PGE 338 Lecture 0:
Introduction to Data Analytics and (Geo)statistics

Lecture outline . . .

- **Who am I?**
- **Course Objectives**
- **Setting Up**
- **Some Concepts**
- **Grade Distribution and Course Expectations**
- **My Online Resources**

Introduction

Probability

Univariate

Bivariate

Time Series Analysis

Spatial Analysis

Machine Learning

Uncertainty Analysis

Michael Pyrcz, The University of Texas at Austin

My Goal this Semester

Enhance your engineering or science career with new Data Science:

Data Analytics, Geostatistics, and Machine Learning capabilities.

Demystify data analytics, geostatistics and machine learning with accessible content, lectures, demonstrations and hands-on experience.

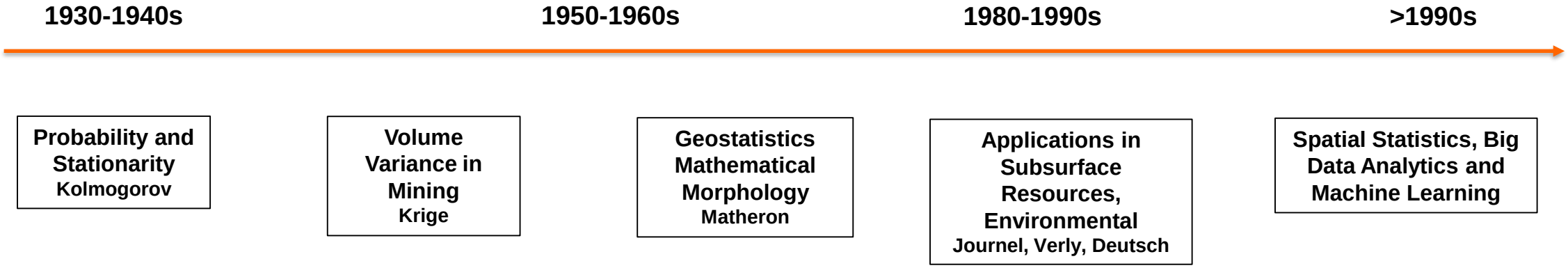
Provide you with a data-driven modeling toolkit, **change the way you see data and data related challenges**

Open the door for new opportunities for success in the digital revolution – ‘get ready to update your CV’!

I’m confident that we can do this.

An Historical Perspective

'We, subsurface engineers and geoscientists, are the original data scientists; we have been big data long before tech learned about big data!'



Complicated, heterogeneous, sparsely sampled, vast systems with complicated physics and high value decisions.

PGE 338 Lecture 0:
Introduction to Data Analytics and (Geo)statistics

Lecture outline . . .

- **Who am I?**

Introduction

Probability

Univariate

Bivariate

Time Series Analysis

Spatial Analysis

Machine Learning

Uncertainty Analysis

Michael Pyrcz, The University of Texas at Austin

Who Am I?

Spring 2018 Class of Introduction to Data Analytics and Geostatistics

Oil and Gas University, Florence, Italy

Anadarko, Midland, TX

Michael Pyrcz

1. Pyrcz: is pronounced “perch”

2. **I’ve Done This**: over 17 years of experience in consulting, teaching and industrial R&D in statistical modeling, reservoir modeling and uncertainty characterization. I left industry to join UT Austin PGE Aug. 20217.

3. I Left Industry to Teach: “I want to give you a competitive edge in your careers with geostatistics, data analytics and statistical / machine learning.”

Who Am I?

Fall 2018 Class of Introduction to Data Analytics and Geostatistics

Fall 2017 PGE 383

Fall 2022 PGE 383

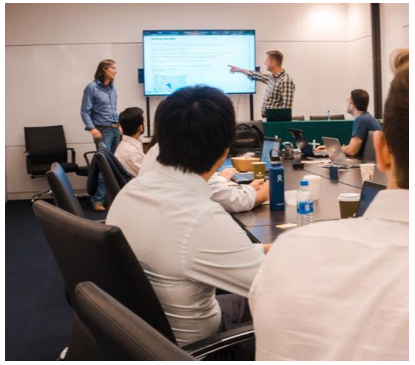
Michael Pyrcz

4. **I'm open to feedback**: got ideas, feedback to improve the learning opportunities or manage course load let me know.

5. **I'm an Engineer and a Geoscientist**: I often teach engineers, geoscientists and data scientists.

Who Am I?

Frequent Industry Courses



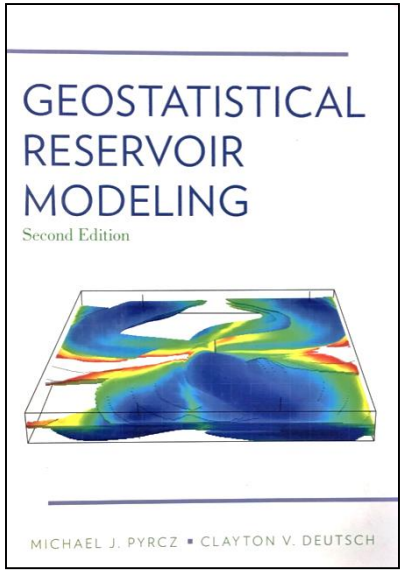
Data Science Bootcamps

Data Analytics and Machine Learning Consortium

Michael Pyrcz

6. **I Known the State of the Art and Trends**: frequently teach and consult in our industry, have a broad network and recognized for my contributions to statistics, modeling, data analytics and geostatistics theory and practice.

I'm preparing you to meet the demand and succeed.



My Book

Who Am I?

PGE Hackathon 2022

AAPG SEPM 2018 Panel Discussion on Modeling

CPGE Webinar 2017 on Big Data

2018 Cockrell Convocation

Michael Pyrcz

7. **Active in Outreach, Social Media and Professional Organizations** associate editor with Computers and Geosciences, editorial board of Mathematical Geosciences for the International Association of Mathematical Geosciences, program chair for SPE Data Analytics Technical Section

Check out GeostatsGuy on Twitter, GitHub, GeostatsGuy Lectures on YouTube

Office Hours

Michael Pyrcz

Office hours:

Wednesday 12:30 PM – 2:00 PM
Location: CPE 5.174

Friday 3:00 PM – 4:30 PM
Location: CPE 5.174

Nataly Chacon-Buitrago

Ph.D. Student
Office Hours:
Tuesday 5:00 PM - 6:30 PM
Thursday 8:00 AM - 9:30 AM
Location:
3rd Floor CPE, Caudle Center TA Room

First assignment given on Friday, January. 19th, office hours schedule starts January 19th.

Hackathon!

More information and sign up!

Information session and mixer to help make teams.

Today (Jan. 17) 5:15 pm in CPE 2.210.

My Course Notes

There is **no required textbook** for the class. I provide course notes in .pdf on the CANVAS website, along with recorded lectures on my YouTube channel.

- You may not copy nor post my course content.

Here are some books for further reading:

Books (I've asked the library to stock a couple of extras):

- Statistics for Petroleum Engineers and Geoscientists, Jensen, J. R., Lake, L. W., Corbett, P.M.W. and Goggin, D.J., Prentice Hall PTR, Upper Saddle River, New Jersey, 1997.

- Geostatistical Reservoir Modeling, Pyrcz, M.J., and Deutsch, C.V., Oxford University Press, New York, 2014.

- James, G., Witten, D., Hastie, T., Tibshirani, R., 2013, An Introduction to Statistical Learning with Applications in R, Springer (free online).

Journel Papers and additional reading assignments and references

- On canvas and links to online resources in lecture materials.

PGE 338 Lecture 0:
Introduction to Data Analytics and (Geo)statistics

Lecture outline . . .

- **Course Objectives**

Introduction

Probability

Univariate

Bivariate

Time Series Analysis

Spatial Analysis

Machine Learning

Uncertainty Analysis

Michael Pyrcz, The University of Texas at Austin

Course Learning Objectives

By the end of this course, you will be able to:

- **calculate** probabilities with frequentist and Bayesian methods
- **identify** the statistical significance of your results
- **develop** and communicate robust uncertainty models
- **quantify** and **mitigate** spatial data sampling bias
- **calculate** and **model** spatial correlation for spatial features
- **apply** spatial estimation and simulation methods to support decision making
- **summarize** and **check** your models and make optimum decisions in the presence of uncertainty
- **demystify** machine learning and **apply** inferential and predictive methods
- **design** data science workflows with open-source Python packages

To increase your impact, value added in industry as an engineer or scientist!

Course Learning Objectives

By the end of this course, you will be able to:

Given a positive X-ray test, what is the probability that this equipment has a crack?

***A* = BOP has cracks** **P(A|B) = ?**
***B* = positive X-ray test**

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)} = \frac{P(B|A)\,P(A)}{P(B|A)\,P(A) + P(B|A^c)\,P(A^c)}$$

$$\frac{\text{True Positive}}{(0.99)(0.001) + (0.02)(0.999)} = 0.047$$

True Positive **False Positive**

Blow out preventer

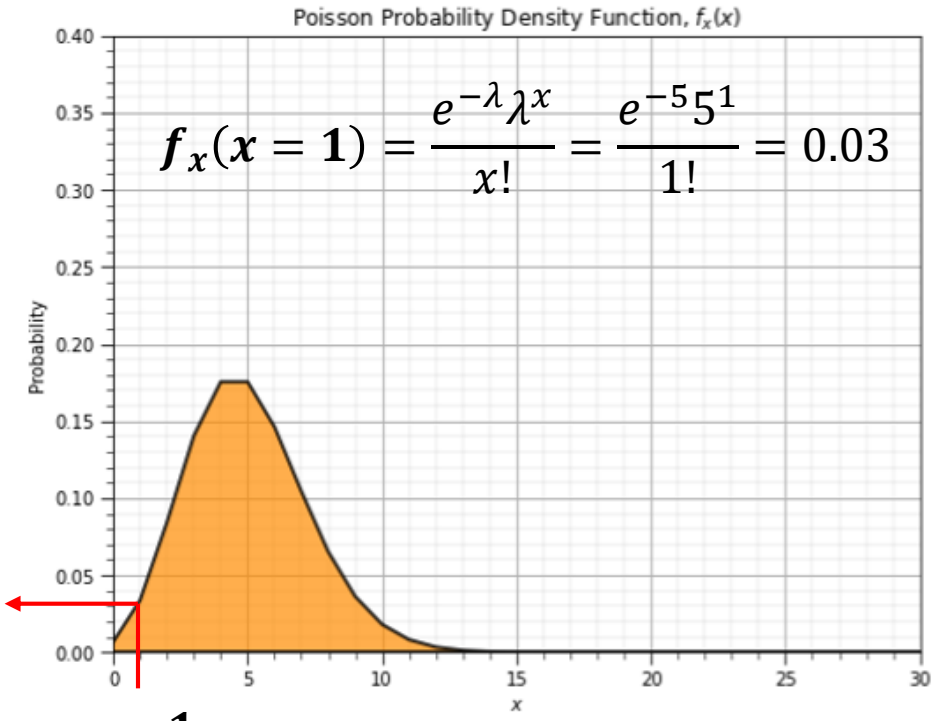
Is this a good test?

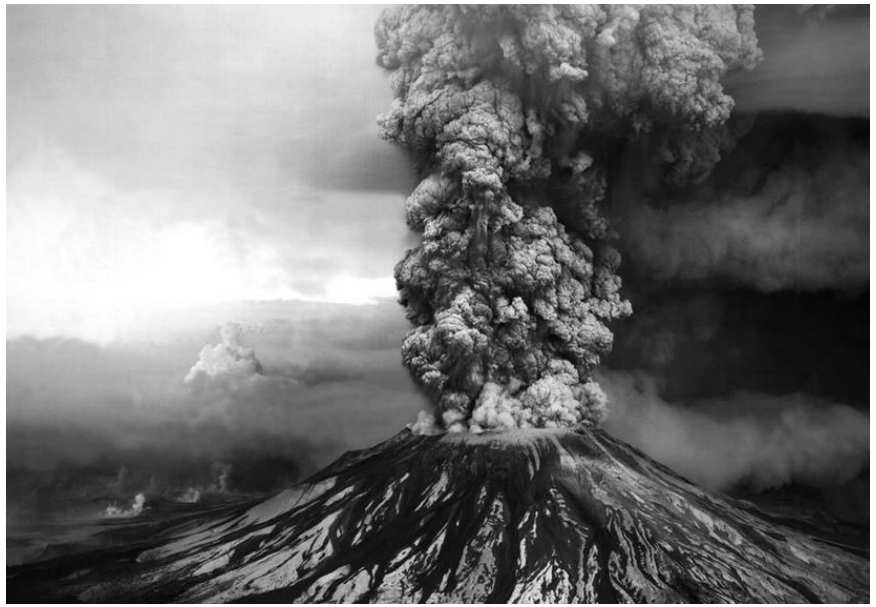
Image from https://www.arescotx.com/photos-videos/what-is-a-blowout-preventer/

Course Learning Objectives

By the end of this course, you will be able to:

What is the probability of 1 volcanic eruption in the next 1 million years, given an expectation of 5?





