Data Analytics (PETR 6397)

Introduction

Dr. Ahmad Sakhaee-Pour

Petroleum Engineering Department

University of Houston

Spring 2025

General information

Instructor: Dr. Ahmad Sakhaee-Pour

Office: ERP 9, Rm 162

Email: asakhaee@central.uh.edu

Office hours: Thursday, 8:30–9:30 pm (by appointment)

Textbook

Required

1. Geron, A., Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly, 2022
2. Lake, L. W., Petroleum Engineering Handbook, 2007 (https://petrowiki.spe.org/)

Optional

1. Chollet, F., Deep learning with Python. Simon and Schuster, 2021
2. Bishop, C. M., Pattern recognition and machine learning. Springer, 2006
3. Peters, E. J., Advanced Petrophysics: Geology, porosity, absolute permeability, heterogeneity, and geostatistics (Vol. 1). Greenleaf Book Group, 2012
4. Peters, Ekwere J. Advanced petrophysics: Dispersion, interfacial phenomena (Vol. 2). Greenleaf Book Group, 2012
5. Published articles in petroleum engineering: https://onepetro.org/

Evaluation

Basis for grading

- Homework: 20%
- Midterm: 40%
- Final exam: 40%

*Minimum final exam grade required to pass the course: 50

Grading

- ≥ 90% : A
- ≥ 80% : B
- ≥ 70% : C
- ≥ 60% : D

Class attendance

- You are encouraged to attend the lectures

- Slides will be shared before the beginning of each lecture

Midterm date and policy

First part

- Written: Thursday, Mar 6, 5:30 – 7:00 pm, open notes but no access to the internet or laptop
- Location: ERP 9 135, regular classroom

Second part

- Code: 24 hours
- Submission: 7:00 pm, on Friday, Mar 7, via Canvas
- Location: online

Final exam date and policy

- Date: Thursday, May 1, 6:00 – 8:00 pm (campus-wide schedule)
- Rules will be announced later

Exam policies

- No make-up exam
- Nothing short of true verifiable emergency will be accepted as an excuse
- Disputed grades must be resubmitted for regrading within 72 hours of their return to students

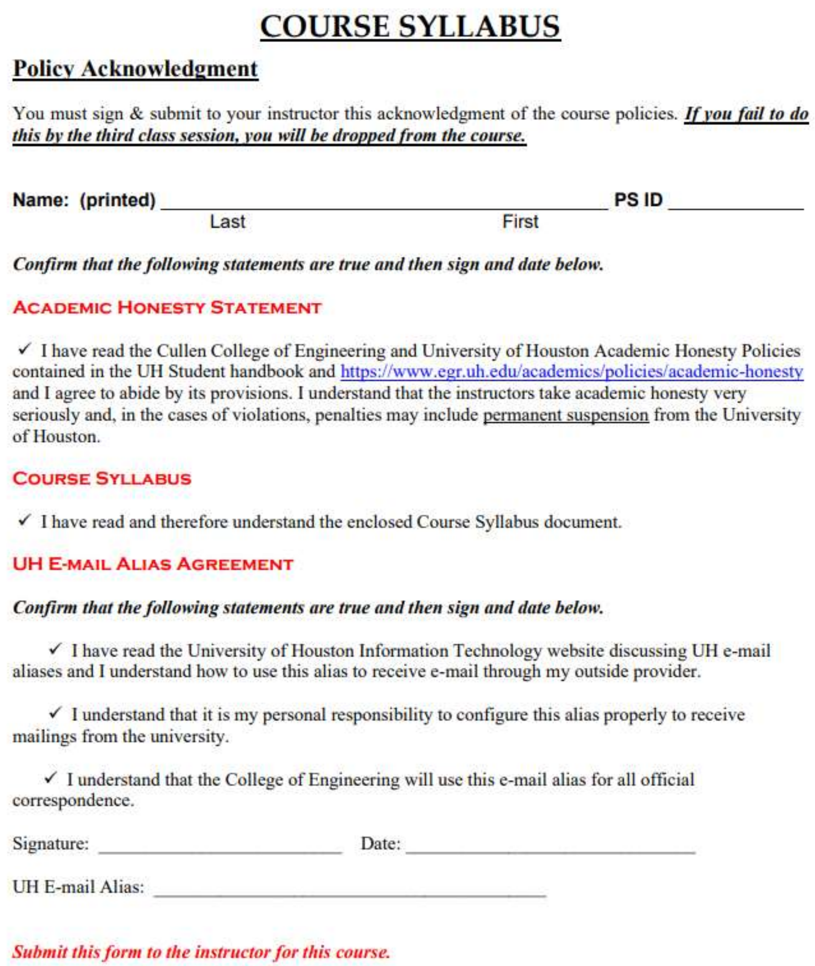
Homework

- Homework will be due at the beginning of the class on the assigned due date

- Soft copy should be submitted through Canvas

Homework 1 (more info on the last slide)

- Please turn in the acknowledgement page after reading the syllabus



<u>COURSE SYLLABUS</u>

<u>Policy Acknowledgment</u>

You must sign & submit to your instructor this acknowledgment of the course policies. ***<u>If you fail to do this by the third class session, you will be dropped from the course.</u>***

Name: (printed) ______________________________________ **PS ID** ____________
Last First

Confirm that the following statements are true and then sign and date below.

ACADEMIC HONESTY STATEMENT

✓ I have read the Cullen College of Engineering and University of Houston Academic Honesty Policies contained in the UH Student handbook and https://www.egr.uh.edu/academics/policies/academic-honesty and I agree to abide by its provisions. I understand that the instructors take academic honesty very seriously and, in the cases of violations, penalties may include <u>permanent suspension</u> from the University of Houston.

COURSE SYLLABUS

✓ I have read and therefore understand the enclosed Course Syllabus document.

UH E-MAIL ALIAS AGREEMENT

Confirm that the following statements are true and then sign and date below.

✓ I have read the University of Houston Information Technology website discussing UH e-mail aliases and I understand how to use this alias to receive e-mail through my outside provider.

✓ I understand that it is my personal responsibility to configure this alias properly to receive mailings from the university.

✓ I understand that the College of Engineering will use this e-mail alias for all official correspondence.

Signature: ______________________ Date: __________________________

UH E-mail Alias: ___________________________________

Submit this form to the instructor for this course.

Topics

- Introductory terms in data science and petroleum engineering, and online resources
- Coding environment (Jupyter Notebook, Google Colab, TensorFlow vs PyTorch) and assisting applications
- Regression
- Classification
- Support Vector Method, Decision Tree
- Dimensionality reduction
- Clustering

More recent developments

- Deep learning
- Convnet
- Generative Adversarial Network (GAN)
- Variational autoencoder (VAE) and Diffusion model

