What Does a Geoscientist Need to Know About Geostatistics?

and Why It Would Be Helpful?

Michael Pyrcz, Associate Professor
Texas Center for Geostatistics
Hildebrand Department of Petroleum and Geosystems Engineering
Bureau of Economic Geology, Jackson School of Geosciences
The University of Texas at Austin

Lecture Goals

A Introduction and Warm-up with Critical Geostatistics Concepts

High grade and **cover** those **fundamental concepts** that can impact your daily work.

- ✓ **New Concepts** → **New Opportunities**
- ✓ **Cross Discipline Expertise** → **Communication / Integration**
- ✓ **Impact Your Work** → **New Workflows and Research**

Outline

1. **Definitions**
   - geostatistics / reservoir modeling
   - big data
   - complexity

2. **Fundamental concepts.**
   - big data and machine learning
   - stationarity
   - bias
   - uncertainty
   - facies
   - spatial continuity

3. **Some Motivational Examples**

(Geo)statistics
Some Definitions

Statistics is concerned with mathematical methods for collecting, organizing, and interpreting data, as well as drawing conclusions and making reasonable decisions on the basis of such analysis.

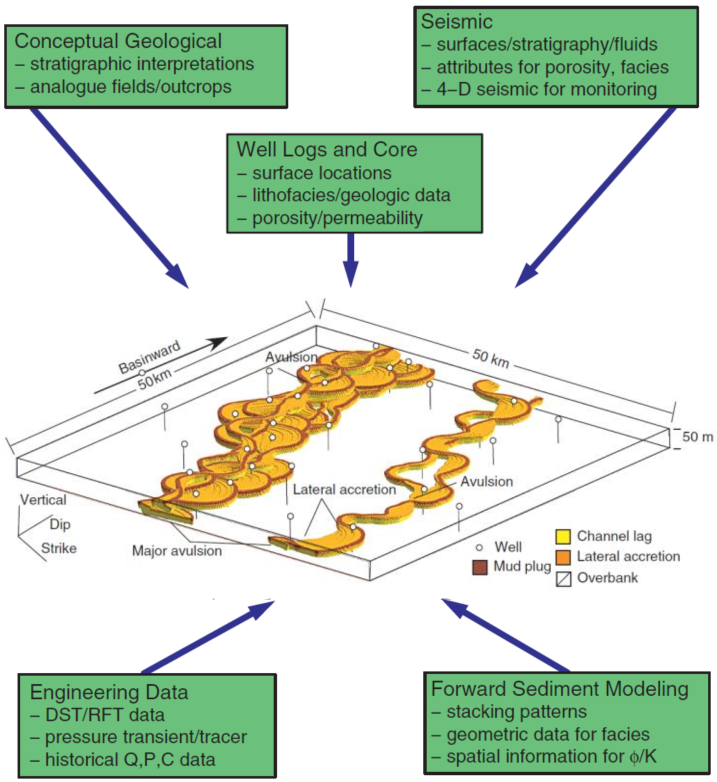
Geostatistics is a branch of applied statistics that emphasizes (1) the geological context of the data, (2) the spatial relationship between data, (3) spatial uncertainty and (4) the different volumetric support and precision of the data.

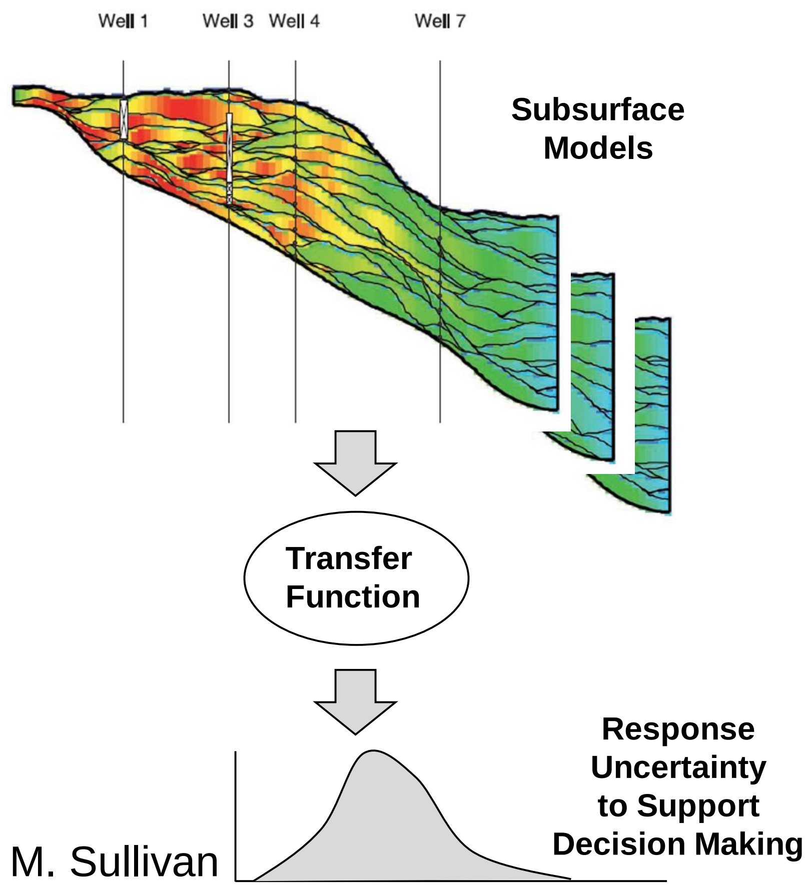
Why do we work with geostatistics in Geosciences?

- ✓ **Geological Context**
- ✓ **Spatial Relationships**
- ✓ **Variable Scale of Data**
- ✓ **Variable Data Precision**
- ✓ **Highly Multivariable**

(Geo)statistics
Some Definitions

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





‘If it doesn’t get in the model, it doesn’t matter!’ – M. Sullivan

(Geo)statistics
Some Definitions

Geostatistics developed from **practice of subsurface estimation and modeling** in mining, theory added later.

TABLE 2.1. RESERVOIR CONCEPTS AND ASSOCIATED GEOLOGICAL AND GEOSTATISTICAL EXPRESSIONS

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 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

[a]Most geostatistical constructs can be directly mapped to geological constructs that describe the reservoir.

Common concepts, it all translates!

Big Data, Machine Learning

Big Data Criteria:

- Volume
- Velocity
- Variety
- Veracity

We have this!

We've been big data before there was big data.

image from https://www.humansatsea.com

Machine / Statistical Learning

Training a computer detect features, find complicated relationships with complicated, multivariate, large datasets

We've been doing that too.

It is still (geo)statistical modeling…

Webinar - Big Data Analytics for Petroleum Engineering: Hype or Panacea?

December 8, 2017: Big Data Analytics for Petroleum Engineering: Hype or Panacea? Little Data + Simple Model = Big Data?

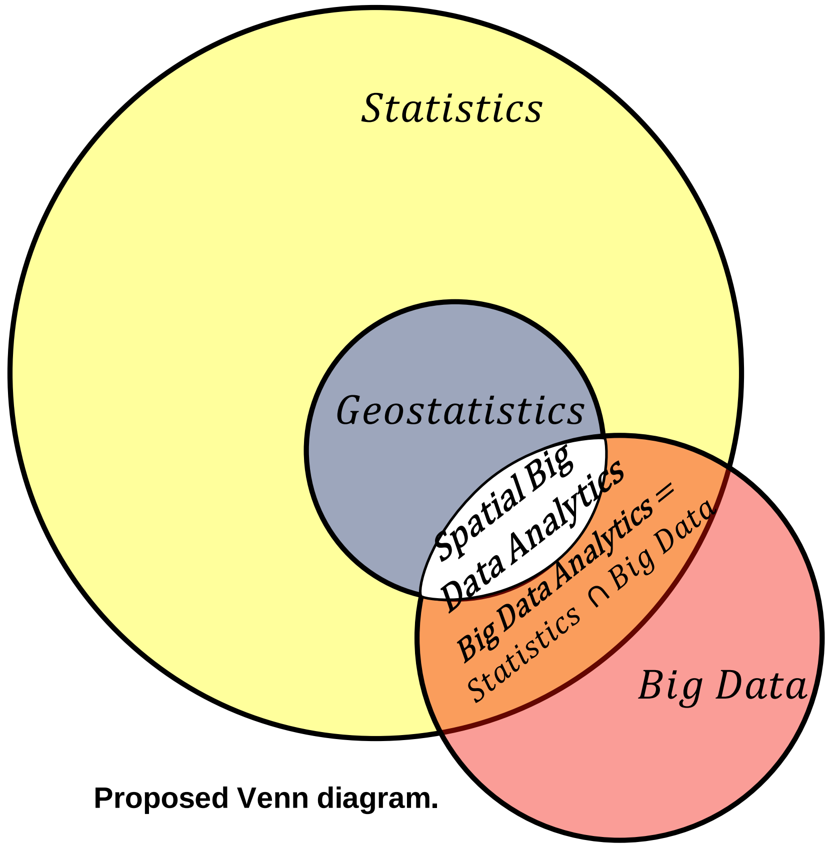
Webinar at http://www.cpge.utexas.edu/?q=node/385

Big Data, Machine Learning and Geostatistics

Statistics is concerned with mathematical methods for collecting, organizing, and interpreting data, as well as drawing conclusions and making reasonable decisions on the basis of such analysis.

Geostatistics is a branch of applied statistics that emphasizes: (1) the spatial (geological) context of the data, (2) the spatial relationship between data, (3) the different volumetric support and precision of the data, and (4) spatial and data uncertainty.

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



Proposed Venn diagram.

Given this :

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

Model Accuracy and Complexity

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

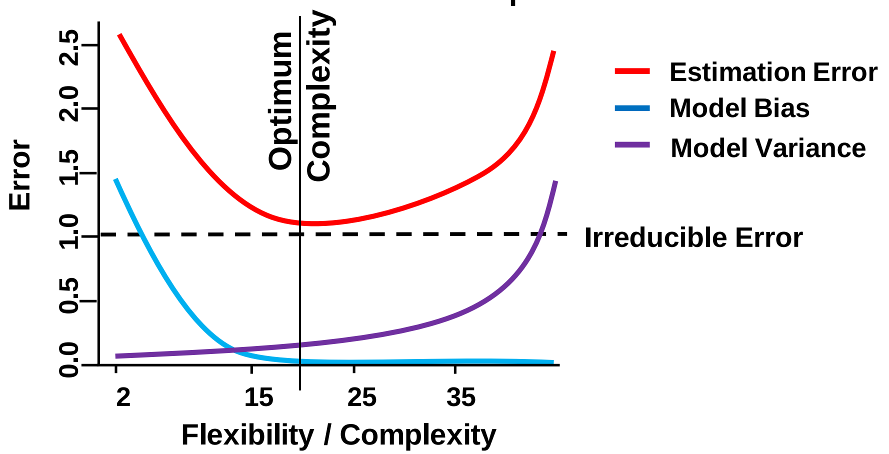
$$\underbrace{E\left[\left(y_0 - \hat{f}(x_1^0, \ldots, x_m^0)\right)^2\right]}_{\text{Estimation Error / Model Goodness}} = \underbrace{Var\left(\hat{f}(x_1^0, \ldots, x_m^0)\right)}_{\text{Model Variance}} + \underbrace{\left[Bias\left(\hat{f}(x_1^0, \ldots, x_m^0)\right)\right]^2}_{\text{Model Bias}}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

Model Variance - variance due to limited data (simpler models ⬇ lower variance)

Model Bias – error due to simple model (simpler models ⬆ higher bias)

Irreducible error - due to missing variables and limited samples

There are **trade-offs,** resulting in an **optimum level of complexity.**



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

Stationarity
Substituting time for space

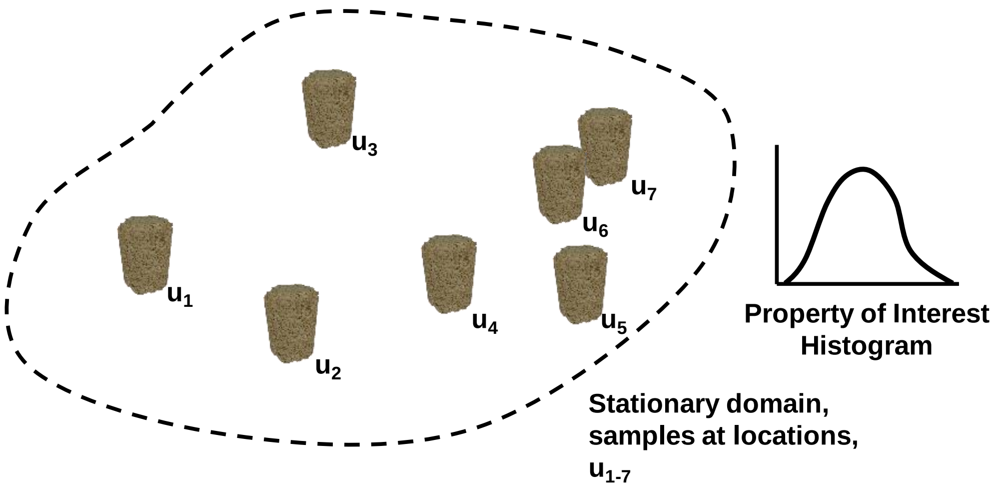
Any statistic requires replicates, repeated sampling (e.g. air or water samples from a monitoring station). In our geospatial problems repeated samples are not available at a location in the subsurface.