Mount St. Helens, May 18th,1980.

Can we relate the geologic record to contemporary risk?

Image from https://rarehistoricalphotos.com/eruption-mount-st-helens-1980/

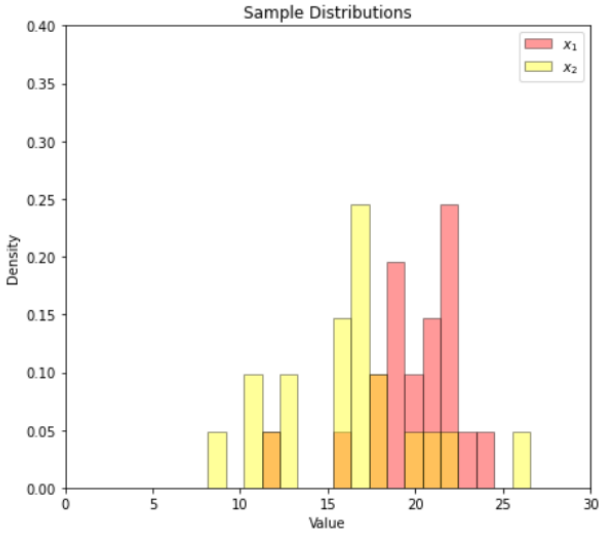
Course Learning Objectives

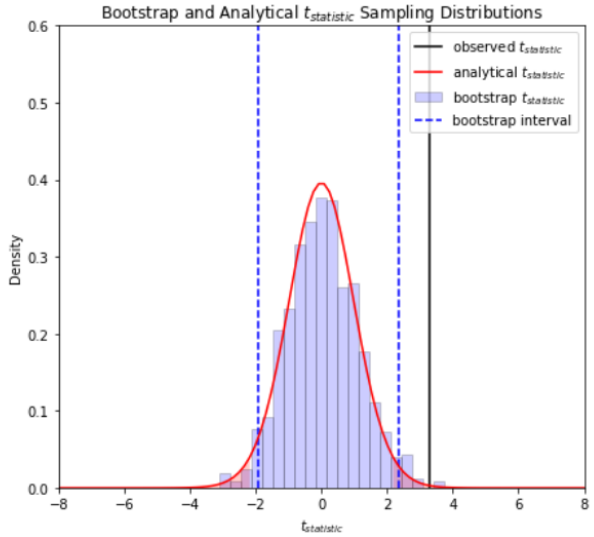
By the end of this course, you will be able to:

Is the grain size (mm) different in core x_1 and x_2?

$$\hat{t} = \frac{|16 - 20|}{\sqrt{\left(\frac{1}{20} + \frac{1}{20}\right)\left(\frac{(20-1)9.0 + (20-1)16.0}{20 + 20 - 2}\right)}} = 3.58 > t_{crit}$$

Various core samples.





Yes

Go from difference to statistically significantly different.

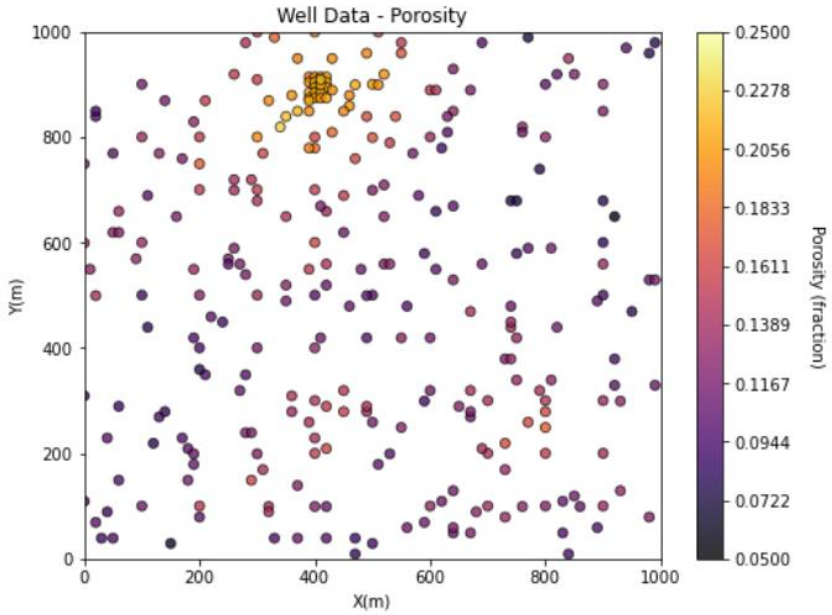
Analytical and bootstrap-based hypothesis testing, interactive Python (Interactive_Hypothesis_Testing.ipynb).

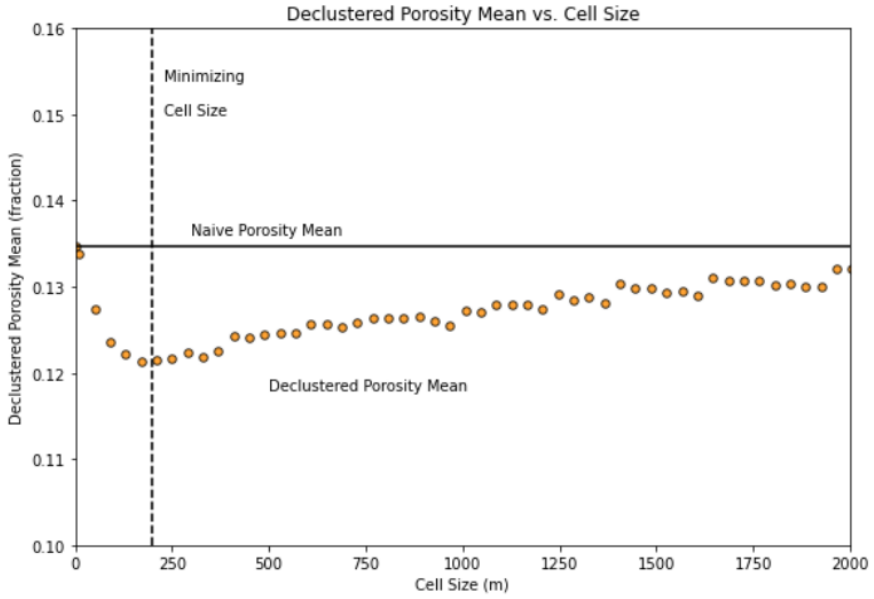
Image from https://kocurekindustries.com/core-sampling-services.

Course Learning Objectives

By the end of this course, you will be able to:

Is this data sampled in a biased or representative manner?





Cell-based spatial data declustering (GeostatsPy_declustering.ipynb).

Water well drilling on a slope.

Yes, 10% inflated.

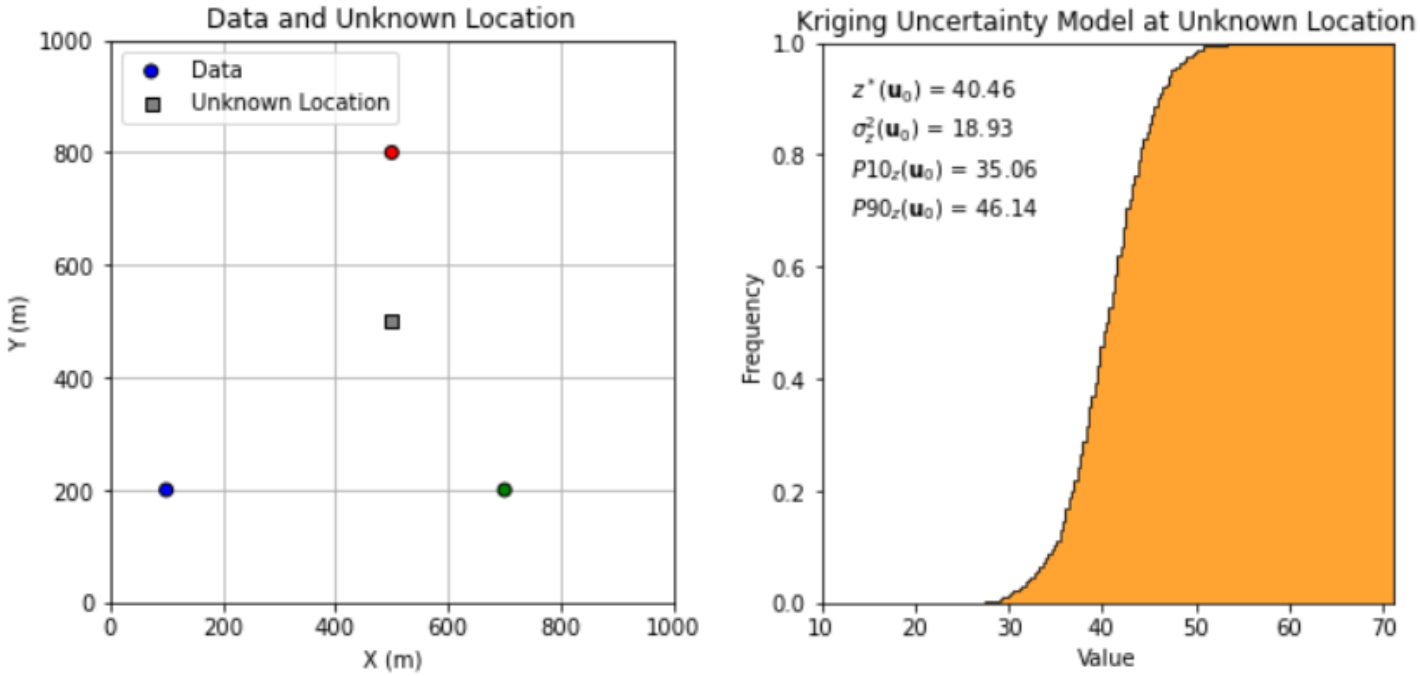
Detect and mitigate in your spatial data bias.

Image from http://www.lakevalleywell.com/.

Course Learning Objectives

By the end of this course, you will be able to:

What is the best estimate of gold grade at an unsampled location?



Simple kriging uncertainty model (Interactive_Simple_Kriging.ipynb).

Super Pit gold mine Western Australia.

40.46 g/tonne
[35.1 – 46.1] g/tonne

Spatial uncertainty models to support resource development decision making.

Image from https://en.wikipedia.org/wiki/Gold_mining

The Topics That We Should Cover

Structure for the Course

- Introduction to Geostatistics and Course Policies
- Fundamentals of Statistics and Probability
  - Frequentist and Bayesian Probability
  - Statistical Methods and Metrics
- Univariate Analysis
  - Basic Data Analysis and Display
  - Confidence Test and Hypothesis Testing
  - Measures of Heterogeneity
- Multivariate Analysis
  - Multivariate Statistics and Workflows
- Spatial Data Analysis and Modeling
  - Spatial Analysis
  - Spatial Estimation and Simulation
- Machine Learning
  - Inference and Prediction Theory and Methods

All Data Analytics and Machine Learning Builds From Probability

New Lens to See the World, Quantification

The Topics That We Should Cover

Structure for the Course

- Introduction to Geostatistics and Course Policies
- Fundamentals of Statistics and Probability
  - Frequentist and Bayesian Probability
  - Statistical Methods and Metrics
- Univariate Analysis
  - Basic Data Analysis and Display
  - Confidence Test and Hypothesis Testing
  - Measures of Heterogeneity
- Multivariate Analysis
  - Multivariate Statistics and Workflows
- Spatial Data Analysis and Modeling
  - Spatial Analysis
  - Spatial Estimation and Simulation
- Machine Learning
  - Inference and Prediction Theory and Methods

Communication with Management, Stakeholders

Integrate Uncertainty to Everything You Calculate!