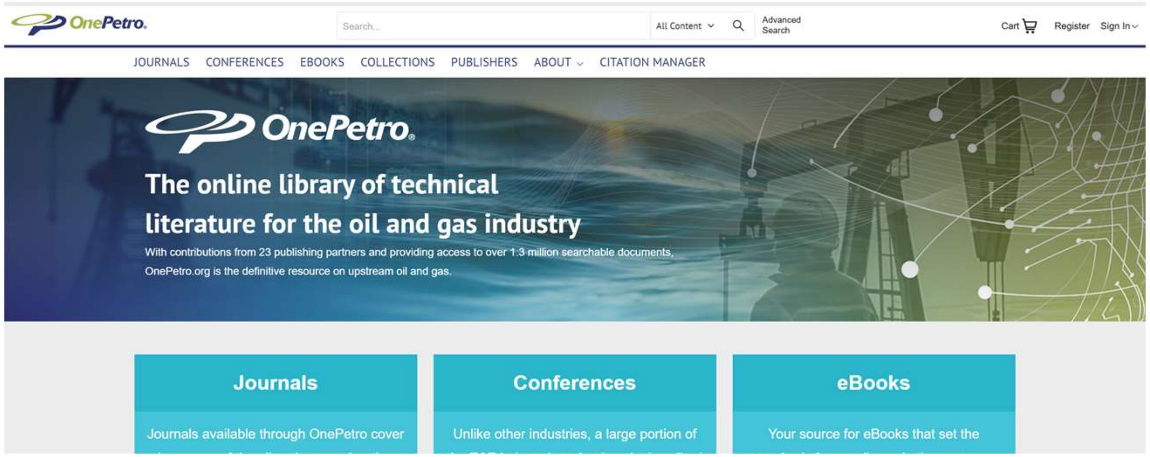
Introduction

Online resources

- OnePetro, InterPore, and Elsevier
- PetroWiki

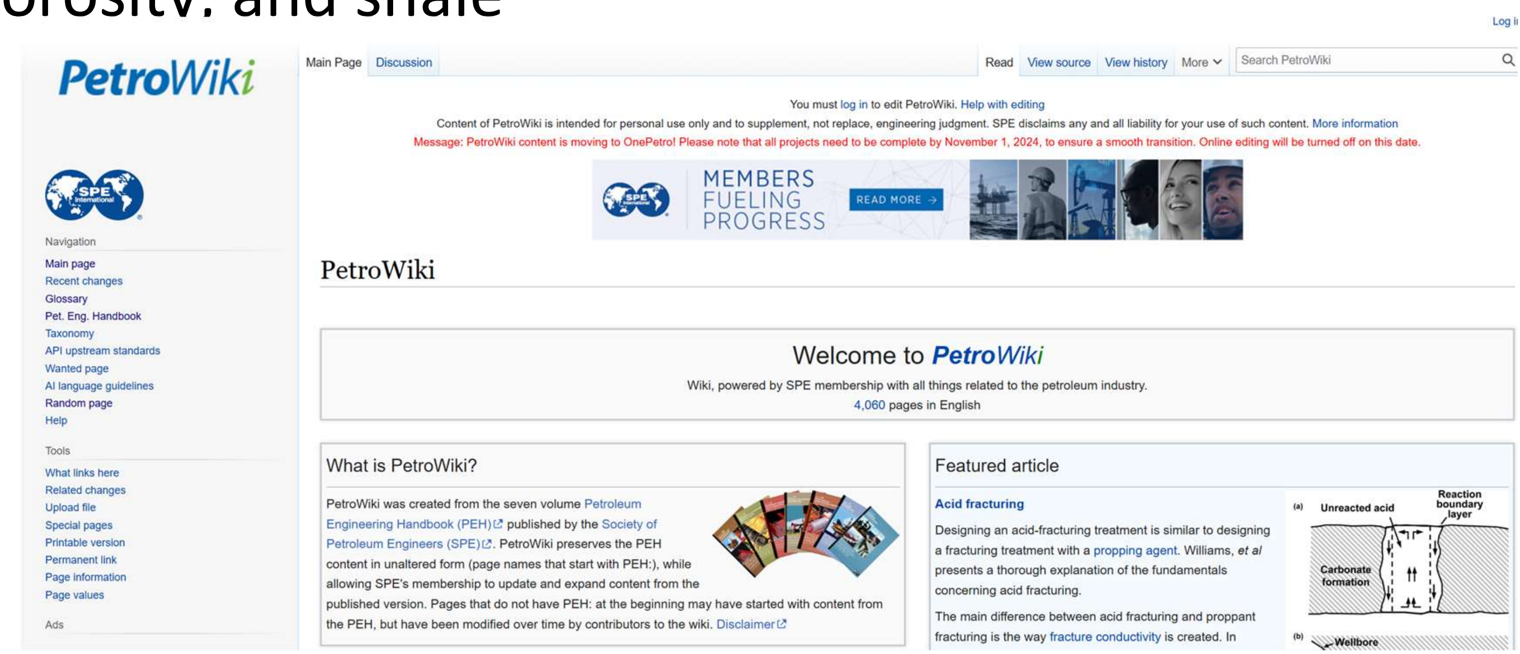
OnePetro, InterPore, Elsevier

- OnePetro publishes conference and journal articles of petroleum engineering
- Researchers also release their findings through other venues (InterPore, Elsevier)
- Generative AI and, to some extent, deep learning are in the early stages of deployment in petroleum engineering



PetroWiki

- Information (if available) is more accurate than common search engines
- First check whether the basic definition is available from the glossary
- In class: Try finding the definitions of permeability, porosity, and shale



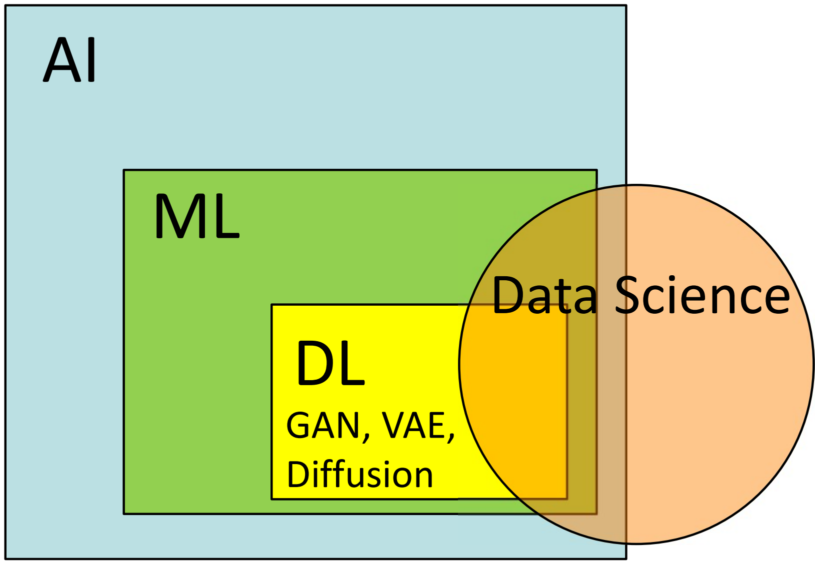
Importance of linear algebra

- Linear algebra plays a major role in this course
- We will review some of the relevant concepts quickly
- It would be great if your refresh your linear algebra knowledge

Data Science

- Data Science is at the intersection of computer science and statistics

- It deals with data collection, preparation, analysis, visualization, management, and preservation of information to extract knowledge from data