Extracted Sample at Location, u_1 Hole Remains at Location, 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".

Stationarity
Substituting time for space

The decision of the stationary domain for sampling is an expert choice. Without it we are stuck in the "hole" and **cannot calculate any statistics** nor say anything about the behavior of the subsurface **between the sample data**.

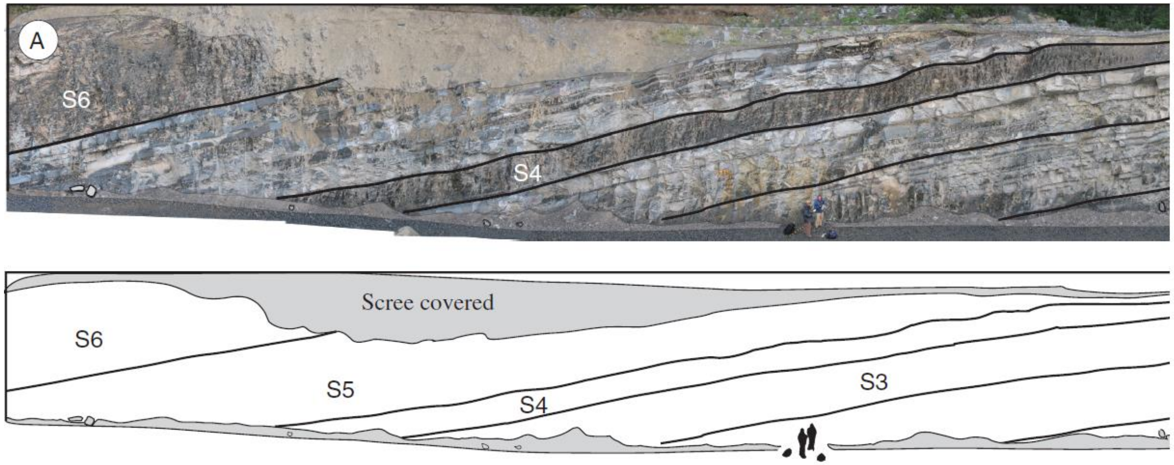


Import License: choice to pool specific samples to evaluate a statistic.

Export License: choice of where in the subsurface this statistic is applicable.

Stationarity
Definition 1: Geologic

Geological Definition: e.g. **‘**The rock over the stationary domain is sourced, deposited, preserved, and postdepositionally altered in a similar manner, the domain is map-able and may be used for local prediction or as information for analogous locations within the subsurface; therefore, it is useful to pool information over this expert mapped volume of the subsurface.’



Photomosaic, line drawing Punta Barrosa Formation sheet complex (Fildani et al. (2009).

Stationarity
Definition 2: Statistical

Statistical Definition: The metrics of interest are invariant under translation over the domain. For example, one point stationarity indicates the that histogram and associated statistics do not rely on location, **u**. Statistical stationarity for some common statistics:

Stationary Mean: $E\{Z(\mathbf{u})\} = \mathrm{m}, \ \forall \ \mathbf{u}$

Stationary Distribution: $F(\mathbf{u}; z) = F(z), \ \forall \ \mathbf{u}$

Stationary Semivariogram: $\gamma_z(\mathbf{u}; \mathbf{h}) = \gamma_z(\mathbf{h}), \ \forall \ \mathbf{u}$

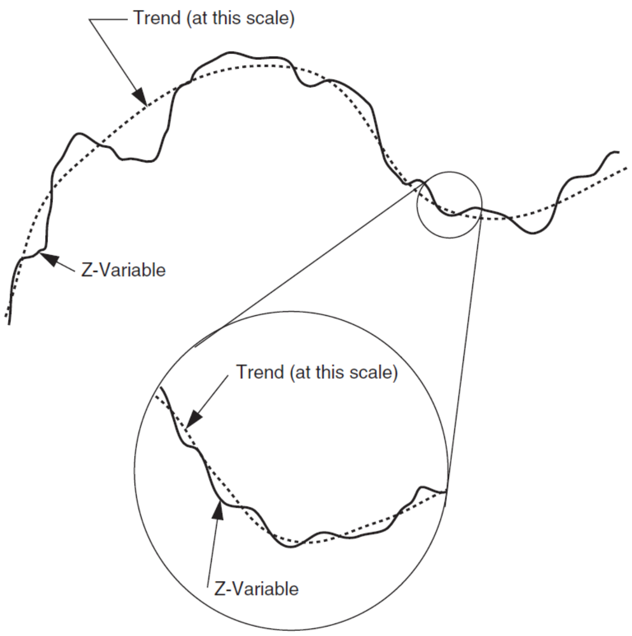
Stationarity: *What metric / statistic? Over what volume?*

May be extended to any statistic of interest including, facies proportions, bivariate distributions and multiple point statistics.

Stationarity
Comments on Stationarity

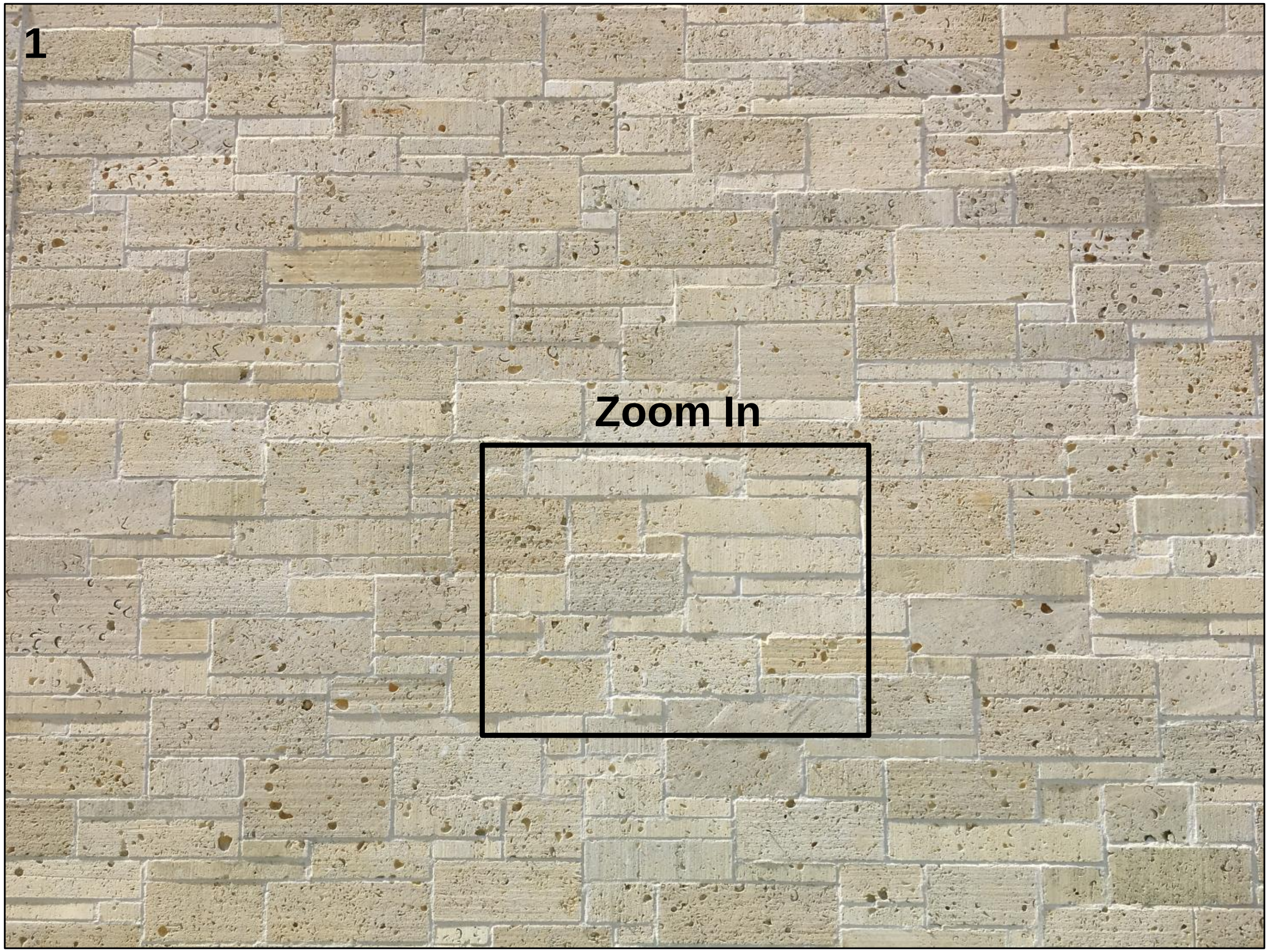
Stationarity is a decision, not an hypothesis; therefore it cannot be tested. Data may demonstrate that it is inappropriate.

The stationarity assessment depends on scale. This choice of modeling scale(s) should be based on the specific problem and project needs.



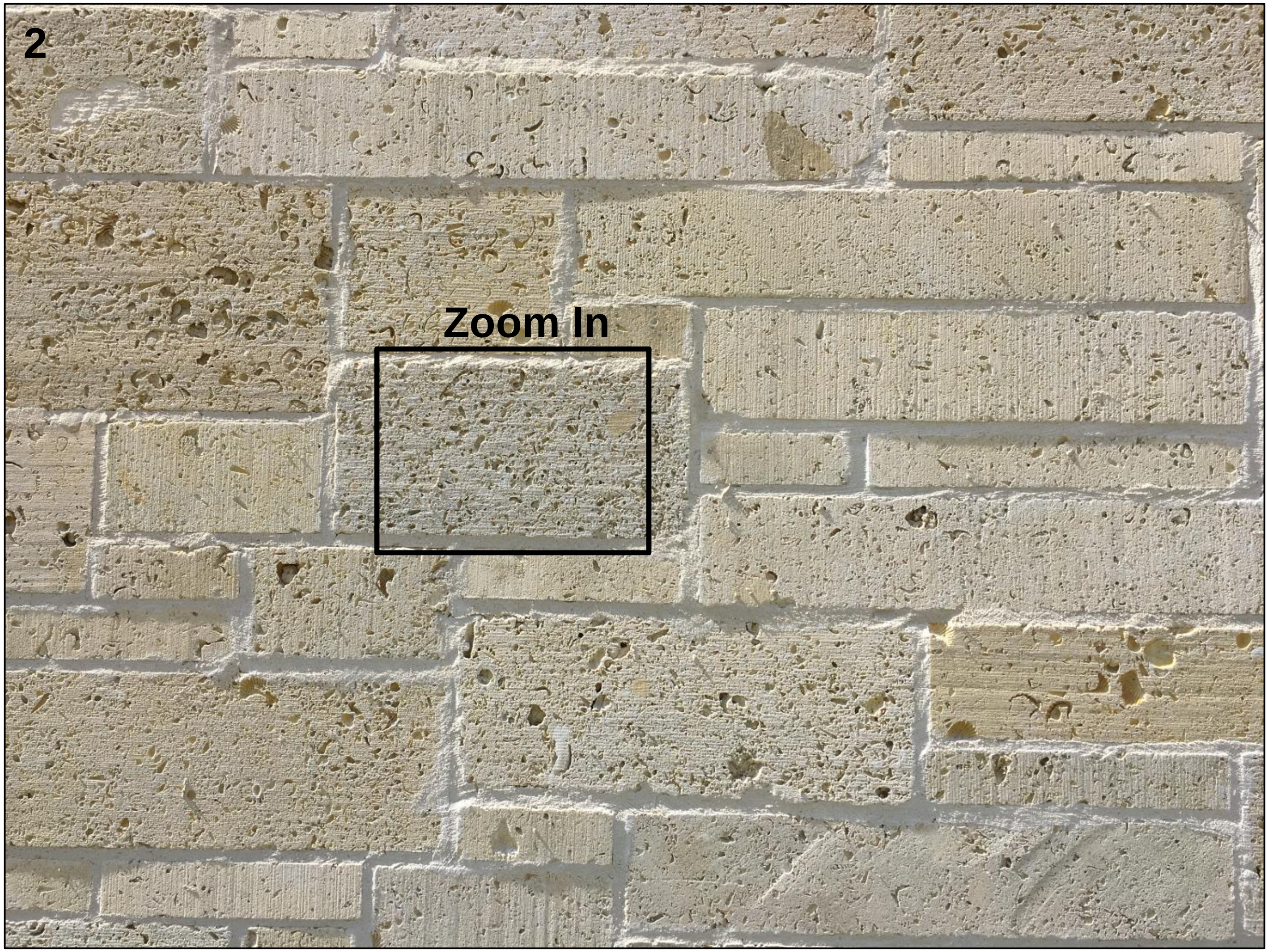
Scales of stationarity from Pyrcz and Deutsch (2014).