Reservoir Summarization Support Development Decision Making

The Topics That We Should Cover

Structure for the Course

- Introduction to Geostatistics and Course Policies
- Fundamentals of Statistics and Probability
  - Frequentist and Bayesian Probability
  - Statistical Methods and Metrics
- Univariate Analysis
  - Basic Data Analysis and Display
  - Confidence Test and Hypothesis Testing
  - Measures of Heterogeneity
- Multivariate Analysis
  - Multivariate Statistics and Workflows → **Most Problems Are Multivariate!**
- Spatial Data Analysis and Modeling
  - Spatial Analysis
  - Spatial Estimation and Simulation
- Machine Learning
  - Inference and Prediction Theory and Methods

The Topics That We Should Cover

Structure for the Course

- Introduction to Geostatistics and Course Policies
- Fundamentals of Statistics and Probability
  - Frequentist and Bayesian Probability
  - Statistical Methods and Metrics
- Univariate Analysis
  - Basic Data Analysis and Display
  - Confidence Test and Hypothesis Testing
  - Measures of Heterogeneity
- Multivariate Analysis
  - Multivariate Statistics and Workflows
- Spatial Data Analysis and Modeling
  - Spatial Analysis → **Quantify Spatial Continuity**
  - Spatial Estimation and Simulation → **Building Models for Reservoir Production Forecasting**
- Machine Learning
  - Inference and Prediction Theory and Methods

The Topics That We Should Cover

Structure for the Course

- Introduction to Geostatistics and Course Policies
- Fundamentals of Statistics and Probability
  - Frequentist and Bayesian Probability
  - Statistical Methods and Metrics
- Univariate Analysis
  - Basic Data Analysis and Display
  - Confidence Test and Hypothesis Testing
  - Measures of Heterogeneity
- Multivariate Analysis
  - Multivariate Statistics and Workflows
- Spatial Data Analysis and Modeling
  - Spatial Analysis
  - Spatial Estimation and Simulation
- Machine Learning
  - Inference and Prediction Theory and Methods

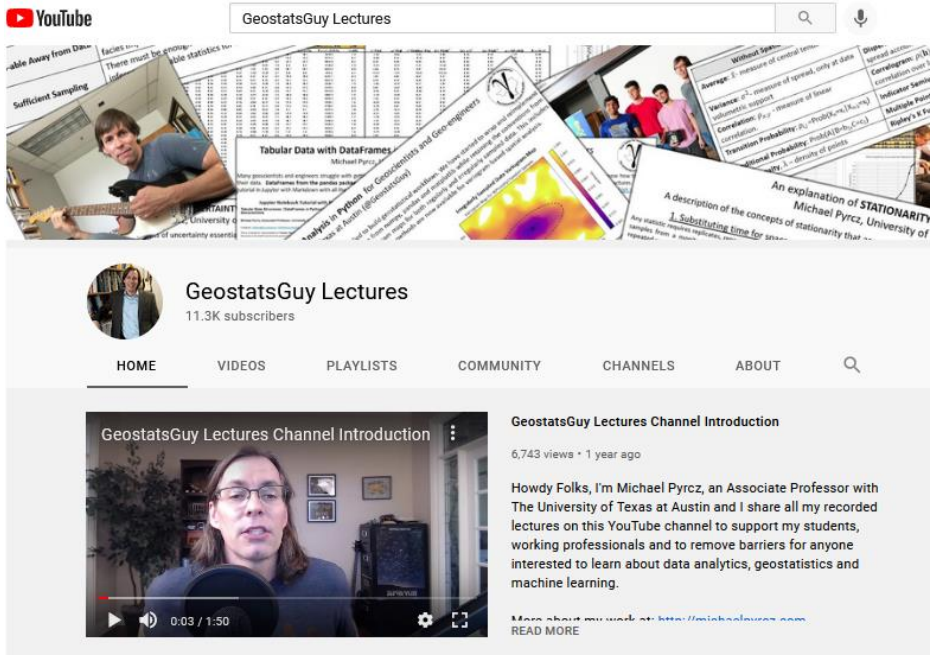
The Basics for Building Machines!

Really, How Will I Do That?

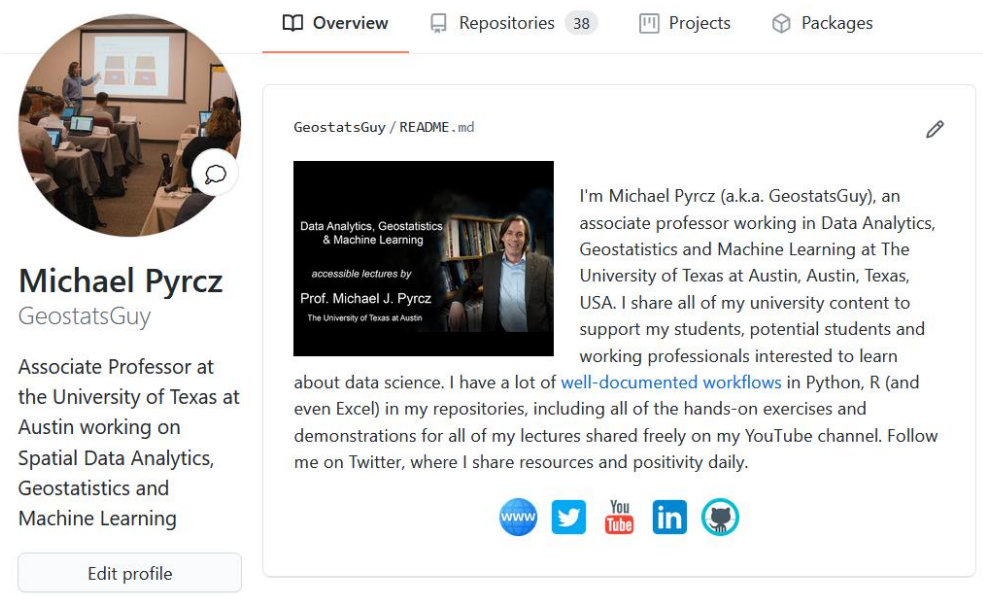
Resources:

1. In-class lectures will cover the necessary theory and methods, and we will walk though examples together.

2. All lectures, demos and workflows are available to you online on Canvas, YouTube and GitHub for review.



'GeostatsGuy Lectures' Channel on YouTube



'GeostatsGuy Repositories' on GitHub

Really, How Will I Do That?

Working with Discrete Outcomes? Binomial Distribution Demo, Michael Pyrcz, University of Texas at Austin, @GeostatsGuy

For discrete outcomes with indepence between trials and stationary probability we can appy the binomial distribution. Here's an example for a 20 well exploration program with 20% (red) and 40% (blue) probability of success.
The resuls show (1) the binomial probability density functions with the probabilities for number of successful wells and (2) the probabilities of streaks of consecutive failures

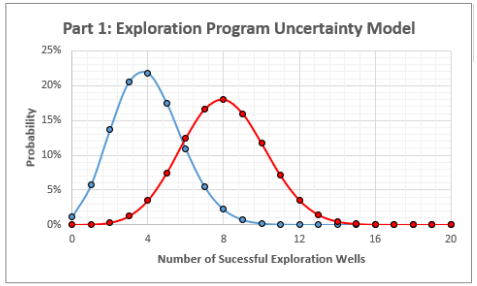
Number of Wells	Case 1: Probability of Success	Case 2: Probability of Success
20	20%	40%

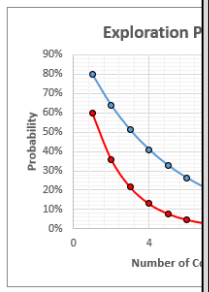
Part 1: Binomial PDF

Number of Success	Case 1: Probability	Case 2: Probability
0	1%	0%
1	6%	0%
2	14%	0%
3	21%	1%
4	22%	3%
5	17%	7%
6	11%	12%
7	5%	17%
8	2%	18%
9	1%	16%
10	0%	12%
11	0%	7%
12	0%	4%
13	0%	1%
14	0%	0%
15	0%	0%
16	0%	0%
17	0%	0%
18	0%	0%
19	0%	0%
20	0%	0%

Part 2: Probability of All Failures

Number of Failures	Number of Trials	Case 1: Probability	Case 2: Probability
1	1	80%	60%
2	2	64%	36%
3	3	51%	22%
4	4	41%	13%
5	5	33%	8%
6	6	26%	5%
7	7	21%	3%
8	8	17%	2%
9	9	13%	1%
10	10	11%	1%
11	11	9%	0%
12	12	7%	0%
13	13	5%	0%
14	14	4%	0%
15	15	4%	0%
16	16	3%	0%
17	17	2%	0%
18	18	2%	0%
19	19	1%	0%
20	20	1%	0%





Demonstration Instructions

Modify the case 1 and case 2 probablity of exploration success and observed the number of exploration su
failures.

What Should You Observe?

The binomial PDF is centered on the expectation (Prob(Success) x Number of Trials) and the variance is:
the number of trials x probability of success x probability of failure.
Streaks of consecutive failures are quite possible when failure probability is high.

Hypothesis testing, Difference in Variances, Michael Pyrcz, University of Texas at Austin, @GeostatsGuy on Twitter

This is the one tailed, f-test for difference in variances. = $H_0: \sigma_1^2 = \sigma_2^2$, and $H_1: \sigma_1^2 > \sigma_2^2$, where we constrain σ_1^2 to be the larger variance.
Sample sets X1 and X2 are assumed to come from a Gaussian distribution.

Parameters

X1: mean 0.05, stdev 1.2

X2: mean 0, stdev 1.5

Sample Data

X1	X2
0.83	-0.18
1.25	0.82
0.41	0.26
0.15	2.57
0.80	0.55
-0.04	-0.32
3.15	1.02
-0.45	2.47
-0.39	-1.95
-0.75	-0.26
2.56	-0.57
1.35	1.14
2.01	-0.57
0.86	0.24
1.90	1.34
1.11	-0.59
-0.32	-1.12
-1.22	-0.35
0.04	-1.41
0.66	2.01
-1.79	-0.37
-1.38	0.60
1.52	-3.45
-0.16	-1.26
1.41	2.56
-0.77	3.05
-0.80	1.71
0.51	-4.19
-0.81	1.33
0.36	-1.15
1.43	-2.27
-0.63	0.31
-0.95	-0.36
1.38	-0.78
0.36	-1.76
-1.48	0.63
0.60	-1.73
-0.89	-1.20

Histograms

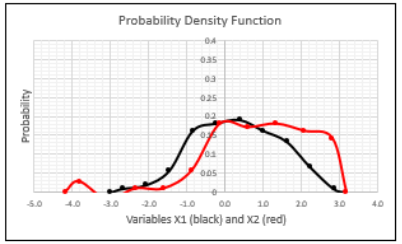
X1: min -3.0, max 3.1

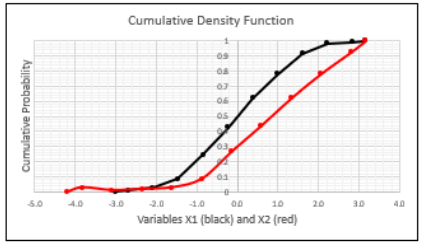
Bins	-3.00	-2.39	-1.77	-1.16	-0.54	0.07	0.69	1.30	1.92	2.53	
	-2.39	-1.77	-1.16	-0.54	0.07	0.69	1.30	1.92	2.53	3.15	
X1	-3.0	-2.69	-2.08	-1.46	-0.85	-0.23	0.38	1.00	1.61	2.23	2.84	3.1
Prob	0	0.01	0.02	0.06	0.16	0.18	0.19	0.16	0.13	0.07	0.01	0.00
Cum. Prob	0	0.01	0.03	0.09	0.25	0.43	0.62	0.78	0.91	0.98	0.99	1.00

Histograms

X2: min -4.2, max 3.2

Bins	-4.19	-3.57	-2.96	-2.34	-1.73	-1.11	-0.50	0.12	0.73	1.35	
	-3.57	-2.96	-2.34	-1.73	-1.11	-0.50	0.12	0.73	1.35	1.96	
X2	-4.2	-3.82	-3.08	-2.35	-1.61	-0.88	-0.14	0.59	1.33	2.07	2.80	3.17
Prob	0	0.03	-0.02	0.01	0.01	0.06	0.18	0.17	0.18	0.16	0.14	0.00
Cum. Prob	0	0.03	0.01	0.02	0.03	0.09	0.27	0.44	0.62	0.78	0.92	1.00





f-test for difference in variances

1. Sample Statistics

X1: mean 0.32, var 1.32, count 105

X2: mean -0.07, var 2.64, count 105

2. Specify Alpha Level

Level 0.05

3. Calculate f_{stat} and $f_{critical}$

fstat 2.00

fcritical 1.38 F_{α,N_1-1,N_2-1}

P(F<=f) one tail 0.02%

4. Check criteria

f_{stat}		$f_{critical}$	Test
2.005	<	1.383	$F > F_{\alpha,N_1-1,N_2-1}$

H1: Reject the null hypothesis.

5. Evaluate p-value.

p-value 0.02% < 5.0%

Reject if probability is less than confidence level.

6. Check with EXCEL built-in F-test

p-value 0.02%

Excel demonstrations available at:

https://github.com/GeostatsGuy/ExcelNumericalDemos

Really, How Will I Do That?