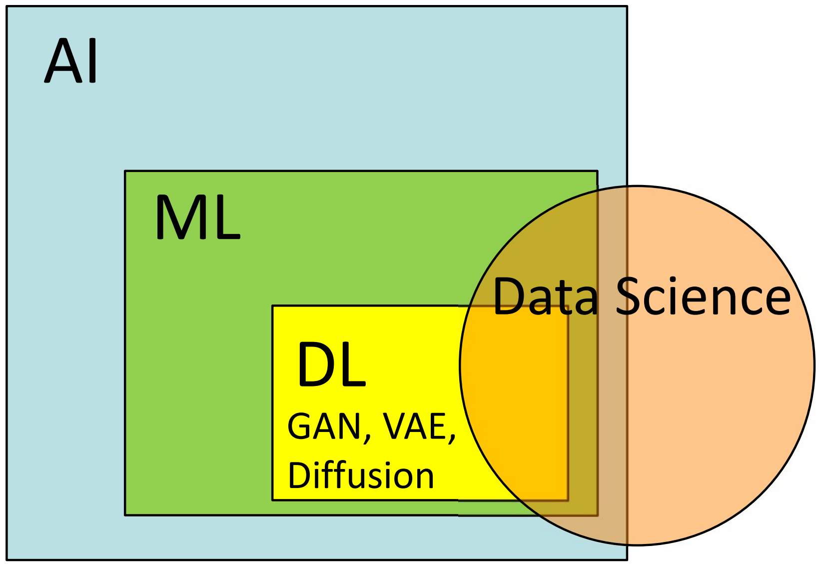
- You do not have to use AI, ML, or DL to solve your problem

- Using GAN or Stable Diffusion Model is not necessarily a good choice even if it is a hot topic

Artificial Intelligence (AI)

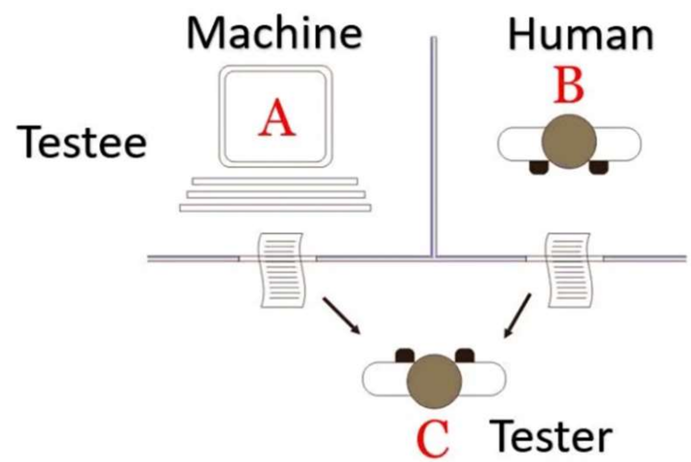
- Artificial intelligence builds machines and computers to learn, reason, and act in ways that would usually require human intelligence

- The Turing test defines the intelligence



Turing test (imitation game, 1950)

- Turing suggests we should ask if the machine can win a game called the Imitation Game
- A human evaluator (C) judges a text transcript of a natural-language conversation between a human (B) and a machine (A)
- The machine passes (and has intelligence) if the evaluator cannot reliably tell them apart



AI differs from other computer-based methods

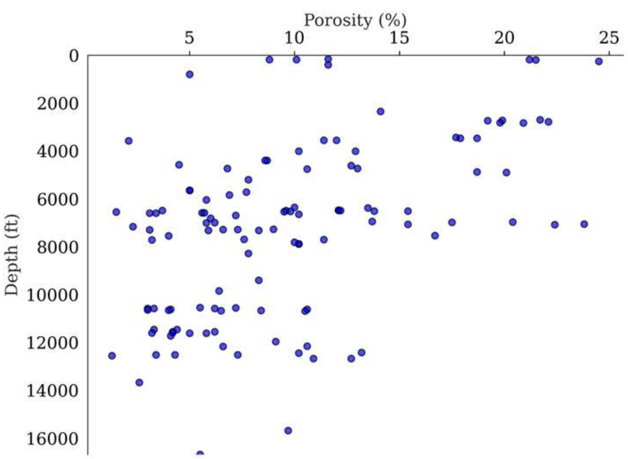
- Intelligence is not memorizing

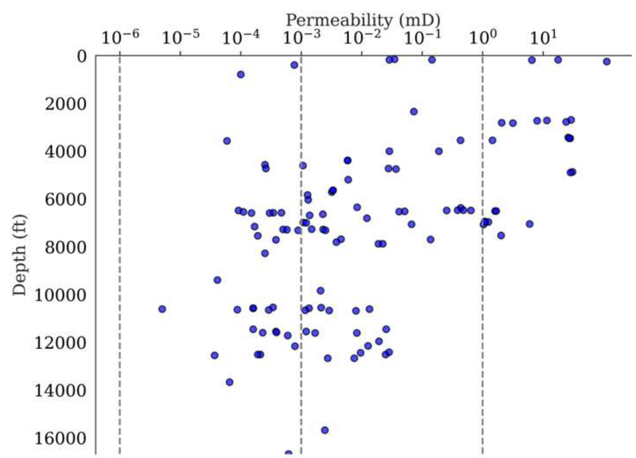
- Intelligence is not logical thinking based on a structured and rule-based approach

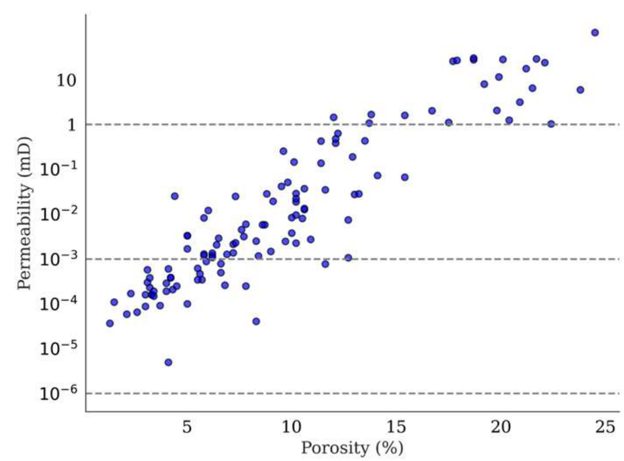
- AI uses data patterns and statistical analysis to decide

Basic definitions of permeability and porosity

- Permeability and porosity are two fundamental properties of rock that control multiphase flow (oil, water, carbon dioxide) in the subsurface
- Permeability indicates the flow rate for a given pressure difference (higher permeability means a higher rate)
- Porosity controls the void fraction of porous medium (higher porosity means more oil stored in the subsurface)

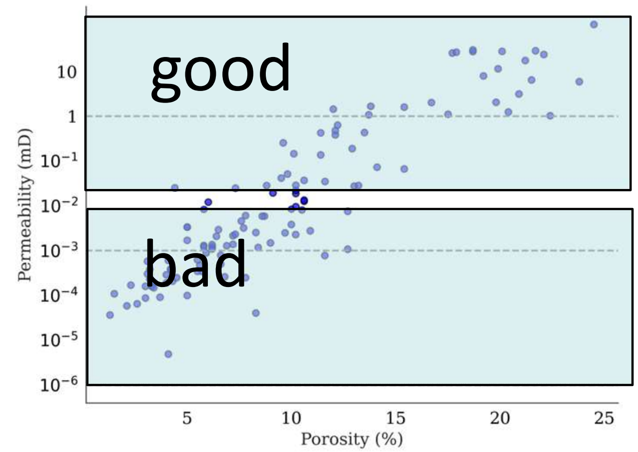




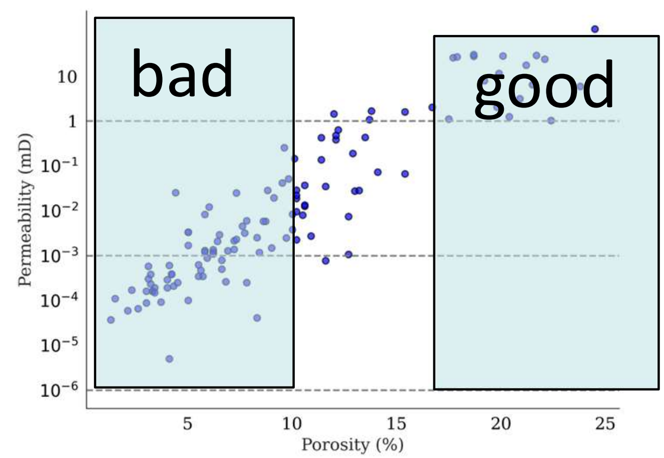


Question: Which one is AI?