Stationarity and Scale Example



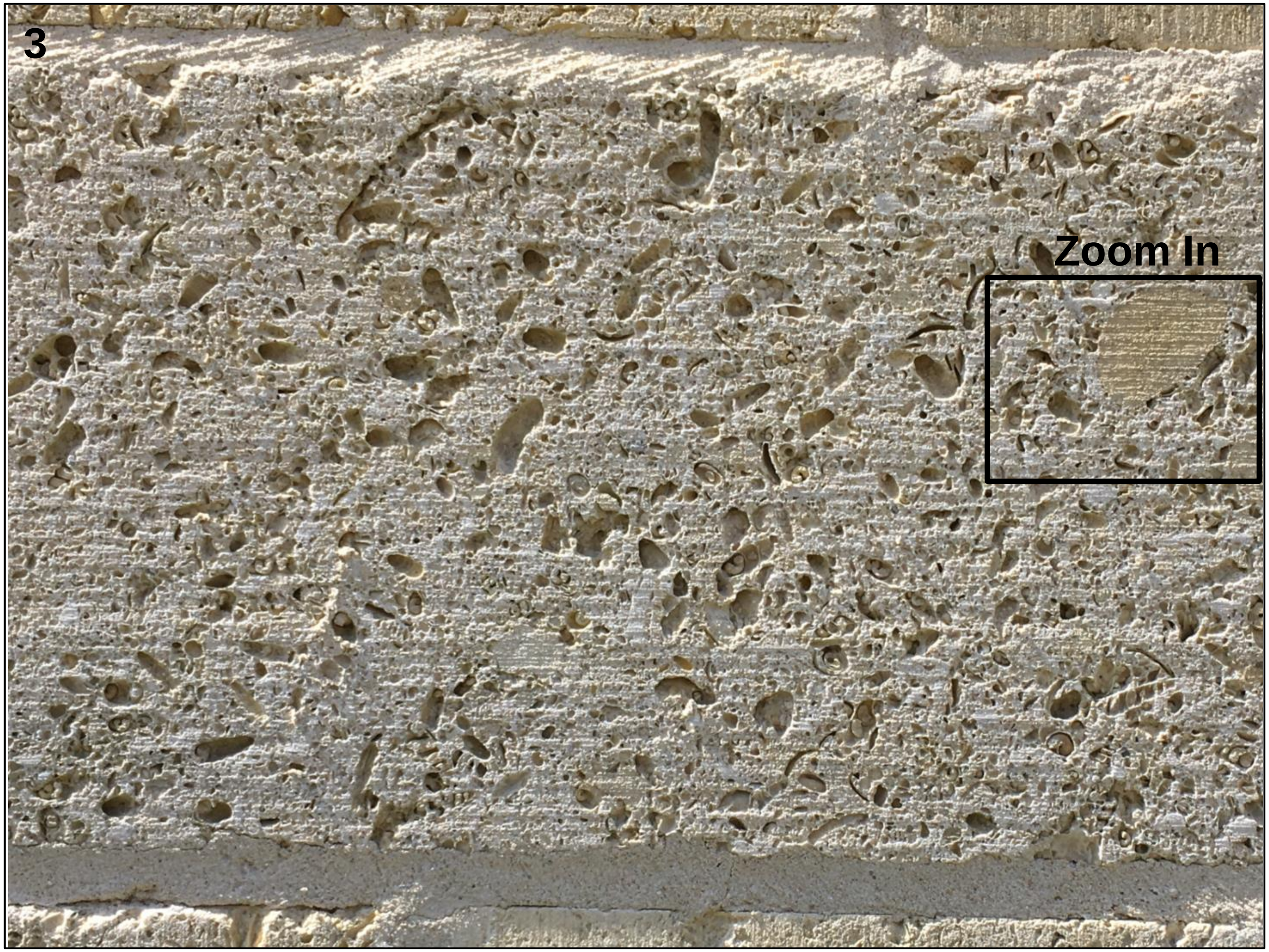
- Is this image stationary? What metric do you consider?

Stationarity and Scale Example



- A smaller group of bricks?

Stationarity and Scale Example

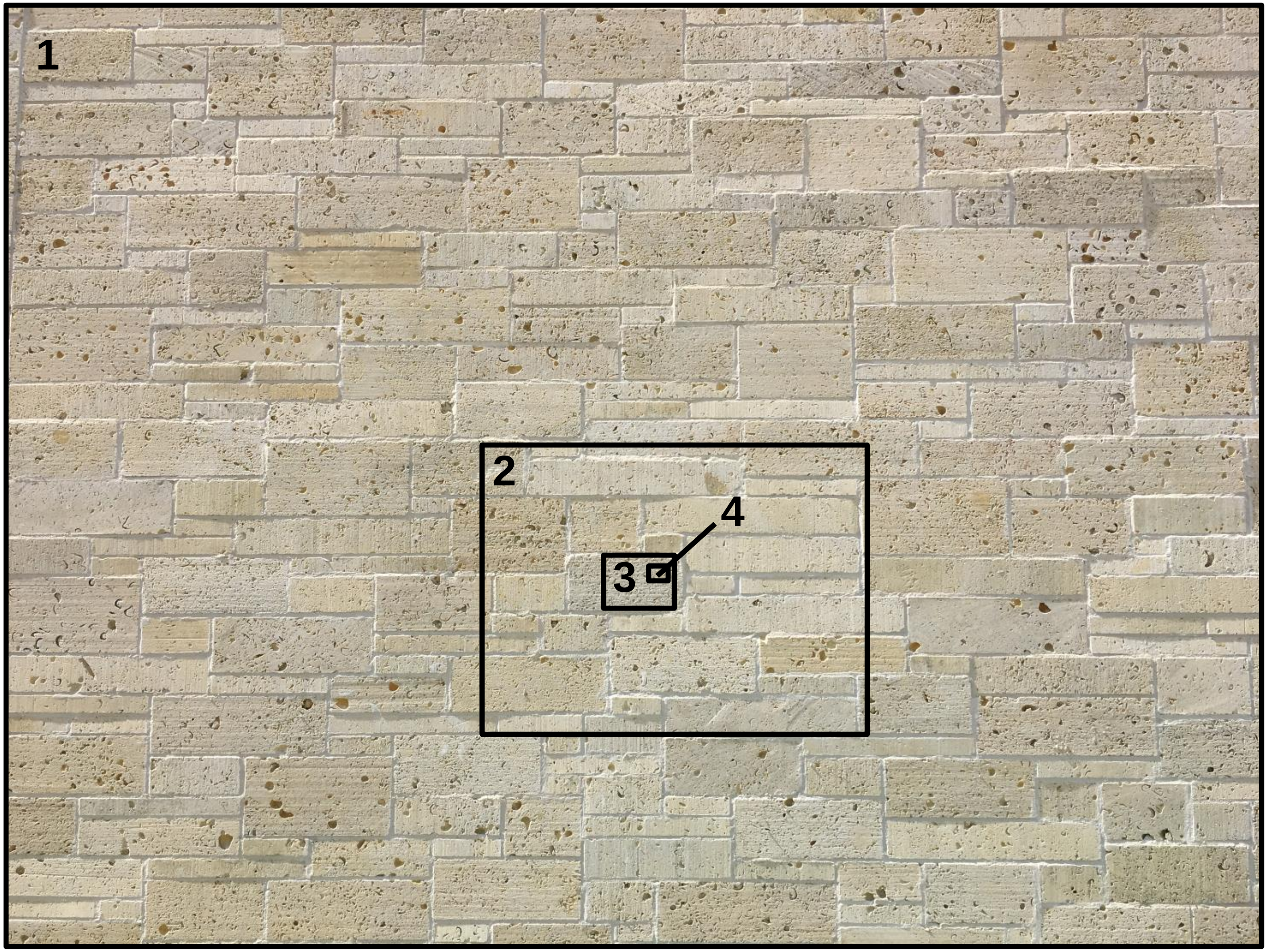


- A single brick?

Stationarity and Scale Example

- Small part of a brick?

Stationarity and Scale Example



- Is this image stationary? What metric do you consider?

Comments on Stationarity

We cannot avoid a decision of stationarity. No stationarity decision and we cannot move beyond the data. Conversely, assuming broad stationarity over all the data and over large volumes of the earth is naïve.

Geomodeling stationarity is the decision: (1) over what region to pool data (import license) and (2) over what region to use the resulting statistics (export license).

Nonstationary trends may be mapped and the remaining stationary residual modelled statistically / stochastically, trends may be treated uncertain.

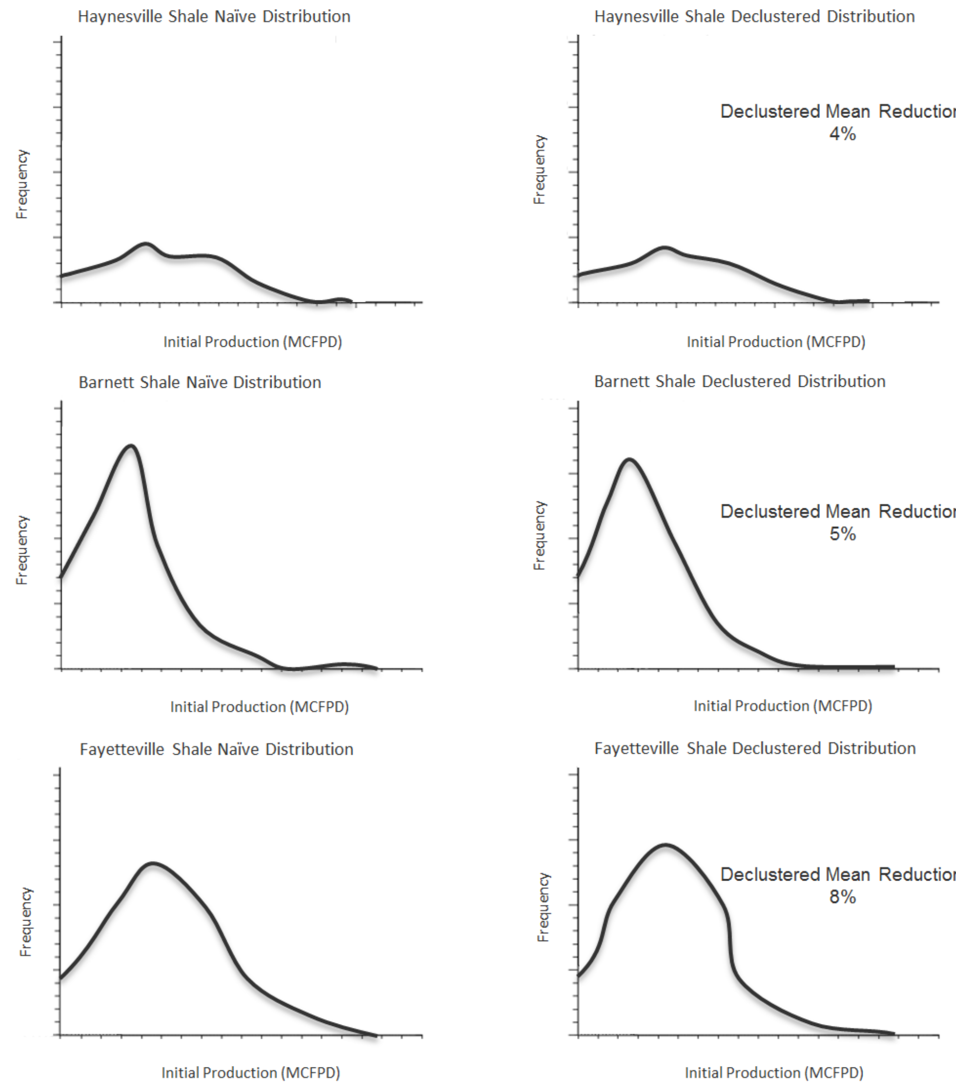
Good geological mapping and data integration is essential!

it is the framework of any subsurface model.

Bias is Ubiquitous

Representative Statistics

- Compiled IP datasets for domestic shale plays
  - Filtered datasets to reduce influence of completions
- Representativity an issue even with large datasets and relatively good coverage
  - Observed changes in naïve to declustered means of 4 – 8%



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

One Source of Bias
Data Collection

Data is collected to answer questions:

- how far does the contaminant plume extend? – *sample peripheries*
- where is the fault? – *drill based on seismic interpretation*
- what is the highest mineral grade? – *sample the best part*
- who far does the reservoir extend? – *offset drilling*

and to maximize NPV directly:

- maximize production rates

Data Collection

There are also limits to our data collection:

- accessibility to the sample – obstruction, reliable drilling, subsalt imaging
- inability to process the sample – may not be able to recover shale core samples
- can't run permeability evaluation on low permeability rock

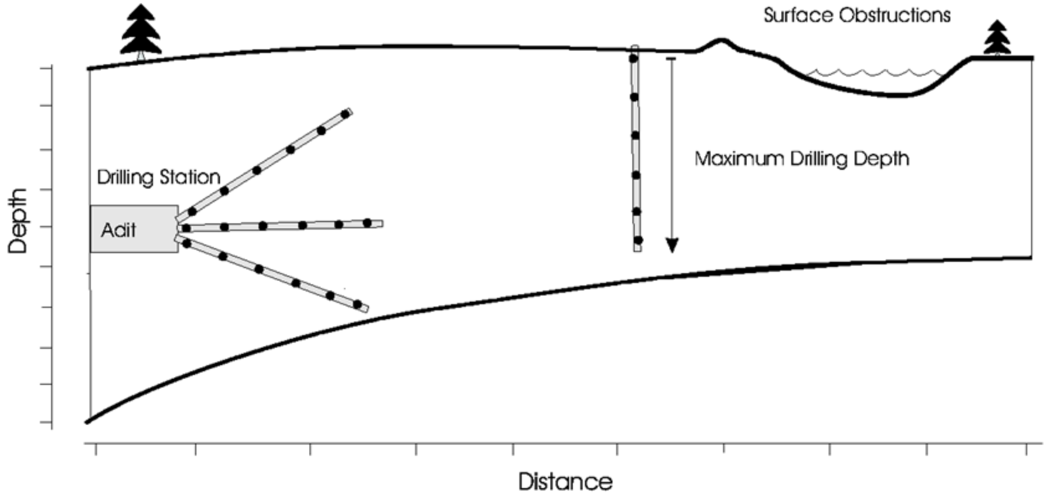


Image from Pyrcz and Deutsch (2003) http://gaa.org.au/pdf/DeclusterDebias-CCG.pdf

Data Collection

If we were sampling for representativity of the sample set and resulting sample statistics, by theory we have 2 options:

1. random sampling
2. regular sampling (as long as we don't align with natural periodicity)

What would happen if you proposed random sampling in the Gulf of Mexico at $150M per well?