GeostatsPy: Confidence Intervals and Hypothesis Testing for Subsurface Data Analytics in Python

Michael Pyrcz, Associate Professor, University of Texas at Austin

Twitter | *GitHub* | *Website* | *GoogleScholar* | *Book* | *YouTube* | *LinkedIn*

Reporting Uncertainty and Significance

With confidence intervals and hypothesis testing we have the opportunity to report uncertainty and to report significance in our sta
on standard methods with their associated limitations and assumptions. See the lecture and workflow on Bootstrap https://git.io/fh
approach to assess uncertainty in statistics.

This is a tutorial / demonstration of **Confidence Intervals and Hypothesis Testing in Python** for Subsurface Modeling. In Pyth
the Stats functions (https://docs.scipy.org/doc/scipy/reference/stats.html) provide excellent tools for efficient use of statistics.

I have previously provided these examples worked out by-hand in Excel (https://github.com/GeostatsGuy/LectureExercises/blob/m
/Lecture7_CI_Hypoth_eg_R.xlsx) and also in R (https://github.com/GeostatsGuy/LectureExercises/blob/master/Lecture7_CI_Hypoth
dataset available as a comma delimited file (https://git.io/fxLAt).

This tutorial includes basic, typical confidence interval and hypothesis testing methods that would commonly be required for Engine

1. Student-t confidence interval for the mean
2. Student-t hypothesis test for difference in means (pooled variance)
3. Student-t hypothesis test for difference in means (difference variances), Welch's t Test
4. F-distribution hypothesis test for difference in variances

Caveats

I have not included all the details, specifically the test assumptions in this document. These are included in the accompanying cours

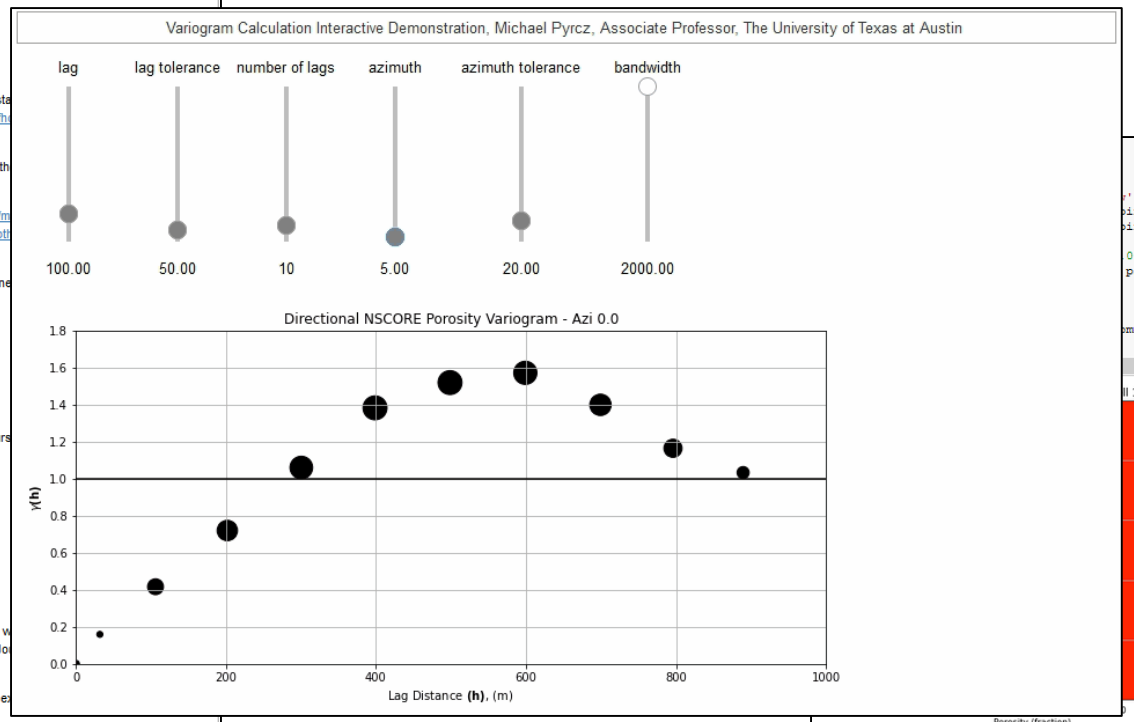
Workflow Goal

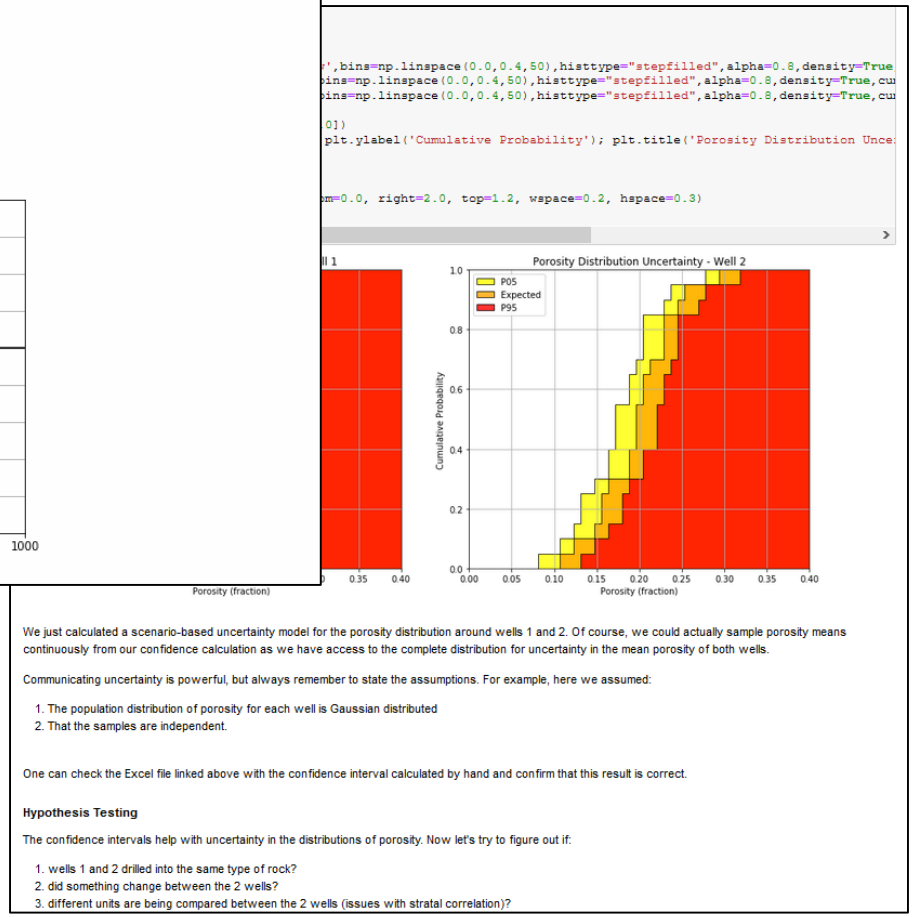
0. Introduction to Python in Jupyter including setting a working directory, loading data into a Pandas DataFrame.
1. Learn the basics for working with confidence intervals and hypothesis testing in Python.
2. Demonstrate the efficiency of using Python and SciPy package for statistical analysis.
3. Learn how to quantify uncertainty and significance in samples.

Objective

In the PGE 383: Stochastic Subsurface Modeling class I want to provide hands-on experience with building subsurface modeling w
excellent vehicle to accomplish this. I have coded a package called GeostatsPy with GSLIB: Geostatistical Library (Deutsch and Jo
provides basic building blocks for building subsurface modeling workflows.

The objective is to remove the hurdles of subsurface modeling workflow construction by providing building blocks and sufficient ex
per se, but we need the ability to 'script' workflows working with numerical methods.





We just calculated a scenario-based uncertainty model for the porosity distribution around wells 1 and 2. Of course, we could actually sample porosity means continuously from our confidence calculation as we have access to the complete distribution for uncertainty in the mean porosity of both wells.

Communicating uncertainty is powerful, but always remember to state the assumptions. For example, here we assumed:

1. The population distribution of porosity for each well is Gaussian distributed
2. That the samples are independent.

One can check the Excel file linked above with the confidence interval calculated by hand and confirm that this result is correct.

Hypothesis Testing

The confidence intervals help with uncertainty in the distributions of porosity. Now let's try to figure out if:

1. wells 1 and 2 drilled into the same type of rock?
2. did something change between the 2 wells?
3. different units are being compared between the 2 wells (issues with stratal correlation)?

Python Workflows in Jupyter notebook demonstrations available at:

Canvas/Files/Python

and more available at:

https://github.com/GeostatsGuy/PythonNumericalDemos

Software

GeostatsPy

- Set of Functions in Python
  - GeostatsPy is a set of Python functions for most of the required geostatistical methods
  - Much is reimplemented in Python.
  - This wrapper is a very simple, write out the parameter file and the data, run GSLIB executable, and read in the results
  - Written by myself, we will use for the geostatistics.

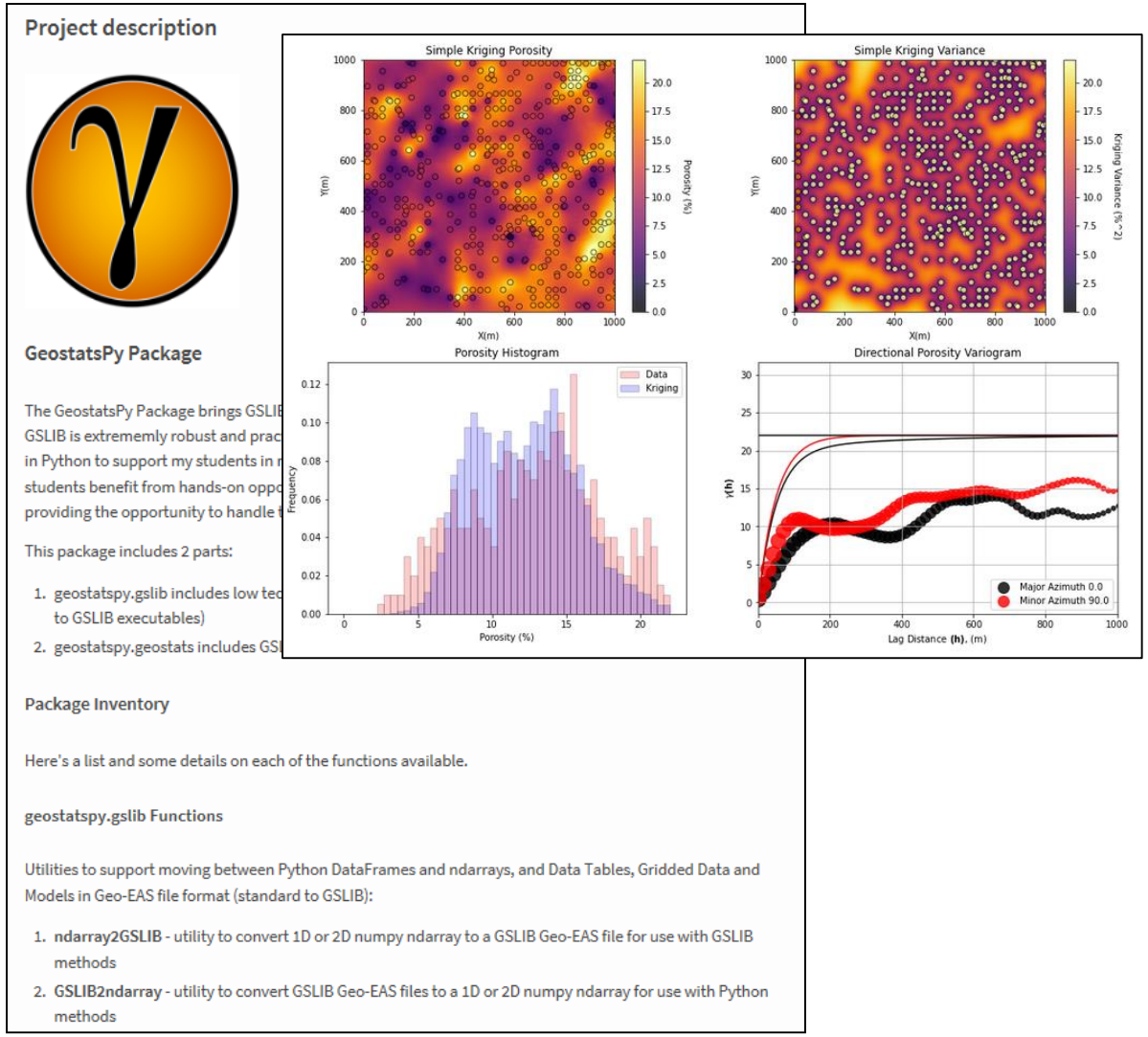
- I wrote this package to support my courses.

GeostatsPy is on PyPI:

https://pypi.org/project/geostatspy/

and source code is on GitHub:

https://github.com/GeostatsGuy/GeostatsPy



PGE 338 Lecture 0:
Introduction to Data Analytics and (Geo)statistics

Lecture outline . . .