A) Higher permeability is better because we can produce hydrocarbon faster

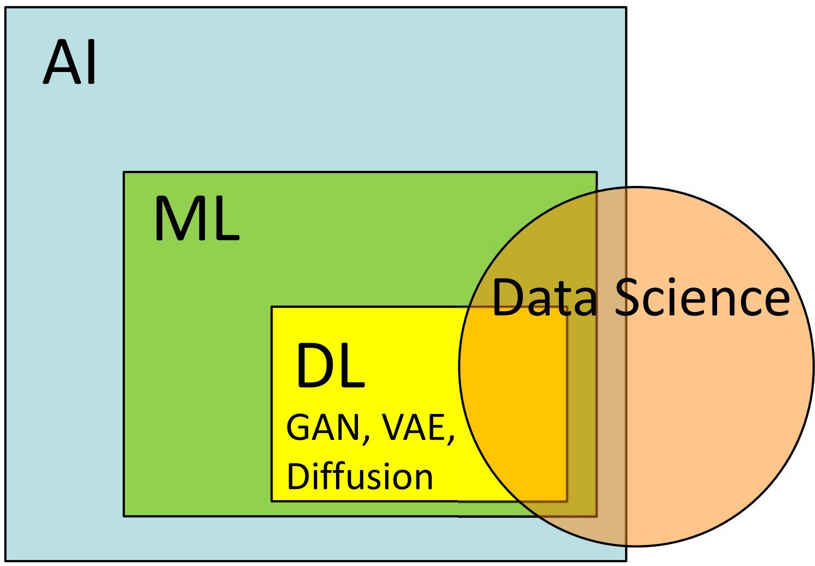


B) Higher porosity is better because we can recover more hydrocarbon volume



Machine learning

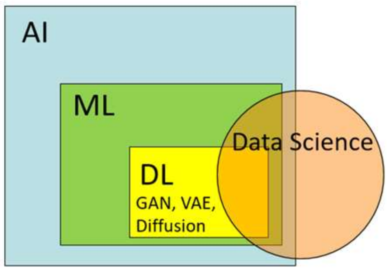
- Machine learning (ML) focuses on creating algorithms to use inputs to refine and improve their capabilities for dealing with future inputs.



- ML has predictive analytics

- ML figures out patterns of incoming information and deals with future inputs

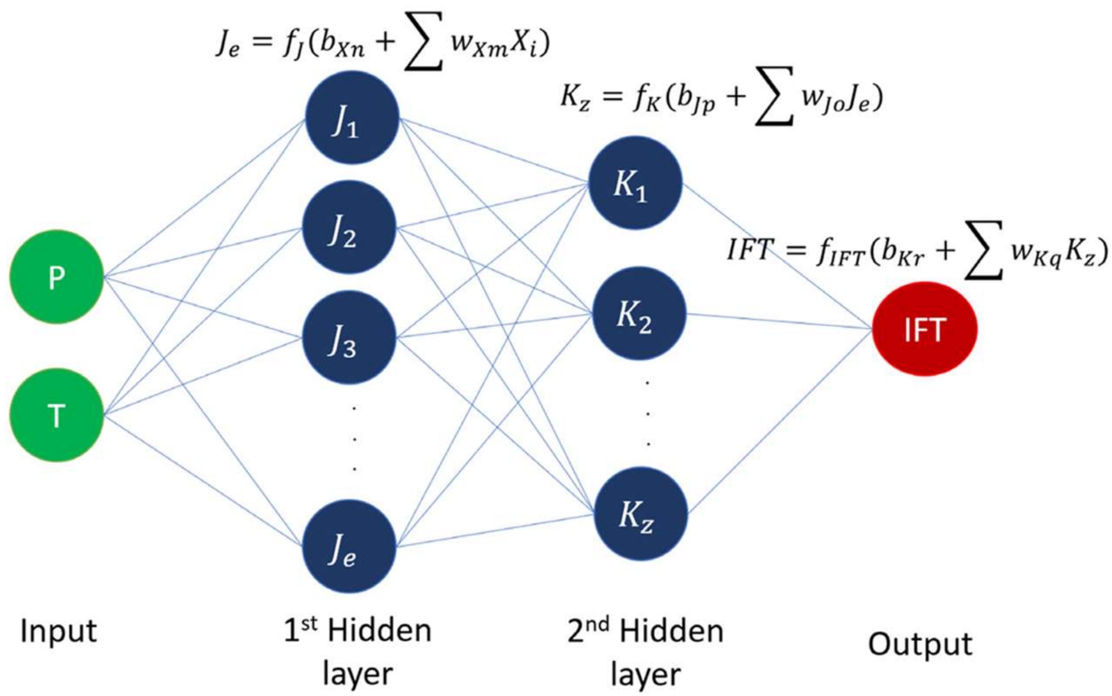
- Arthur Samuel: The field of study that gives computers the ability to learn without explicitly being programmed



Deep learning

• Deep learning uses multilayered neural networks to mimic the decision-making of the human brain

• Deep means at least two layers between the input and output layers. Sometimes, people say there must be at least three layers between the input and output.



Alessa et al. 2022

Neural networks have been around for a long time

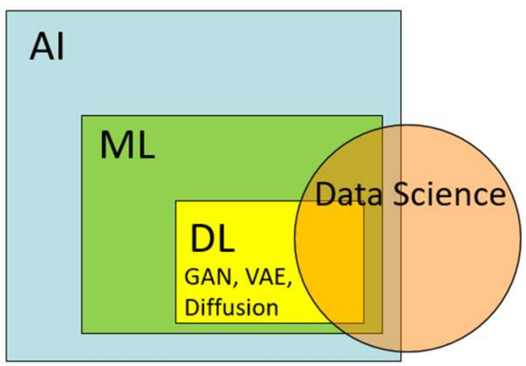
- 1943: McCulloch and Pitts first discussed how neurons might work
- 1957: Rosenblatt simulated the perceptron on an IBM 704
- …

IBM 704

- The difference is that now we have access to powerful computers using GPUs
- IBM 704 executed ~12,000 floating-point operations per second (FLOPS)
- iPhone 12 performs ~11 trillion FLOPS

Generative models and their applications in petroleum engineering

- GAN, VAE, and Diffusion models are based on deep neural networks
- GAN, VAE, and Diffusion models generate synthetic data by analyzing the latent space. That is why we call them generative models
- Where is ChatGPT in the diagram?
- Why do we need to generate new data in the subsurface?
- Which areas have more potential in petroleum engineering?



Supervised vs unsupervised learning

A. Supervised Learning:

We know the correct answer during the training

Goal: Create the “correct” outputs for new inputs by learning from the training examples

Relevant topics include classification and regression

B. Unsupervised Learning:

We DO NOT know the correct answer during the training

Goal: Discover the underlying structure (pattern, trend) by analyzing the inputs

Relevant topics include clustering and probability density estimation

Question: What is the main difference?

