SPATIAL DATA ANALYTICS

Introductory Lecture Outline

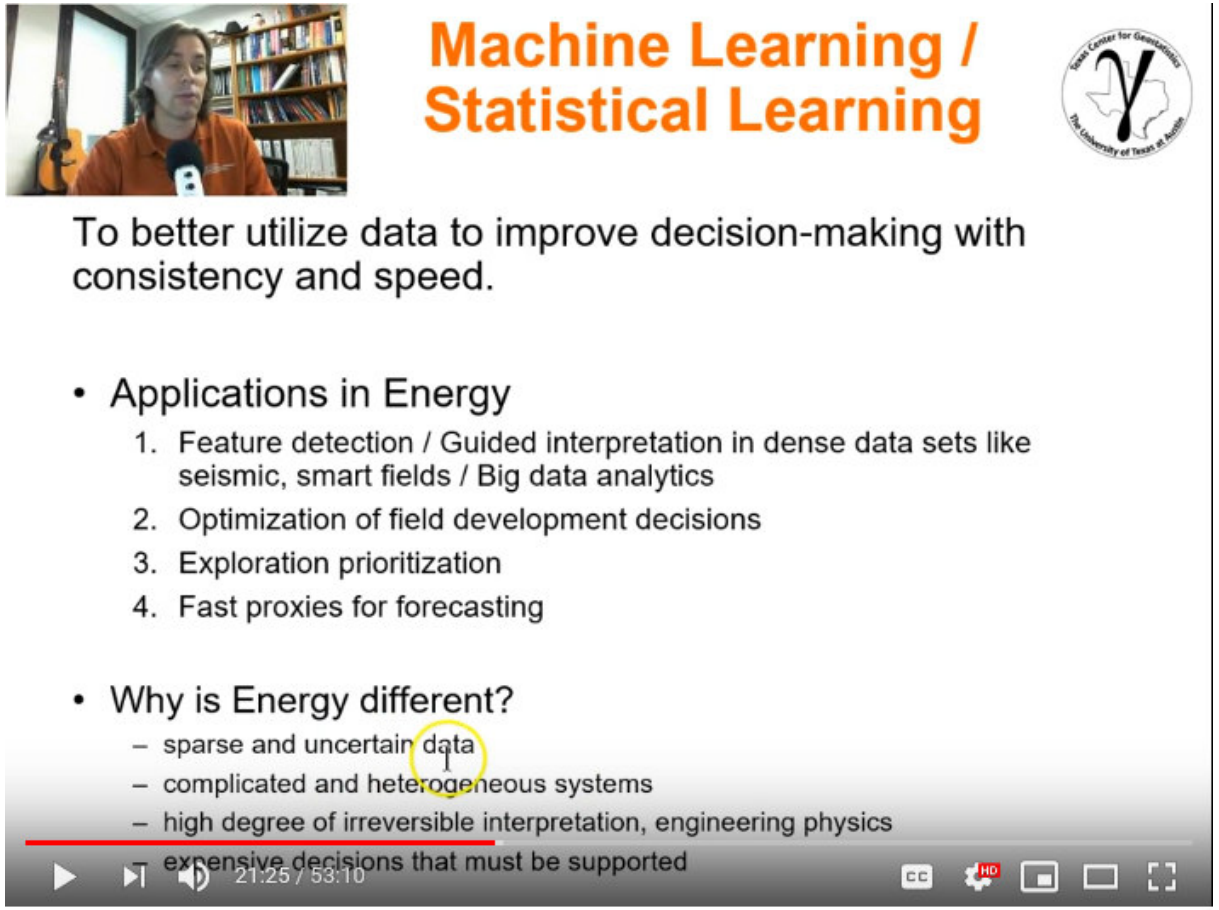
- General Comments
- Data Analytics / Geostatistics
- Machine / Statistical Learning
- Prediction and Inference

SPATIAL DATA ANALYTICS

Introduction

Other Resources:

- Recorded Lecture Statistical / Machine Learning

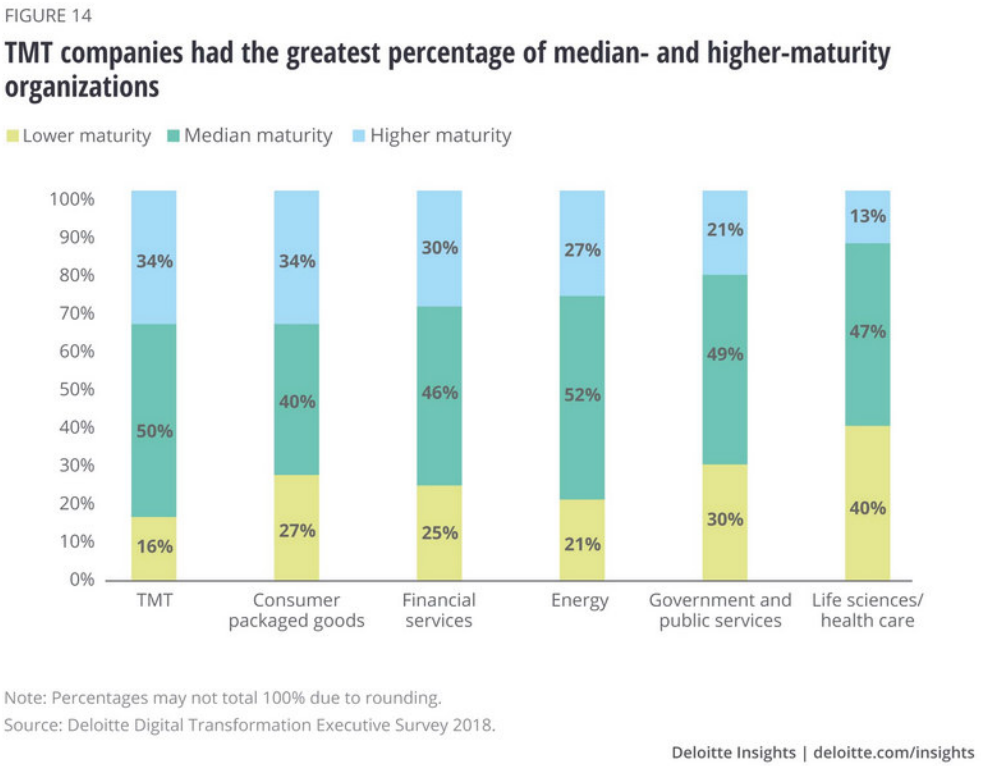


GOALS OF THIS LECTURE

- 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

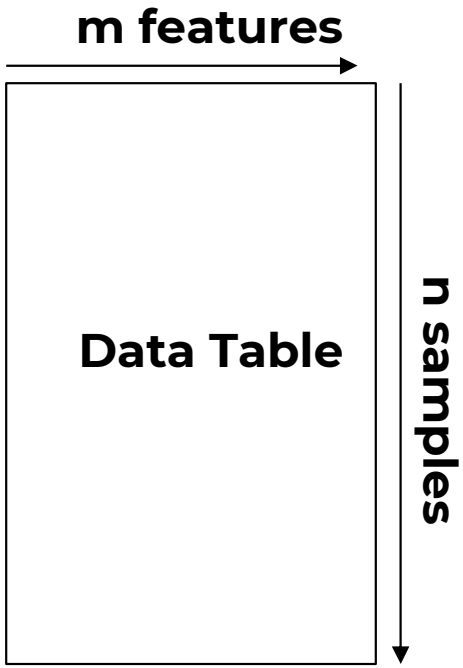
DATA ANALYTICS

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

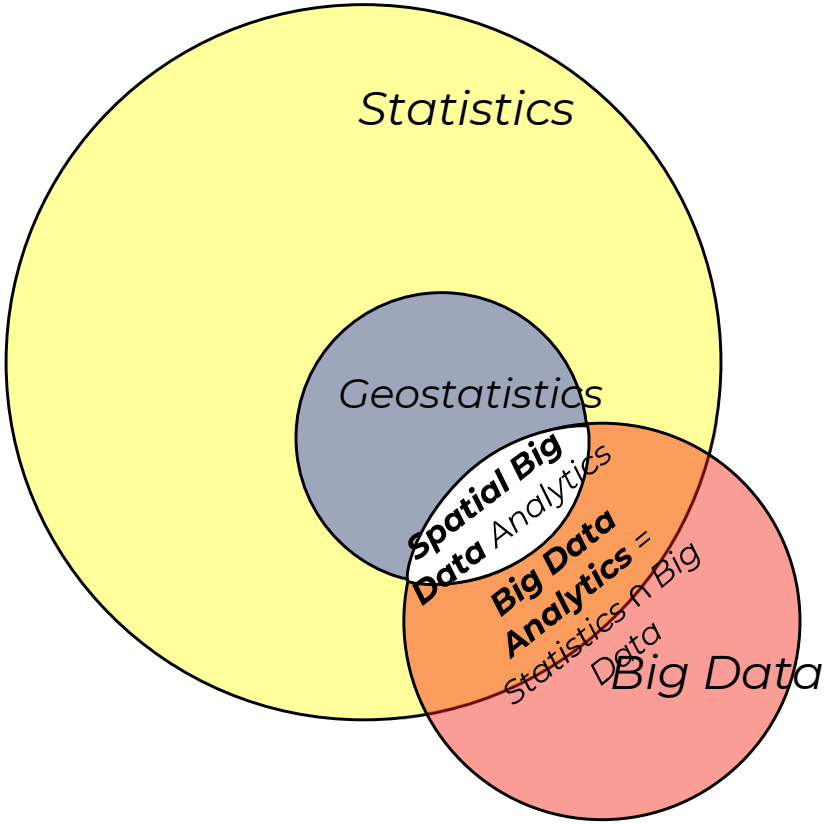
  "Energy has been big data before tech learned about big data." – Michael Pyrcz

- **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.