- **Setting Up**

Introduction

Probability

Univariate

Bivariate

Time Series Analysis

Spatial Analysis

Machine Learning

Uncertainty Analysis

Michael Pyrcz, The University of Texas at Austin

Getting Setup on Your Laptop

Time to Get Technical, For Next Class

1. Install Anaconda on your Laptop, that's Python!

2. Install the GeostatsPy, my geostatistics/spatial data analytics python package

3. Bring your laptop to class from now on

4. Have the Jupyter notebook workflows loaded up in class. See the links on the Canvas topics pages.

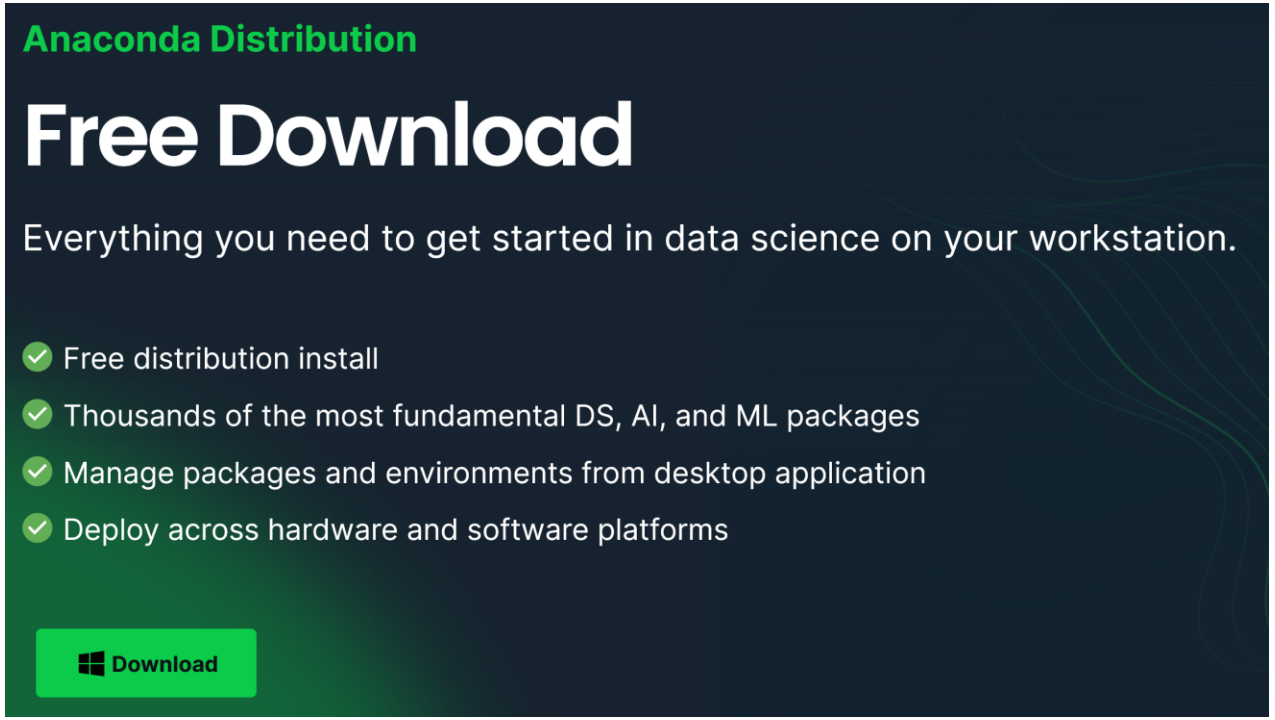
We are going to start working through methods / workflows in class.

Getting Setup on Your Laptop

Time to Get Technical, For Next Class Please

- We are going to start working through demonstrations in class

Install Anaconda with Python on your Laptop



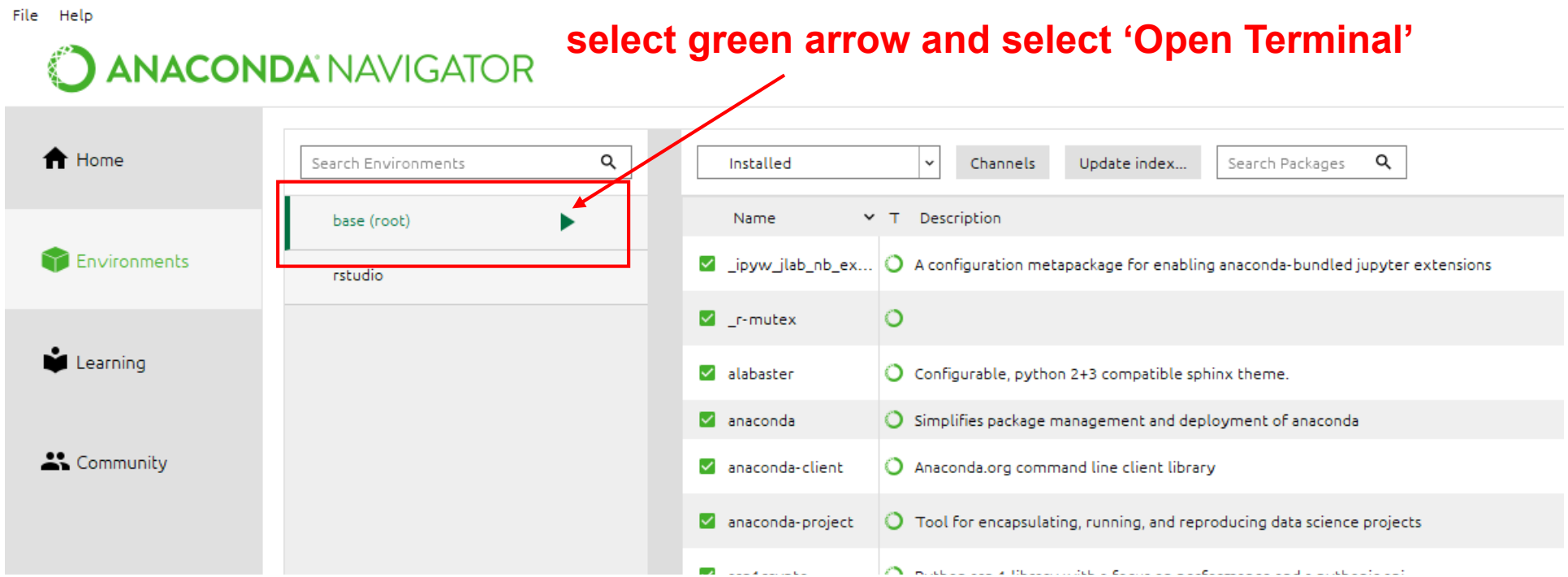
If you have an older version of Anaconda, please uninstall and install current version.

Getting Setup on Your Laptop

Time to Get Technical, For Next Class Please

Install GeostatsPy package:

Then open Anaconda Navigator (one of the programs with Anaconda3)



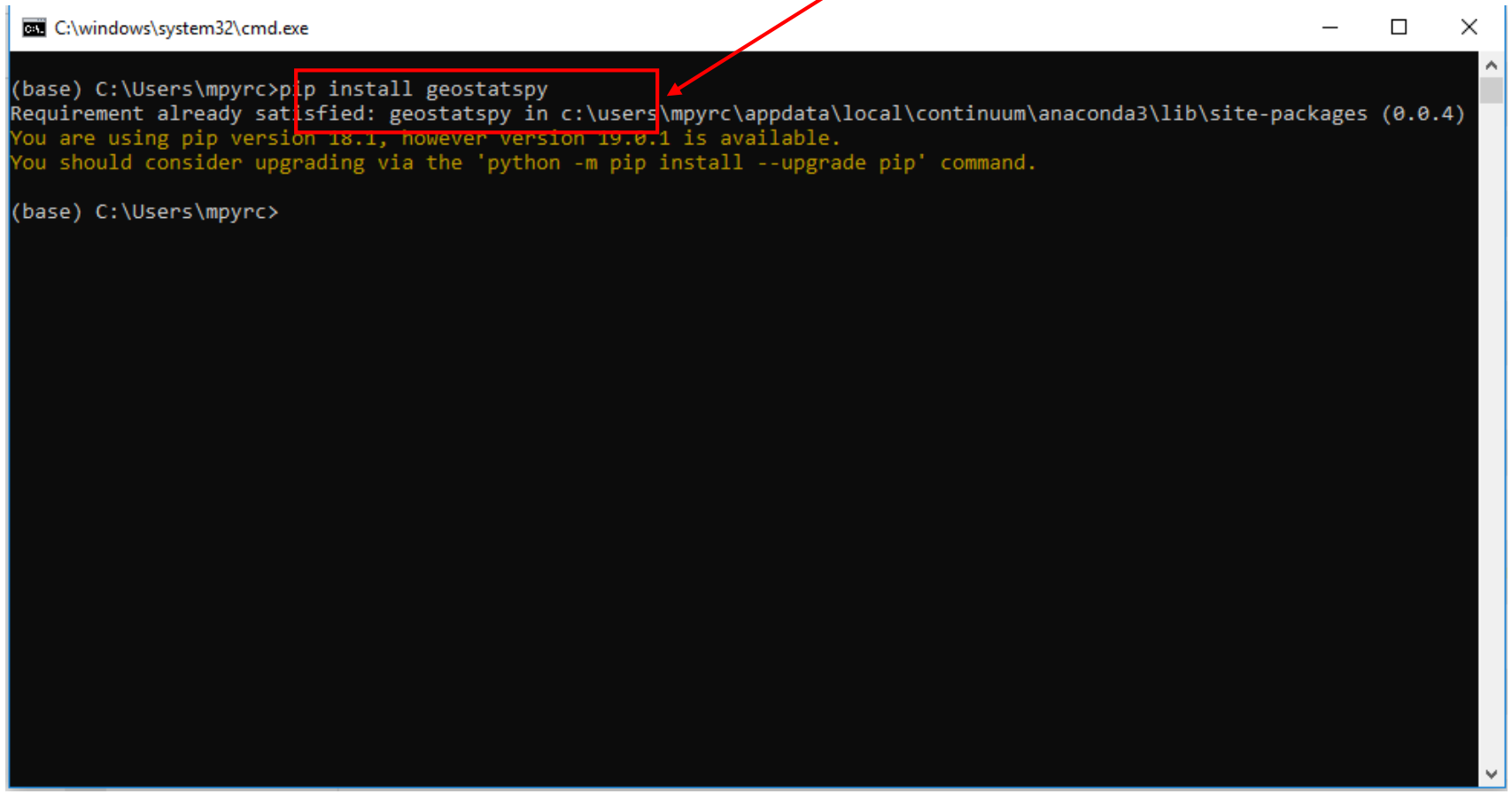
Anaconda Navigator is an application included with the Anaconda installation to manage environments and packages.

Getting Setup on Your Laptop

Time to Get Technical, For Next Class Please

Install GeostatsPy package:

type ‘pip install geostatspy’



Terminal window launch on the previous slide.

PGE 338 Lecture 0:
Introduction to Data Analytics and (Geo)statistics

Lecture outline . . .

- **Some Concepts**

Introduction

Probability

Univariate

Bivariate

Time Series Analysis

Spatial Analysis

Machine Learning

Uncertainty Analysis

Michael Pyrcz, The University of Texas at Austin

(Geo)statistics
What Will You Learn in This Course?

Class Goals

*Geostatistics **provide quantitative lens for a new perspective** on subsurface data and to better understand subsurface development uncertainty.*

1. **Deductive Statistics** – pooling data for the purpose of quantification of univariate, multivariate and spatial phenomenon.
2. **Inferential Statistics** – methods to make inferences concerning the population from a sample. In the subsurface we only directly sample about 1 trillionth of the reservoir; therefore, essential!
3. **Frequentist Statistics** – drawing conclusions based on frequencies or proportions. So many problems can be solved by counting!
4. **Bayesian Statistics** – drawing conclusions based on updating belief with new information. Bayesian theorem can be solved difficult problems such as the probability of a subsurface depositional setting, given well core data observations.
5. **Statistical Representativity** – sampling for representativity and treatment of bias. All subsurface datasets are biased!

(Geo)statistics
What Will You Learn in This Course?

