Introduction to Machine Learning in Geosciences
Introduction
GEO371T/GEO391

Mrinal K. Sen
Geosciences
UT Austin

August 26, 2021

Content

- Course Logistics
- Motivation
- Quest in Geosciences and other fields
  - Basic Questions in Sciences
  - Basic Questions in Geosciences
  - Prediction (of Unknowns)/remote sensing
  - Interpolation
  - Methods for answering these questions: Physics Based, Statistical
- What is new now? Automation; Big Data (analytics)
- How do we learn to learn?
- What is ML? Types of ML - ML background linear algebra statistics computing
- Learning under uncertainty!

Course Logistics

- **Course Name**: Introduction to Machine Learning in Geosciences
- **Course Number**: GEO371T/GEO391
- **Unique Numbers**: 28182 28324
- **Timing**: Tuesday/Thursday 9:30-11:00 am
- **Course website**: Notes/slides will be posted in Canvas
- **Course website**: Mrinal K. Sen mrinal@utexas.edu
- PLEASE mention GEO371T in the email subject!
- **Office hours**: Wed 3-4:30 PM via zoom and by appointment at any other time

Teaching Assistant

Dimitri Voytan

dvoytan@utexas.edu

Practice Sessions, Assignments, Project & Grading

- We will have one lecture and one class exercise every week.
- Homework will be assigned once in 2 weeks.
- One mid-term exam in October
- Each student is required to work on a project that uses some aspect of machine learning.
  - Please discuss project ideas with the instructor and TA.
  - Short presentation of project ideas in the 3rd week of September
  - Need to submit a final project report and workflow, and a presentation.
  - Undergrads may choose not to do a project in which case an assignment will be provided.
- Grading: 30% homework + 30% mid-term + 40% project

Resources

- https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf
- https://github.com/amueller/introduction_to_ml_with_python
- https://jakevdp.github.io/PythonDataScienceHandbook/
- https://www.cs.ubc.ca/~murphyk/MLbook/
- http://ciml.info/ A course in Machine Learning by Hal Daume III
- http://www.charuaggarwal.net/neural.htm Aggarwal, C. C., 2018, Neural Network and Deep Learning, Springer
- Ma, Y, Z., 2019, Quantitative Geoscienes: Data Analytics, Geostatistics, Reservoir Characterization, and Modeling, Springer

Learning Outcomes

- Understand how various machine leaning algorithms work.
- Formulate your own ML application.
- Implement on your own/come up with noel ideas for improving.
- Identify algorithms appropriate for a particular application.
- Get motivated to learn more!
- **Yes, ML uses calculus, linear algebra, optimization, statistics ... but you can learn all!**

Lecture Plan

1. L1 – Introduction:Course Introduction Inference in Geosciences (In-verse Method, Geostatistics, NN)
2. Lab 1: Introduction to python
3. L2 – Introduction: Big Data Analytics and ML Overview
4. Lab 2: Python libraries
5. L3 -Optimization methods and Sampling: I (Gradient descent,Newton, SGD, minibatch SGD)
6. Lab 3: Optimization practice
7. L4 – Background Statistics (Descriptive and Inferential Statstics) data/housingdata)
8. Lab 5: Regeression practice
9. L6 – Perceptrons and Neurons : A simple NN model 6: building a simple NN model for classification (well log – faciesclassification)

Lecture Plan

1. L7 -Nearest Neighbor, KNN
2. Lab 7:NN, KNN (geoscience data)
3. L8 – Decision Tree
4. Lab 8: Decision tree
5. L9 -Random forest
6. Lab 9 Random forest (geoechemical data)
7. L10– Feature Engineering
8. Lab 10 Feature Engineering
9. L11 – OVER-FITTING. UNDERFITTING, VARIANCE, BIAS
10. L12 – Dimensionality Reduction
11. Lab 11 Dimensionality Reduction
12. L13 – Naive Bayes
13. Lab 12 – NB
14. L14 – Deep Learning: Convolutional Neural Networks
15. Lab 14 – CNN
16. L15 – Auto-encoders

Lecture Plan

1. Lab 15 – Auto-encoders
2. L16 – Recurrent Neural Networks
3. **Summary**
4. **Class Presentations**

Scientific Method in Research

- Used in all fields of science, from biology to physics and chemistry, the scientific method allows researches and scientists to test their theories in a standard manner that is uniform to everyone globally.
- One of the main advantages of the scientific method is repeatability. Since the experiment and all the details are expected to be recorded clearly, it allows others the ability to replicate the experiment and accept the hypothesis widely.

https://study.com/academy/answer/what-are-the-7-steps-of-the-scientific-method.html

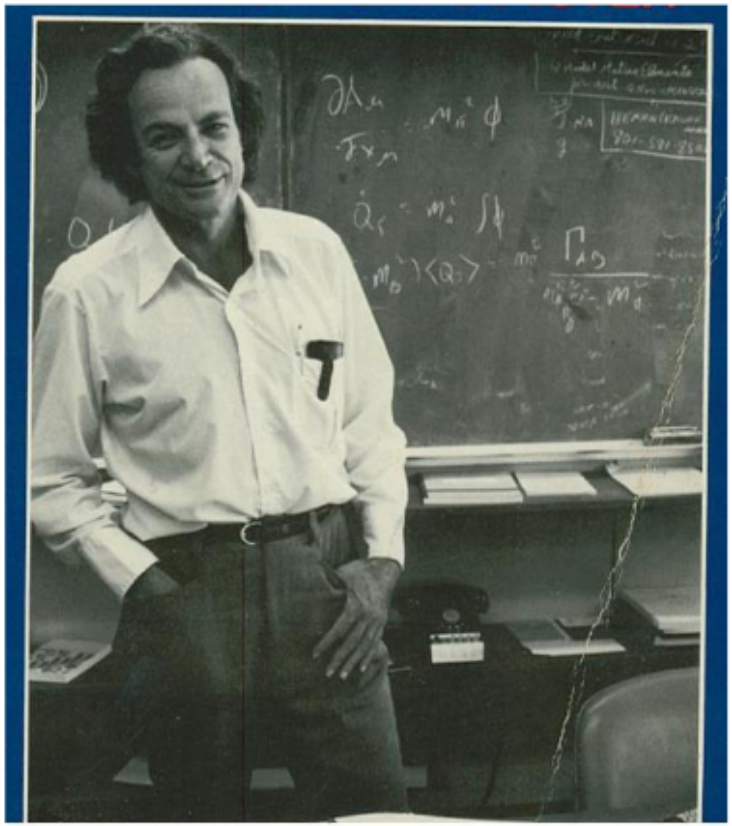
Scientific Thinking

Steps

- Question: identifying the problem. What information is unknown or missing.
- Research: information about the problem. What is already known and what isn't.
- Hypothesis: prediction or an educated guess of a possible outcome.
- Experiment: testing the hypothesis.
- Observation: data collected while performing the experiment. All crucial details that can have even a minimal impact on the outcome or the question of the original problem.
- Result/Conclusion: determining if the hypothesis is correct and what impacted the outcome
- Communicate: presenting the result and date through various media, usually a lab report.

https://study.com/academy/answer/what-are-the-7-steps-of-the-scientific-method.html

Scientific Thinking

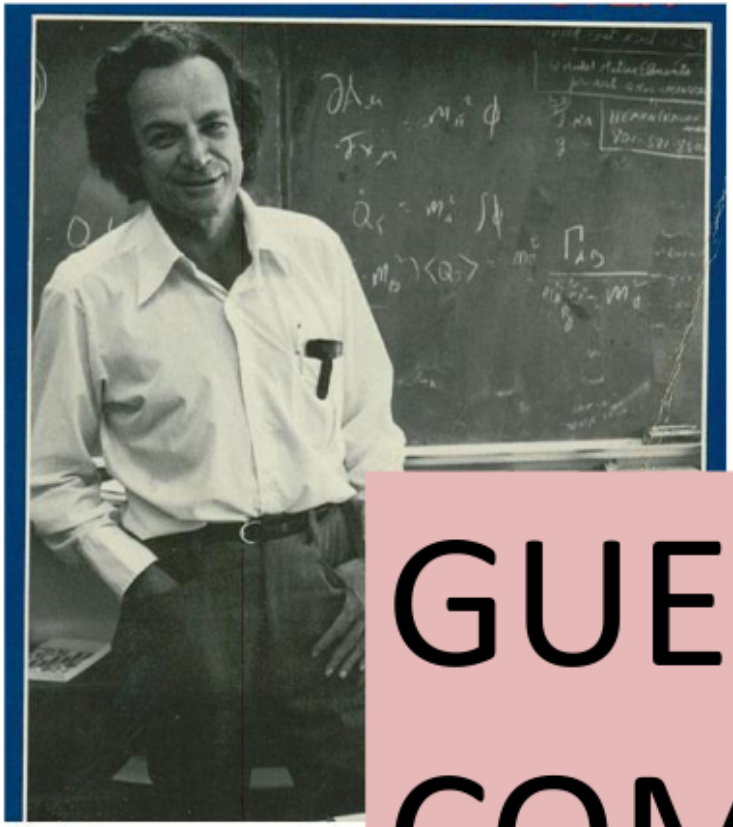


In general, we look for a new law by the following process: First we guess it. Then we compute the consequences of the guess to see what would be implied if this law that we guessed is right. Then we compare the results of the computation to nature, with experiment or experience, compare it directly with observations to see if it works. If it disagrees with experiment it is wrong.

In that simple statement is the key to science.