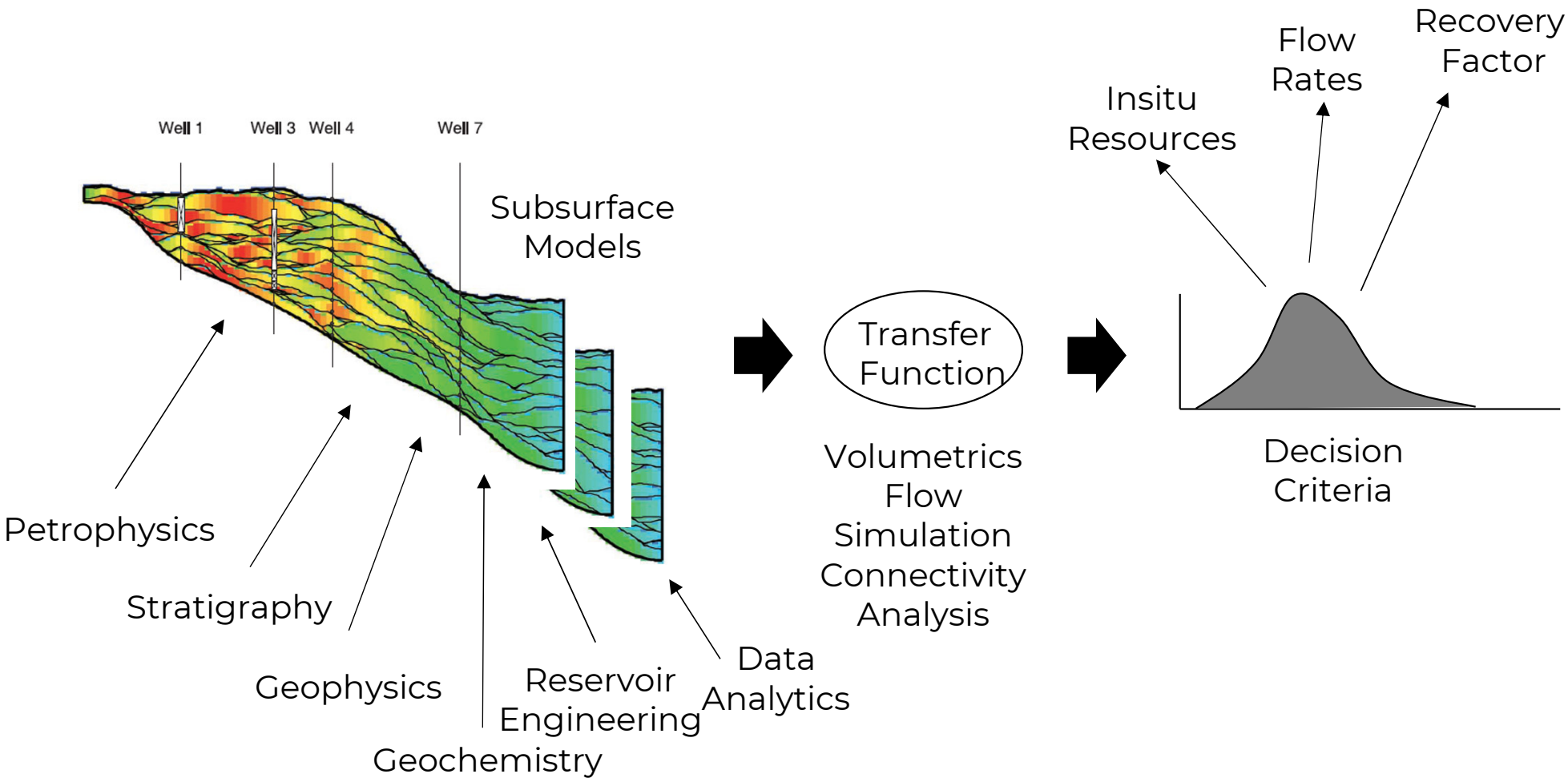


Proposed Venn diagram for spatial big data analytics.

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.

