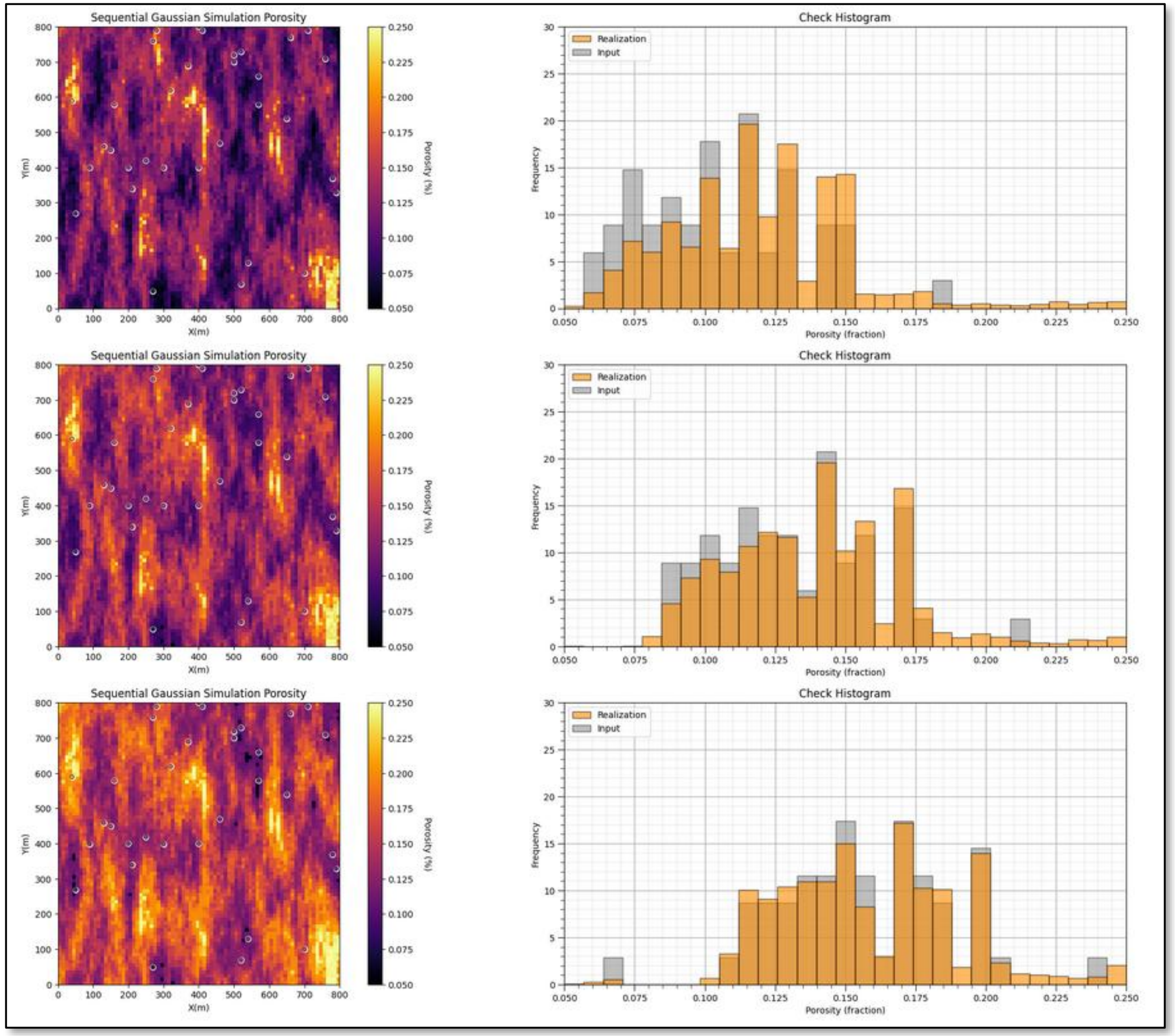
We must take ownership of all the model inputs, they must be modelled!

Sequential Gaussian simulation with 3 reference distributions, file is SGSIM_inputs.ipynb.

Biased Low

Correct Distribution

Biased High



Debias Before Machine Learning

Commonly applied methods included:

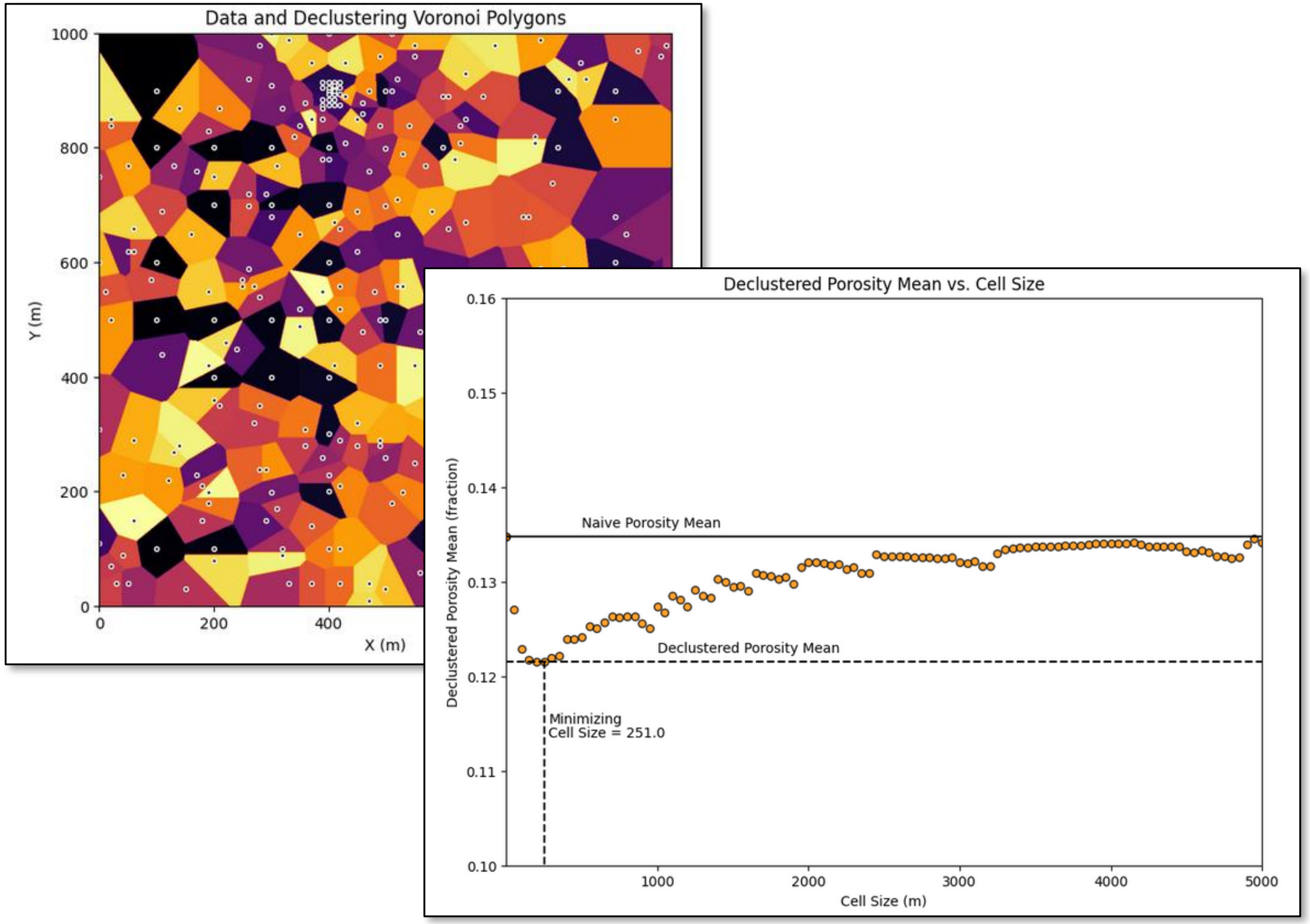
- cell-based (Journel, 1983)
- polygonal (Thiessen, 1911; David, 1977)
- kriging-based (Frykman and Deutsch, 1998)
- soft data debiasing (Deutsch, Frykman and Xie, 1997)

This is the standard procedure in geostatistics.

Debias and Then,

- weight data, balance the training, and / or correct predictions

We assume data is biased unless proven otherwise.



Polygonal declustering (above) and cell-based declustering (below) examples, file is GeostatsPy_declustering_all.ipynb.

What Did We Learn?

Garbage in, garbage out.

Bias in, bias out!

For machine learning we must take ownership of our data and model inputs and debias them and ensure consistency.