Class Goals

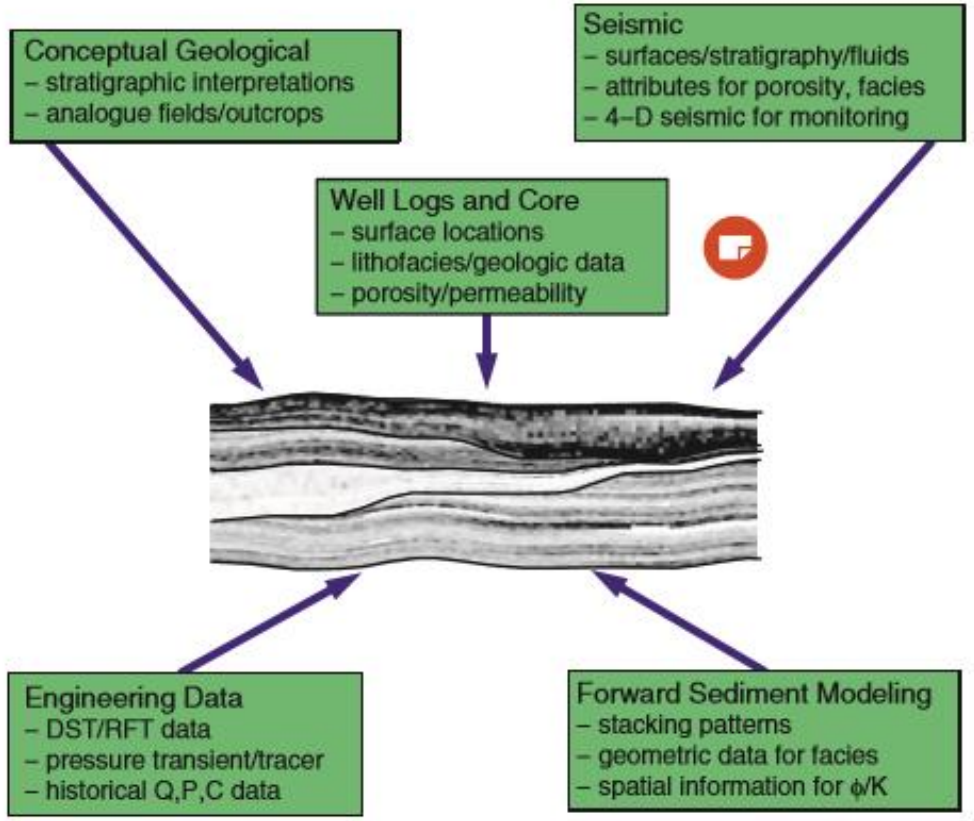
*Geostatistics **provide quantitative lens for a new perspective** on subsurface data and to better understand subsurface development uncertainty.*

6. **Statistical Significance** – testing the significance of results with hypothesis testing and confidence intervals. Does the result matter?
7. **Spatial Modeling** – the subsurface has spatial structures; therefore, by capturing them we get much better models.
8. **Statistical Modeling** – data-driven modeling and prediction. Regression to machine learning, let's join the data-driven paradigm!
9. **Uncertainty Models** – modeling, summarizing and making decisions in the presence of uncertainty. Uncertainty is ubiquitous and due to our ignorance.
10. **Big Data Analytics** – working with high volume, variety and velocity data inherent to subsurface exploration. We have big data too!

(Geo)statistics
What Will You Learn in This Course?

Why should you have a greater proficiency in geostatistics?

1. Most reservoir asset subsurface teams develop **geostatistical models**. If you work with the subsurface, you will work with statistical, stochastic reservoir models!
2. Geostatistical modeling sits in the **middle of the subsurface team and integrates all available engineering**, geological and geophysical information. Improved reservoir modeling capability results in improved communication and integration in the subsurface team.
3. Geostatistical models are directly applied for **forecasting that support decision making**. If your expertise does NOT impact the model, you may NOT impact the development decision!
4. Black box, naïve geostatistical modeling will result in **bad decisions**. You'll learn enough about what is going on "under the hood" to critically evaluate and improve the models!



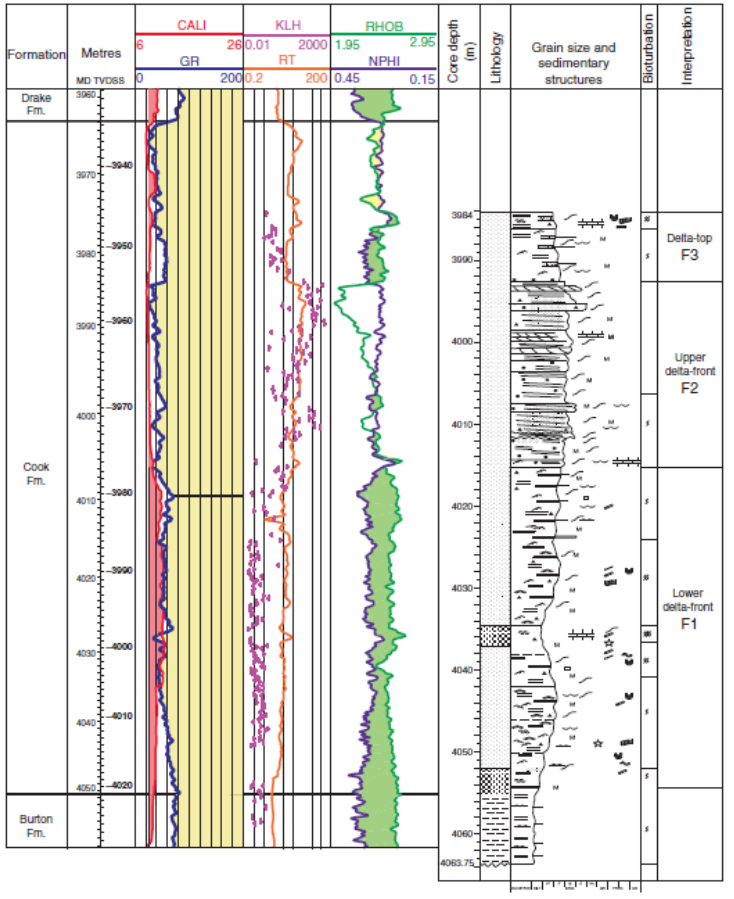
Data integration to build geostatistical reservoir models.

Geostatistics, why?

Why should you have a greater proficiency in geostatistics?

There is so much more than reservoir modeling with geostatistics.

- Analyze induced seismicity
- Assess impact of completions
- Well bore stability
- Detect features in seismic
- Exploration basin analysis
- Uncertainty in well logs



Suite of Well Logs with Interpreted Structures from the Core Data and Stratigraphic Units Form the Cook Formation, a Shallow Marine Sandstone Reservoir from the North Sea. The core has been interpreted as a fluvial/deltaic depositional setting with general progradation upward (note the general coarsening upward) and used to calibrate the log response (Folkestad et al. 2012).

Data Analytics, why?

Why should you have a greater proficiency in data analytics?

There is demand for engineers, scientists with data analytics skills.

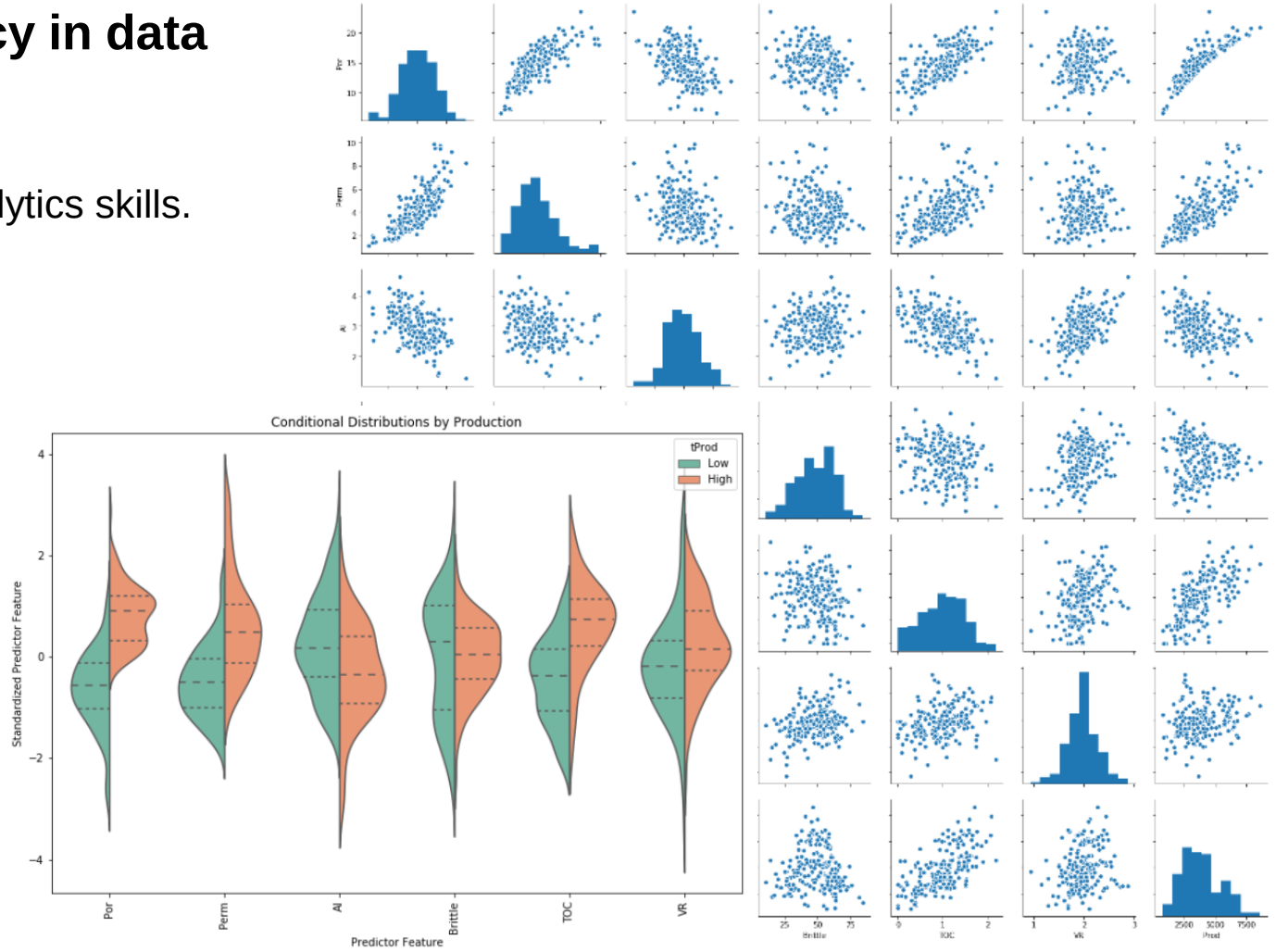
- Data preparation, debiasing, transforms
- Detect patterns
- Extract information

Before we said:

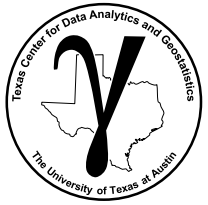
Those with the best data win!

Now we say:

Those with the best data and do the best with their data win!



Subsurface data analytics examples.



My Goal

I provide a basic data analytics and geostatistics foundation that will impact your work during your career.

- A way of thinking / new lens.
- Collection of concepts and methods.
- Exposure to statistical workflows and software - Empower you!

We will go for broad coverage, but not depth.

- Resources available
- Basic concepts
- Enable you to build out your own workflows with existing tools
- Set you up to think “geostatistically” and like a “data scientist” when you need to!

Learning Not Schedule Driven

- Lectures will span multiple sessions. We will take multiple runs at difficult topics to ensure success.

PGE 338 Lecture 0:
Introduction to Data Analytics and (Geo)statistics

Lecture outline . . .

- **Grade Distribution and Course Expectations**

Introduction

Probability

Univariate

Bivariate

Time Series Analysis

Spatial Analysis

Machine Learning

Uncertainty Analysis

Michael Pyrcz, The University of Texas at Austin

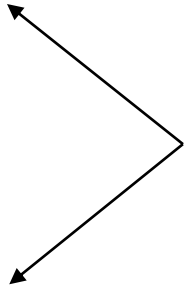
Class Expectations

Grading Distribution:

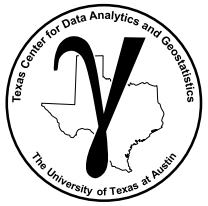
Item	Weight (%)
Homework and Project Assignments	18
Quizzes	15
Midterm Exam No. 1	20
Midterm Exam No. 2	20
Final Exam	25
Participation	2
Total	100

Course Percentage to Grades:

Grade	Threshold
A	90%
A-	85%
B+	80%
B	75%
B-	70%
C+	60%
C	50%
F	Otherwise



Check this out!



Class Expectations

No Risk, Optional Final:

If a student completes both midterms, then the final is an optional, no risk final.

Students will receive their in-class grade after the 2nd midterm and may choose to write the final to improve their grade.

If the grade on the final (if written) does not exceed the in-class mark, then it is discarded. If either midterm is not written then the final is required (the optional, no risk option above is not available).

If a midterm is missed or excused, then the final is required and will be weighted as indicated in the syllabus. This ensures that all course concepts are tested.