- Richard P. Feynman, *The Character of Physical Law*

Scientific Thinking



GUESS,
COMPUTE &
COMPARE

Four Paradigms

- **Experimentation**: Beginning in ancient Greece and China, people tried to explain their observations through natural laws instead of supernatural causes. EMPIRICAL
- **Theory**: By the 17th century, scientists like Isaac Newton tried to make predictions for new phenomena and would verify hypotheses by conducting experiments.
- **Computation and Simulation**: The advent of high-performance computers in the latter half of the 20th century allowed scientists to explore regimes inaccessible to experiment and theory, such as climate modeling or galaxy formation, by numerically solving systems of equations on a large scale and in fine detail.
- **Data Mining**: Using more-powerful computers, scientists begin with the data and direct programs to mine enormous databases for relationships. In essence, they use computers to discover the rules by studying the data. [DATA INTENSE DISCOVERY]

https://hbr.org/2010/11/the-big-idea-the-next-scientific-revolution

The Four Paradigms of Science

- **Experi**
  tried t
  superr
- **Theory**
  make
  by con
- **Compu**
  compu
  to exp
  climat
  systen
- **Data M**
  the da
  relatio
  by stu

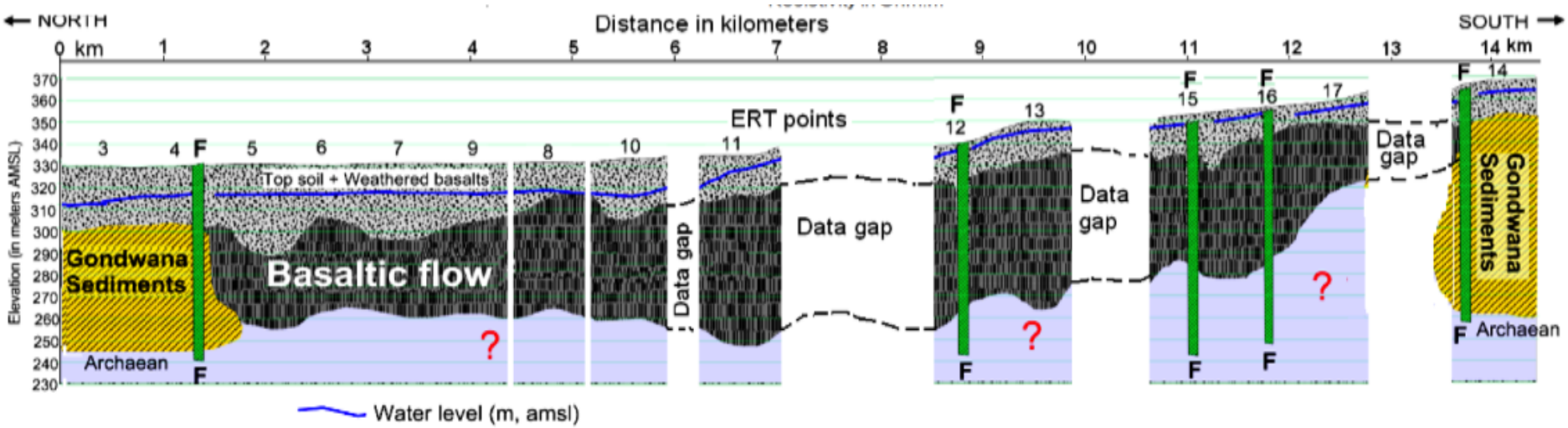
Use of Technology to do high impact Science

Geoscience Questions

Earth Science problems or questions are easy to understand!

- What causes mountain building?
- Why are the rock types different in different parts of the earth?
- How old is the earth?
- What causes earthquakes and volcanic eruption?
- What causes earth's magnetic field?
- Oil, natural gas and minerals: how are they formed and how are they distributed?
- ..
- ..

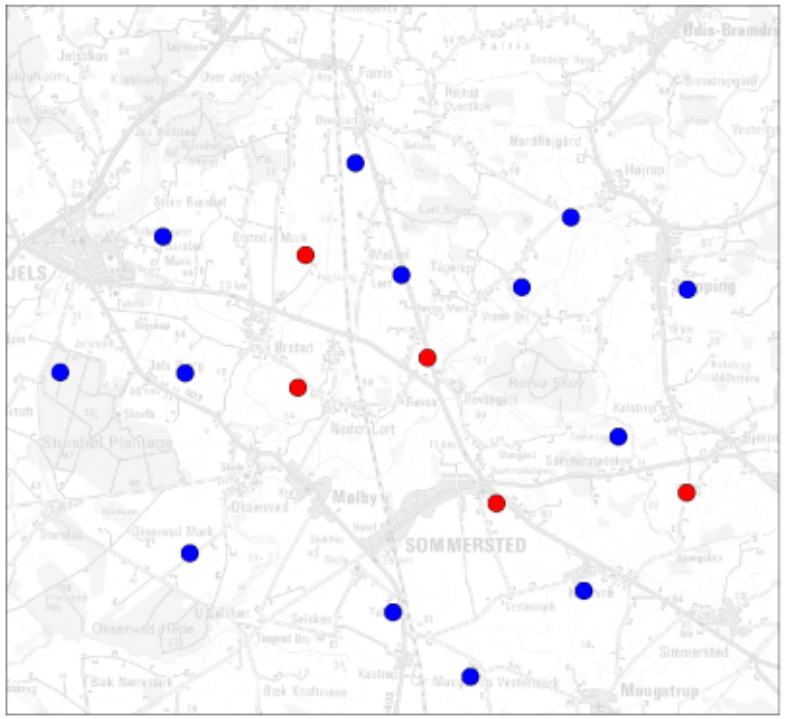
Direct Observation by Drilling



Spatial data density

High data density is crucial for reliable mapping results !

Suppose you have indications of buried valleys visible in some of your boreholes (red color)

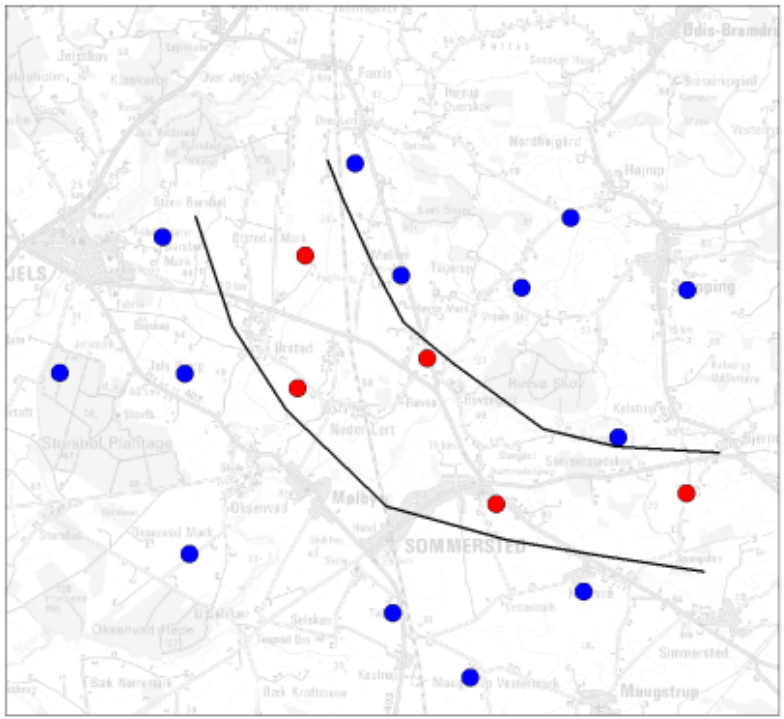


Spatial data density

High data density is crucial for reliable mapping results !

Suppose you have indications of buried valleys visible in some of your boreholes (red color)

– possible conclusion 1



Spatial data density

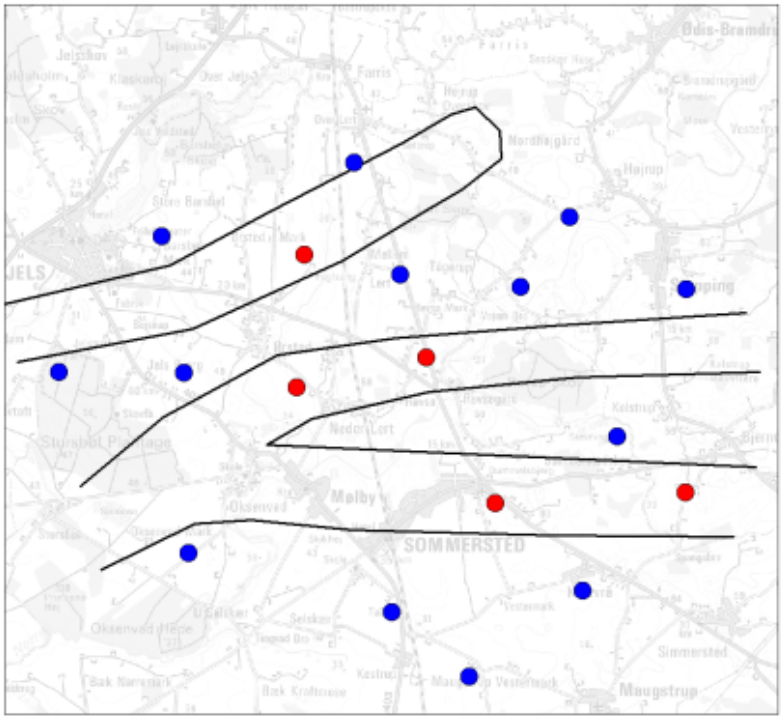
High data density is crucial for reliable mapping results !

Suppose you have indications of buried valleys visible in some of your boreholes (red color)

– possible conclusion 1

– possible conclusion 2

Which one is the most likely?



Spatial data density

High data density is crucial for reliable mapping results !

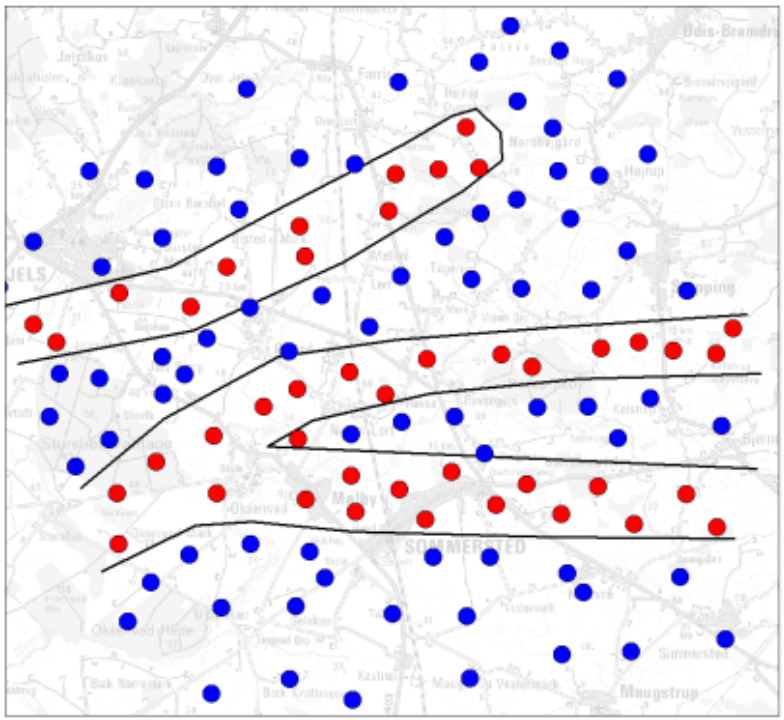
Suppose you have indications of buried valleys visible in some of your boreholes (red color)