PGE 383 Subsurface Machine Learning

Lecture 19: Dr. Pyrcz's Secrets for Success with Machine Learning

Lecture outline:

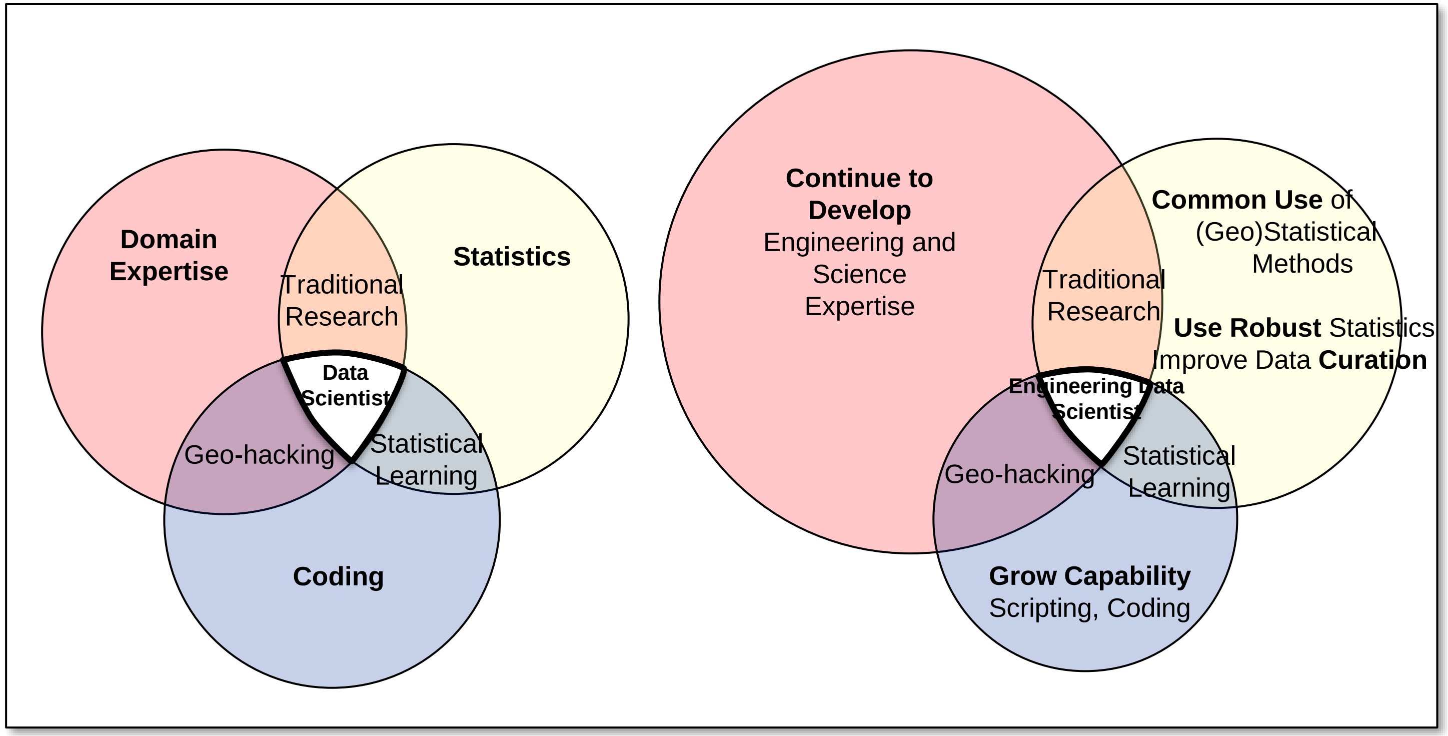
- **Use Our Domain Knowledge**

A Possible Career Path

My Suggestions for You to Add / Improve Your Data Science

An alternative to the data scientist, the 4th paradigm-ready engineer or scientist

- Continue to grow our domain knowledge, engineering
- Build from our knowledge in data analytics and (geo)statistics
- Grow scripting and coding with open-source data analytics and machine learning



The data scientist Venn diagram and a proposed alternative for 4th paradigm ready engineers and geoscientists.

We are building on our geoscience and engineering strengths.

Professional Standards

For Example, From a Canadian Provincial Association's Competancy Exam Outline

22 Key Competencies & Indicators include:

- **Risk Management for Technical Work** – identify risks and impact of risks, risk mitigation *stress testing, black swans, failure analysis*

- **Application of Theory** – ability to apply engineering theory *math, statistics, physics*

- **Solution Techniques & Results Verification** – understand the engineering principles in computer aided solutions *never just run the program*

Watch it here,

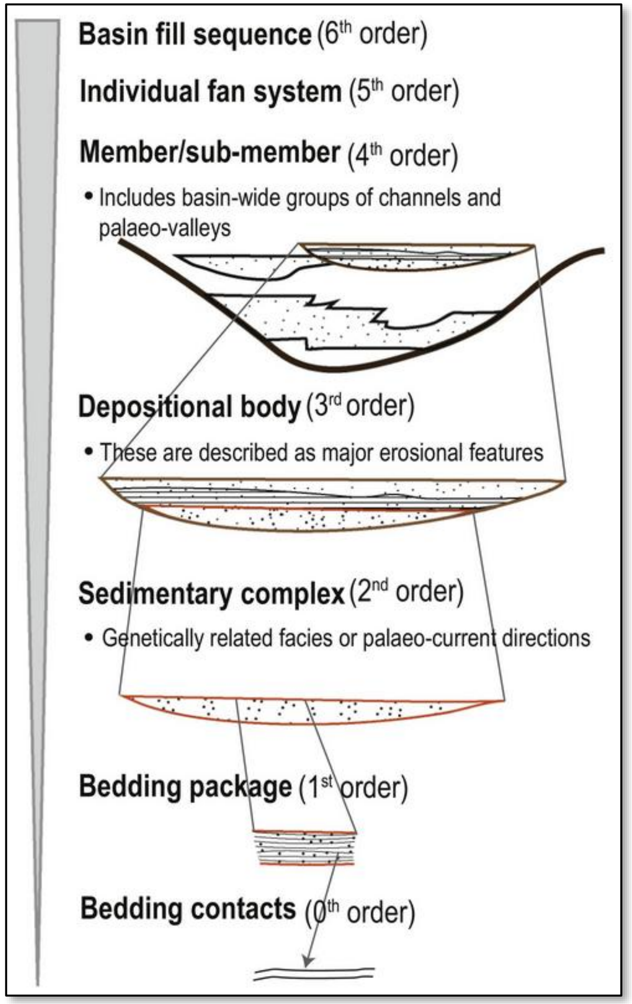
Canadian engineering ring (above) and geoscience ring (below).

Dr. Pyrcz's lecture, "Applying Machine Learning as a competent Engineer or Geoscientist" https://youtu.be/W_ZDg1Wb2vM?si=COVW2z5hksw7IWXP

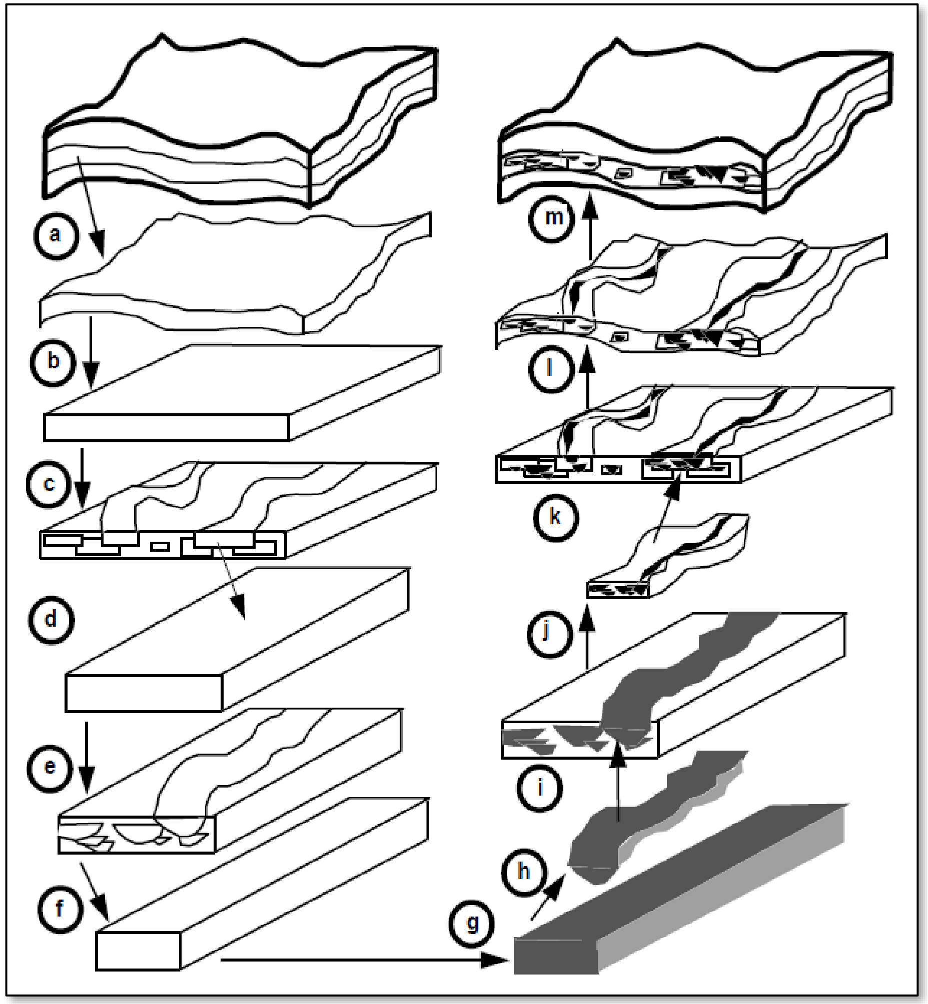
Geological Knowledge, Hierarchical Modeling

Interacting Allogenic and Autogenic Forces

- Result in complicated, multiscale heterogeneities.
- We typically cannot just model one scale of the subsurface.
- Our machine learning models must be multiscale.