Class Expectations

Working with Software / Code:

There is no programming prerequisites, but in class exercises and assignments will require use of the following:

1. **Excel Spreadsheet for e.g., Probability, Distributions, Hypothesis Testing**
   - A lot can be done in Excel built-in functions

2. **Python**
   - Various packages to support advanced statistical methods, machine learning etc.
   - Building and modifying existing workflows, producing and interpreting results
   - Install Anaconda Python
   - Install GeostatsPy

For coding we focus on workflow development.

- I provide a lot of examples and workflows to assist.
- You won't need to start from scratch.

Class Expectations

General Structure for Lectures:

1. Class notes in .pdf, code examples, and datasets will be posted by the day before.
2. A stats moments at start of class (student ‘volunteers’, everyone will do one).

We will have an online sign up

3. Lecture material with attempt for interactivity.
4. Walk-through examples.
5. Hands-on practice (bring your laptops with required software installed).

Additional Advice for Success

Exams and Quizzes

1. Definitions of key terms that describe fundamental concepts
   - You do not have to use my words exactly, demonstrate knowledge of the concepts
2. Short answer, explain steps of a standard workflow
   - Concepts and logical steps, understand how to solve problems
3. Simplified problems, demonstrate how to solve them
   - No need for a calculator

Assignments

1. For numerical methods cut and paste output, plots to word documents and convert to PDFs to upload.
2. Be concise. Include very short explanations, answers to questions, labels on figures.
3. Include requested executive summaries (1 paragraph)
4. If the work seems onerous talk to me!

Assignment Submissions

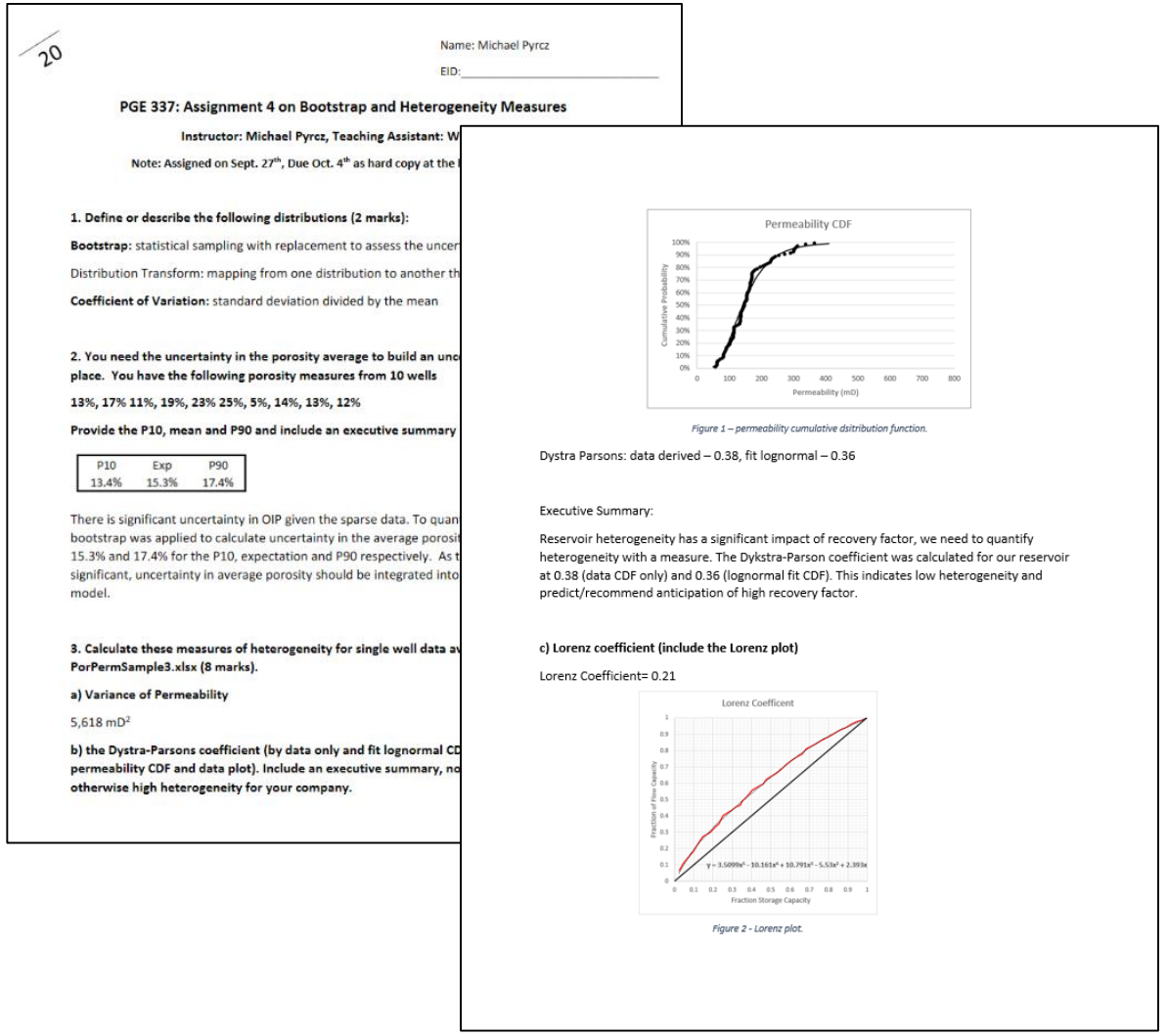
Assignments:

- Professional, written communication
  - Brief, concise and clear answers to questions
  - Demonstrate knowledge from lecture content.
  - Use efficient, creative figures to communicate
  - 1-2 pages as a PDF
  - No code, nor Jupyter notebooks

Executive Summary:

- A concise high-level explanation for your manager
- 1 paragraph, 4 sentences

1. What is the issue / challenge?
2. What was done to address this issue / challenge?
3. What was the result of the completed work?
4. What is the recommendation going forward?



20

Name: Michael Pyrcz

EID:

PGE 337: Assignment 4 on Bootstrap and Heterogeneity Measures

Instructor: Michael Pyrcz, Teaching Assistant: W

Note: Assigned on Sept. 27th, Due Oct. 4th as hard copy at the

1. Define or describe the following distributions (2 marks):

Bootstrap: statistical sampling with replacement to assess the uncer

Distribution Transform: mapping from one distribution to another th

Coefficient of Variation: standard deviation divided by the mean

2. You need the uncertainty in the porosity average to build an unc place. You have the following porosity measures from 10 wells

13%, 17% 11%, 19%, 23% 25%, 5%, 14%, 13%, 12%

Provide the P10, mean and P90 and include an executive summary

P10	Exp	P90
13.4%	15.3%	17.4%

There is significant uncertainty in OIP given the sparse data. To quan bootstrap was applied to calculate uncertainty in the average porosit 15.3% and 17.4% for the P10, expectation and P90 respectively. As t significant, uncertainty in average porosity should be integrated into model.

3. Calculate these measures of heterogeneity for single well data av PorPermSample3.xlsx (8 marks).

a) Variance of Permeability

5,618 mD^2

b) the Dystra-Parsons coefficient (by data only and fit lognormal CD permeability CDF and data plot). Include an executive summary, no otherwise high heterogeneity for your company.

Figure 1 – permeability cumulative dsitribution function.

Dystra Parsons: data derived – 0.38, fit lognormal – 0.36

Executive Summary:

Reservoir heterogeneity has a significant impact of recovery factor, we need to quantify heterogeneity with a measure. The Dykstra-Parson coefficient was calculated for our reservoir at 0.38 (data CDF only) and 0.36 (lognormal fit CDF). This indicates low heterogeneity and predict/recommend anticipation of high recovery factor.

c) Lorenz coefficient (include the Lorenz plot)

Lorenz Coefficient= 0.21

Figure 2 - Lorenz plot.

Assignment template, Canvas/files/ PGE338_assignment_template.pdf

Other Topics

Academic Integrity Expectations

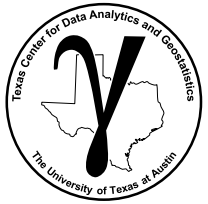
See syllabus and Student Conduct & Academic Integrity website. Here's some additional points to assist:

1. acknowledge the contributions of other sources to your scholastic efforts, e.g., cite code blocks and text from others that you use in your work

2. teamwork is encouraged but complete your assignments independently. If you have a laptop issue and must temporarily work with other student(s) and will have same results, all students must indicate this on their submitted assignments

3. examinations are all closed book, and collaboration or use of any content on quizzes, midterms and the final are not permitted

4. course content, including lecture notes, demonstrations in Jupyter notebooks and Excel spreadsheets are copyrighted, copying or posting online is not permitted. I share it online, so share the link not the content!



Other Topics

Assignments:

- Contact me early if going to be excusably absent and unable to complete an assignment or contact me as soon as possible in case of an emergency.

- Late assignments / miss quizzes will be accepted with reasonable excuse and timely notification when possible (illness, family emergency, interviews, conferences etc.).

Other Topics

Accessible and Inclusive:

- The university is committed to creating an accessible and inclusive learning environment consistent with university policy and federal and state law. Please let me know if you experience any barriers to learning so I can work with you to ensure you have equal opportunity to participate fully in this course. If you are a student with a disability, or think you may have a disability, and need accommodations please contact Disability and Access (D&A). Please refer to D&A's website for contact and more information: http://diversity.utexas.edu/disability/
- If you are already registered with D&A , please deliver your Accommodation Letter to me as early as possible in the semester so we can discuss your approved accommodations and needs in this course.

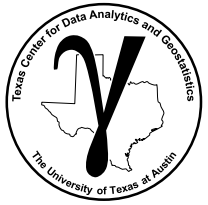
Getting Help:

- I welcome course feedback at anytime. Let me know if something could be improved to help you learn.
- e-mail is the best way to contact me: mpyrcz@austin.utexas.edu. Please include 'PGE 383' in the subject or send the e-mail through the Canvas site.

Other Topics

Classroom Environment:

- Positive Professional Conduct
  - No cell phones out
  - Laptops only for course related activities
  - No disruptive behaviors – negative, disrespectful comments, ongoing discussions



Other Topics

- Please read the syllabus carefully and let me know if you have any questions or concerns.
- Would you be able to attend the office hours of the instructor or TA? If not, please let me know.
- Please let me know, if there is anything, I can do to improve the learning environment.
- If you are struggling, see me early.

PGE 338 Lecture 0:

Introduction to Data Analytics and (Geo)statistics

Lecture outline . . .

- **My Online Resources**

Introduction

Probability

Univariate

Bivariate

Time Series Analysis

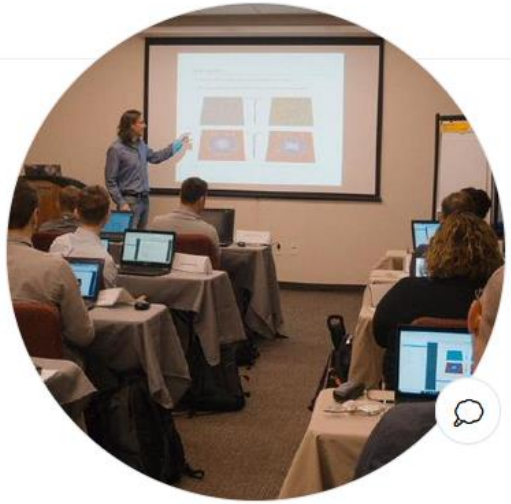
Spatial Analysis

Machine Learning

Uncertainty Analysis

Michael Pyrcz, The University of Texas at Austin

GitHub/GeostatsGuy



Michael Pyrcz

GeostatsGuy

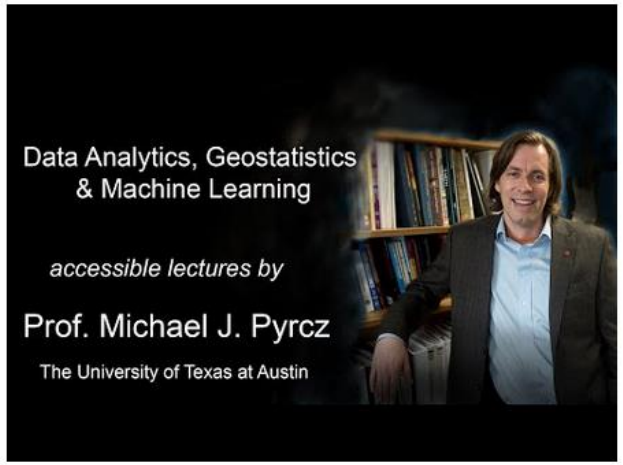
Associate Professor at the University of Texas at Austin working on Spatial Data Analytics, Geostatistics and Machine Learning

Edit profile

1.4k followers · **8** following · ☆ **0**

Overview | Repositories 38 | Projects | Packages

GeostatsGuy / README.md



I'm Michael Pyrcz (a.k.a. GeostatsGuy), an associate professor working in Data Analytics, Geostatistics and Machine Learning at The University of Texas at Austin, Austin, Texas, USA. I share all of my university content to support my students, potential students and working professionals interested to learn about data science. I have a lot of well-documented workflows in Python, R (and even Excel) in my repositories, including all of the hands-on exercises and demonstrations for all of my lectures shared freely on my YouTube channel. Follow me on Twitter, where I share resources and positivity daily.