– possible conclusion 1

– possible conclusion 2

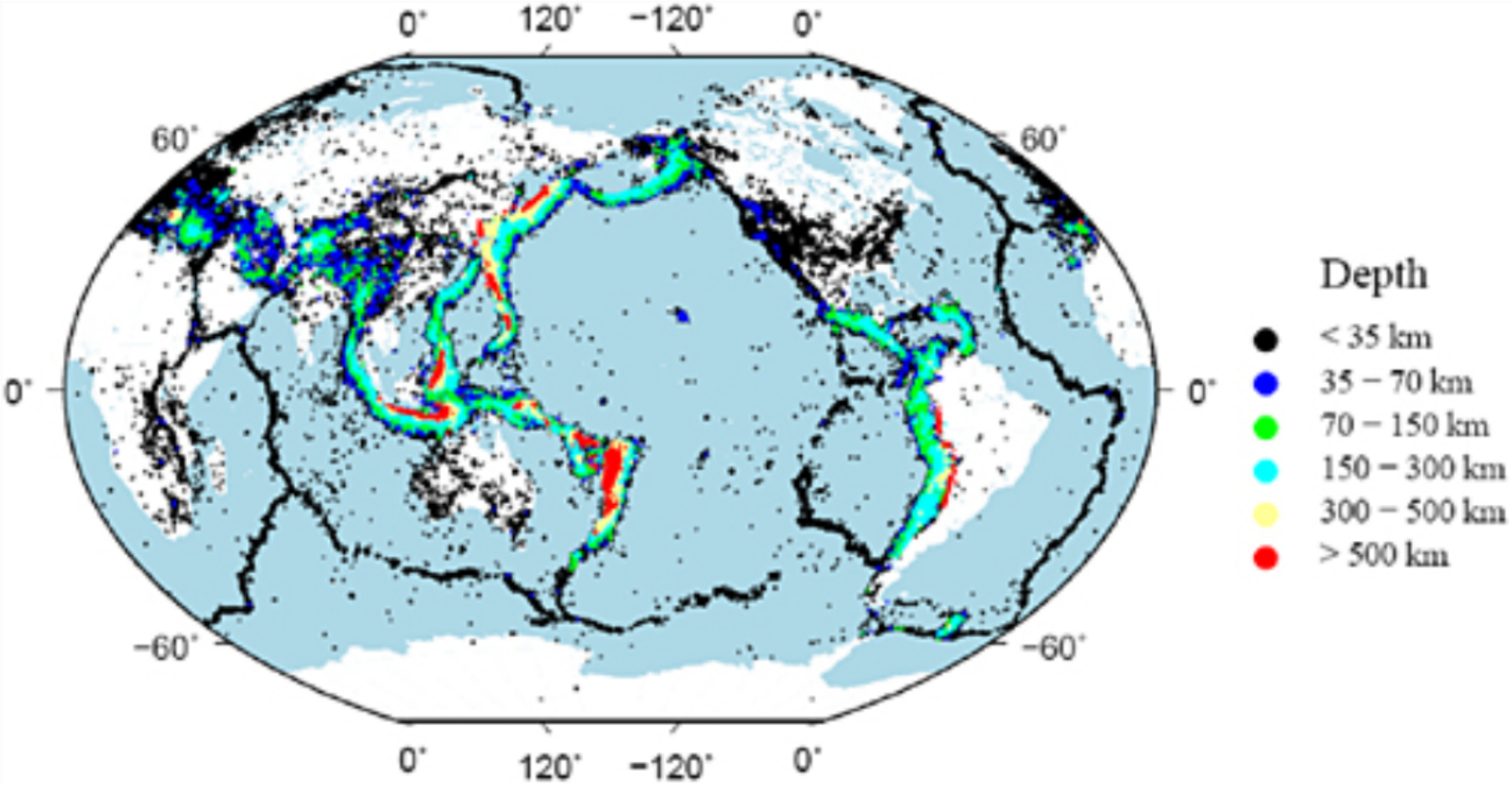
Which one is the most likely?

A good data density makes it easier to decide!



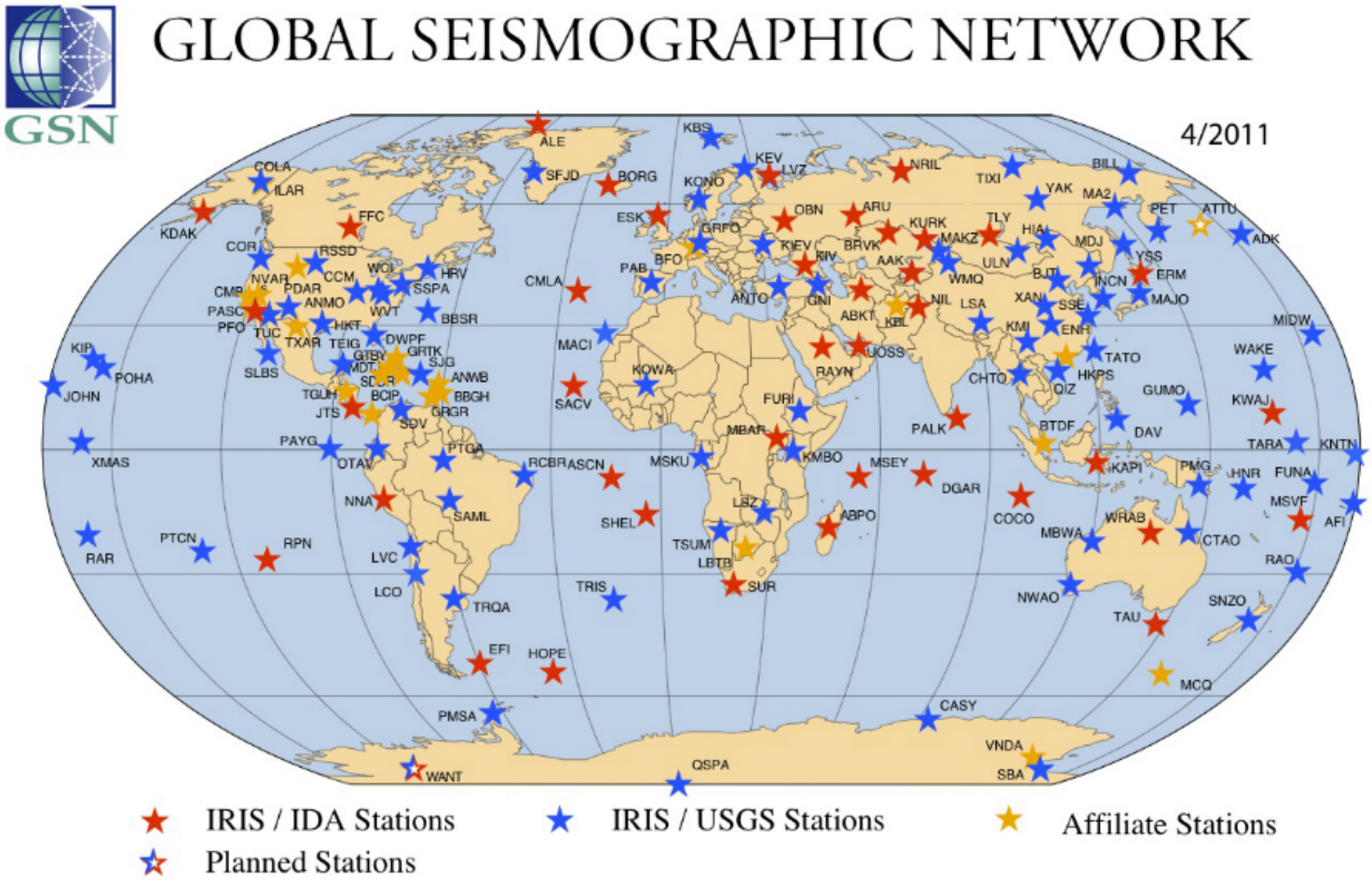
Earthquakes

Our Observations

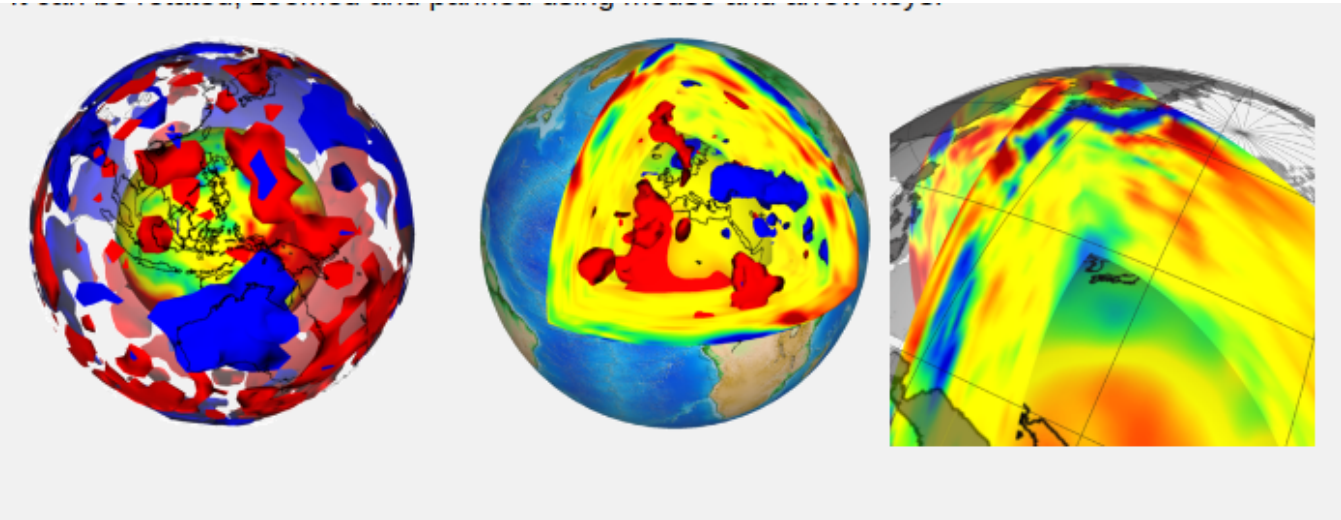


World seismicity map for years 1973-2007, color scale indicates depth range. Data provided by USGS. Figure plotted using GMT.