Deepwater lobes hierarchical architecture (Cullis et al., 2018).



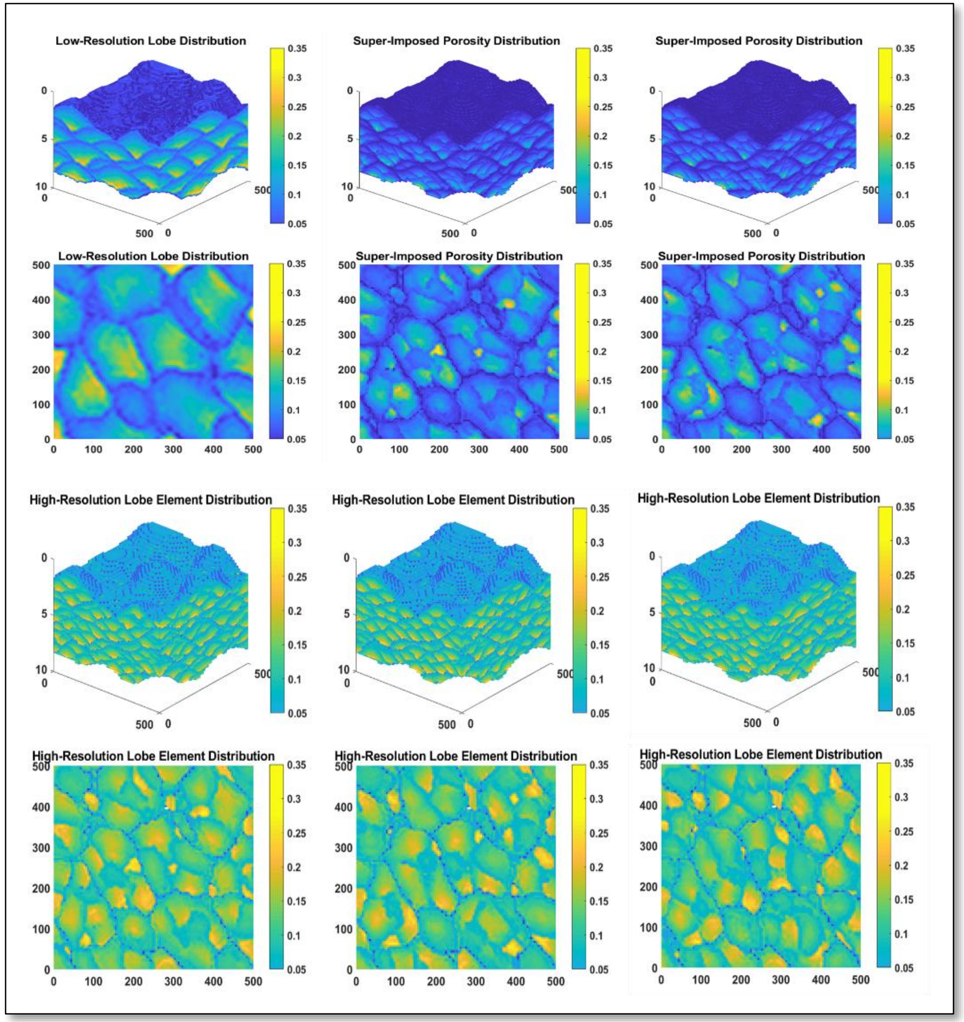
The conceptual approach to hierarchical object-based modeling (Pyrcz and Deutsch, 2004).

Geological Knowledge, Hierarchical Modeling

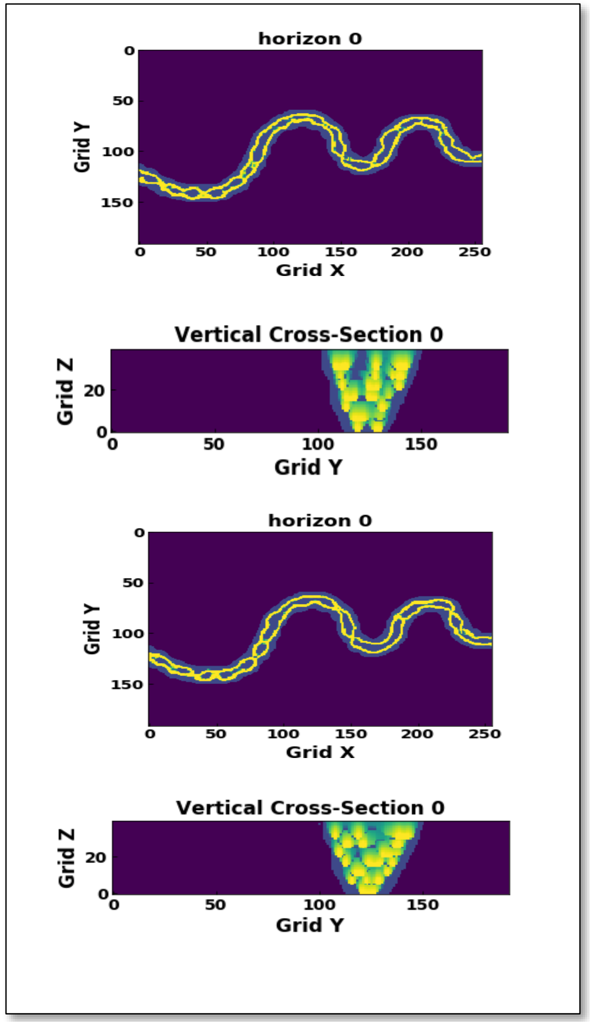
Deep Learning-based Generative Top-Down Models

- start by modeling the largest scale
- fill in next smaller scale conditional to the current scale
- repeat over all required scale

Use Hierarchical Machine Learning Method!



Hierarchical lobe modeling example (after Pan, et al., 2022).



Hierarchical channel modeling example.

Engineering Knowledge, Optimization

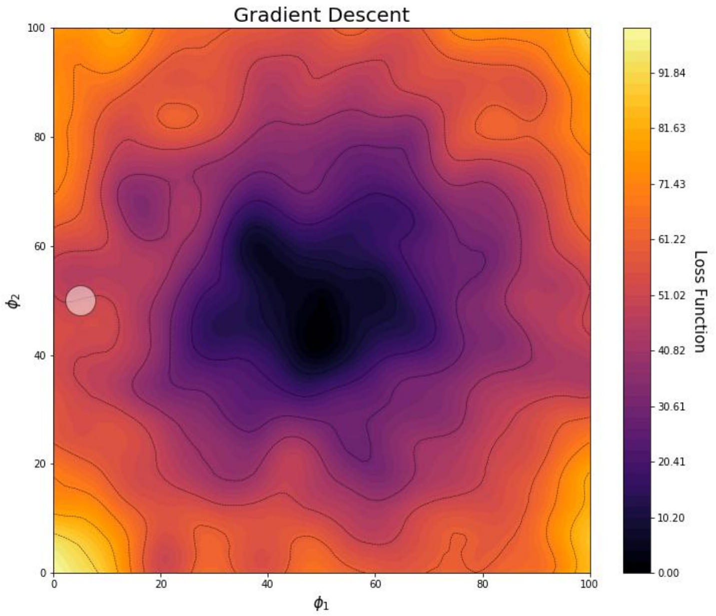
Many Machine Learning Methods are Solved by Optimization.

- Loss functions and norms
- Iterative solution schemes
- LASSO, ANN, deep learning, etc.