We should not change current sampling methods as they result in best economics, we should address sampling bias in the data.

Never use raw spatial data without access sampling bias / correcting.

Solutions to Biased Sampling

- There is a need, however, to adjust the histograms and summary statistics to be representative of the entire volume of interest. We use statistics to make decisions!

1. **Declustering techniques assign each datum a weight** based on closeness to surrounding data
   - $w_i, i = 1, \ldots, n$ (weights are greater than 0 and sum to n)
   - Histogram and cumulative histogram use $w_i, i = 1, \ldots, n$ instead of equal weighted,$w_i = 1.0$.

2. **Debiasing techniques** derive an entirely new distribution based on a secondary data source such as geophysical measurements or expert interpretation

Declustering for Spatial Sampling Bias

- Split up the area of interest with Voronoi partition.
  - Intersected perpendicular bisectors between adjacent data points
  - The result is data weights, all summary statistics can be weighted

$$w(\mathbf{u}_j) = \frac{A_j}{\sum_{j=1}^{n} A_j} \text{ for } \sum_{j=1}^{n} w(\mathbf{u}_j) = 1$$

$$w(\mathbf{u}_j) = n\frac{A_j}{\sum_{j=1}^{n} A_j} \text{ for } \sum_{j=1}^{n} w(\mathbf{u}_j) = n$$

  - This method is sensitive to boundary
  - Commonly applied in a variety of scientific fields for weighted averages of spatial phenomenon with irregular sampling.

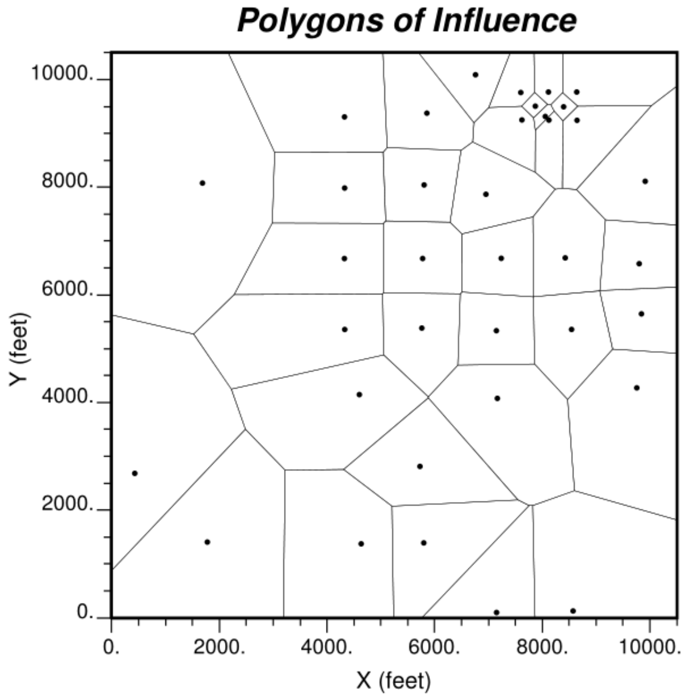
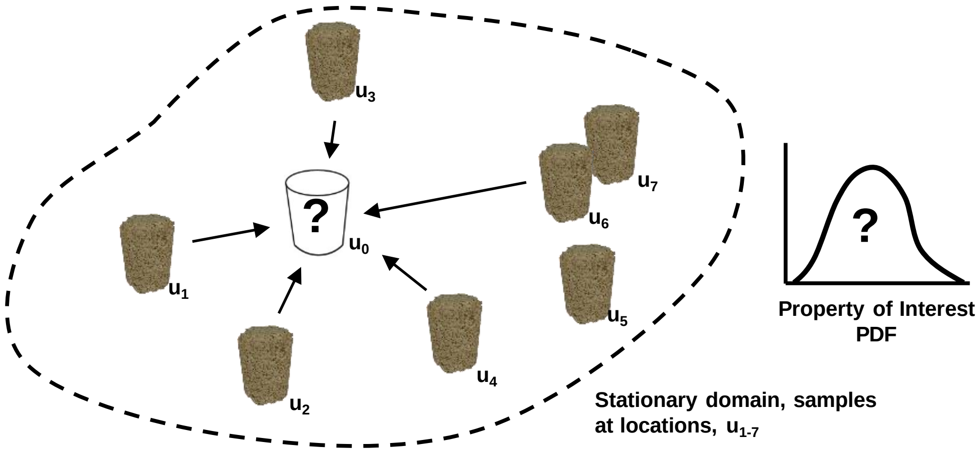


Image from Pyrcz and Deutsch (2014).

Uncertainty
What is uncertainty?



Uncertainty is not an intrinsic property of the subsurface.

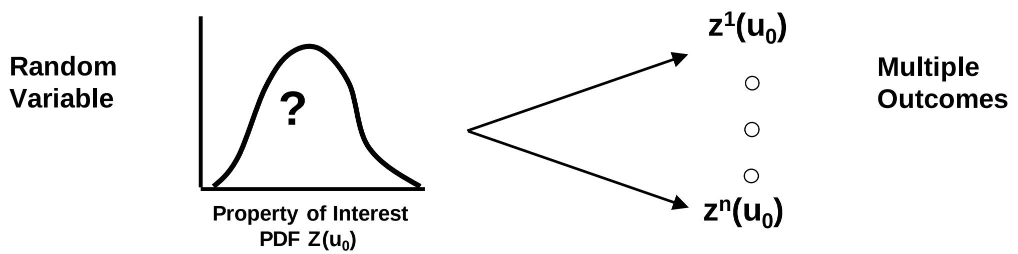
- At every location ($\mathbf{u}_\alpha$) within the volume of interest the true properties could be measured if we had access (facies, porosity etc.).
- **Uncertainty is a function of our ignorance**, our inability to observed and measure the subsurface with the coverage and scale required to support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

sparsity of sample data + heterogeneity = uncertainty

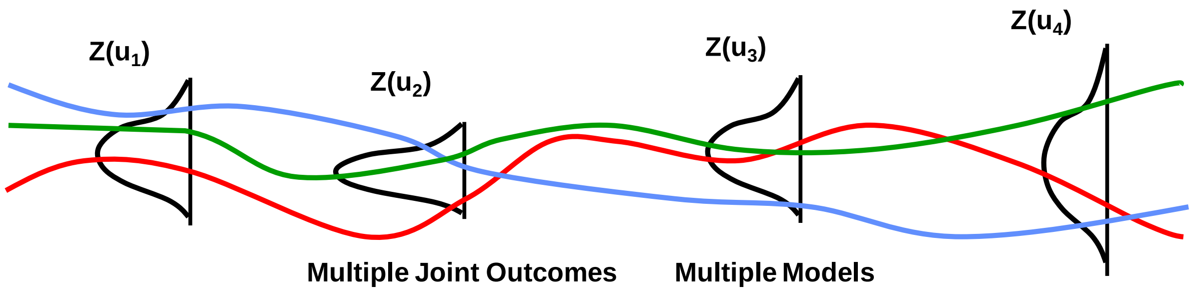
- If the subsurface was homogeneous, with a few measurements uncertainty would be reduced and estimates resolved to a sufficient degree of exactitude.

Uncertainty
How do we represent uncertainty?

Random Variables and Functions: A random variable is a property at a location (u_α) that can take on multiple possible outcomes. This is represented by a probability density function (PDF).



If we take a set of random variables at all locations of interest and we impart the correct spatial continuity between them then we have a **random function**. Each outcome from the random function is a potential model of the subsurface.



Uncertainty
How do we represent uncertainty?

Using Multiple Models: We represent uncertainty with multiple models.

Scenarios: when the input decisions and parameters are changed

Captures interpretation and data uncertainty.

Realizations: when the input decisions and parameters are held constant and only the random number seed is changed

Captures spatial uncertainty.

Working With Multiple Models: It is generally not appropriate to analyze a single or few scenarios and realizations.

Use all the models all the time applied to the transfer function
(e.g. volumetric calculation, contaminant transport, ore grade scale up, flow simulation etc.).

Uncertainty
Comments on uncertainty

Calculating Uncertainty in a Modeling Parameter: Use Bayesian methods, spatial bootstrap etc. You must account for the volume of interest, sample data quantity and locations, and spatial continuity.

If You Know It, Put It In. Use expert geologic knowledge and data to model trends. Any variability captured in a trend model is known and is removed from the unknown, uncertain component of the model. Overfit trend will result in unrealistic certainty.

Types of Uncertainty: (1) data measurement, calibration uncertainty, (2) decisions and parameters uncertainty, and (3) spatial uncertainty in estimating away from data. Your job is to hunt for and include all significant sources of uncertainty.

Uncertainty
Comments on uncertainty

What about Uncertainty in the Uncertainty? Don’t go there! Use defendable choices in your uncertainty model, be conservative about what you known, document and move on.

Uncertainty Depends on Scale. It is much harder to predict a property of tea spoon vs. a house-sized volume at a location (u_{α}) in the subsurface. Ensure that scale and heterogeneity are integrated.

You Cannot Hide From It. Ignoring uncertainty assumes certainty and is often a very extreme and dangerous assumption.

Decision Making with Uncertainty. Apply all the models to the transfer function to calculate uncertainty in subsurface outcome to support decision making in the presence of uncertainty.

Facies
What are the Criteria for Facies?

First some general comments:

1. **Facies / Rock type** is an important decision for subsurface modeling. It should remain a collaborative decision integrating expertise from the project team (Stratigraphers, Reservoir Modelers, Reservoir Engineers, Petrophyscists and Geophysicists).

2. Facies / Rock types **must improve subsurface prediction** away from the data or they do not add value.

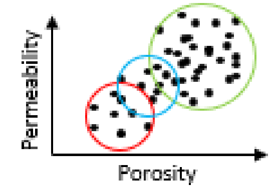
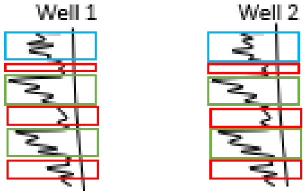
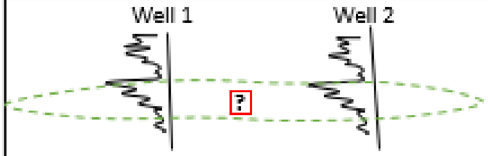
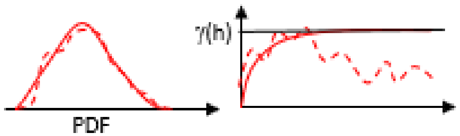
3. **Number of facies** is a balancing act between accuracy of geological concepts and statistical inference, and modeling effort

4. Reservoir modeling is **hierarchical**,
   units *contain* depofacies *contain* lithofacies *contain* por/perm

5. 80-90% of **flow heterogeneity** is captured in the facies models.

Facies
What are the Criteria for Facies?

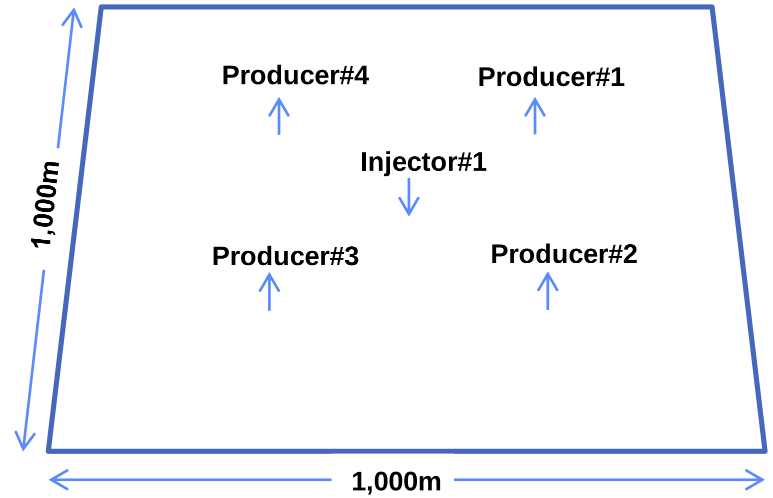
These are the **criteria for facies** (or any categories in reservoir models).

Criteria	Considerations	Example
Separation of Rock Properties	Facies must divide the properties of interest that impact subsurface environmental and economic performance (e.g. grade, porosity and permeability).	
Identifiable in Data	Facies must be identifiable with the most common data available. e.g. facies identifiable only in cores are not useful if most wells have only logs.	
Map-able Away from Data	Facies must be easier to predict away from data than the rock properties of interest directly, facies improves prediction.	
Sufficient Sampling	There must be enough data to allow for reliable inference of reliable statistics for rock properties for each facies.	