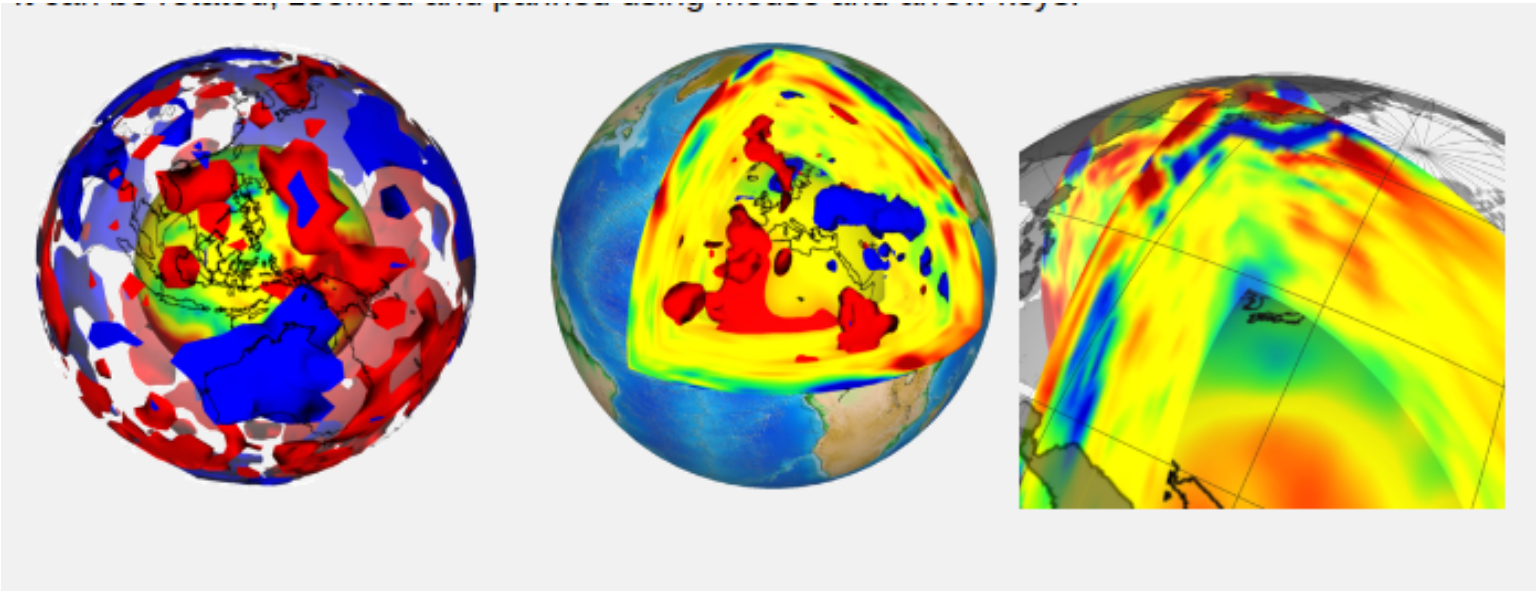
Our Data


GLOBAL SEISMOGRAPHIC NETWORK
GSN
4/2011
IRIS / IDA Stations
IRIS / USGS Stations
Affiliate Stations
Planned Stations