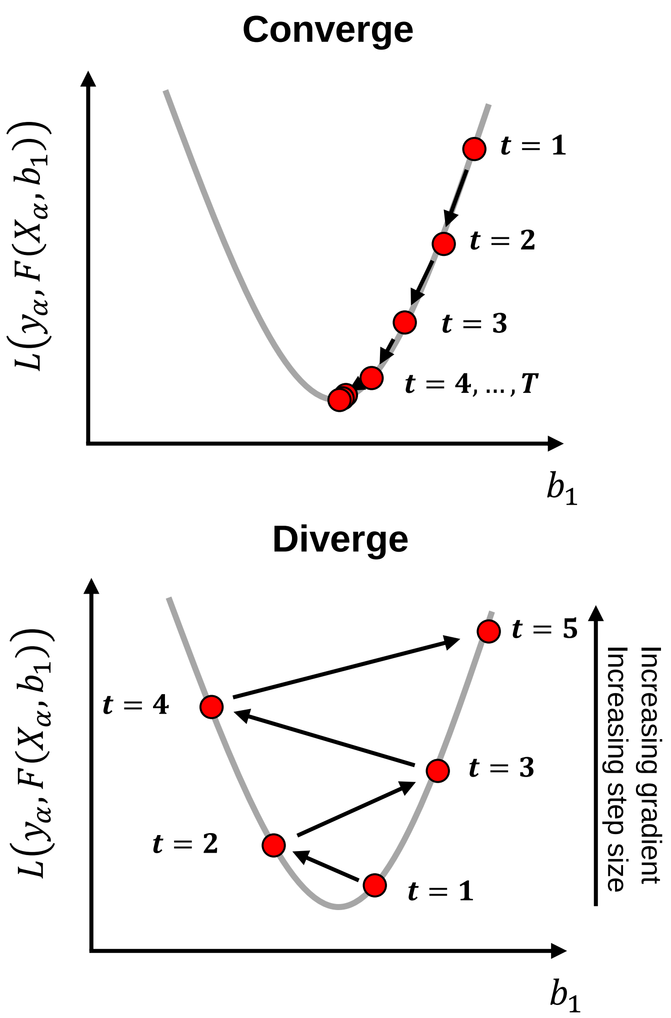
And With Solution Heuristics

- K-means clustering, tree-based methods, etc.

Engineers have a head start with machine learning trade-craft.



Gradient descent demonstration in Python.
Available on GeostatsGuy GitHub
Interactive_Optimization.ipynb



Fundamental Logic, Probability

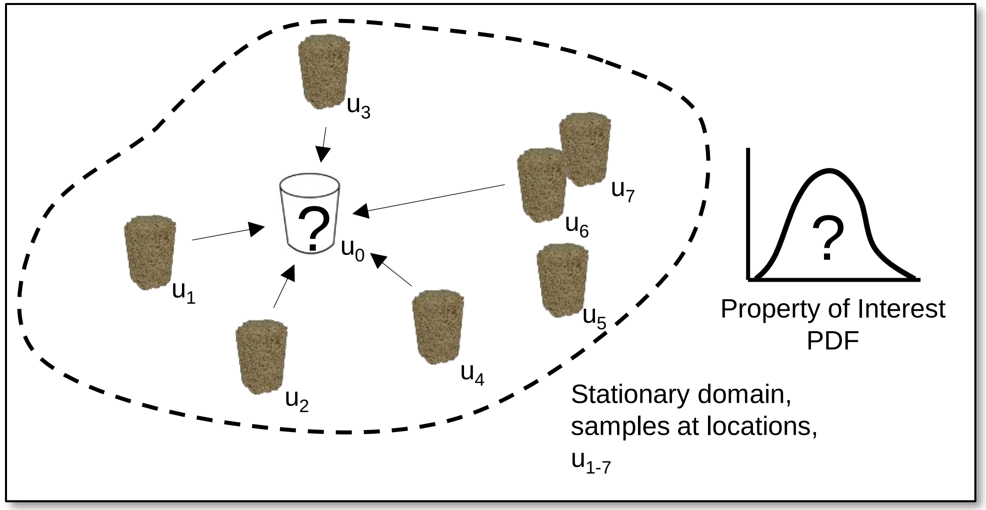
Probability the Language of Uncertainty

- uncertainty is unavoidable
- no objective uncertainty
- uncertainty in the uncertainty?

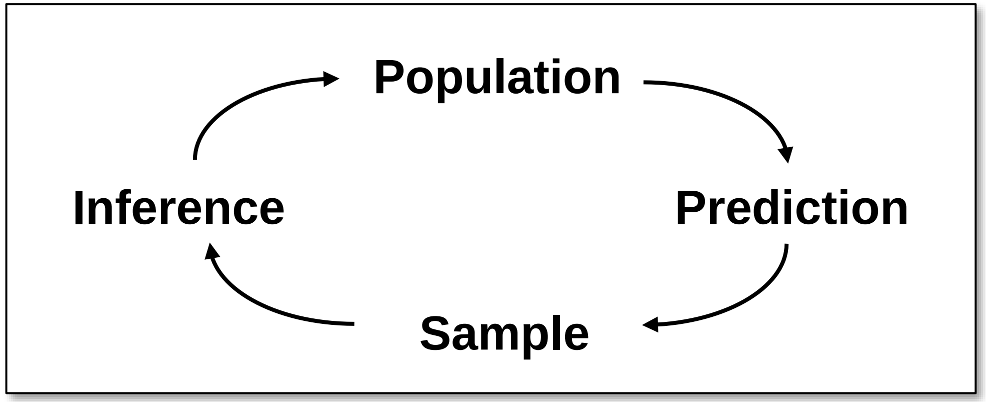
In General Inference Precedes Prediction

Inference - model of the population from the sample

Prediction - predict the next sample from the model of the population



Uncertainty Modeling.



Inference to model the population from a limited sample, prediction to predict the next sample from the model of the population.

What Did We Learn?

Your domain expertise as a scientist or engineer is

the most important foundation for data science.

Integrated, plausible models and,

Machine learning concepts build on our knowledge as engineers and scientists, we are prepared for this!

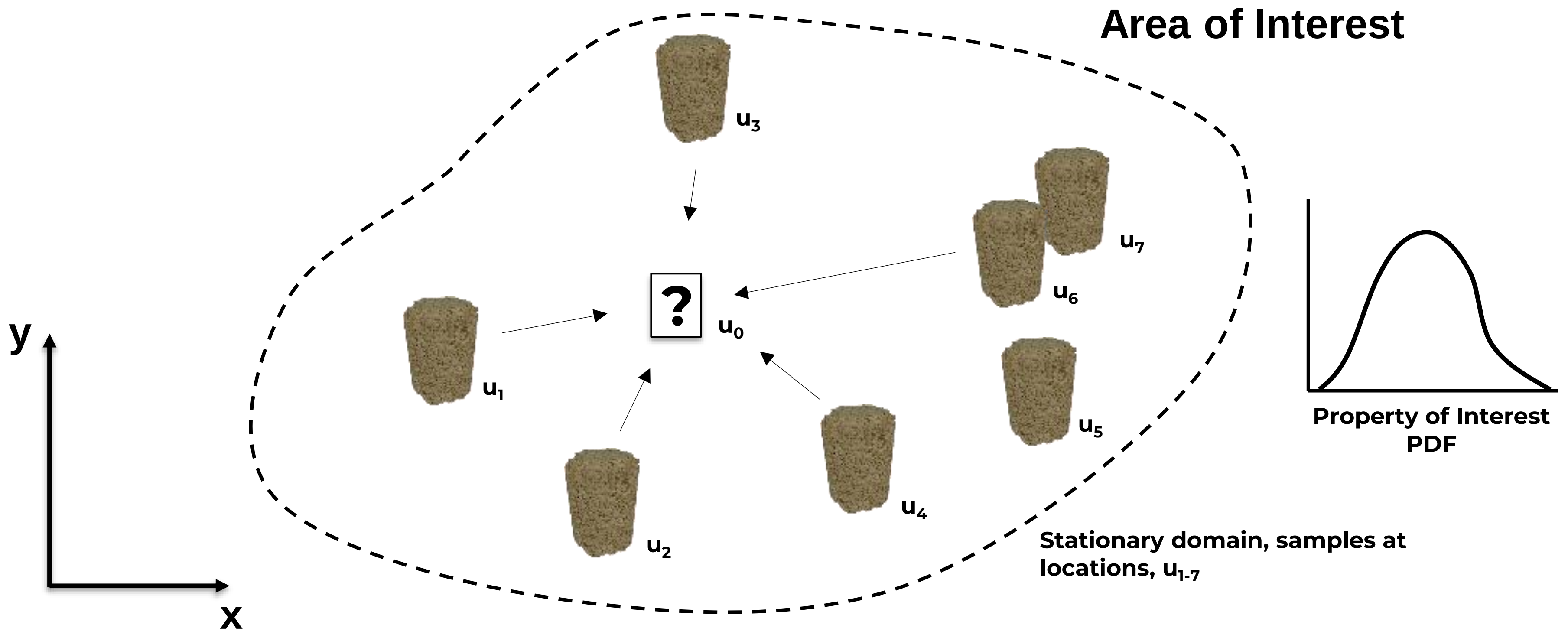
PGE 383 Subsurface Machine Learning

Lecture 19: Dr. Pyrcz's Secrets for Success with Machine Learning

Lecture outline:

- **Uncertainty Always**

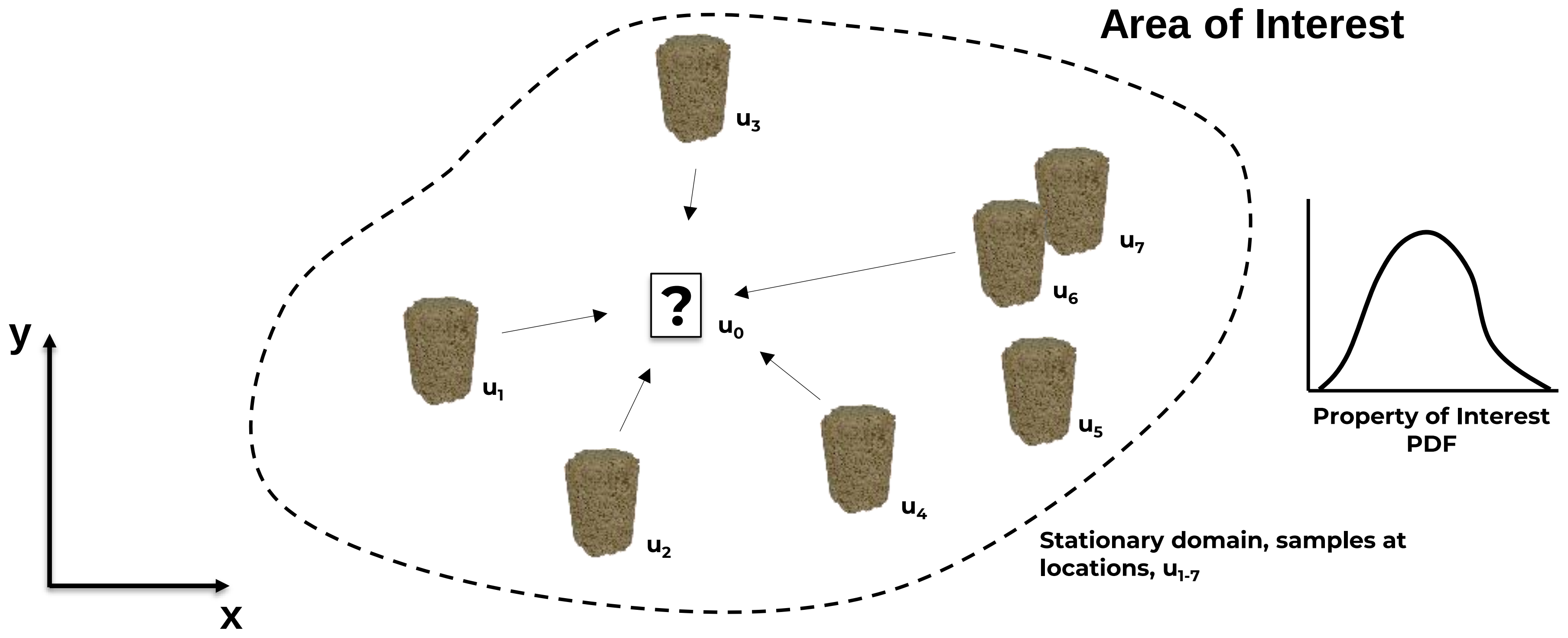
What is Uncertainty?



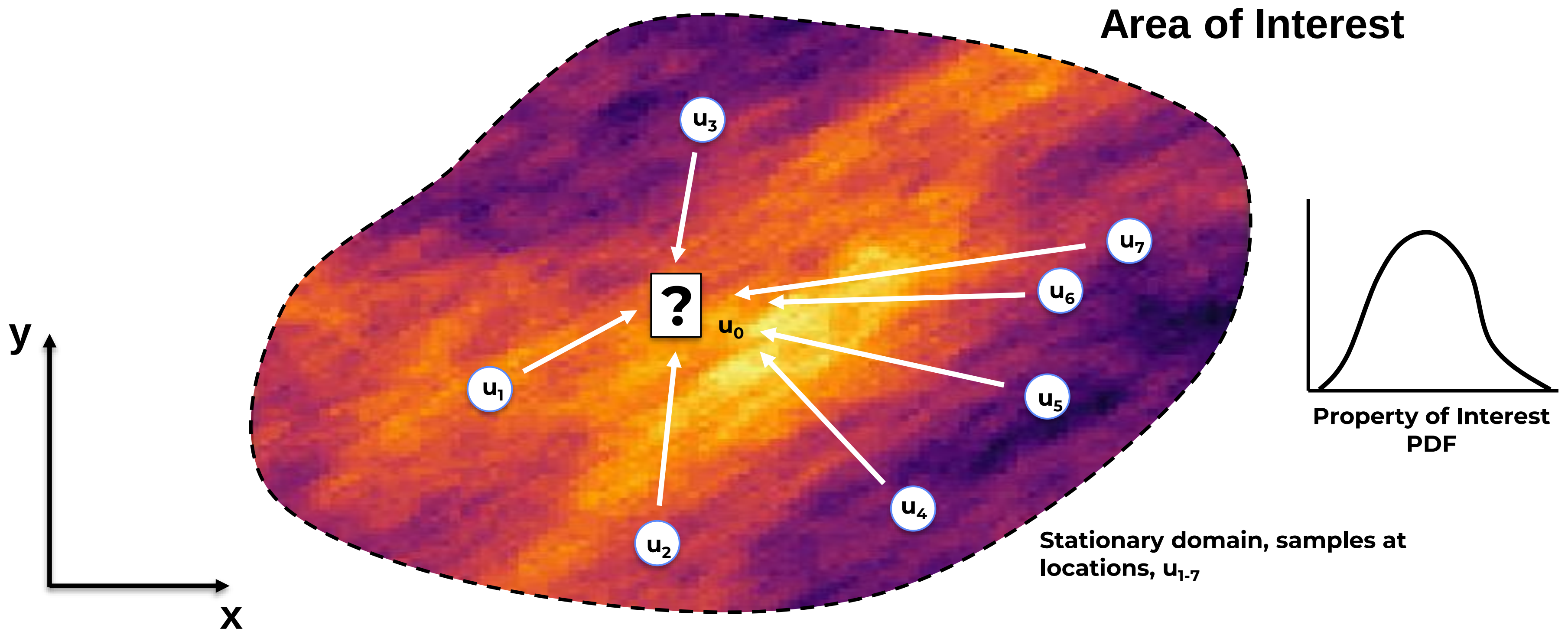
Uncertainty is not an intrinsic property of the subsurface.

- At every location ($\mathbf{u}_\alpha$) within the volume of interest the true properties could be measured if we had access (facies, porosity, etc.).
- **Uncertainty is a function of our ignorance**, our inability to observe and measure the subsurface with the coverage and scale required to support our scientific questions and decision making.

What Causes Uncertainty?



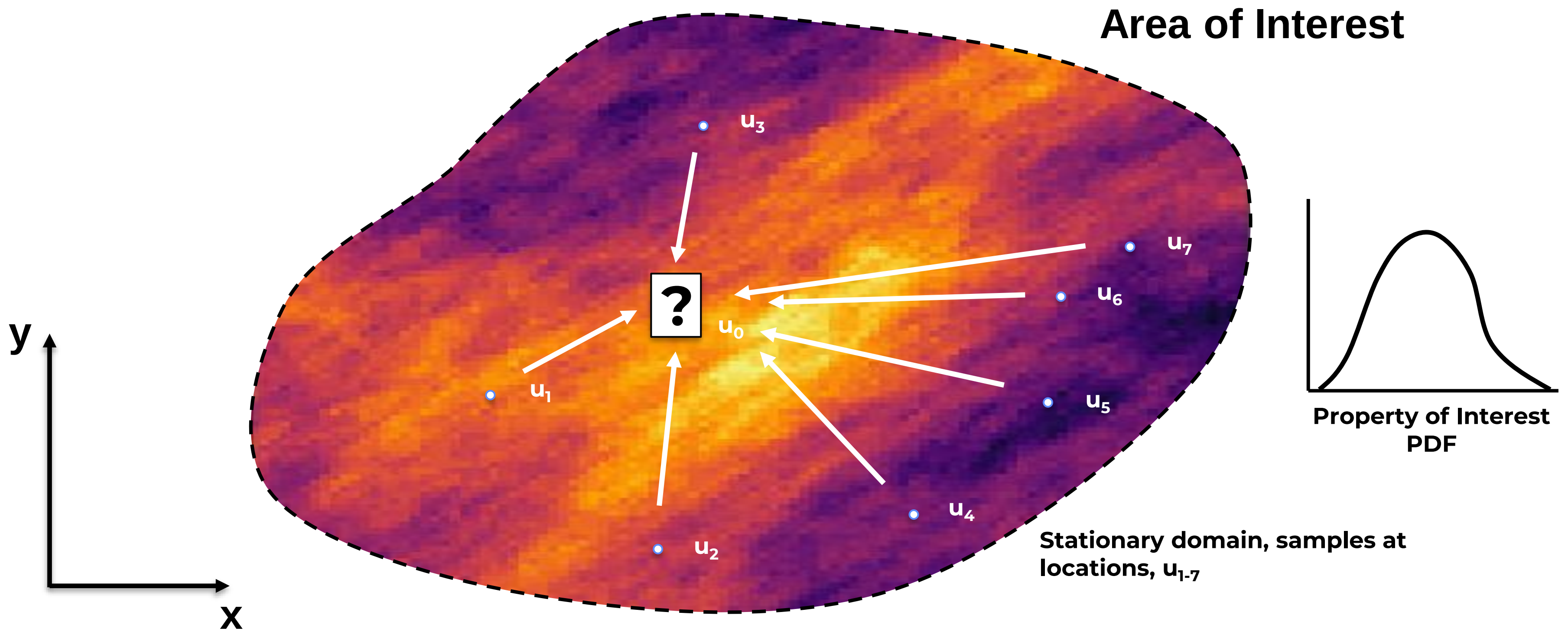
What Causes Uncertainty?