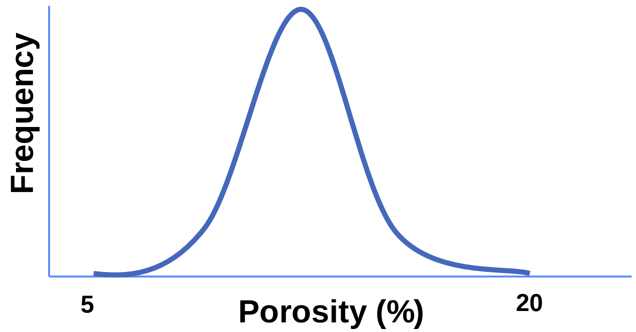
Motivation for Measuring Spatial Continuity

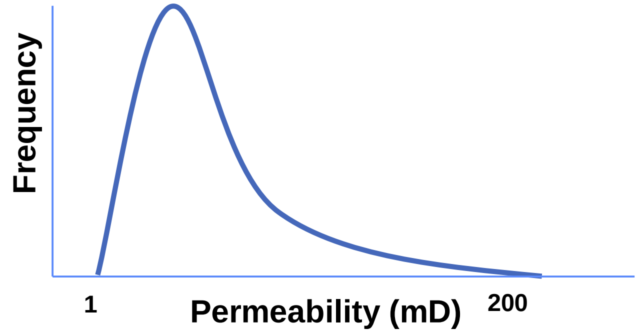
Simple Example

- Area of interest
- 1 Injector and 4 producers



- Porosity and permeability distributions (held constant for all cases)

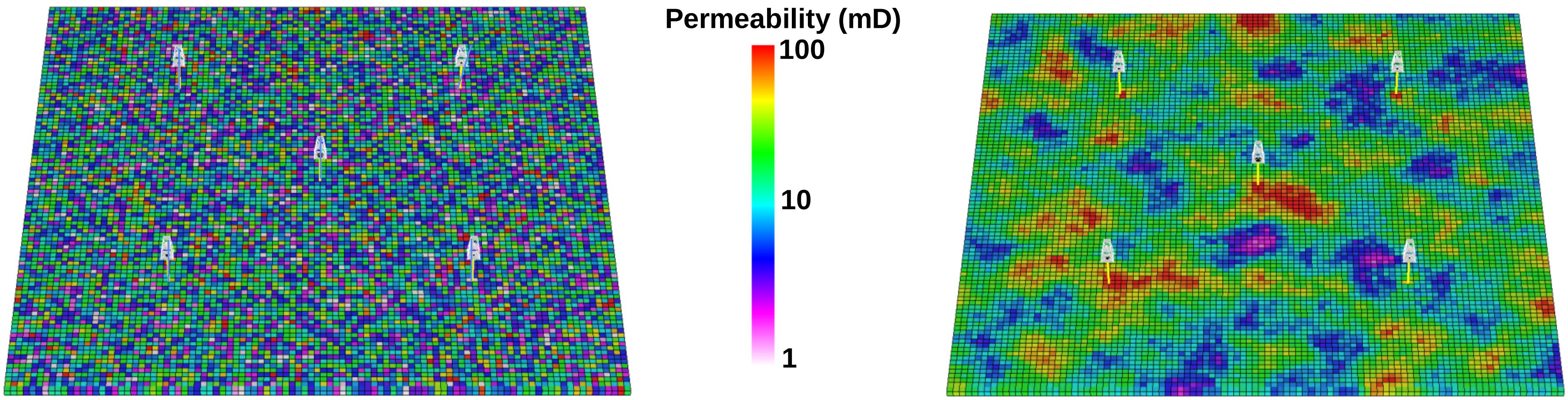




Motivation for Measuring Spatial Continuity

Does spatial continuity of reservoir properties matter?

Consider these models of permeability

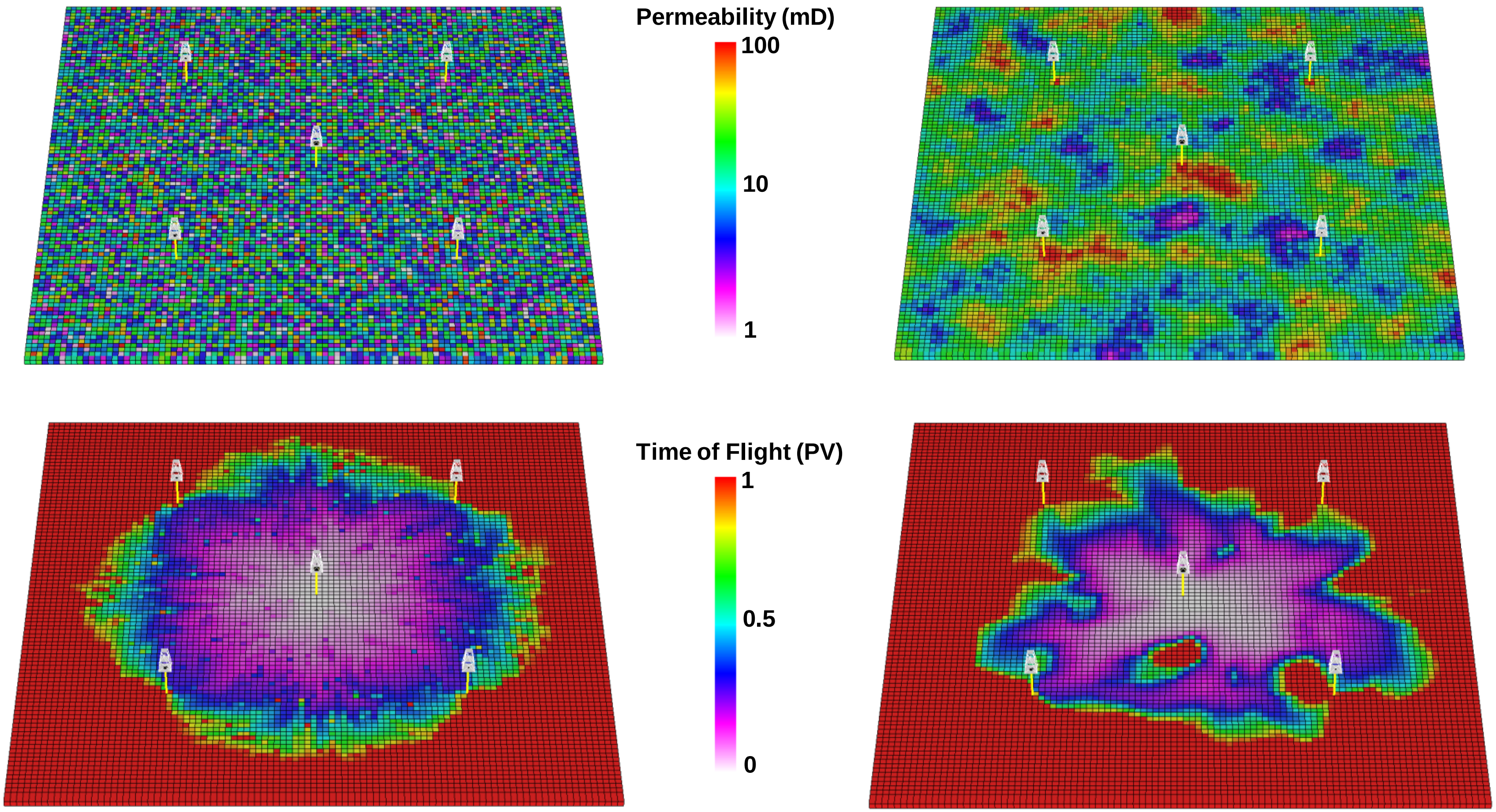


Recall – all models have the same porosity and permeability distributions

- Mean, variance, P10, P90 …
- Same static oil in place!

Motivation for Measuring Spatial Continuity

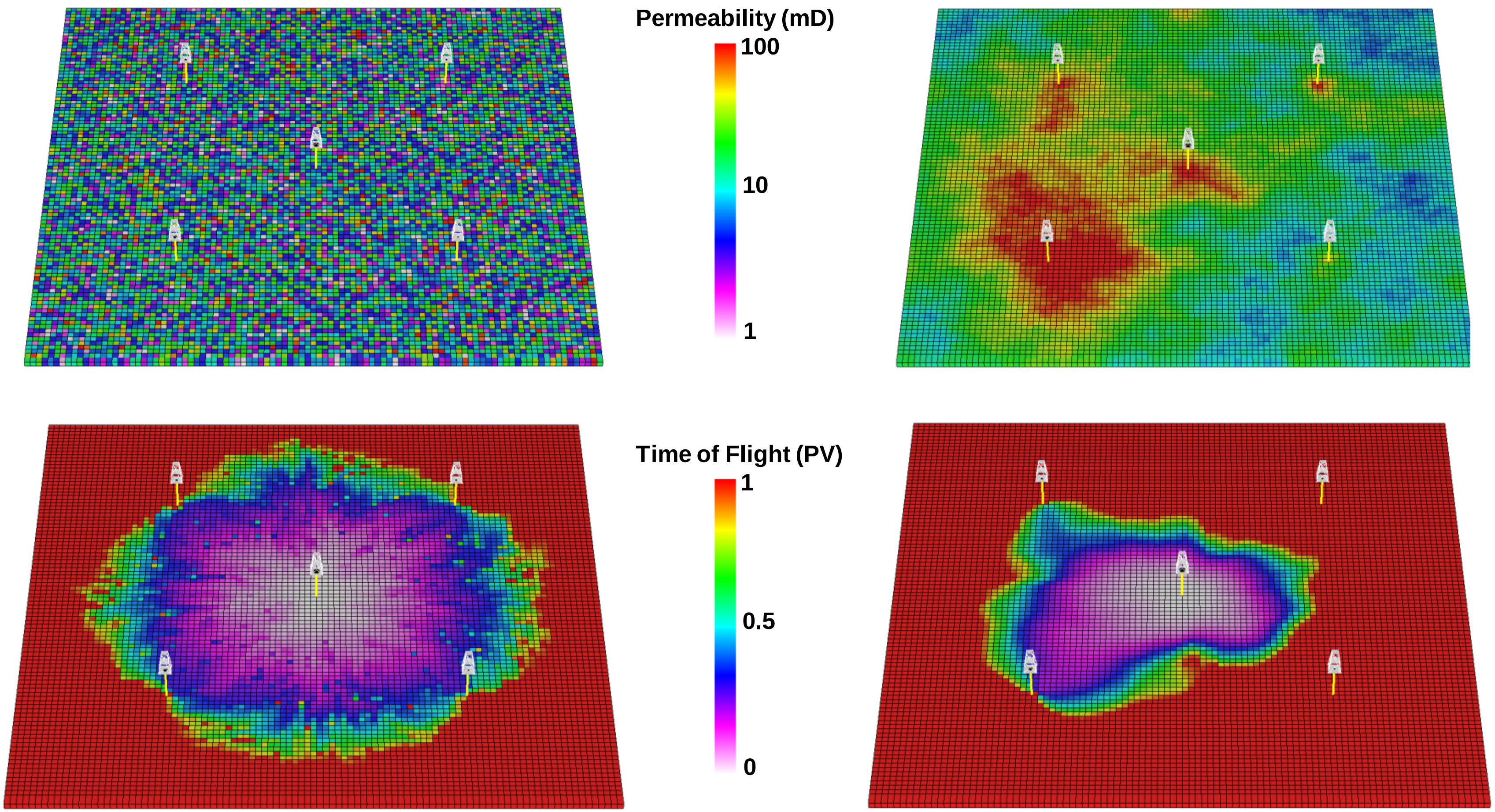
Does spatial continuity of reservoir properties matter?



How does heterogeneity impact recovery factor? Well Estimated Ultimate Recovery?

Motivation for Measuring Spatial Continuity

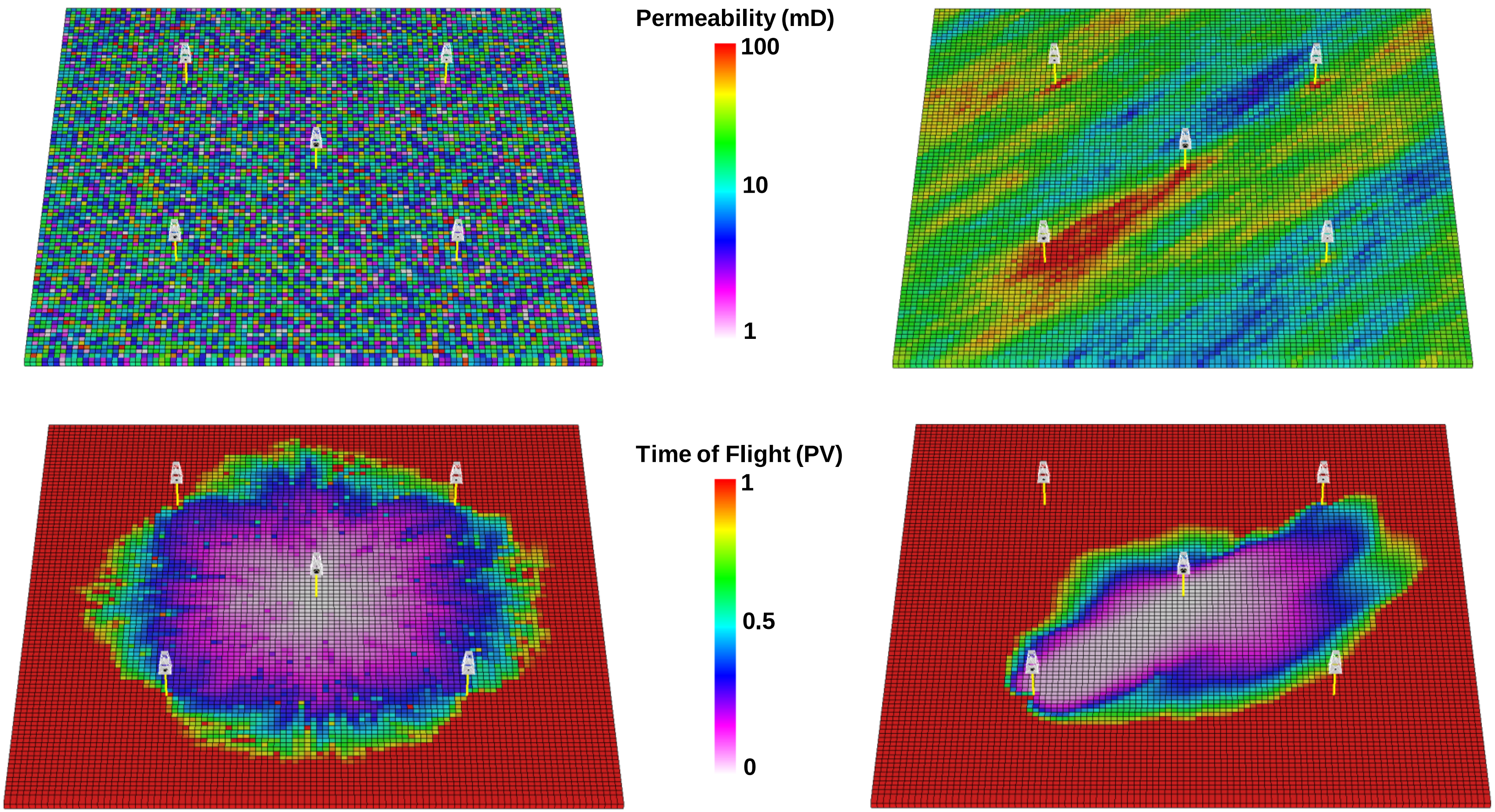
Does spatial continuity of reservoir properties matter?