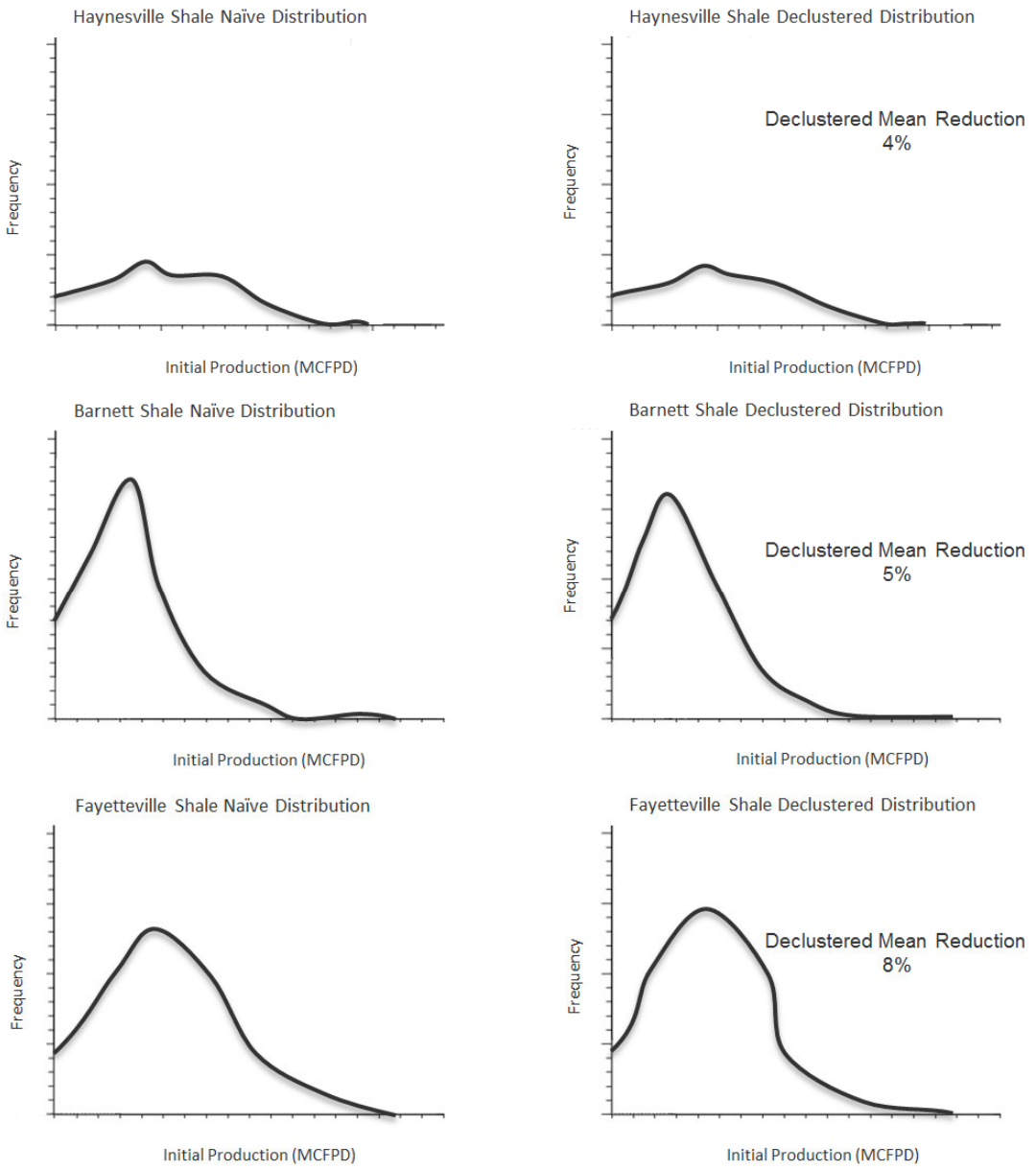
Extracted Sample at Location, $\mathbf{u}_1$

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".