Presence of Spatial and Temporal (e.g., Operational Constraints) Heterogeneity

- The features of interest changes over the area of interest, geology is heterogeneous.
- If the subsurface was homogeneous, with a few measurements, uncertainty would be reduced and estimates resolved to a sufficient degree of certainty.

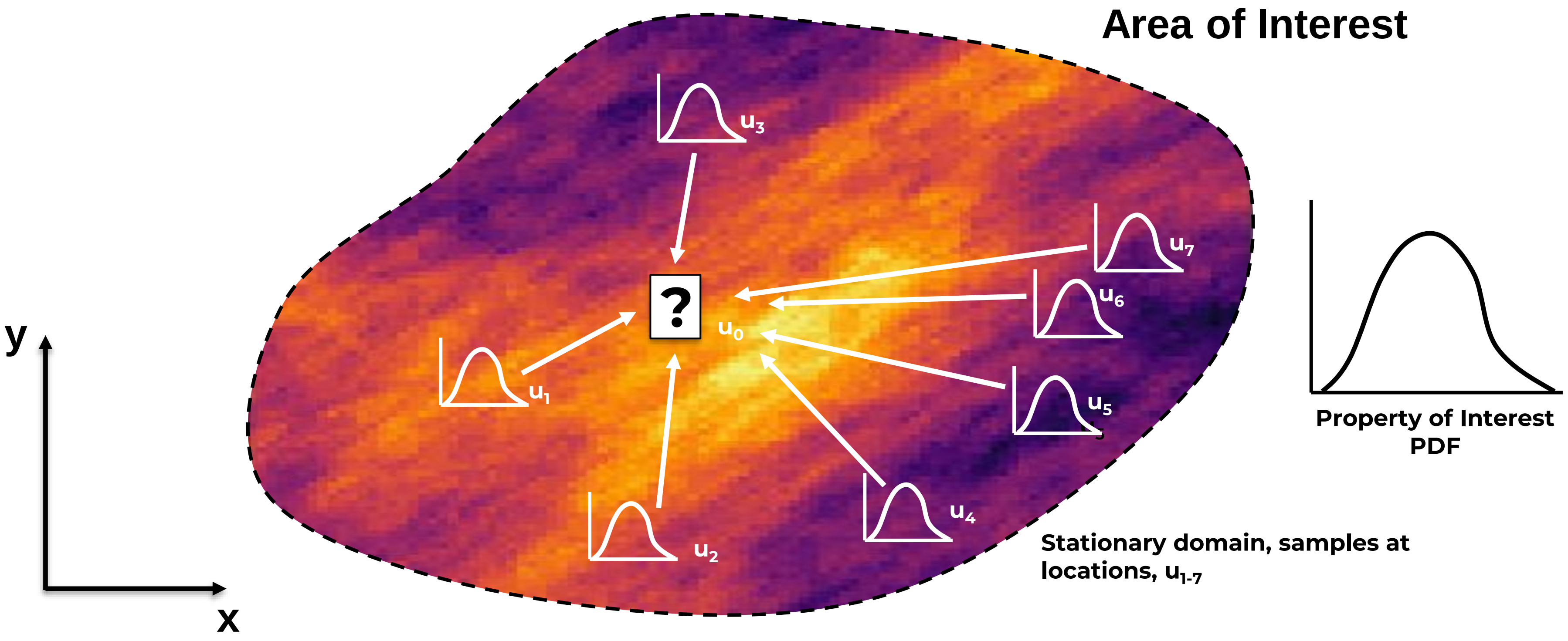
What Causes Uncertainty?



Volume Support / Scale, Missing Scale and Sample Paucity

- Due to the volume support of our direct samples, our coverage of the population is typically, $\frac{1}{100{,}000{,}000{,}000}$!
- The difference between the scale of our direct samples and the volume support of our models results in the **"missing scale"** phenomenon.

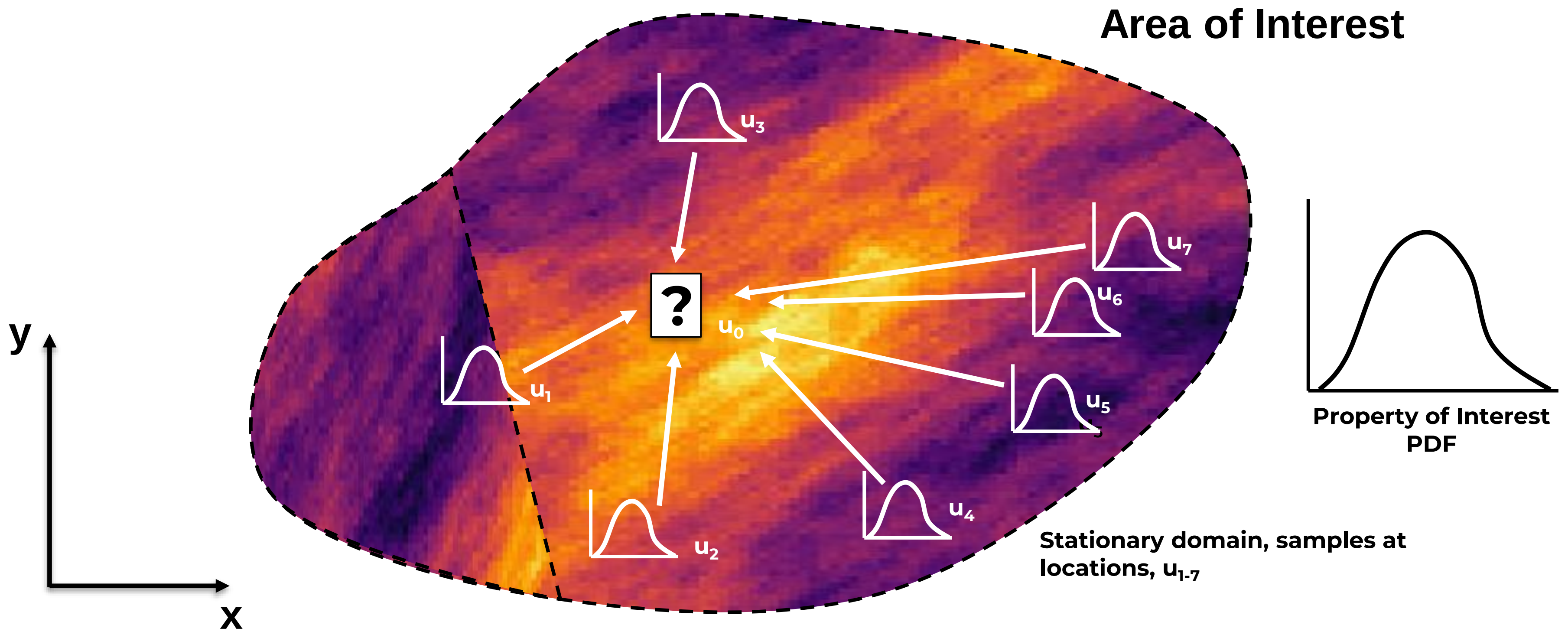
What Causes Uncertainty?



Measurement Error

- There are many sources of error in our measured features, machine precision, sample disturbance, irreducible interpretations, etc.
- As a result, **none of our data is considered "hard data"**!

What Causes Uncertainty?



Nonstationarity, Changes in the Population

- Geological phenomenon are intrinsically nonstationary. Away from our data, the underlying population may change.
- Yet, many of our spatial modeling methods assume stationarity or stationarity residual after detrending.

How Do We Model Uncertainty?

Methods to Model Uncertainty, For Example,

- Diverse training data, prior models, regularization
- Scenarios and cases

Sample Paucity / Global Inference

- Random predictor features

Measurement Error

- Realizations from,
  - bootstrap and Monte Carlo Simulation
  - MCMC
  - deep learning drop out

Model and Interpolation Uncertainty

Determine Sources of Uncertainty, then Integrate Salient Uncertainties

What Did We Learn?

Be an uncertainty sleuth

Discovery uncertainty sources and design workflow

to integrate each salient uncertainty source.

A good uncertainty model is more important than a single, best estimate!