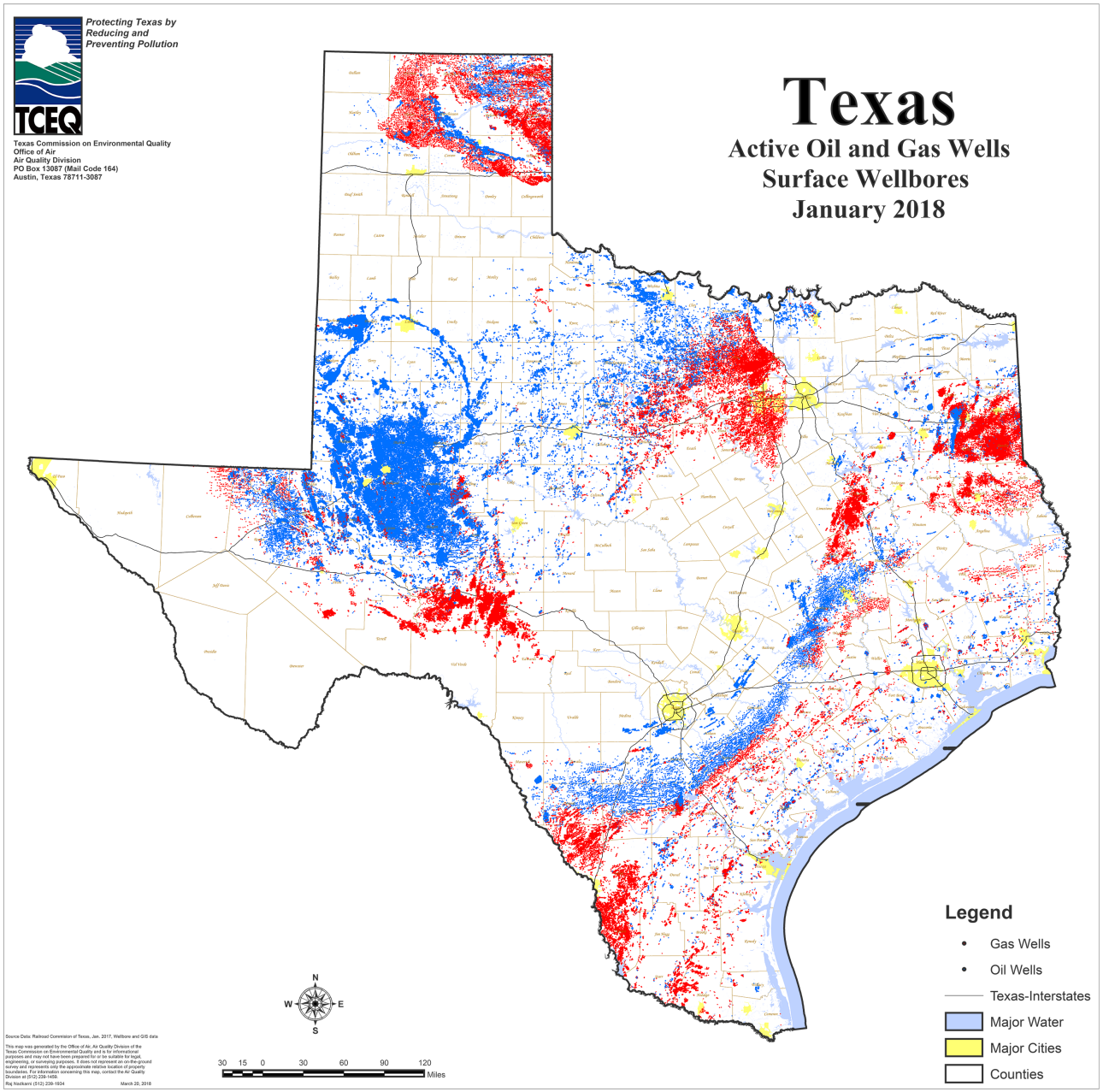
Earthquakes



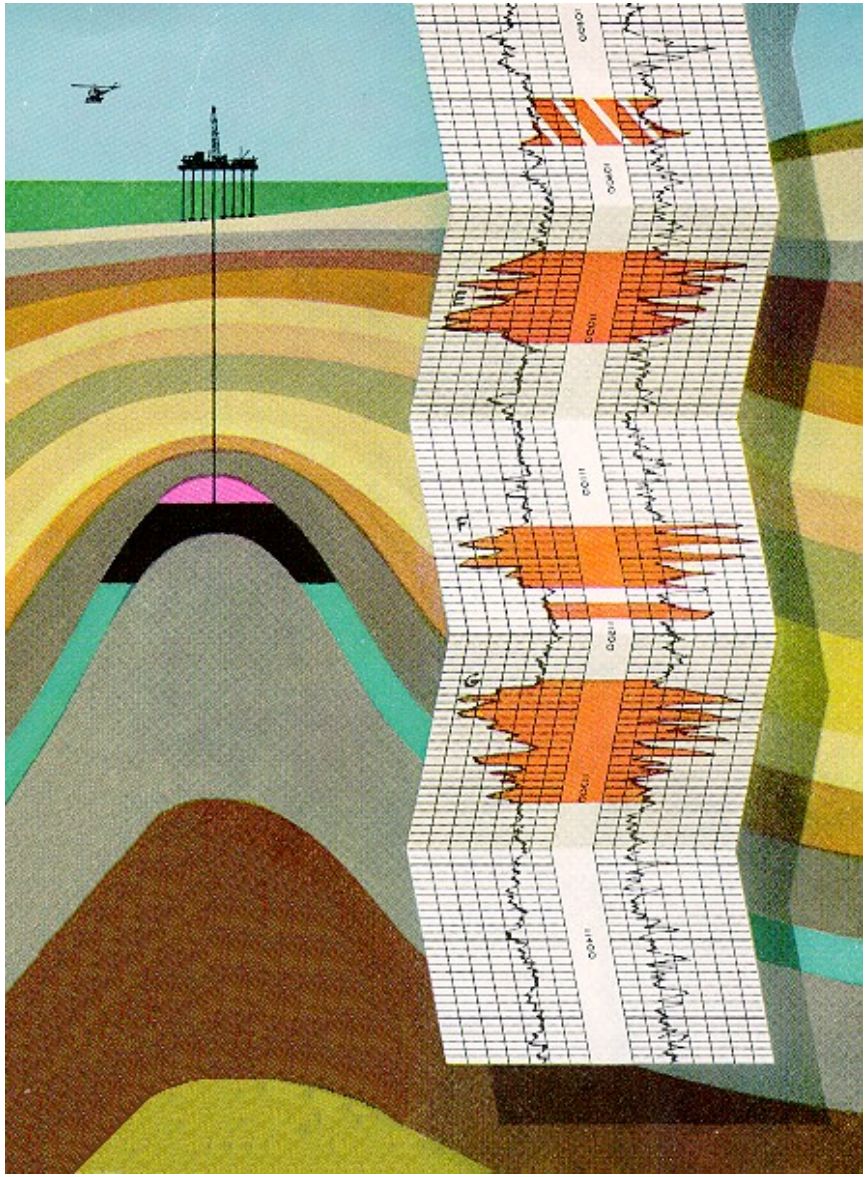
Earthquakes



Oil/Gas Production

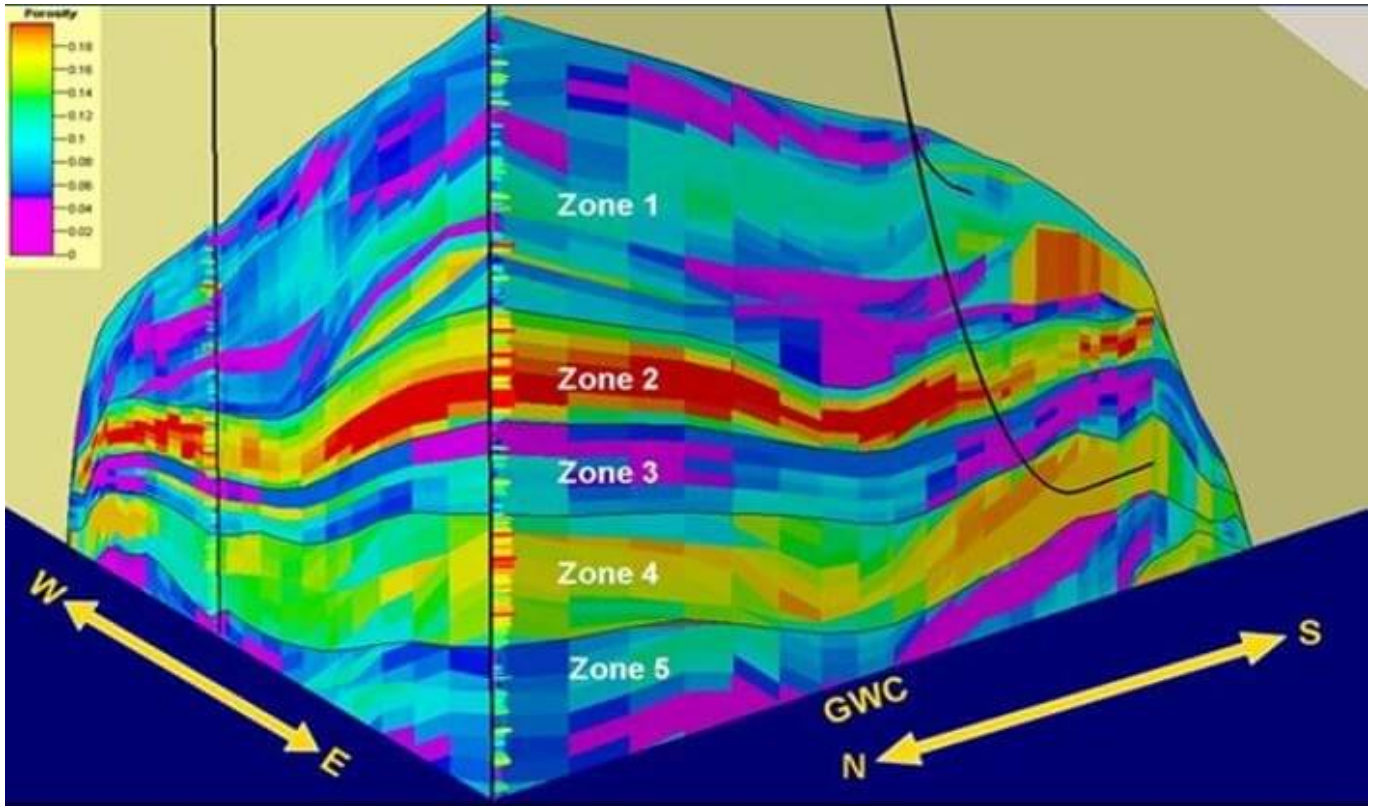


Well-log



http://www.sjvgeology.org/oil/logs.html

Well-log



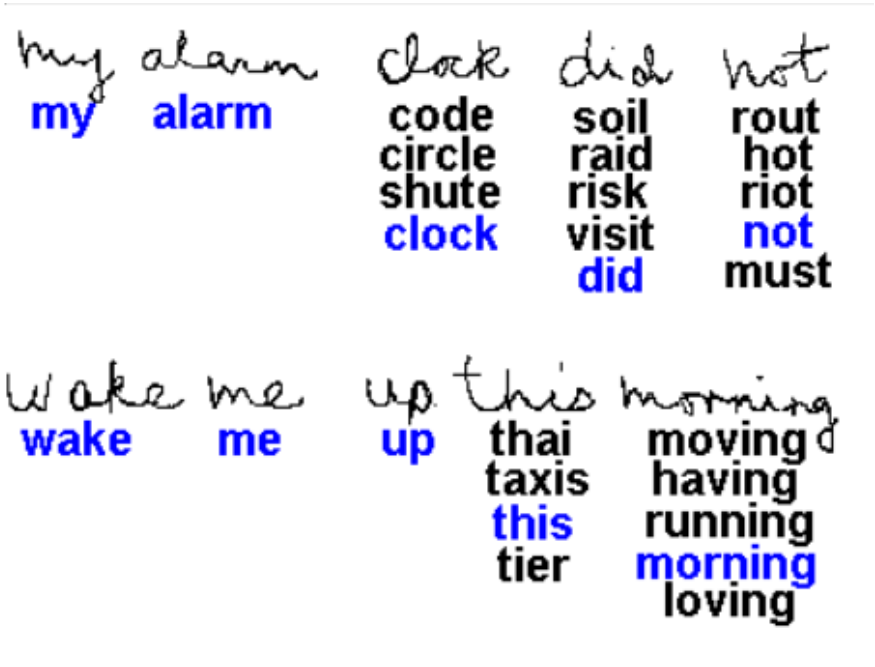
https://www.zpetro.com/Basic-well-logging.php

Weather forecast

Weather forecast

- The prediction of the weather through application of the principles of physics, supplemented by a variety of statistical and empirical techniques.
- Weather forecasts are made by collecting as much data as possible about the current state of the atmosphere (particularly the temperature, humidity and wind) and using understanding of atmospheric processes (through meteorology) to determine how the atmosphere evolves in the future.
- The common instruments of measure are anemometer, wind vane, pressure sensor, thermometer, hygrometer, and rain gauge. The weather measures are formatted in special format and transmit to WMO to help the weather forecast model.

Handwriting



Using language models incorporating collocational (*alarm clock*) and syntactic (**POS**) analysis, we are able to extract the correct sentence:

My alarm clock did not wake me up this morning

https://cedar.buffalo.edu/handwriting/HRoverview.html

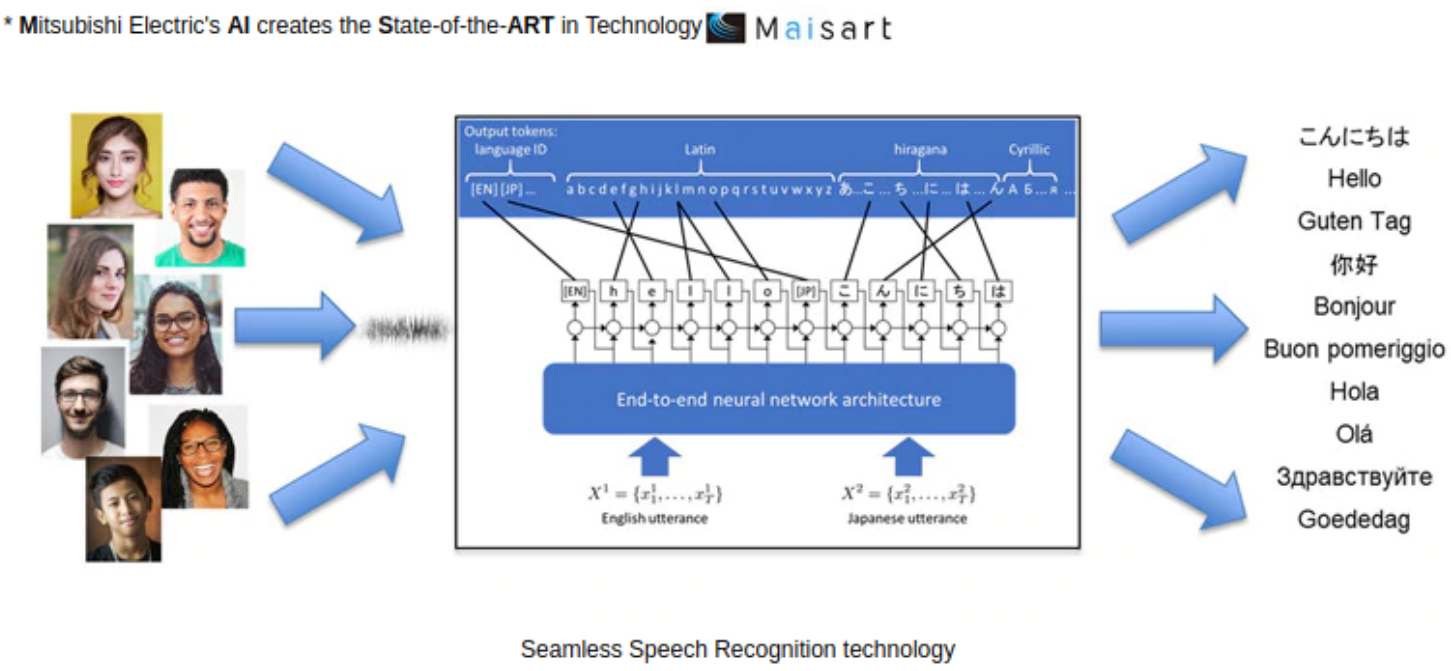
Facial Recognition

Last updated 27 June 2020 - Estimated reading time: 14 minutes

Facial recognition – fascinating and intriguing

https://www.thalesgroup.com/en/markets/digital-identity-and-security/government/biometrics/facial-recognition