Regression

- Determine the output value from the input
- The output is not restricted to discrete values
- Example 1: Estimate the permeability of a formation from its porosity
- Example 2: Approximate the production rate of a formation given its bottom hole pressure, depth, thickness, and location
- Example 3: Determine how much carbon dioxide can be sequestered in a formation from its thickness, porosity, and temperature

Classification

- Determine a decision boundary that separates one class from another
- The output belongs to a limited number of classes (groups)
- Example 1: Decide whether a formation is tight gas sandstone or shale from its permeability
- Example 2: Assess whether drilling in a specific location is economical based on information from adjacent wells, such as recovered hydrocarbon volume and produced fluid (gas vs oil).

Clustering

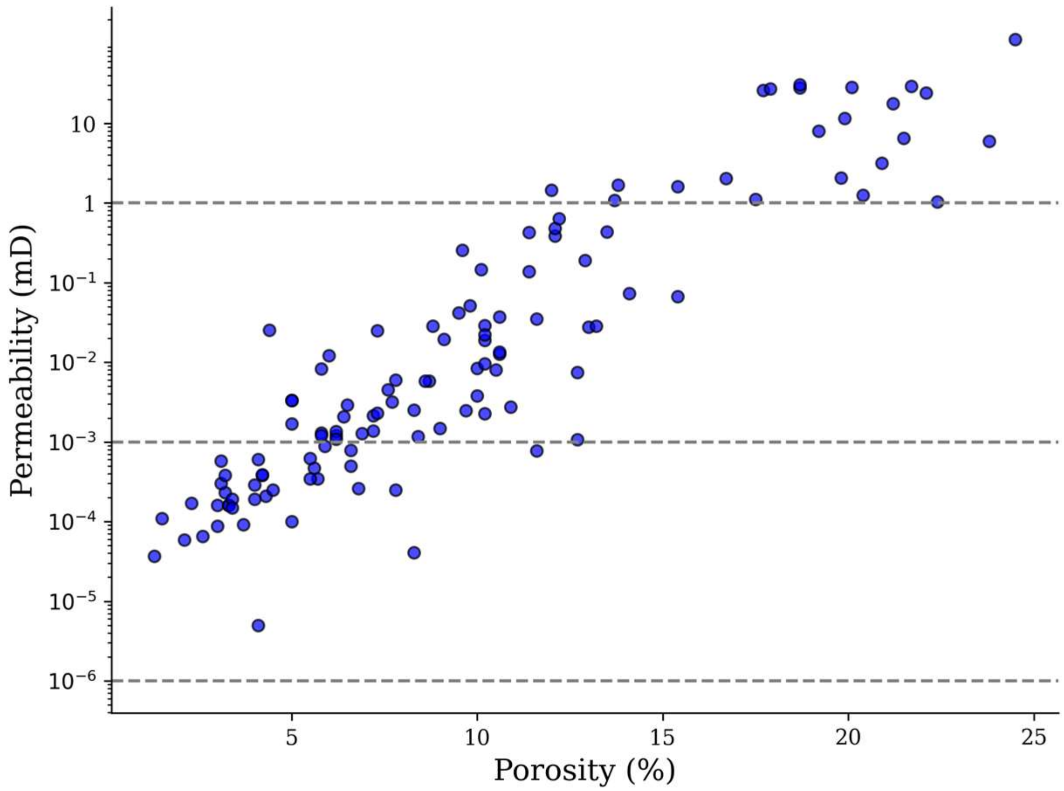
- Cluster (group) the samples together if their inputs are similar
- Clustering is similar to classification in the sense that the samples are divided into groups
- There is no correct answer to check the outcome in clustering
- Example 1: We have access to a large dataset from various reservoirs in West Texas, but the companies did not share the exact location, depth, and hydrocarbon recovery. How can we cluster the data to employ profitable recovery techniques such as flood injection, steam injection, hydraulic fracturing?

Reinforcement learning

- Agent takes actions based on inputs that affect the environment
- The agent gets rewarded or punished
- Goal: Learn to maximize the reward
- This is not our focus in this course
- There is currently limited research in this area compared with generative AI and deep learning

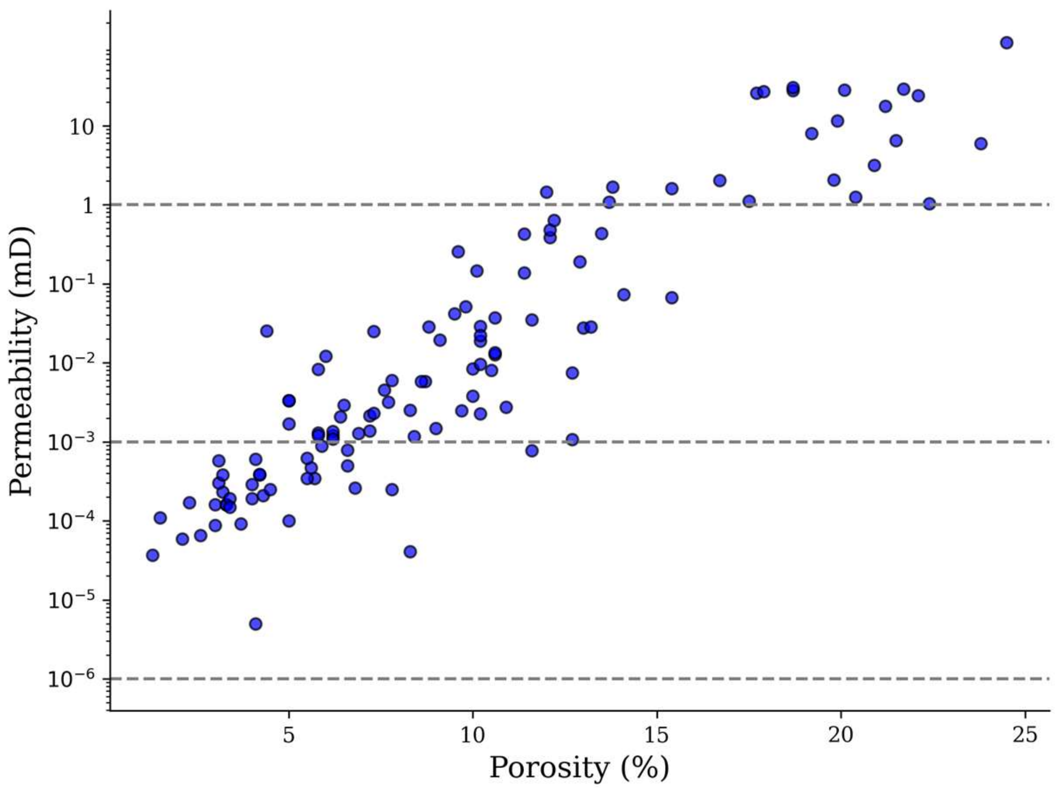
Question 1

- You have access to the following data. Determine the problem type (regression, classification, clustering)