How does heterogeneity impact recovery factor? Well Estimated Ultimate Recovery?

Motivation for Measuring Spatial Continuity

Does spatial continuity of reservoir properties matter?



How does heterogeneity impact recovery factor? Well Estimated Ultimate Recovery?

Motivation for Measuring Spatial Continuity

- For the same reservoir property distributions a wide range of spatial continuities are possible.
- Spatial continuity often impacts reservoir forecasts.
- Need to be able to:

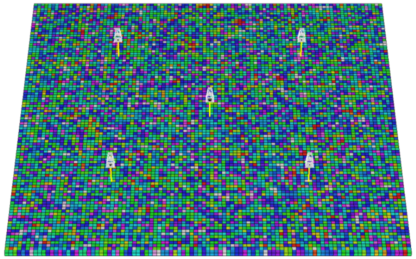
Characterize / Quantify

↓ **Spatial Continuity Measures**

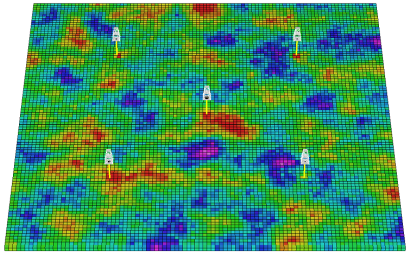
Impose in Reservoir Models

Spatial Continuity

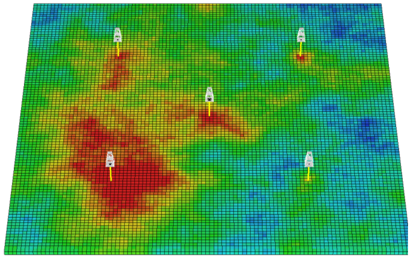
"Very Short"



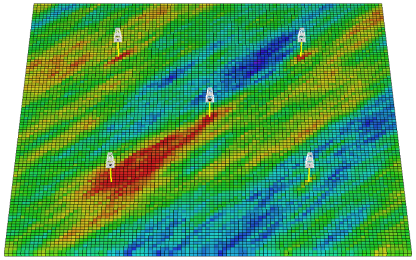
"Medium"



"Long"



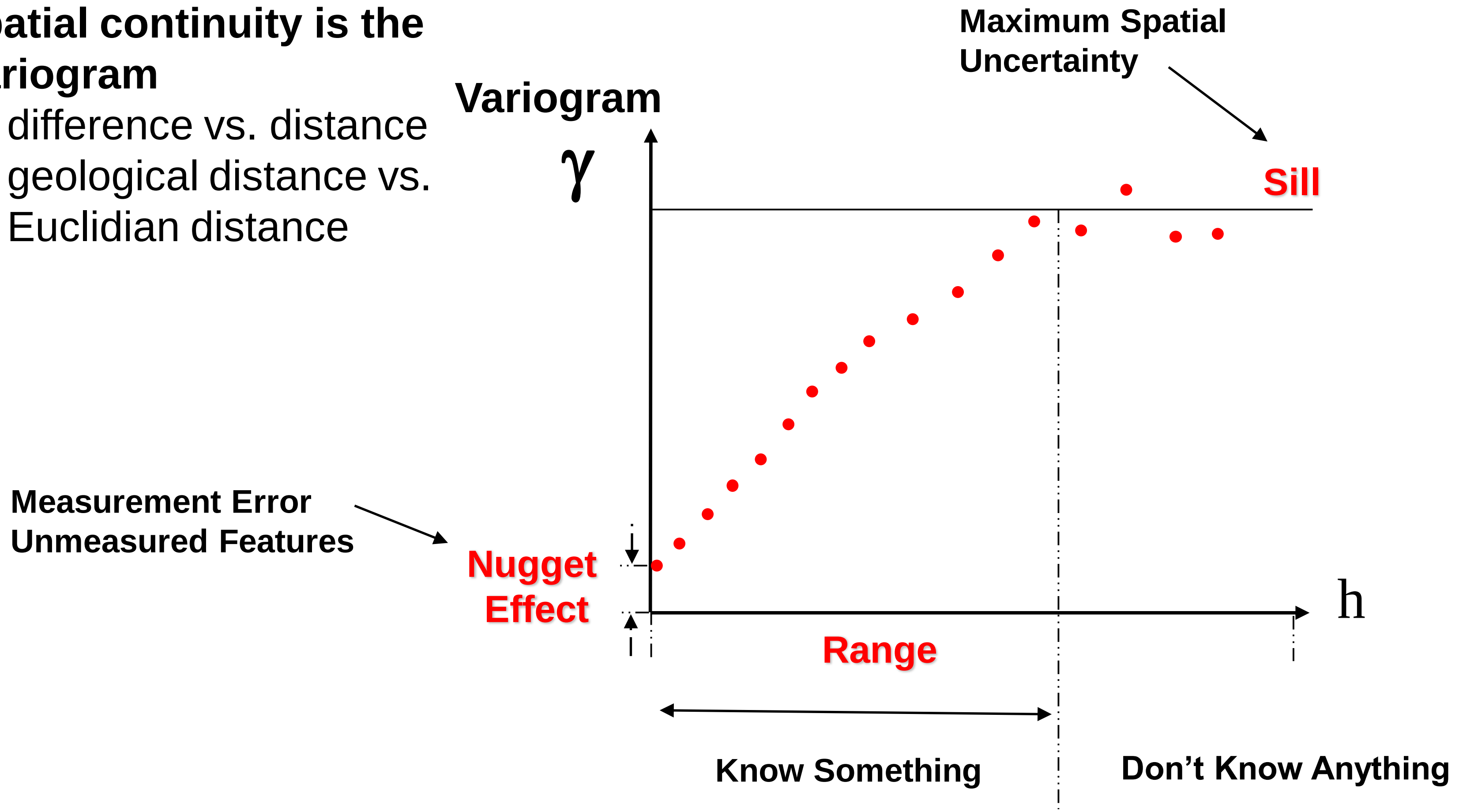
"Anisotropic "



Spatial Continuity

One method to quantify spatial continuity is the variogram

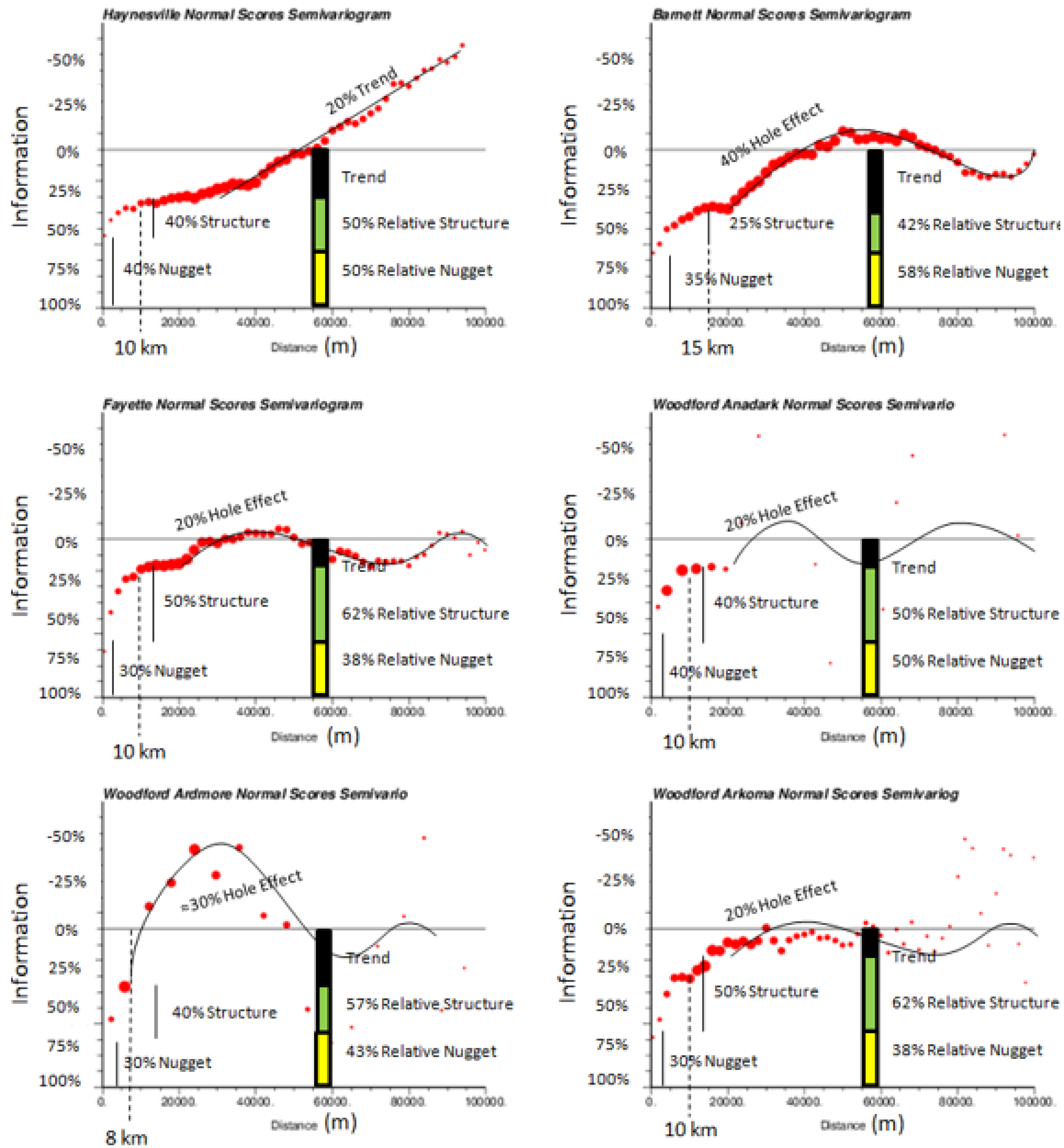
- difference vs. distance
- geological distance vs. Euclidian distance



Spatial Continuity

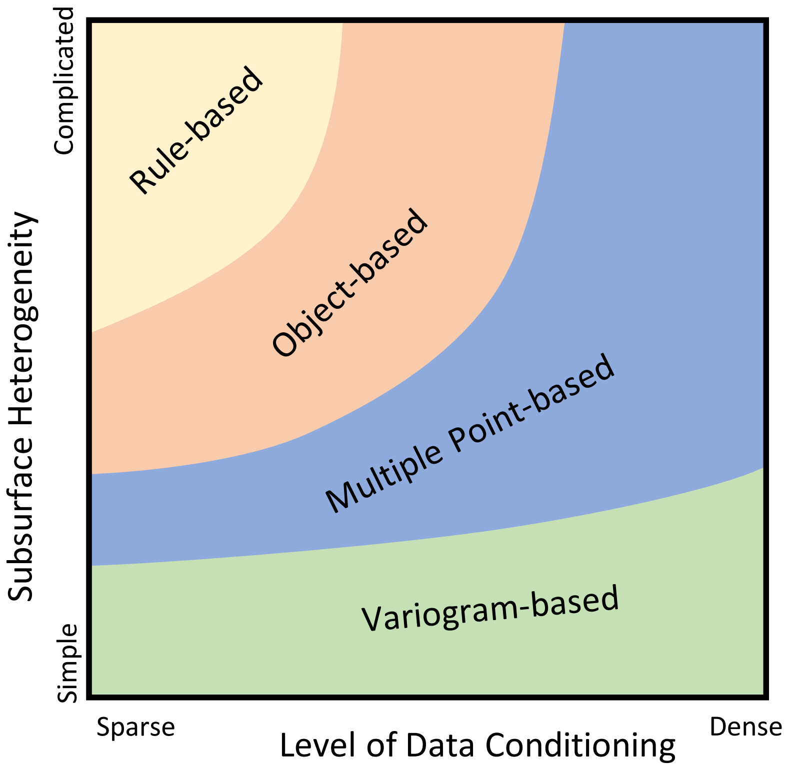
Example for production from shale assets.