Popular repositories

Customize your pins

PythonNumericalDemos

Well-documented Python demonstrations for spatial data analytics, geostatistical and machine learning to support my courses.

GeostatsPy

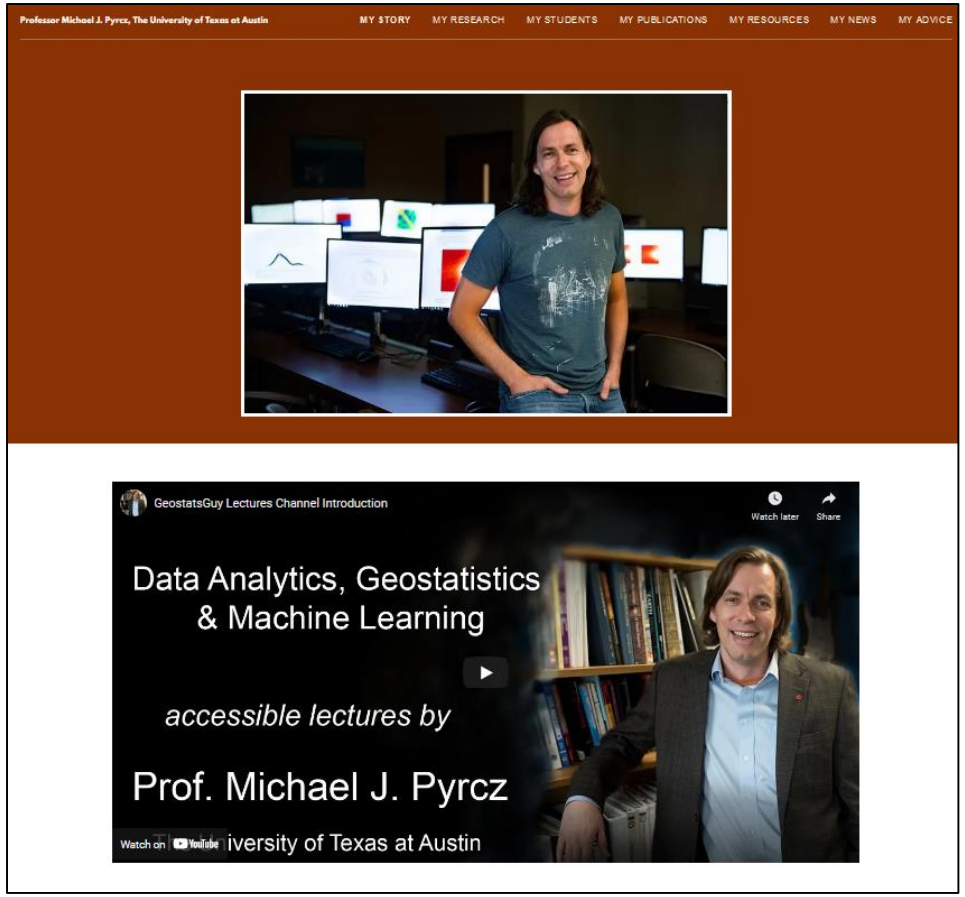
GeostatsPy Python package for spatial data analytics and geostatistics. Mostly a reimplementation of GSLIB, Geostatistical Library (Deutsch and Journel, 1992) in Python. Geostatistics in a Python p...

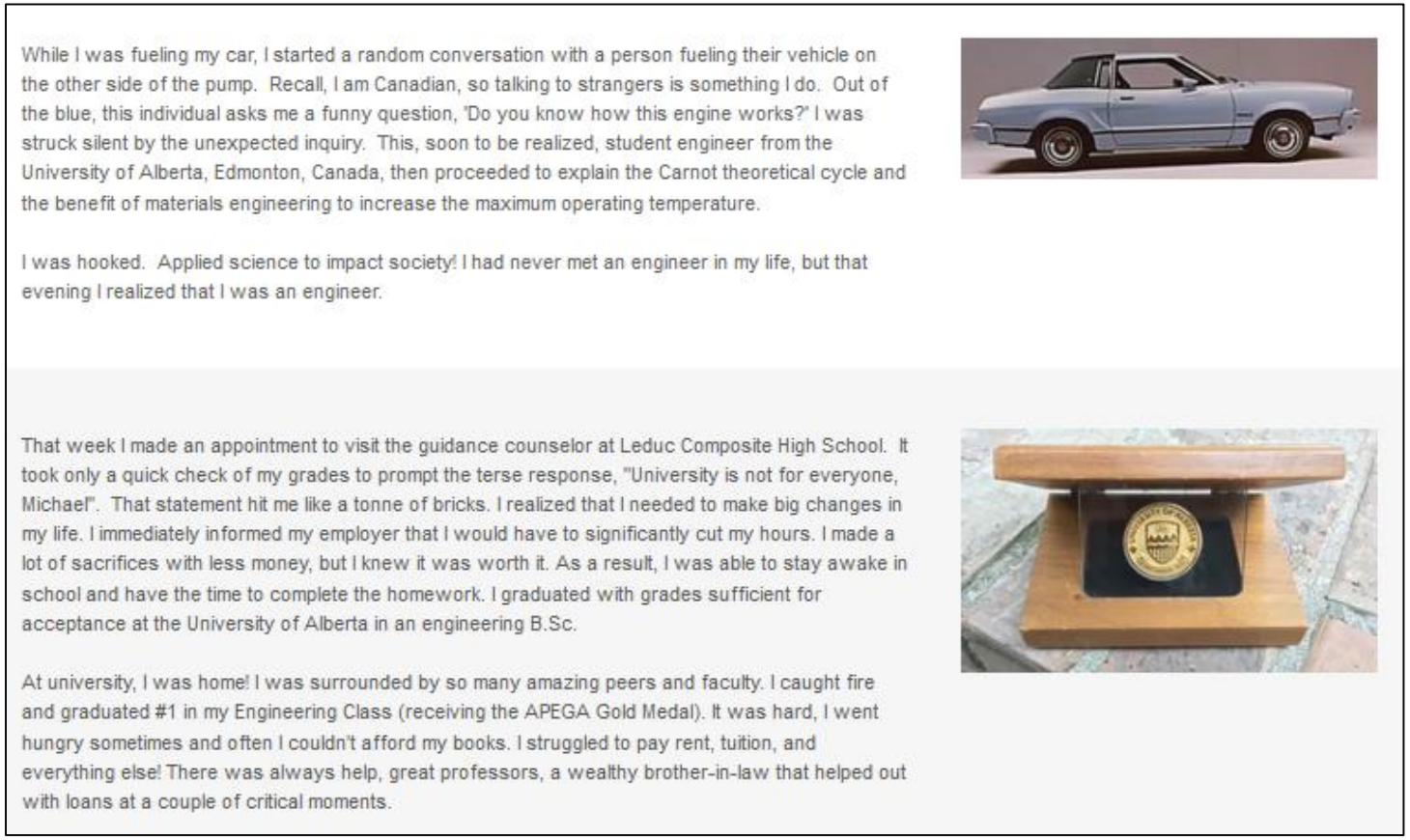
Numerical demonstrations, code examples, synthetic datasets used in class.

@GeostatsGuy on Twitter

Sharing resources on geostatistics, data science, machine learning etc.
Networking with subsurface engineers, geologists, data scientists, software engineers etc.

michaelpyrcz.com



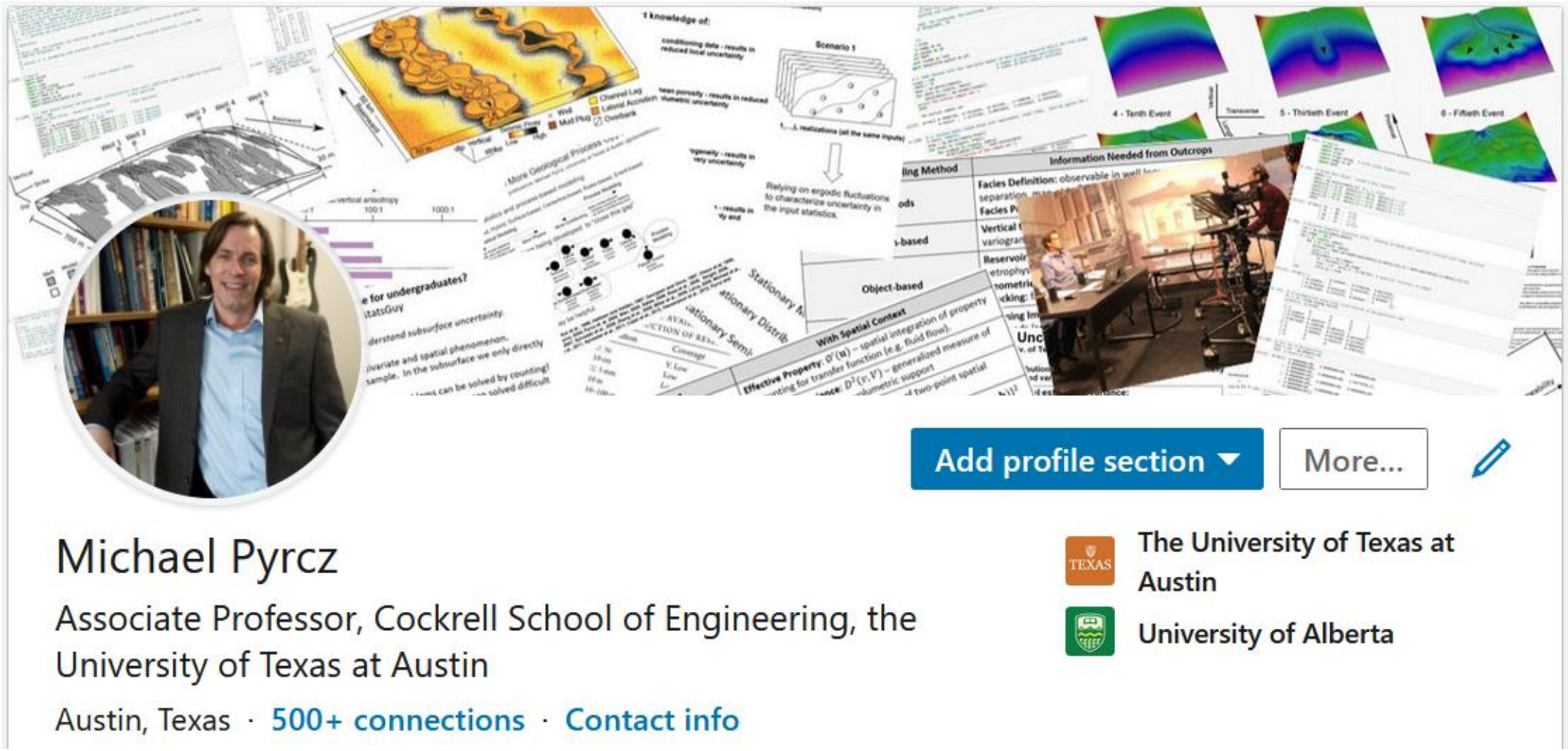


While I was fueling my car, I started a random conversation with a person fueling their vehicle on the other side of the pump. Recall, I am Canadian, so talking to strangers is something I do. Out of the blue, this individual asks me a funny question, 'Do you know how this engine works?' I was struck silent by the unexpected inquiry. This, soon to be realized, student engineer from the University of Alberta, Edmonton, Canada, then proceeded to explain the Carnot theoretical cycle and the benefit of materials engineering to increase the maximum operating temperature.

I was hooked. Applied science to impact society! I had never met an engineer in my life, but that evening I realized that I was an engineer.

My website with information on my research and shared resources, and my advice.

LinkedIn



Consider the opportunity to start building your network online. LinkedIn is often checked to learn about you by professional peers and companies.

PGE 338 Lecture 0:
Introduction to Data Analytics and (Geo)statistics

Lecture outline . . .

- **Who am I?**
- **Course Objectives**
- **Setting Up**
- **Some Concepts**
- **Grade Distribution and Course Expectations**
- **My Online Resources**

Introduction

Probability

Univariate

Bivariate

Time Series Analysis

Spatial Analysis

Machine Learning

Uncertainty Analysis

Michael Pyrcz, The University of Texas at Austin