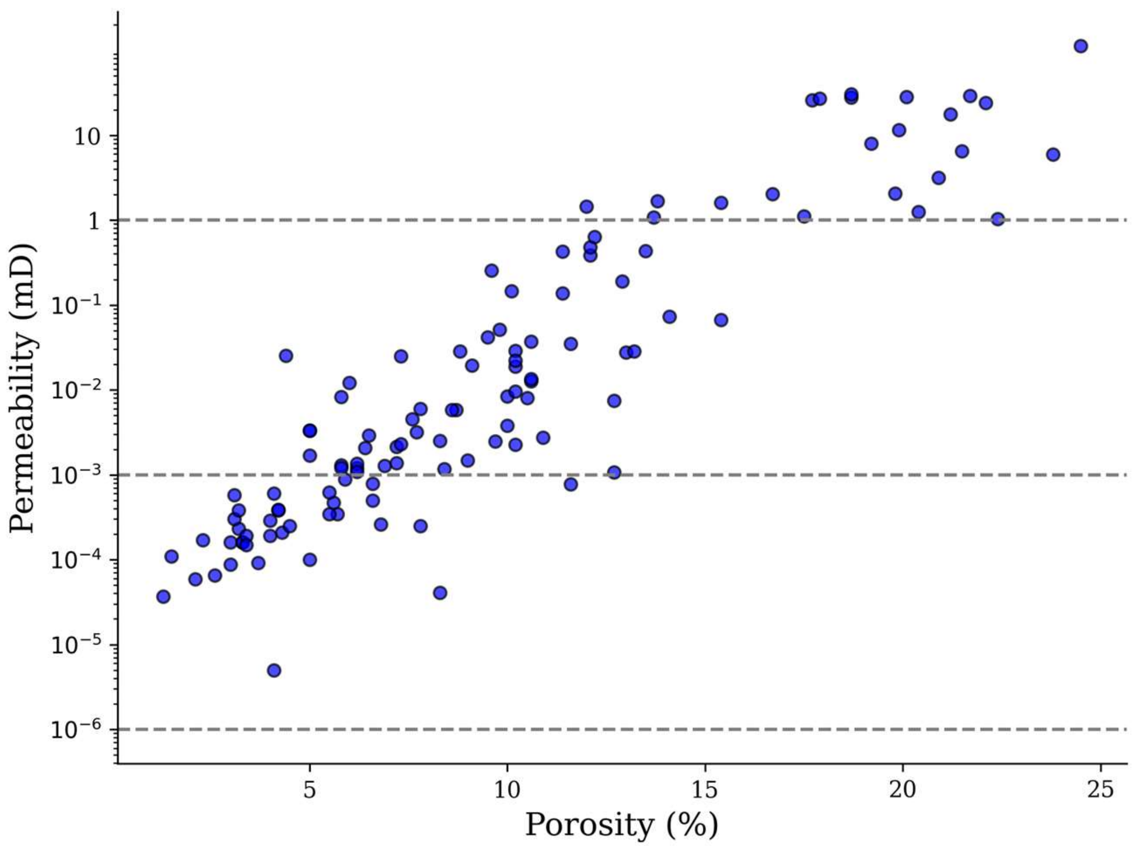
Question 2

- What if your goal is to determine whether each sample was recovered from a profitable formation? Is this regression, classification, or clustering



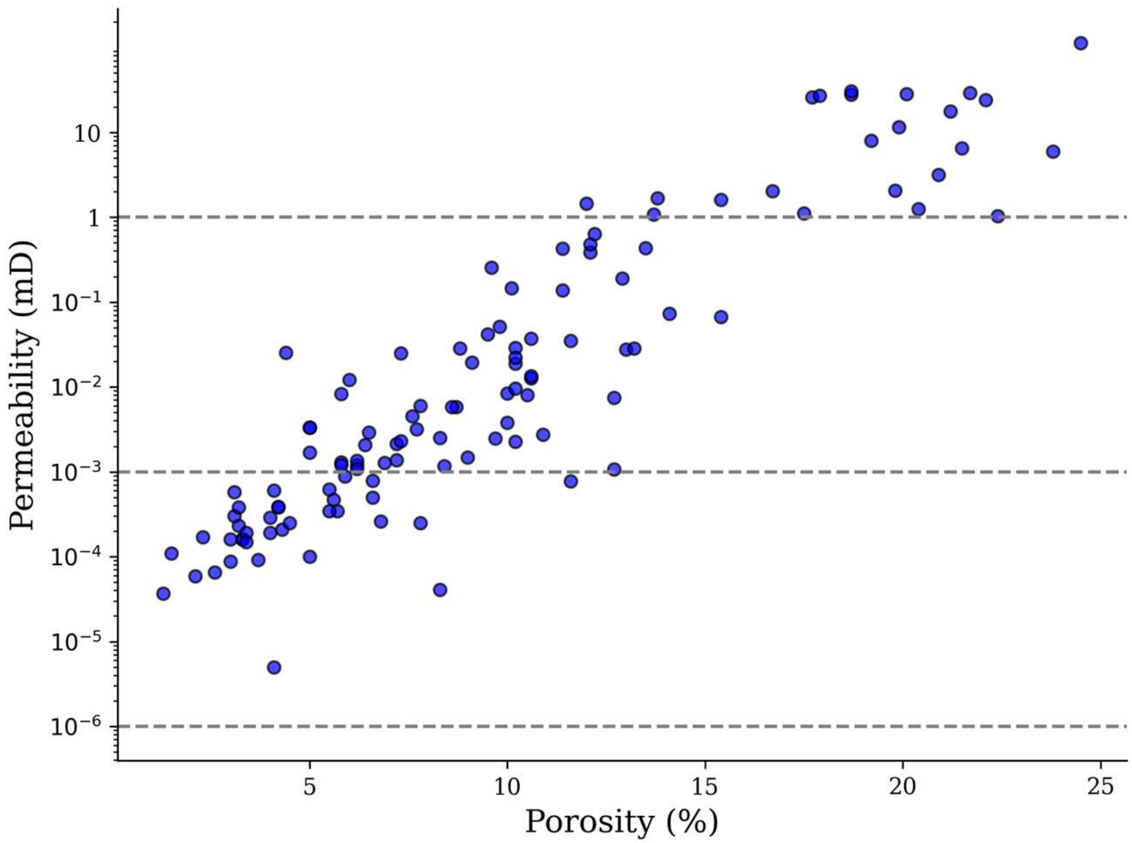
Question 3

- What if your goal is to determine the sample type (shale, tight gas, permeable formation) from porosity? Is this regression, classification, or clustering?