PGE 383 Subsurface Machine Learning

Lecture 19: Dr. Pyrcz's Secrets for Success with Machine Learning

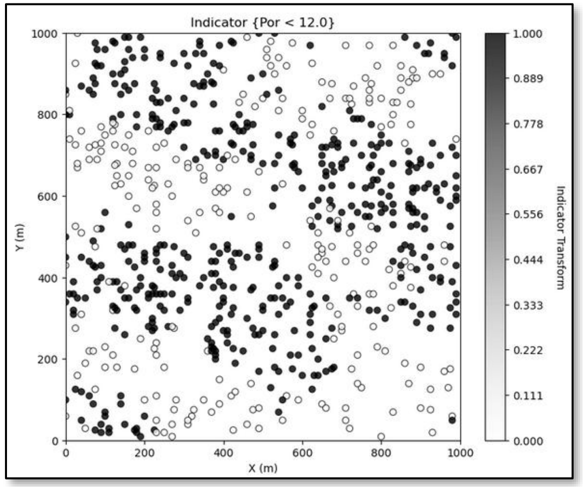
Lecture outline:

- **Make Powerful Plots**

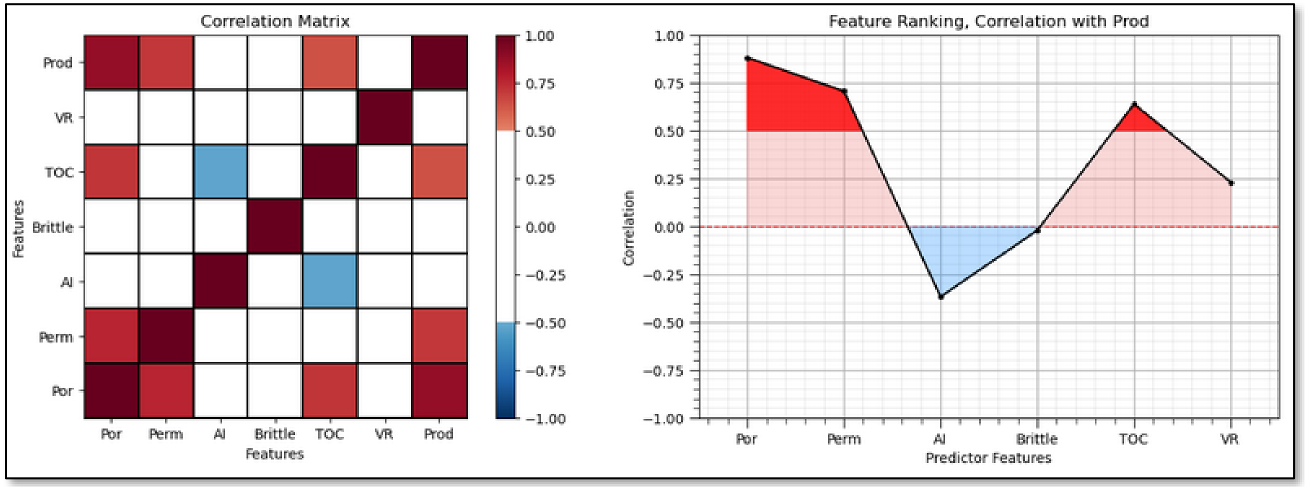
Effective Visualization

Design Your Visualization for Efficient Communication

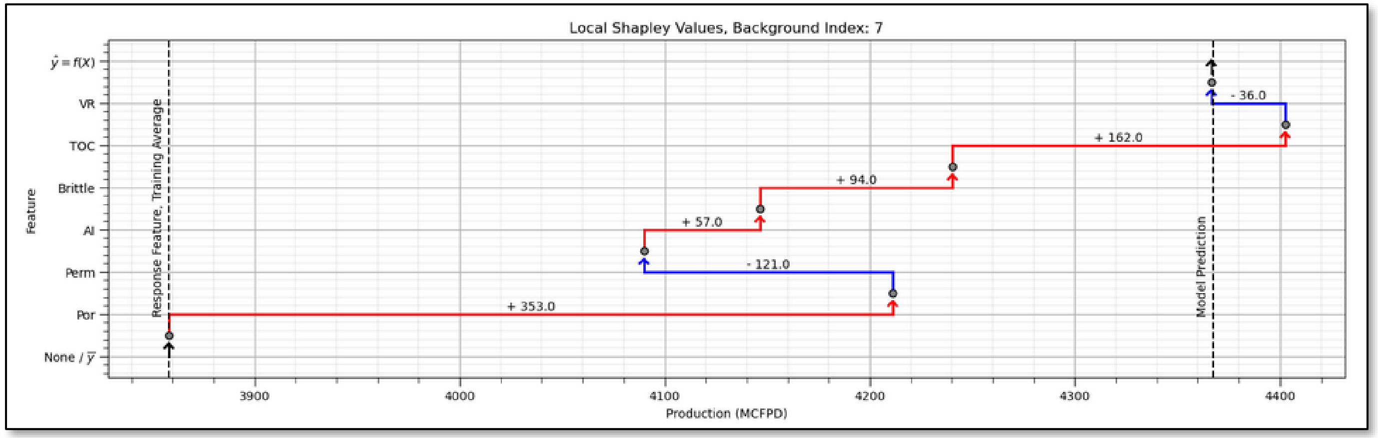
- never use the default plot!
- feature engineering
- modify, highlight, annotate
- even design from scratch



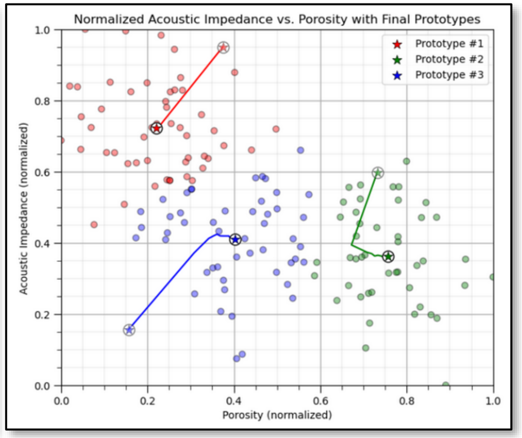
Location map with indicator transform.



Correlation matrix and predictor and response correlation for feature ranking.



Shapley values for a single prediction.



K-means clustering heuristic.

PGE 383 Subsurface Machine Learning

Lecture 19: Dr. Pyrcz's Secrets for Success with Machine Learning

Lecture outline:

- **Educate and Deploy**

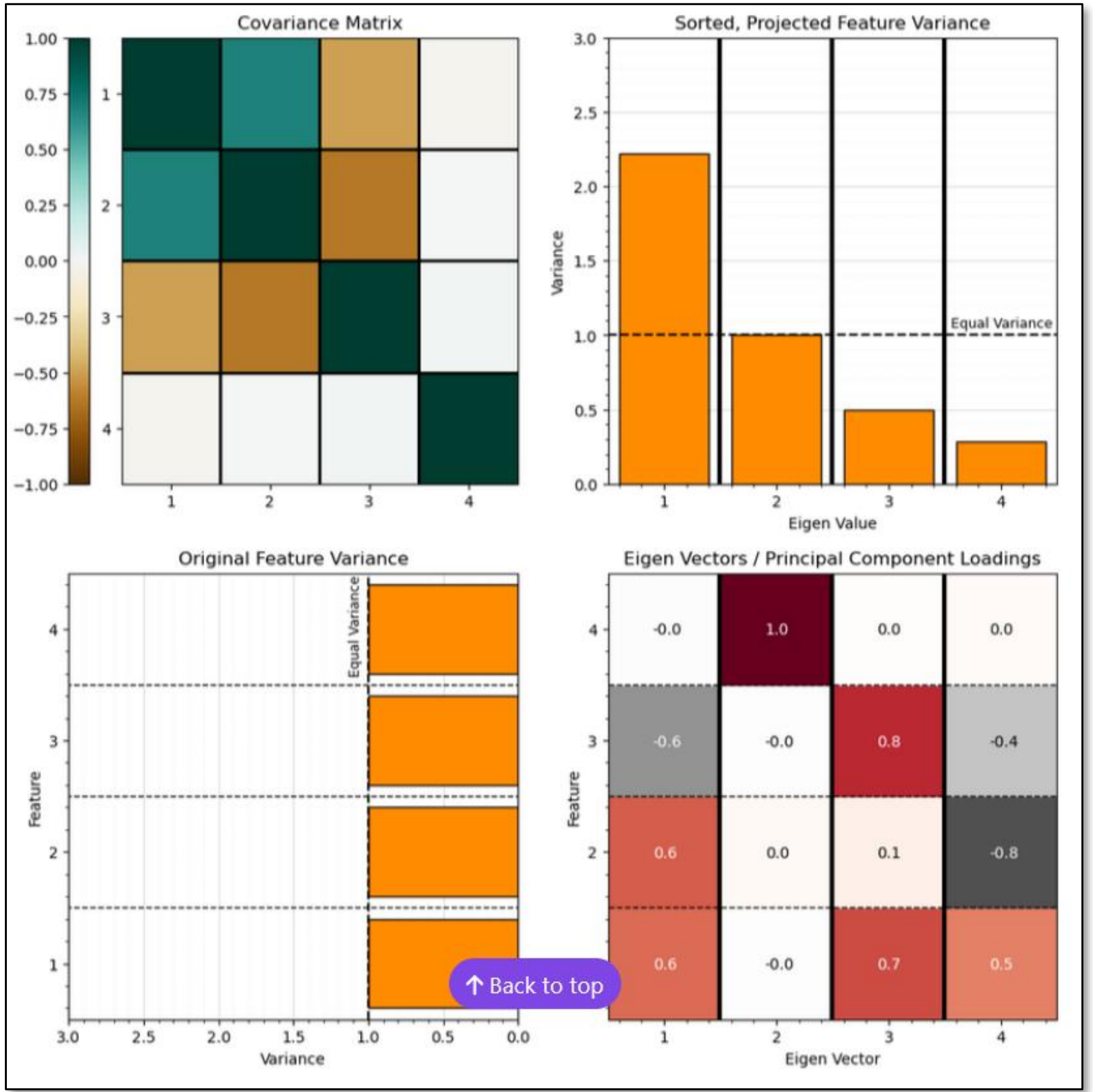
Effective Visualization

Educate Your Audience

- science is a social activity, we experts serve and support others, inform, upskill and lift by sharing your knowledge
- many of us are ‘visual learners’
- use your plots to educate your peers

Deploy your Workflows

- no one uses tools they don’t understand, build you customer based and multiple your efforts!