- spatial continuity of unconventional production
- prediction model
- value of well data



Quantification of spatial continuity of shale gas production rates (Pyrcz et al., 2016).

A Toolbox



Geostatistics includes a tool box of subsurface modeling methods.

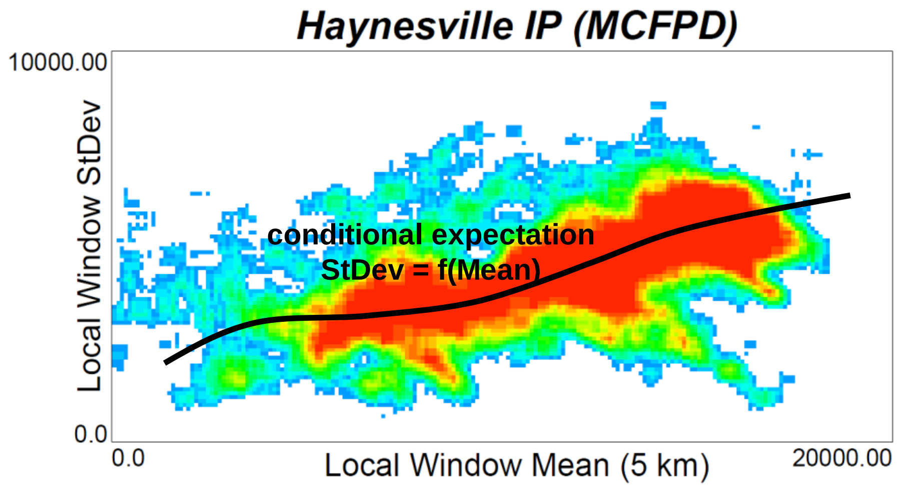
- A scheme for selecting between methods, also consider project goals and resources.
- New methods that model the subsurface with data integration, uncertainty, conditional to data. **Add it to the box.**

Motivational (Geo)statistics Examples

Why Use (Geo)statistics?

Quantification / Summarization / Comparison

- abstraction to a small set of parameters allows us to detect features, learn new insights
- robust measure of significance of differences
- seek opportunities for quantification.



Abstraction allows for efficient characterization and leads to insights, (Pyrcz et al., 2016)

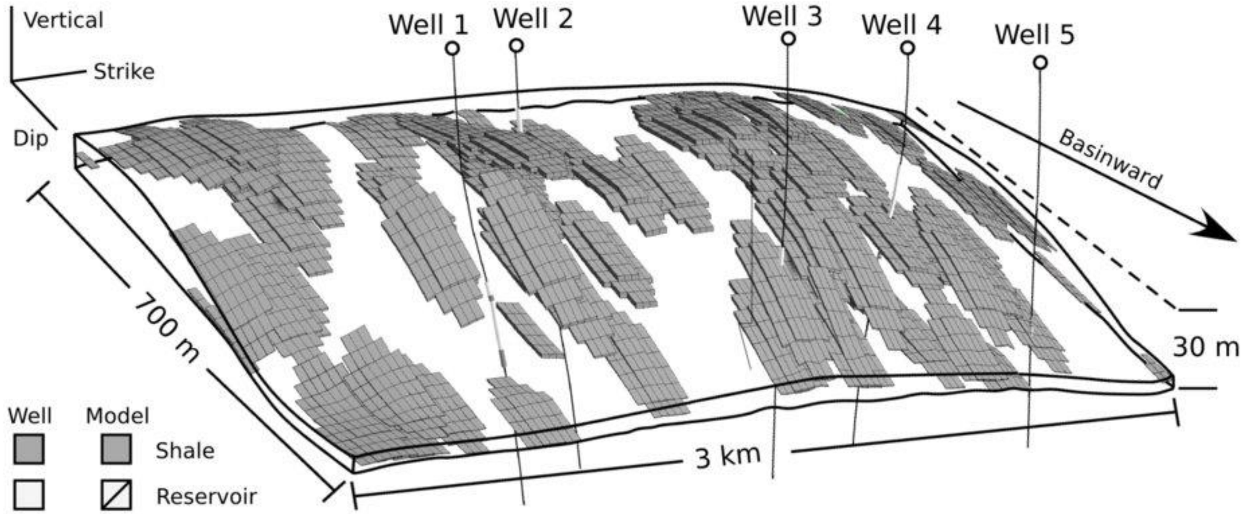
Motivational (Geo)statistics Examples

Why Use (Geo)statistics?

Model of Uncertainty

sparse sampling + heterogeneity = uncertainty

- if we had enough data and understood the phenomenon perfectly there is no uncertainty, no need for a statistical / stochastic approaches.



Can’t know exactly where the shales are from 5 wells and given the shale discontinuity. (Pyrcz and Deutsch, 2014)

Motivational (Geo)statistics Examples

Why Use (Geo)statistics?

Too Big / Too Complicated

Massive Multivariate

- due to the curse of dimensionality we often cannot sample enough to characterize the system
- need to used a statistical multivariate model

Multivariate modeling accounting for complicated relationships (Barnett and Deutsch, 2012)

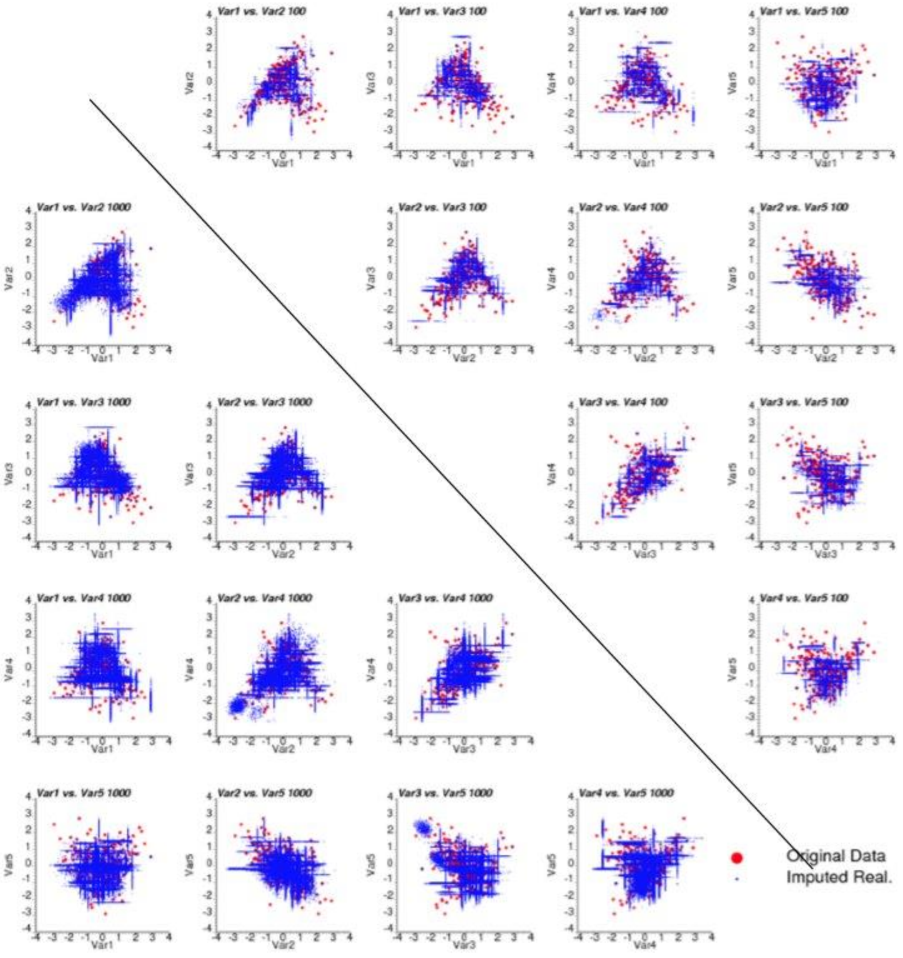


Figure 4: Scatterplots of the sampled observations and imputed observations for 1000 realizations (bottom covariance triangle) and 100 realizations (upper covariance triangle).

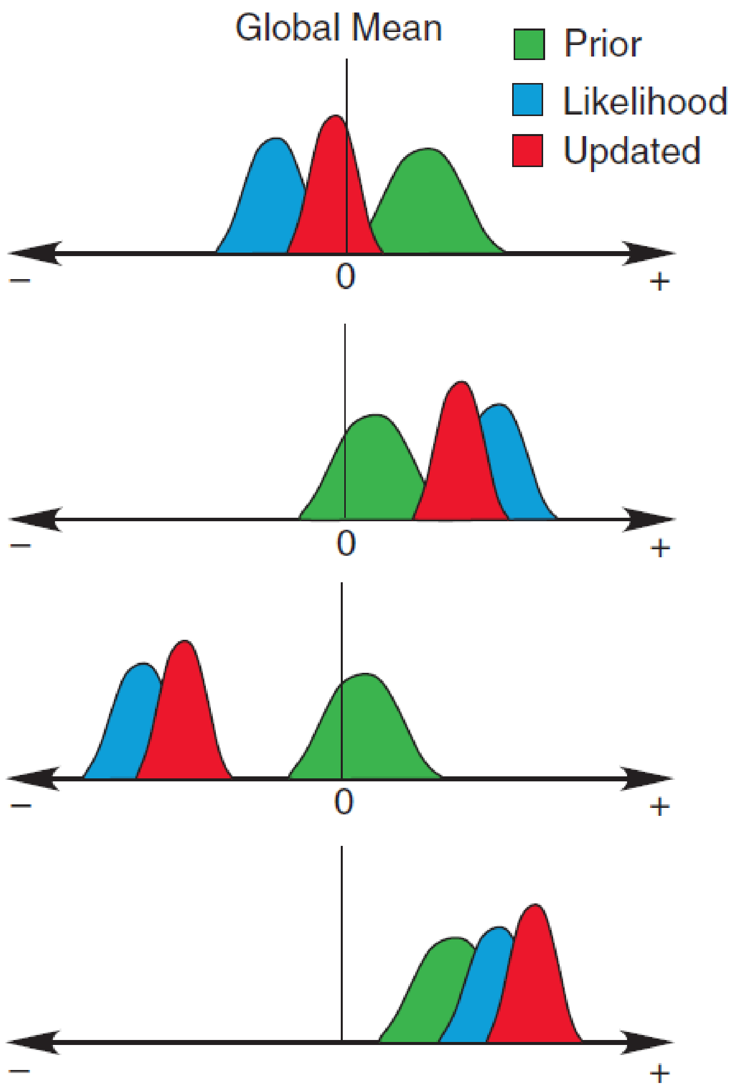
Motivational (Geo)statistics Examples

Why Use (Geo)statistics?

Combining / Updating with New Information

Bayesian Updating

- need statistical models to describe data redundancy



Bayesian updating under the assumption of Gaussianity (Pyrcz and Deutsch, 2012).

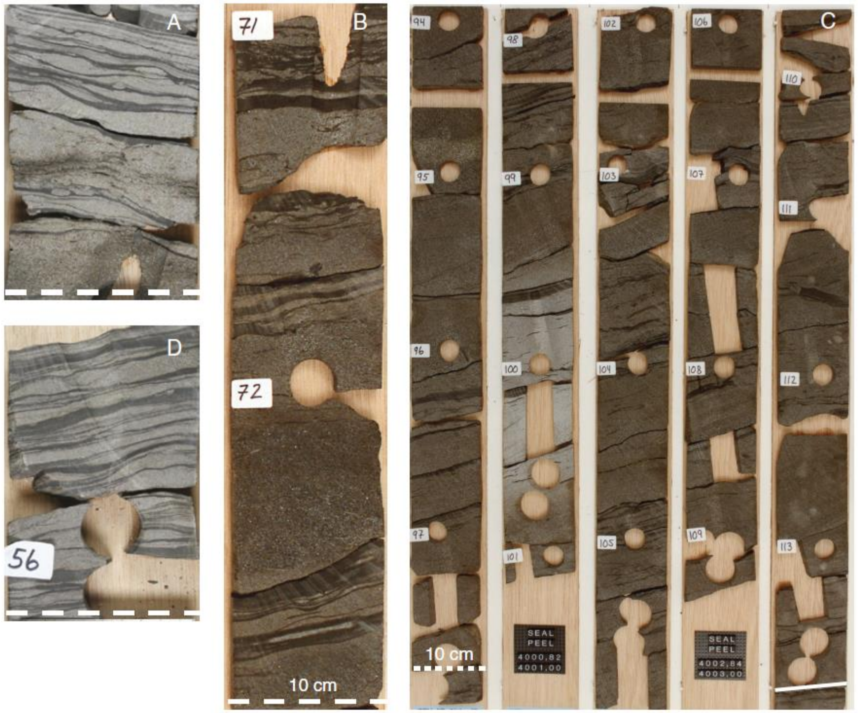
Motivational (Geo)statistics Examples

Why Use (Geo)statistics?