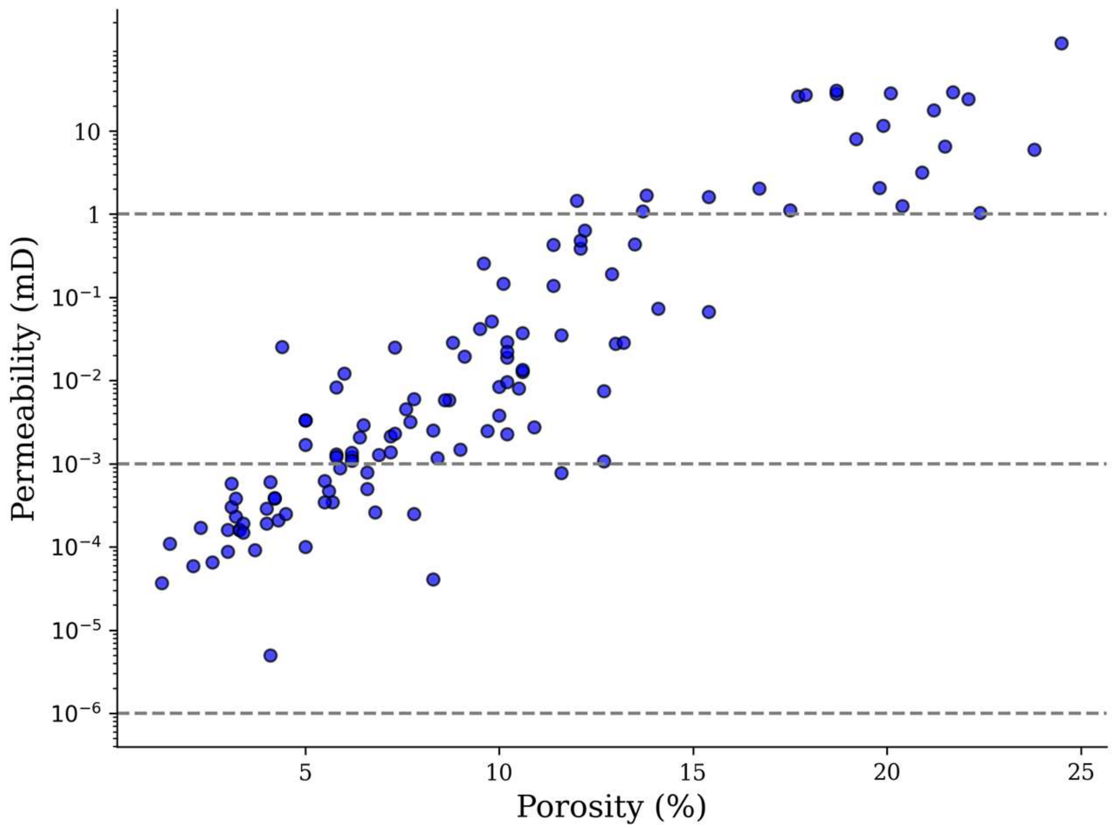
Question 4

- What if your goal is to determine the sample type (shale, tight gas, permeable formation) from permeability? Is this regression, classification, or clustering?



Question 5

- What if your goal is to determine sample permeability from porosity? Is this regression, classification, or clustering



Basics of Data Science

Data Science Cycle: Simple steps that are usually revisited at least a couple of times to reach acceptable results

Problem is translated into analytics question

Step 1. Problem definition

Are data available?
How?
Where?

Step 2. Investigation

Methodology
Model building
Model evaluation

Step 3. Data Analytics

Deployment

Step 4. Implementation

Problem statement

- **Descriptive Analytics** shows what happened in the past
- **Diagnostic Analytics** helps you understand why something happened in the past
- **Predictive Analytics** predicts what is most likely to happen in the future
- **Prescriptive Analytics** recommends what actions you should take to affect outcomes
- **Exploratory Analytics** shows what may be the reason

Data is the most important part

- Data preparation and collection may take a lot of time
- Document the data quality
- Clean the data (missing data, outliers, errors)
- Transform the data
- Combine various datasets to create new views
- Load the data into the target location
- Visualize the data
- Model development should not take long with the existing computer powers and assisting apps when:
  - We define the problem clearly
  - We implement appropriate metrics

Data type

- **Qualitative** (categorical)

    Profitable recovery of hydrocarbon vs. failure

    Good vs. medium vs. poor reservoir

    Onshore vs. offshore reservoir

    Tight gas vs. shale vs. permeable reservoirs

- **Quantitative**

    Rate of hydrocarbon recovery

    Recoverable hydrocarbon volume

    Reservoir pressure

    Reservoir temperature

Basic Definitions in Petroleum Engineering

Introduction

- Basic definitions of petroleum engineering, reservoir engineering, and petrophysics
- Reservoir trap
- Standard units in petroleum engineering and unit conversion

What is Petroleum Engineering?

- Petroleum engineering designs and develops methods to extract oil and gas from the surface

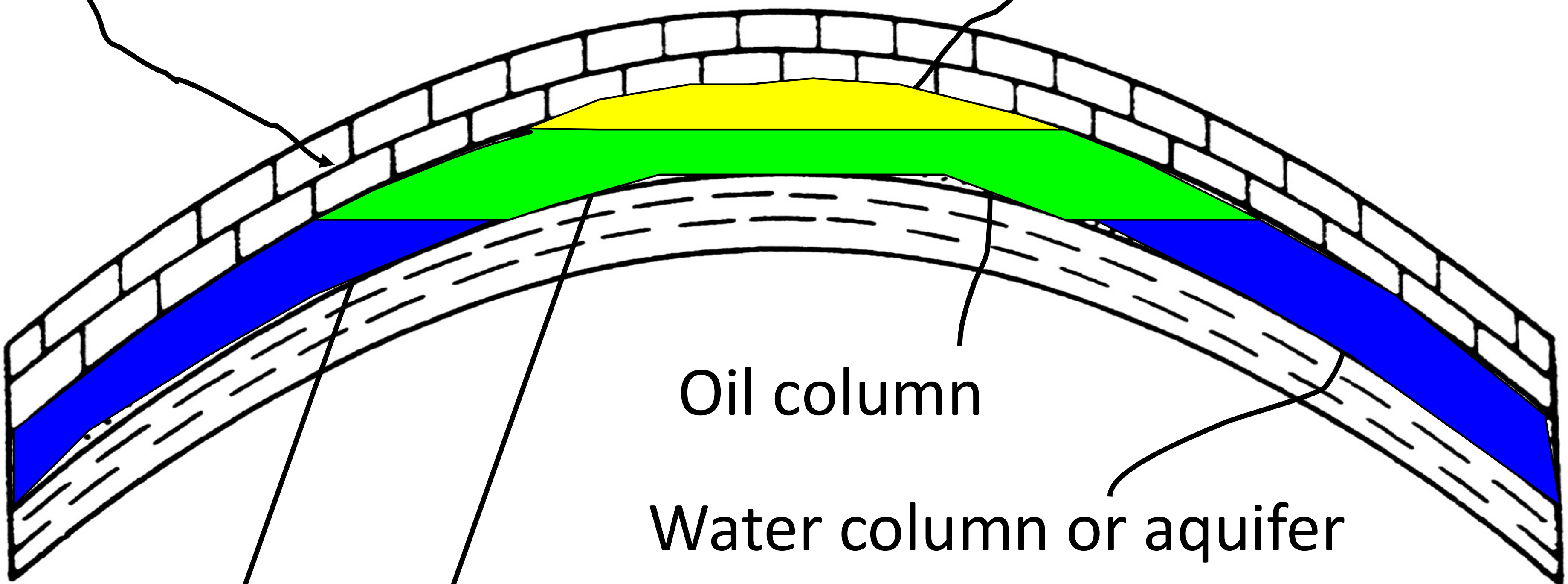
- Reservoir engineering is a sub-discipline that seeks to **determine** and **maximize** the ultimate value of a hydrocarbon, water or storage resource

- Petrophysics is concerned with the physical and chemical properties of rocks and how they interact with fluids

Reservoir Classification

- Location:
  Onshore, arctic, offshore, deepwater
- Predominant mineral type (lithology):
  Sandstone, carbonate, fractured
- Fluid types:
  Oil, gas, water
  Single-phase, two-phase, three-phase
- Type of trap