Principal components analysis, covariance matrix (upper left), original feature variances (lower left), component loadings (lower right), component variances (upper right).

Open Source, Intra-Company Contributions

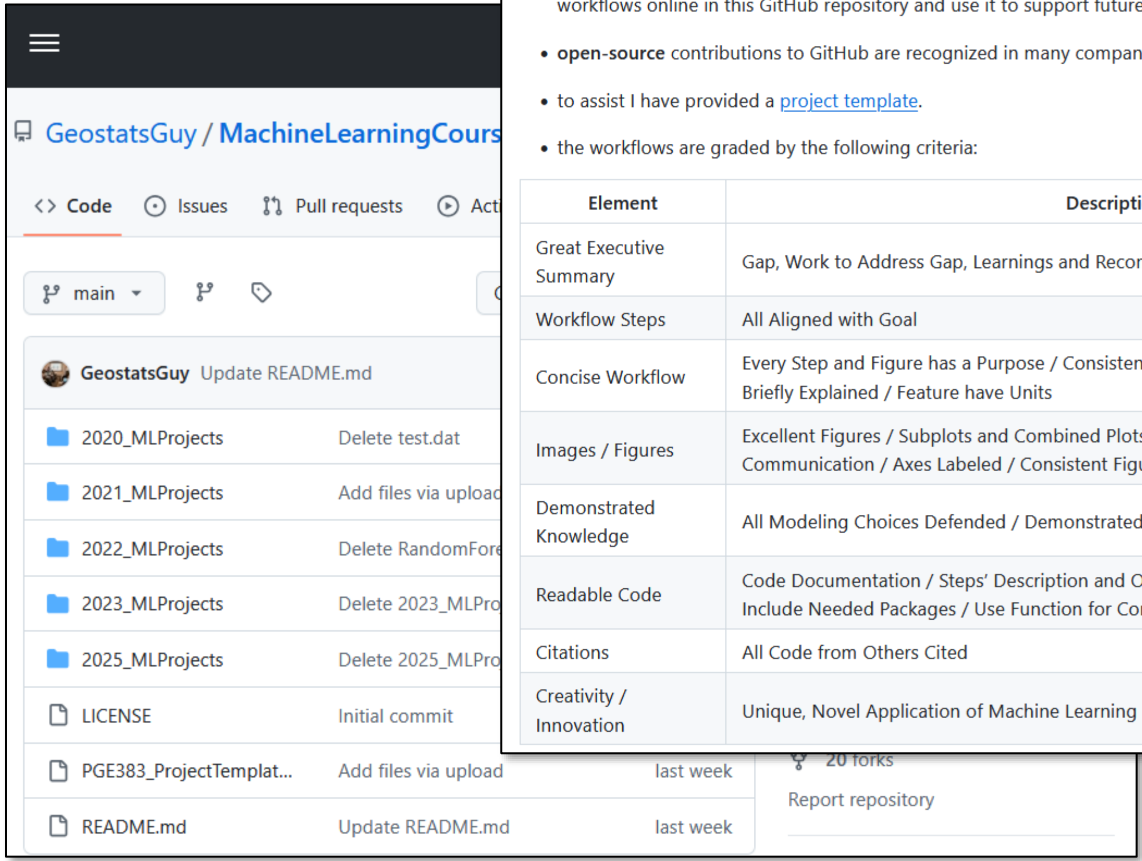
The Final Projects

- yes, you did useful work that I may use on my course with citation to you
- but, I did it for you

Educate and Deploy, Deploy and Educate

- well-documented workflows
- teach others
- build content, open-source if you can

The way you are valued inside the company is proportional to the way you are valued outside the company.



Machine Learning Projects

As part of the course, all students complete a machine learning project.

The challenge: build a well-documented, educational machine learning workflow.

Here's the motivation and more details:

- **produce a comprehensive, concise, well-documented, machine learning workflow in a Jupyter Notebook**. Opportunity to apply course learnings and demonstrate a high level of proficiency. With permission, I post the workflows online in this GitHub repository and use it to support future classes (with credit).
- **open-source** contributions to GitHub are recognized in many companies.
- to assist I have provided a project template.
- the workflows are graded by the following criteria:

Element	Description
Great Executive Summary	Gap, Work to Address Gap, Learnings and Recommendation
Workflow Steps	All Aligned with Goal
Concise Workflow	Every Step and Figure has a Purpose / Consistent with Provided Template / Features Briefly Explained / Feature have Units
Images / Figures	Excellent Figures / Subplots and Combined Plots for Efficient Displays and Communication / Axes Labeled / Consistent Figure Sizes
Demonstrated Knowledge	All Modeling Choices Defended / Demonstrated Extension of Knowledge
Readable Code	Code Documentation / Steps' Description and Observations between Code Blocks / Only Include Needed Packages / Use Function for Concise Code
Citations	All Code from Others Cited
Creativity / Innovation	Unique, Novel Application of Machine Learning

Course GitHub with 5 years of final projects.

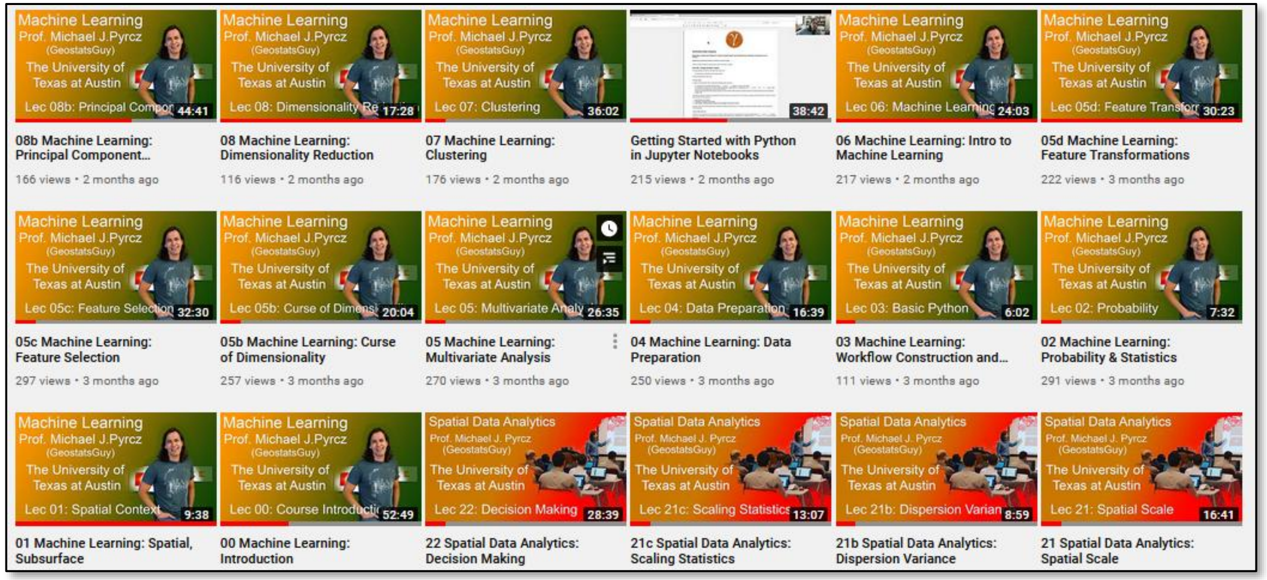
Thank You!

You Very Much Improved this Course with Your Feedback

- Thank you for your patience and your contributions to the course.

The Door was Open, and the Door Remains Open

- Got a project? Need a committee member? Need a reference letter?
- I'll keep putting out resources to support you even after you leave UT Austin!
- I'm on a mission to build new skills at UT and in our industry.



Machine Learning playlist on the 'GeostatsGuy Lectures' YouTube channel.

PGE 383 Subsurface Machine Learning

Lecture 19: Dr. Pyrcz's Secrets for Success with Machine Learning

Lecture outline:

- **Set the Stage**
- **Use Our Geostatistics Domain Knowledge**
- **Use Our Domain Knowledge**
- **Uncertainty Always**
- **Make Powerful Plots**
- **Educate and Deploy**