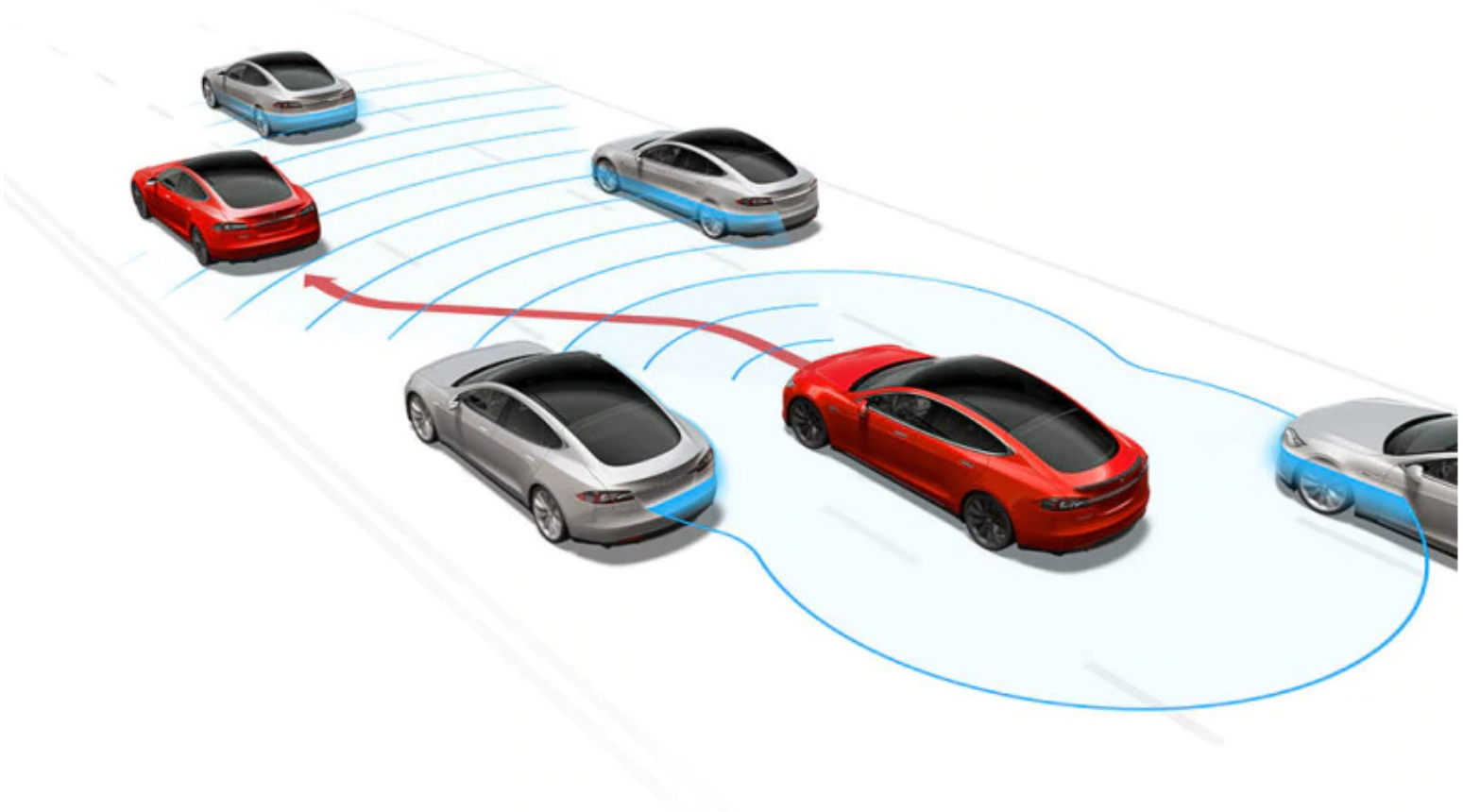
Speech Recognition



Seamless Speech Recognition technology

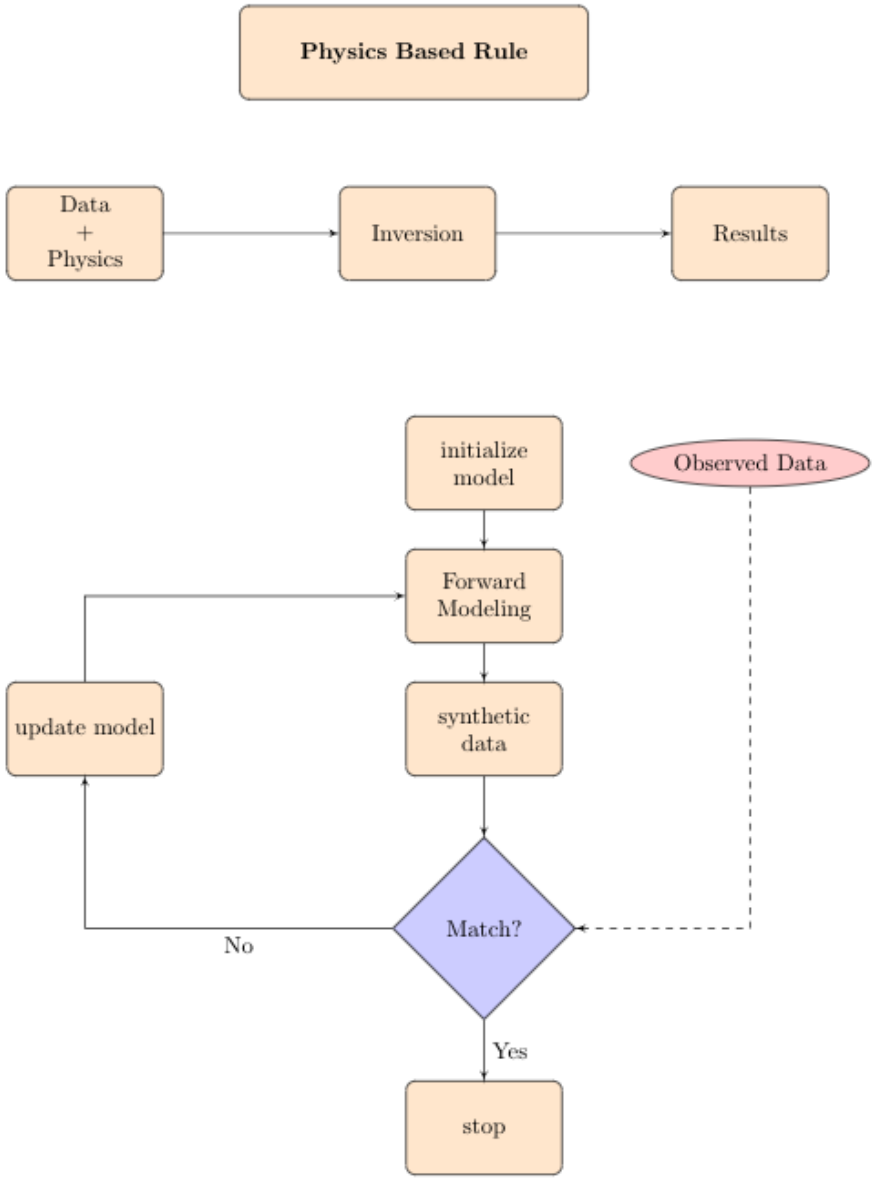
https://us.mitsubishielectric.com/en/news-events/releases/global/2019/0213-g/index.html

Self Driving Cars



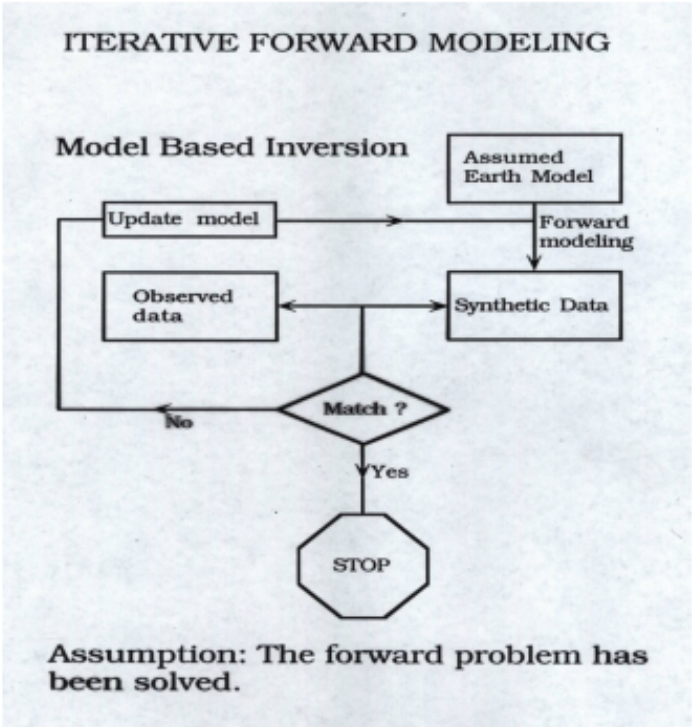
https://www.autotrader.com/car-tech/what-is-autopilot-and-do-other-cars-besides-tesla-offer-it-255973

Physics-Based Approach



Inversion

- Direct Inversion: Reverse physics – VERY UNSTABLE for most applications
- Model Based Inversion: Iterative Model fitting



INVERSION = OPTIMIZATION

Inversion Formulation

- **Data Vector**

$$\mathbf{d} = [d_1, d_2, .., d_N]^T$$

- **Model Vector**

$$\mathbf{m} = [m_1, m_2, .., m_M]^T$$

- **In general**

$$N \neq M.$$

- **Forward Modeling**

$$\mathbf{d} = g(\mathbf{m})$$

- **Linear Problem**

$$\mathbf{d} = \mathbf{Gm}$$

Inversion Formulation

Error/Cost/Objective/Misfit Function

measures differences or similarities between observed and synthetic (numerically computed) data

$$E(\mathbf{m}) = \| \mathbf{d} - g(\mathbf{m}) \|$$

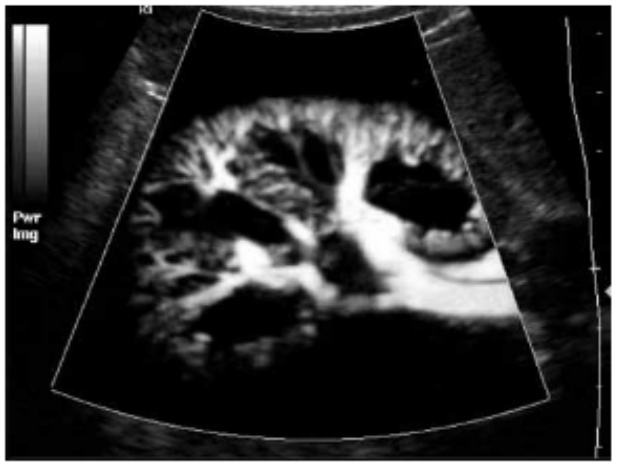
$\| . \|$ is called a 'norm' – most commonly used norm is an L_2 norm or a least squares error measure

$$E(\mathbf{m}) = (\mathbf{d} - g(\mathbf{m}))^T (\mathbf{d} - g(\mathbf{m}))$$

$$E(\mathbf{m}) = \sum_{i=1}^{ND} \left(d^i_{obs} - d^i_{syn} \right)^2$$

Example1

- Ultrasound imaging, also called ultrasound scanning or sonography, involves exposing part of the body to high-frequency sound waves to produce pictures of the inside of the body.