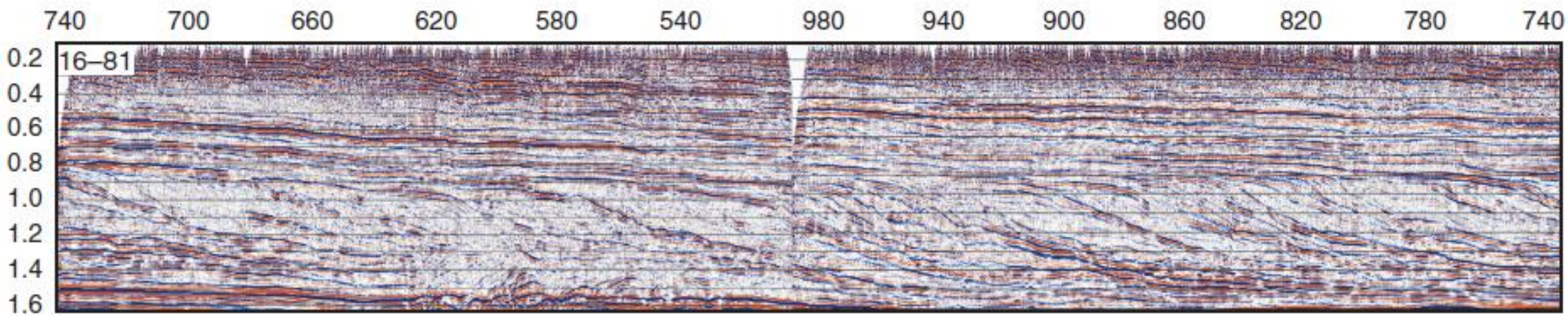
Accounting for Scale

Pores to Production

- statistical models for change of support size



Sectioned Core Photographs of the Cook Formation, a Shallow Marine Sandstone Reservoir from the North Sea. A summary of the core interpretation by Folkestad et al. (2012).



A Seismic Line for the Torok Formation Clinoforms of the Lower Cretaceous in the National Petroleum Reserve of Alaska, USA Just South of the Harrison Bay on the Coast of the Beaufort Sea. Seismic lines are shot by the USGS and are available in public domain. Figure provided in high resolution by Professor Chris Kendall, available at the Society for Sedimentological Research Stratigraphy Website.

Motivational (Geo)statistics Examples

Why Use (Geo)statistics?

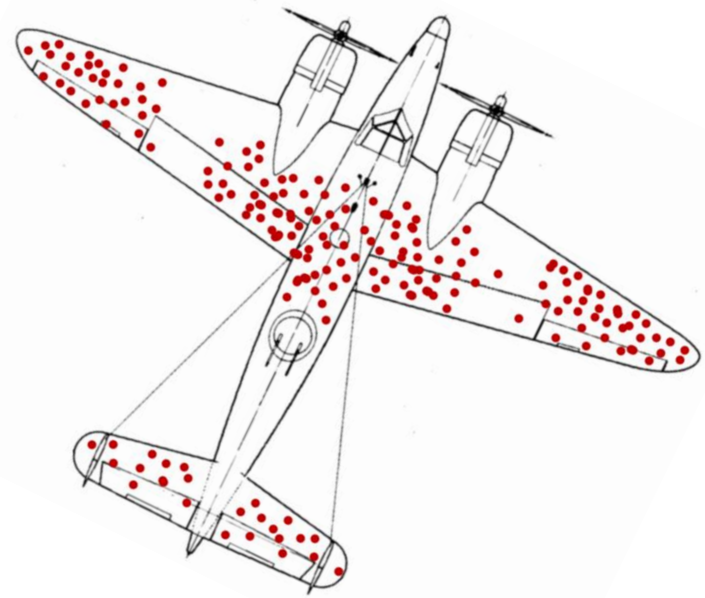
Debias Ourselves

Survivorship Bias

a form of selection bias resulting from selecting samples that "survived" some previous selection process. This often leads to false conclusions.

For example, in WWII the Center for Naval Analyses (@CNA_org Twitter) compiled a dataset of bomber damage to assess where reinforcement was needed.

Statistician Abraham Wald recognized this was a case of survivorship bias. The plans shot in critical locations did not return to base. Wald suggested reinforcement of locations that were not damaged in planes that safely returned to base! (https://en.wikipedia.org/wiki/Survivorship_bias#In_the_military)



Hypothetical dataset of aircraft damage for planes that returned to based. Source https://en.wikipedia.org/wiki/Survivorship_bias#/media/File:Survivorship-bias.png

Is there preselection in our subsurface datasets?

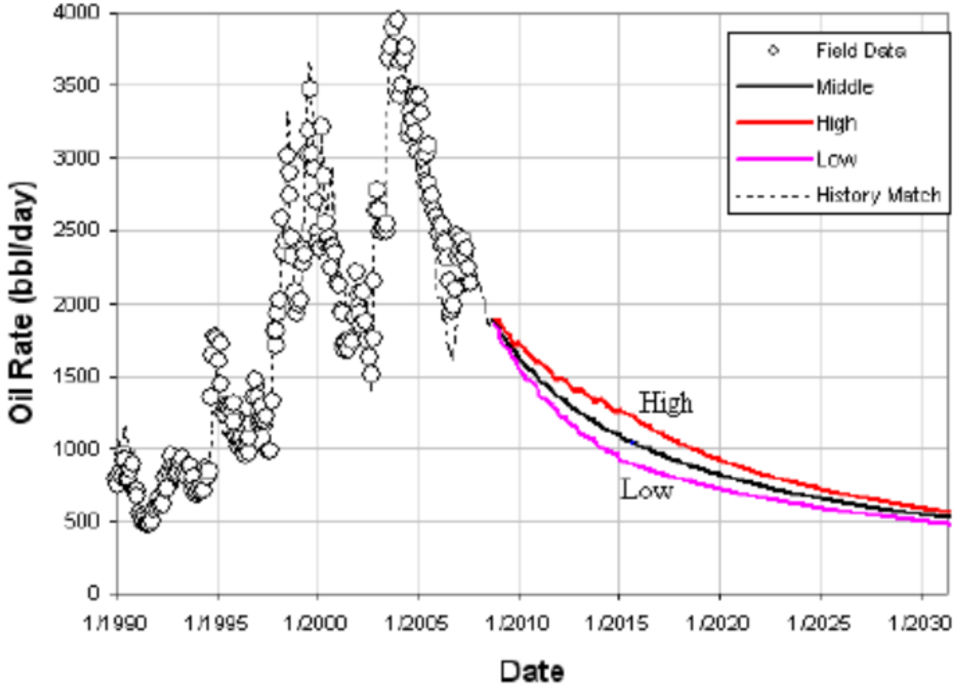
Motivational (Geo)statistics Examples

Why Use (Geo)statistics?

Forecasting / Decision Making

Decision Support

- integrated subsurface models
- to build forecast uncertainty models
- to optimize very expensive project decisions in the presence of subsurface uncertainty



Reservoir forecasting with uncertainty (Yang, 2009).

Let’s Review

Geostatistics is spatial (big) data analytics

We have big data

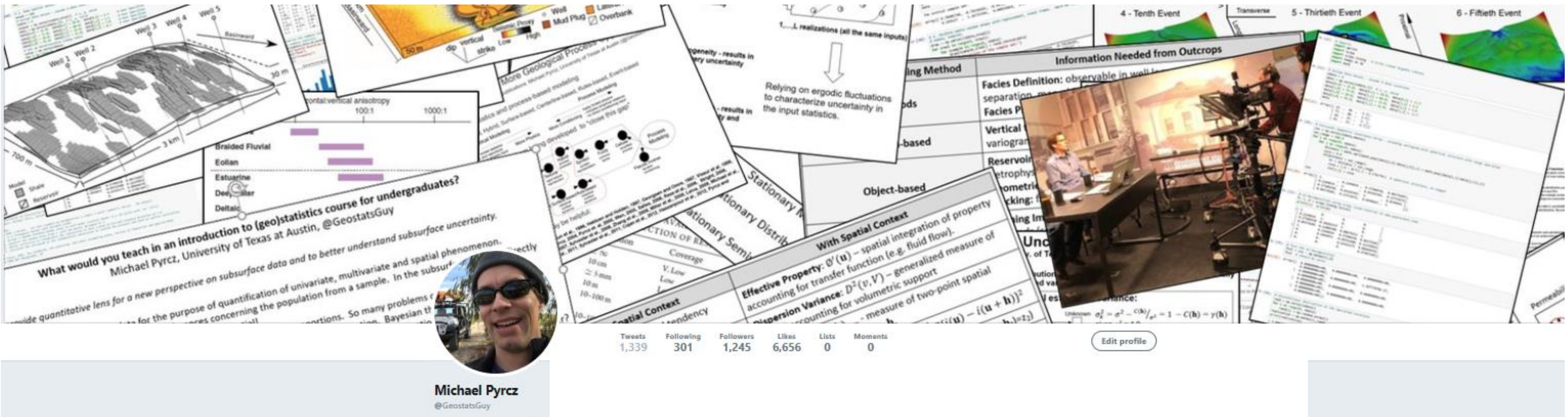
Stationarity is assumed or we are stuck in the well bore

Uncertainty is due to our own ignorance

Facies must help with spatial prediction

We can model spatial continuity

Learn More?



For tweets with Subsurface Geostatistical, Data Analytics and Machine Learning resources -

follow **@GeostatsGuy**

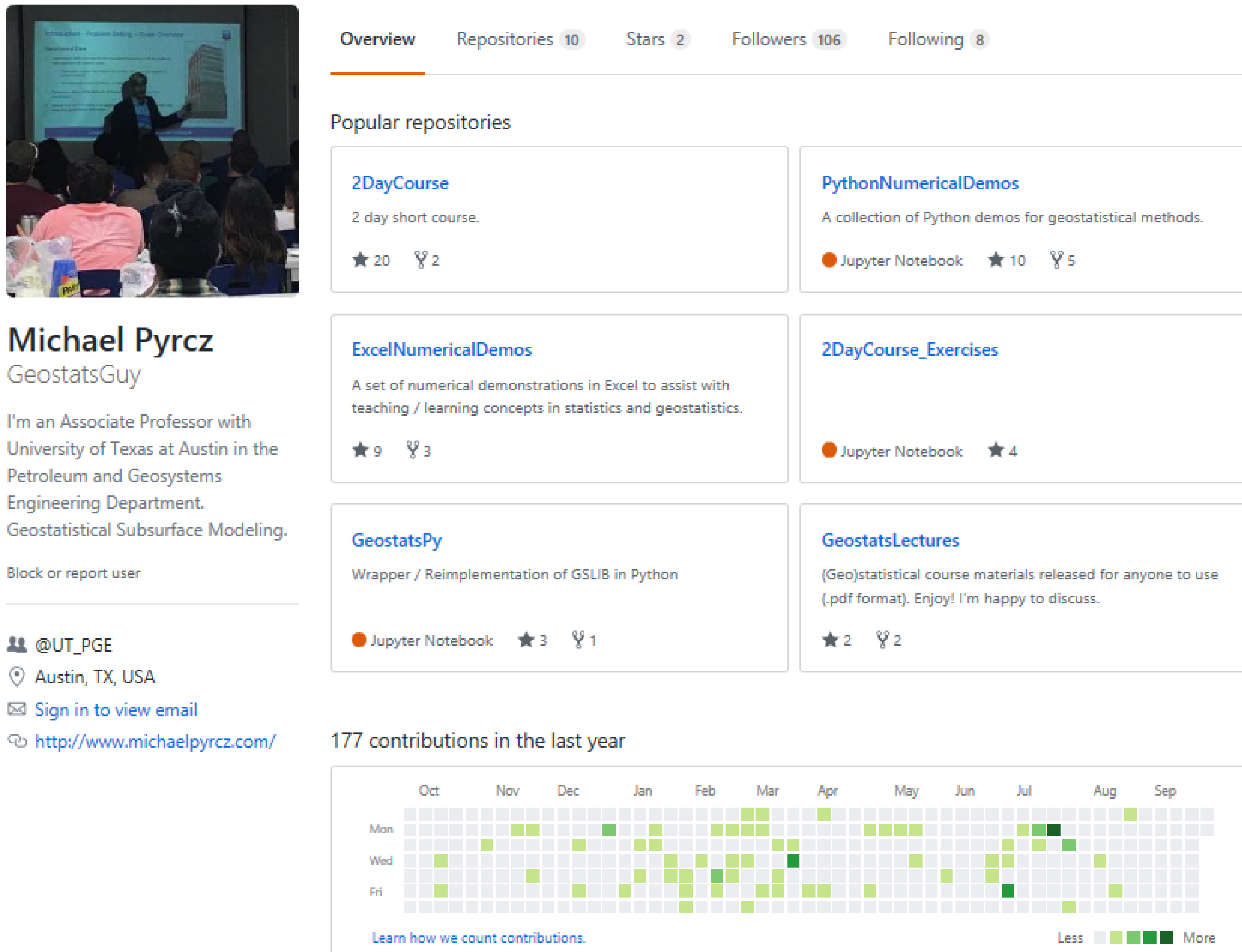
Learn More?

GitHub GeostatsGuy

Demonstration Workflows

Excel, R and Python

- Distributions
- Bootstrap
- Cellular Automata
- Hypothesis Testing
- Lorenz Coefficient
- Decision Making
- Bayesian Updating
- Kriging
- Simulation
- Volume-variance



Learn More?

For my lectures check out my YouTube Channel, ‘GeostatsGuy Lectures’.

Example Topics:

- probability theory
- frequentist vs. Bayesian statistics
- binomial distribution to model exploration success