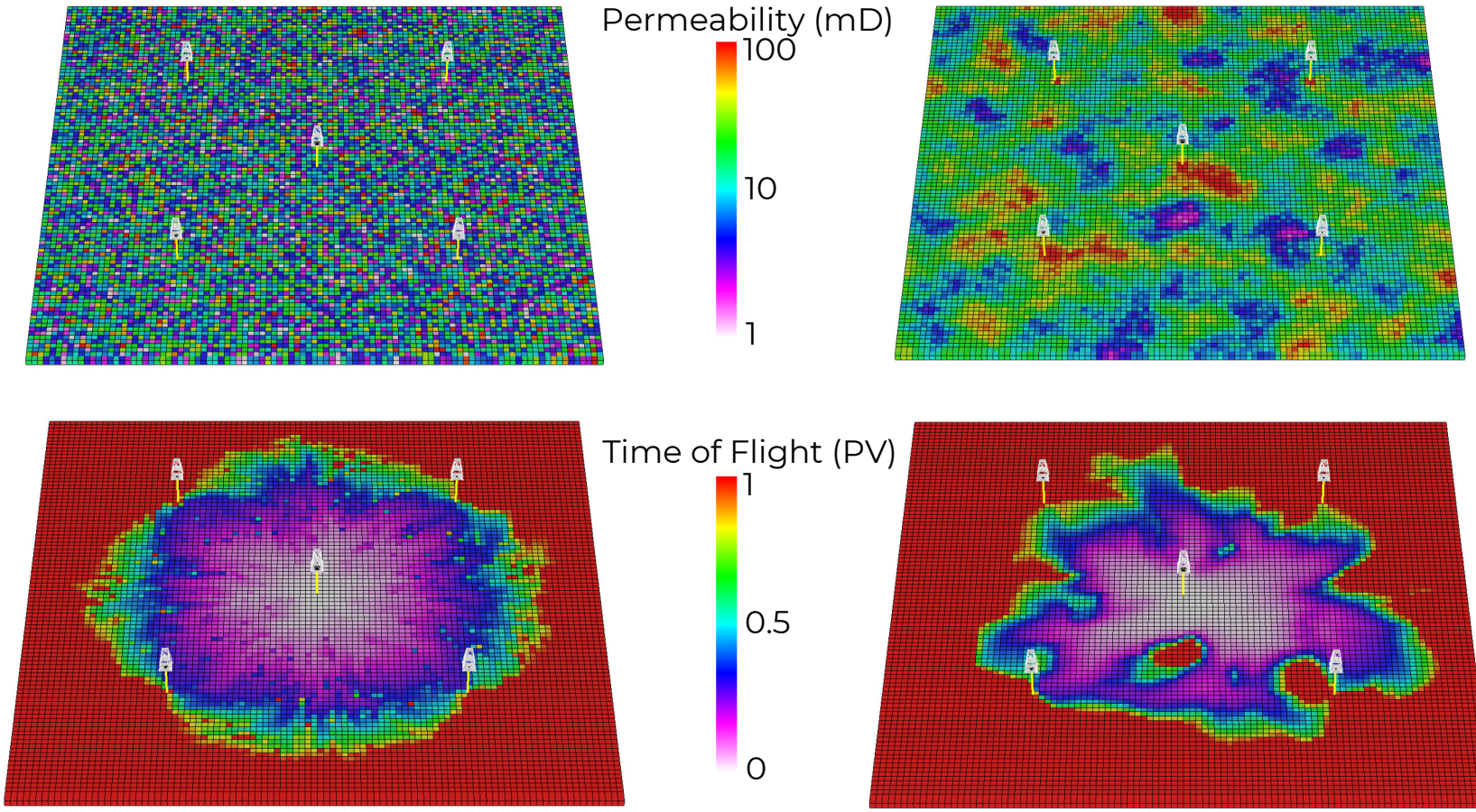
BIAS IN SUBSURFACE SAMPLING

- 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



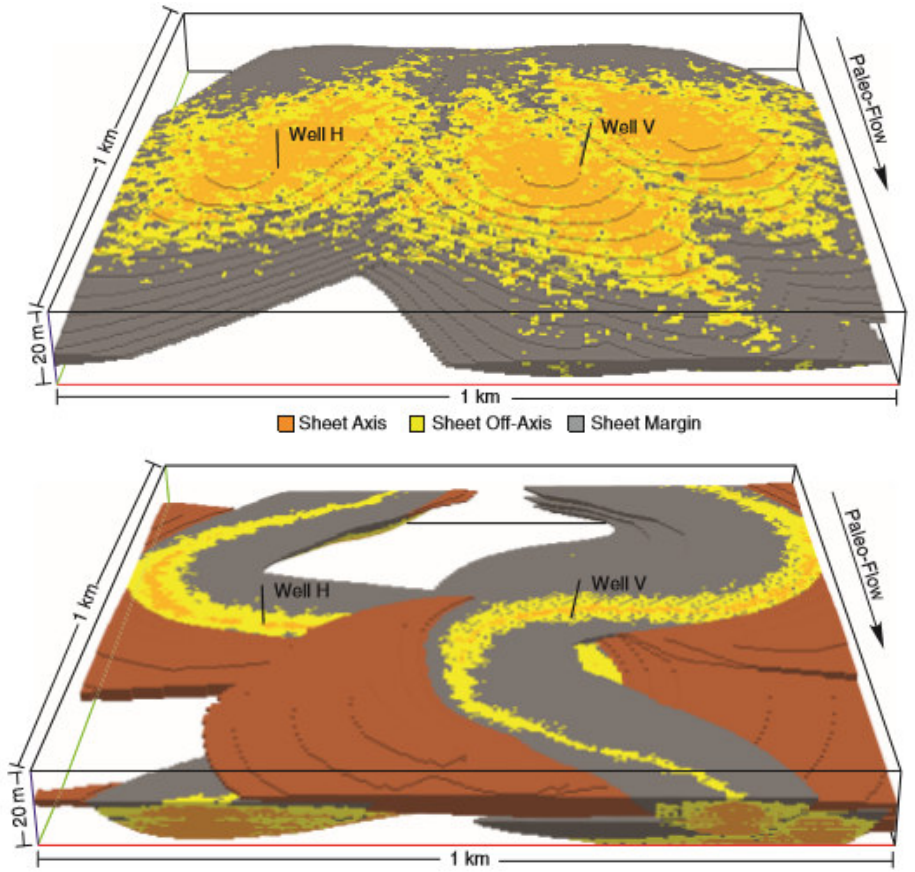
SPATIAL CONTEXT

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



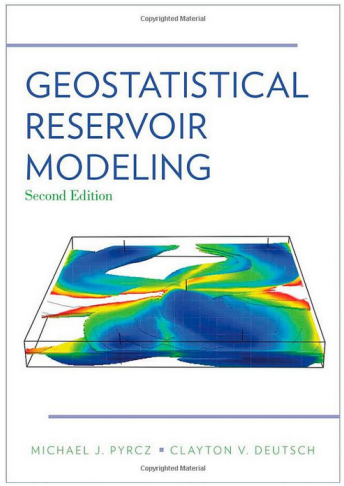
SUBSURFACE UNCERTAINTY

- 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

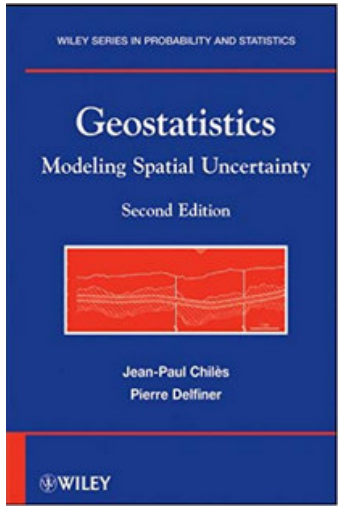


GEOSTATISTICS TEXTBOOKS

- 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*



SPATIAL, SUBSURFACE DATA

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

SPATIAL DATA ANALYTICS

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.*
Skilled Use	*Must understand the theory, assumptions and limitations of our models.* *Black box use is dangerous!*

SPATIAL DATA ANALYTICS

Lecture Outline Recap

- General Comments
- Data Analytics
- Geostatistics