Kidney

Example2

PET Image

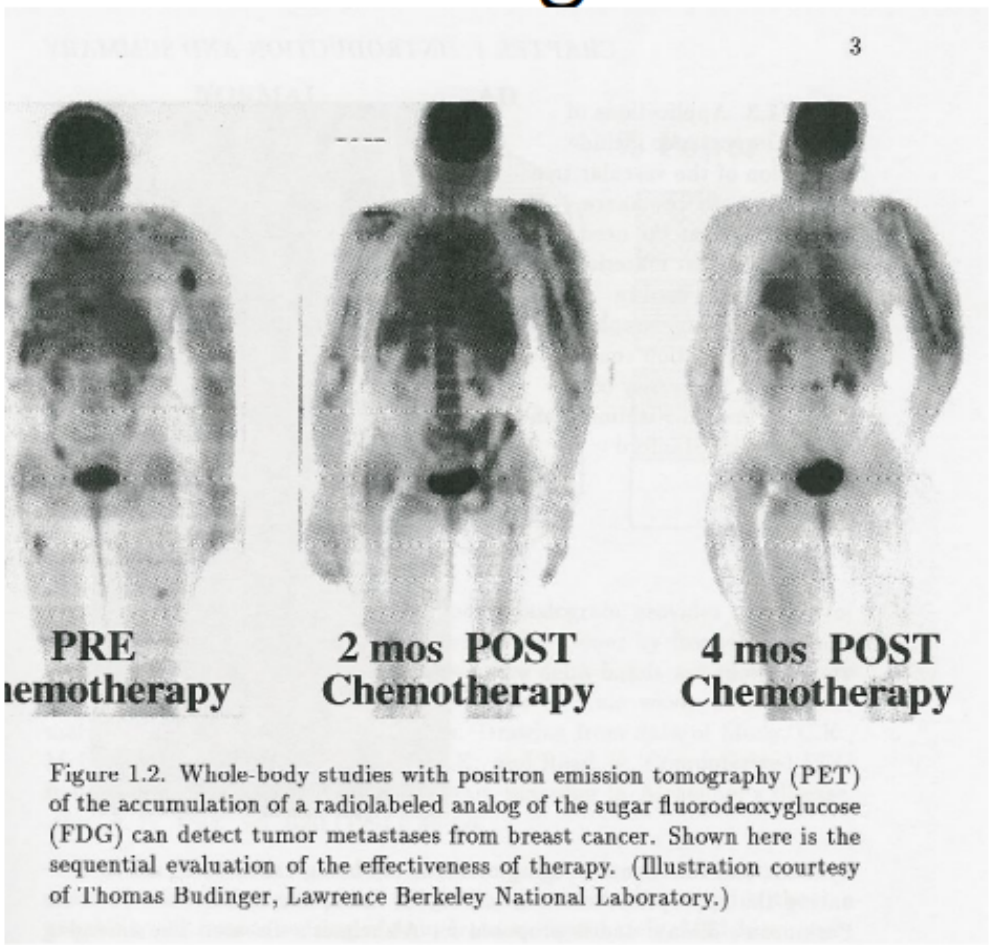


Figure 1.2. Whole-body studies with positron emission tomography (PET) of the accumulation of a radiolabeled analog of the sugar fluorodeoxyglucose (FDG) can detect tumor metastases from breast cancer. Shown here is the sequential evaluation of the effectiveness of therapy. (Illustration courtesy of Thomas Budinger, Lawrence Berkeley National Laboratory.)

Example3

4D seismic image

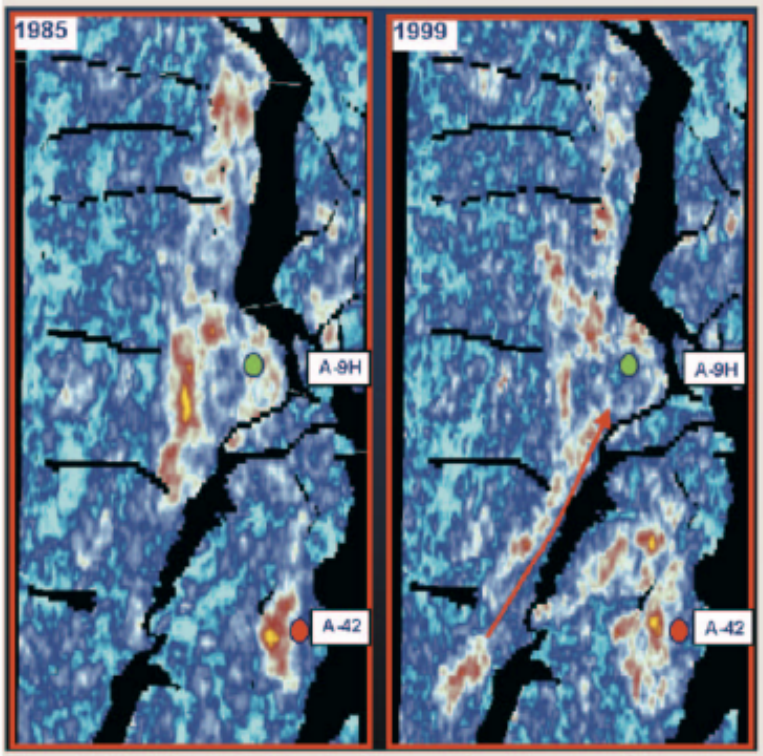
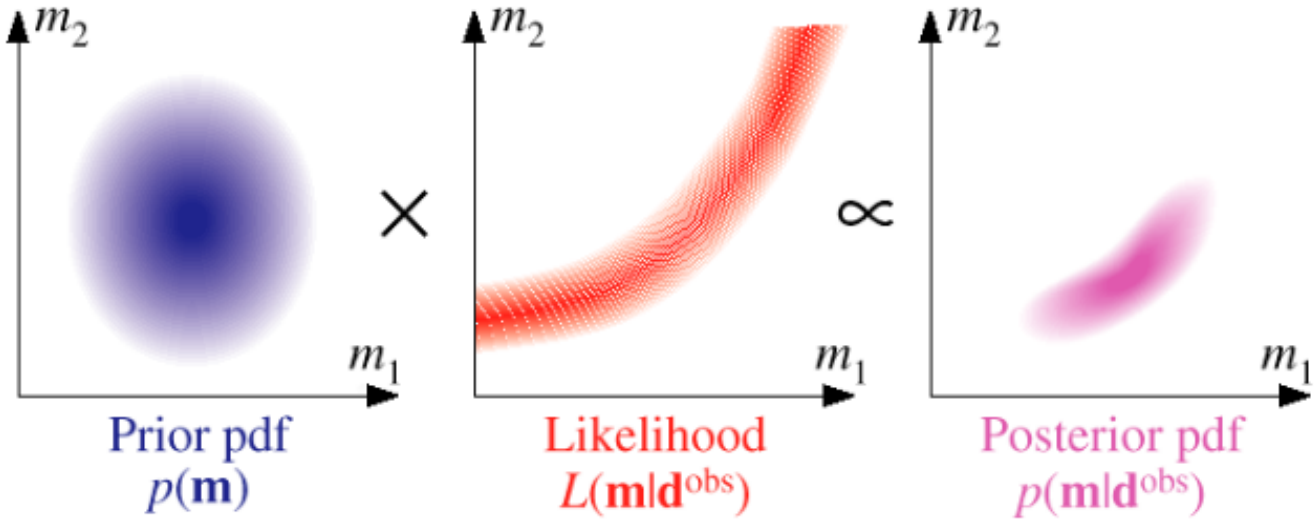


Figure 2. Gullfaks Field, Statoil, Norway. Gas injection well (red) at bottom. Oil producer well (green) above. Time-lapse imaging from 1985 to 1999 shows gas from injector has channeled to the southwest, crossed a major fault block, and flowed northeast along the block edge to reach the producer.

Bayesian Inversion

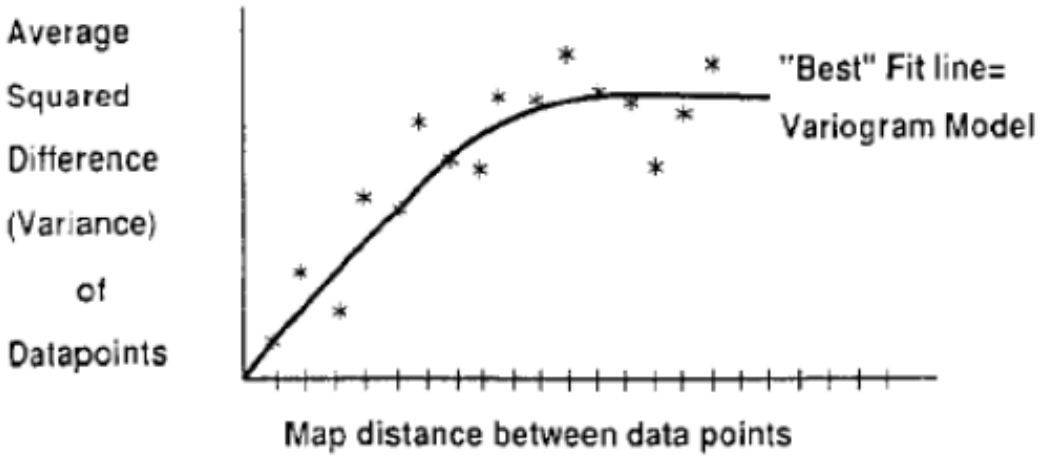


Geostatistics

- Geostatistics is the application of several statistical tools that are used to determine the spatial distribution of geologic variables.
  - Variogram
  - Kriging
  - Co-kriging
  - Conditional Simulation
- A geostatistician would typically derive an "expected value" or average map and some quantitative idea of how accurate (good or bad) the map is.
- LEARN from your data through simple statistical data calculating means, analysis (such as variances, minimum and maximum values and histogram plots) and variogram analysis.
- Find relationships between data sets (if possible) through cross plots. This is usually done by trying to find relationships between sparse well data and relatively dense seismic data.

Variograms

- A Variogram is a graph that is used to express the degree of spatial continuity of a regionalized (mappable) variable.
- It is a cross plot of the average squared difference of the variable of interest between all data pairs a given distance apart versus the distance apart.



Calculating a Variogram

- Make a list of all possible data pairs and compute distance apart and $(Z_i - Z_j)^2$

Well 1 (X,Y,Z)	Well 2 (X,Y,Z)	Distance (1-2)	$(Z_1 - Z_2)^2$
Well 1 (X,Y,Z)	Well 3 (X,Y,Z)	Distance (1-3)	$(Z_1 - Z_3)^2$
Well 1 (X,Y,Z)	Well 4 (X,Y,Z)	Distance (1-4)	$(Z_1 - Z_4)^2$
Well 2 (X,Y,Z)	Well 3 (X,Y,Z)	Distance (2-3)	$(Z_2 - Z_3)^2$
Well 2 (X,Y,Z)	Well 4 (X,Y,Z)	Distance (2-4)	$(Z_2 - Z_4)^2$
Well 3 (X,Y,Z)	Well 4 (X,Y,Z)	Distance (3-4)	$(Z_3 - Z_4)^2$

- Where Z is the variable of interest.

Kriging

How do we use Kriging?

1. **Sample**, preferably at different resolutions
2. **Calculate** the **experimental variogram**
3. **Model** the variogram with one or more **authorized functions**
   - N.b. the variogram model may already be known from other studies or theoretical considereations
4. **Apply** the **kriging system of equations**, with the variogram model of spatial dependence, at each point to be predicted
   - Predictions are often at each point on a **regular grid** (e.g. a raster map)
5. Calculate the **variance** of each prediction; this is based only on the **sample point locations**, *not* their data values.

Kriging

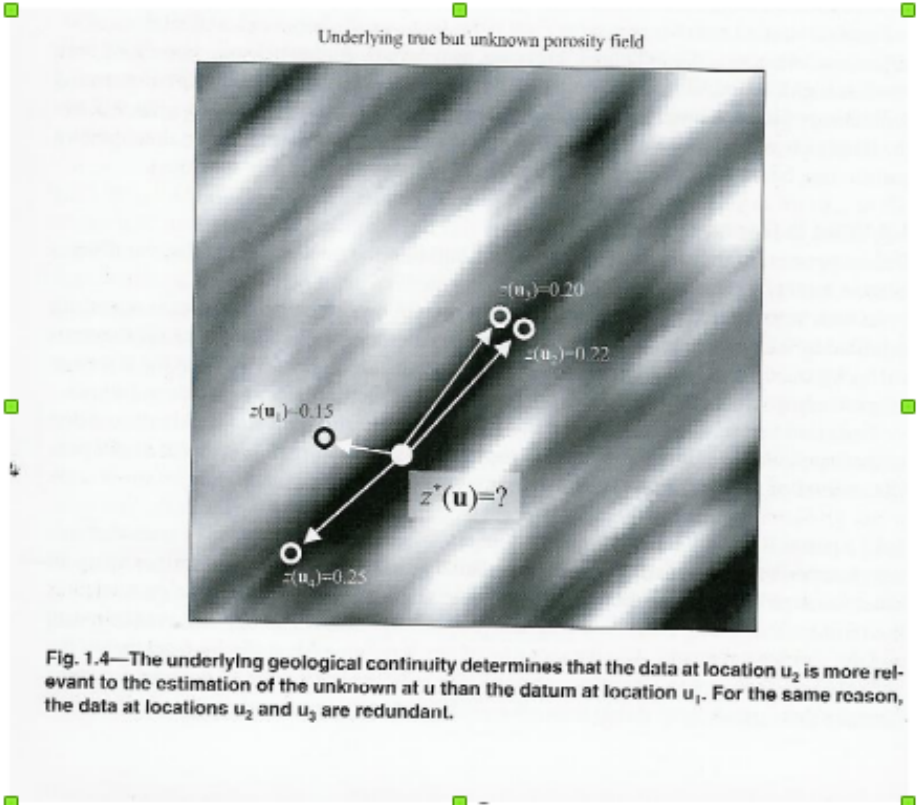


Fig. 1.4—The underlying geological continuity determines that the data at location u_2 is more relevant to the estimation of the unknown at u than the datum at location u_1. For the same reason, the data at locations u_2 and u_3 are redundant.

$$z_1^*(\mathbf{u}) = \sum_{\alpha=1}^{n} \lambda_\alpha z_1(\mathbf{u}_\alpha)$$

Use spatial relationship between data and the unknown to model the unknown

Kriging relies on a measure of spatial continuity (variograms – measures the degree of statitstical dissimilarities between any two values separated by a vector h)

From Caers 2007

Co-Kriging

- Co-kriging is an extension to include data of different type

Z1 – data 1

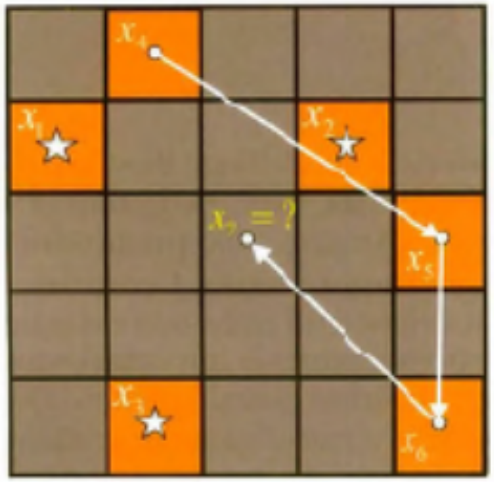
Z2 – data 2

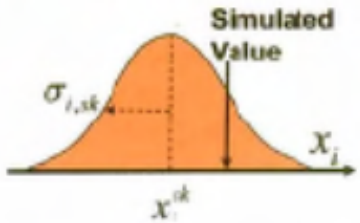
$$z_1^*(\mathbf{u}) = \sum_{\alpha=1}^{n_1} \lambda_\alpha^{(1)} z_1(\mathbf{u}_\alpha) + \sum_{\alpha=1}^{n_2} \lambda_\alpha^{(2)} z_2(\mathbf{u}'_\alpha)$$

$$\begin{aligned}
C_{11}(\mathbf{h}) &= Cov\{Z_1(\mathbf{u}), Z_1(\mathbf{u}+\mathbf{h})\} \\
C_{22}(\mathbf{h}) &= Cov\{Z_2(\mathbf{u}), Z_2(\mathbf{u}+\mathbf{h})\} \\
C_{12}(\mathbf{h}) &= Cov\{Z_1(\mathbf{u}), Z_2(\mathbf{u}+\mathbf{h})\} \\
C_{21}(\mathbf{h}) &= Cov\{Z_2(\mathbf{u}), Z_1(\mathbf{u}+\mathbf{h})\},
\end{aligned}$$

Geostatistics

Sequential Gaussian Simulation





1) Pick non - simulated cell i at random $(i = 1)$

2) Compute kriging estimate and variance

$$x_i^{sk} = m_x - \sum_{j=1}^{i-1} w_j (x_j - m_x)$$

$$\sigma_{i,sk}^2 = \sigma_x^2 - \sum_{j=1}^{i-1} w_j C_{ij}$$

3) Draw simulated value of x_i at random from :

$$p(x_i | x_1, \cdots, x_{i-1}) \propto \exp\left\{ -\frac{[x_i - x_i^{sk}]^2}{2\sigma_{i,sk}^2} \right\}$$

4) Treat simulated x_i as additional control point

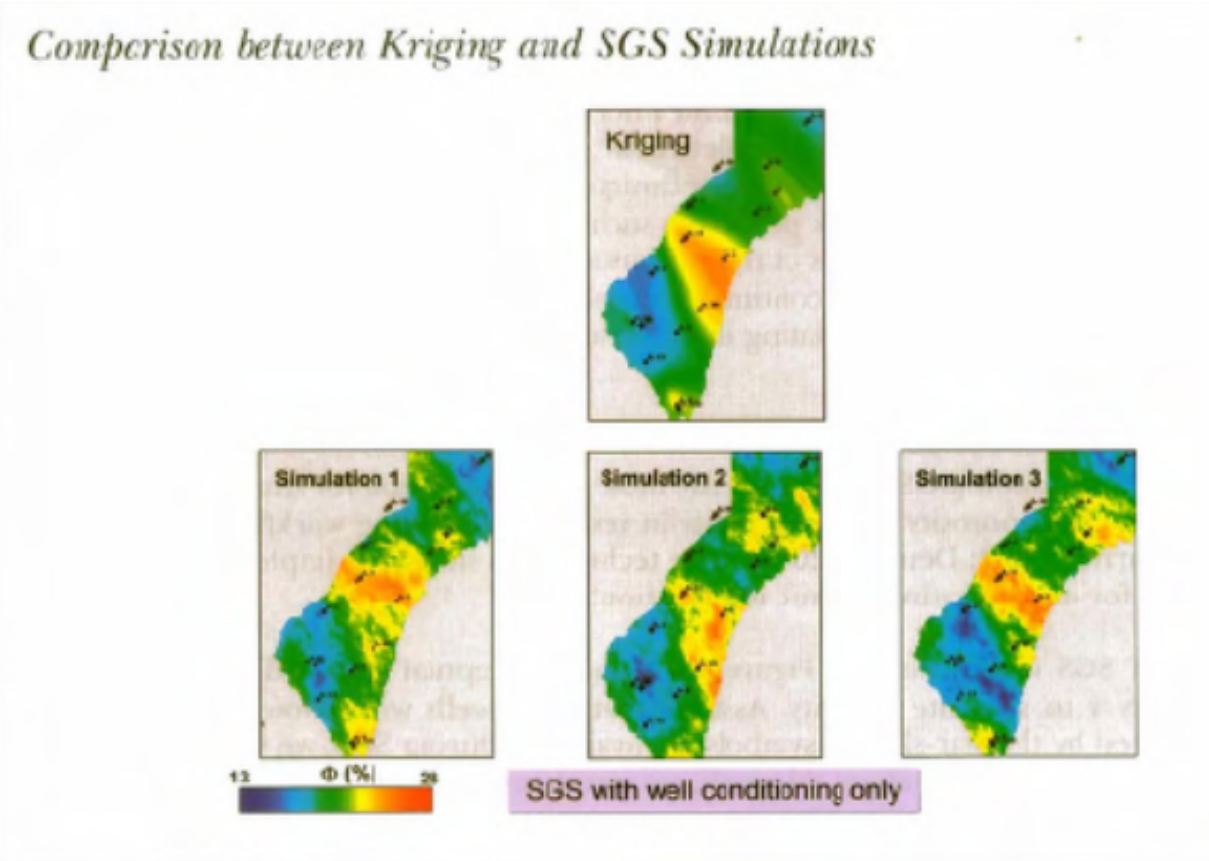
5) Go back to 1) until entire grid is simulated

Fig. 3.2.1

Doyen 2007

Seismic Reservoir Characterization

SGS


Comparison between Kriging and SGS Simulations
Kriging
Simulation 1
Simulation 2
Simulation 3
Φ (%)
SGS with well conditioning only

The Next Steps

- Beyond physics and Geostatistics?
- A combination of the two?
- Machine Learning (Not completely different).
- Big Data, Large memory, Computation time.

Anaconda Installation Instructions

What is Anaconda?

- It includes standard Python libraries used in machine learning such as Numpy, Matplotlib, Pandas, Scikit-learn, and more.
- An open-source distribution of software that is used to manage packages that supplement Python and R coding. Some useful software included with Anaconda includes Spyder (an Integrated Development Environment) and Jupyter Notebooks.
- An installation of Anaconda also includes standard Python libraries used in machine learning such as Numpy, Matplotlib, Pandas, Scikit-learn, and more.
- Anaconda allows you to manage the dependencies of these libraries by creating separate computing environments for different projects. This is done because often times large sets of general purpose libraries will be incompatible with each other.

https://docs.conda.io/projects/conda/en/latest/user-guide/tasks/manage-environments.html.

Practice Sessions using Pyhton

Installation Instructions

1. Go to https://www.anaconda.com/products/individual
2. Click Download
3. Choose the installer that matches your operating system.
4. Download and follow the install wizard
5. After installing, if the anaconda navigator doesn't launch automatically, open a terminal window and type anaconda-navigator. This will open up a window with conda applications. We will be using Jupyter-Notebooks extensively
6. You can access Jupyter Notebooks directly by opening a terminal and typing. **jupyter notebook**
7. If you have any questions, comments, concerns, or complaints just email at dvoytan@utexas.edu