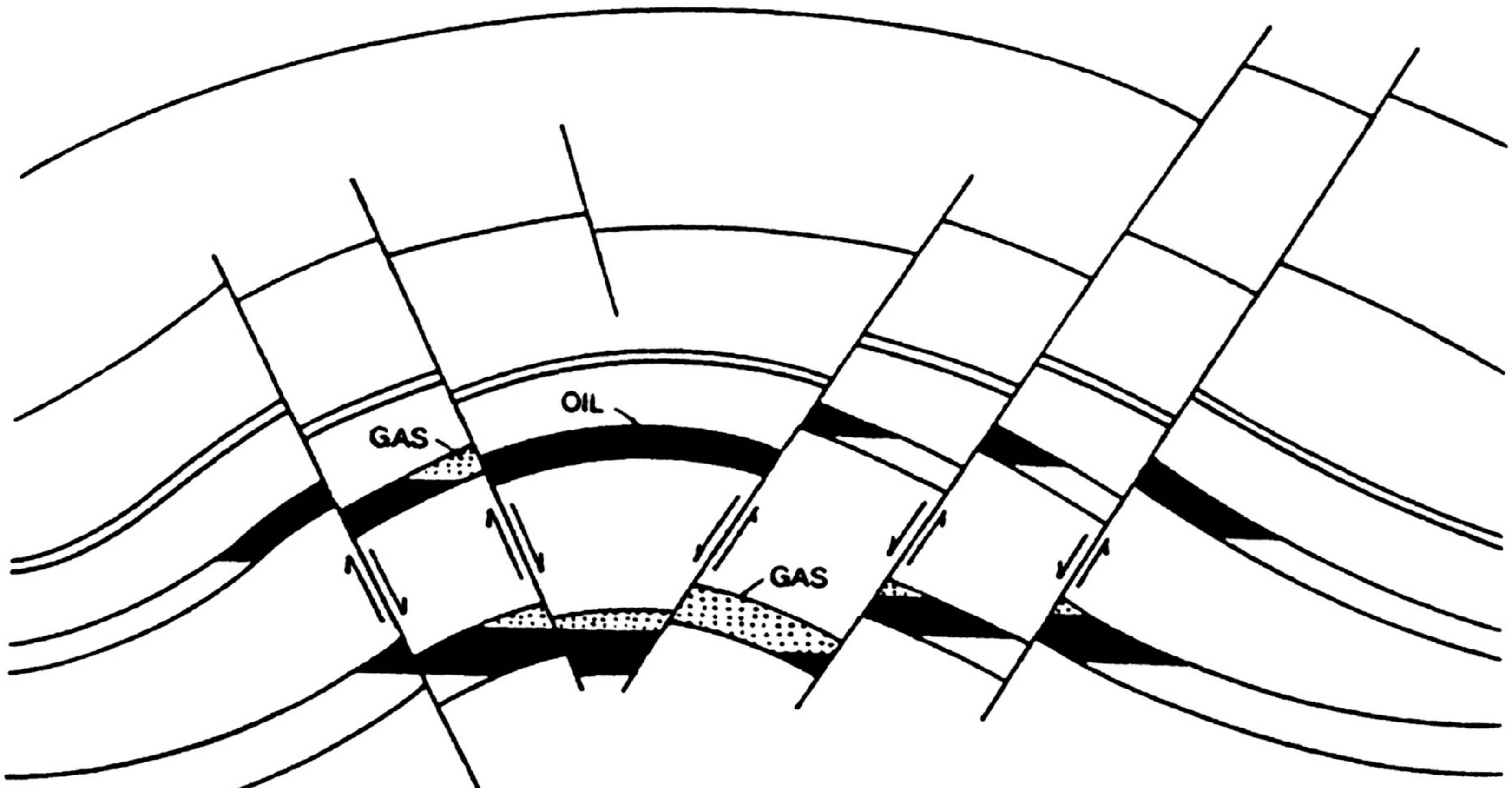
Generic Reservoir



GOC, WOC Change with time

Also: Original water-gas contact (OWGC)

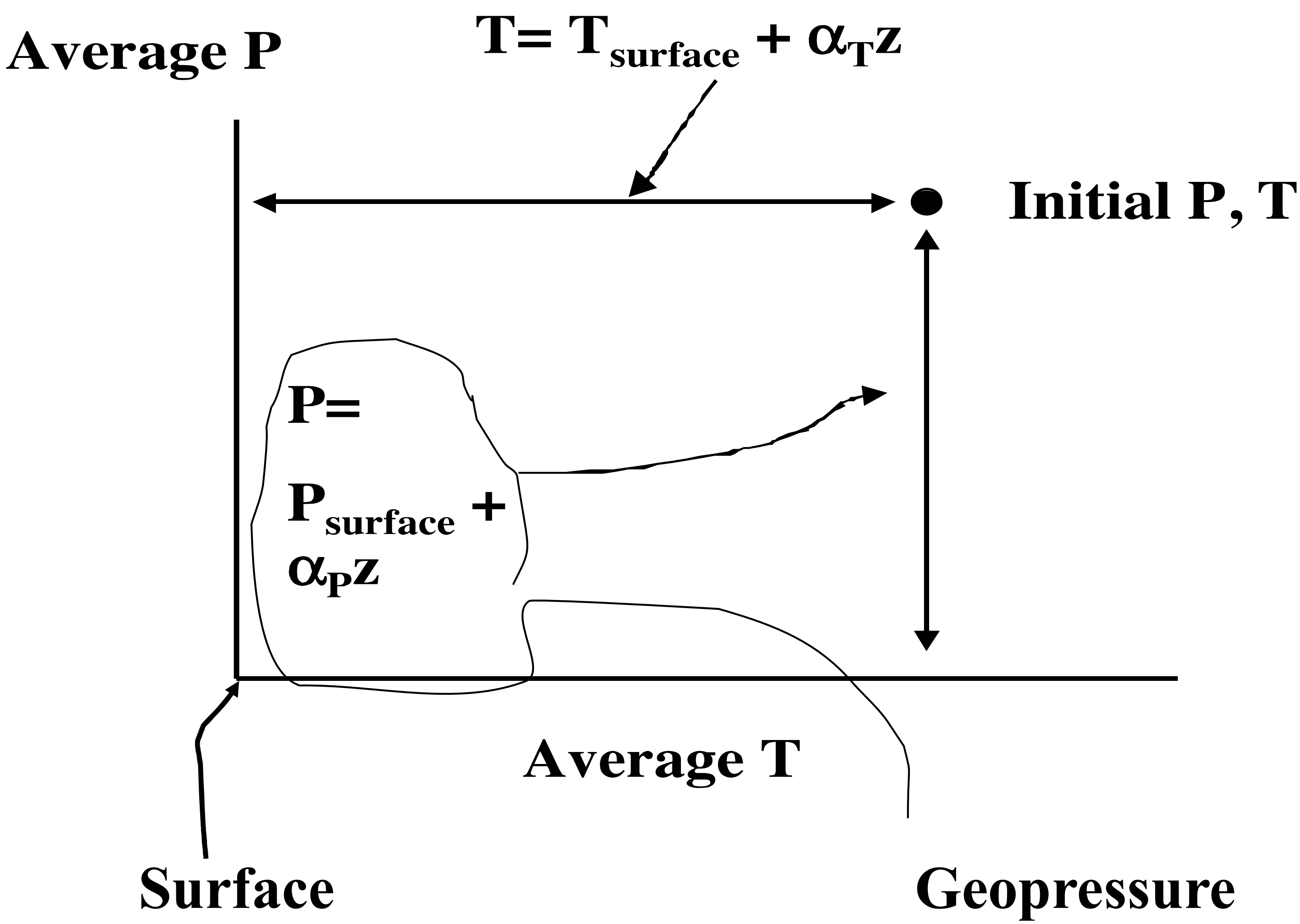
Trap type example



Faulted anticline

Ref. Walsh and Lake (WL)

Reservoir Pressure and Temperature



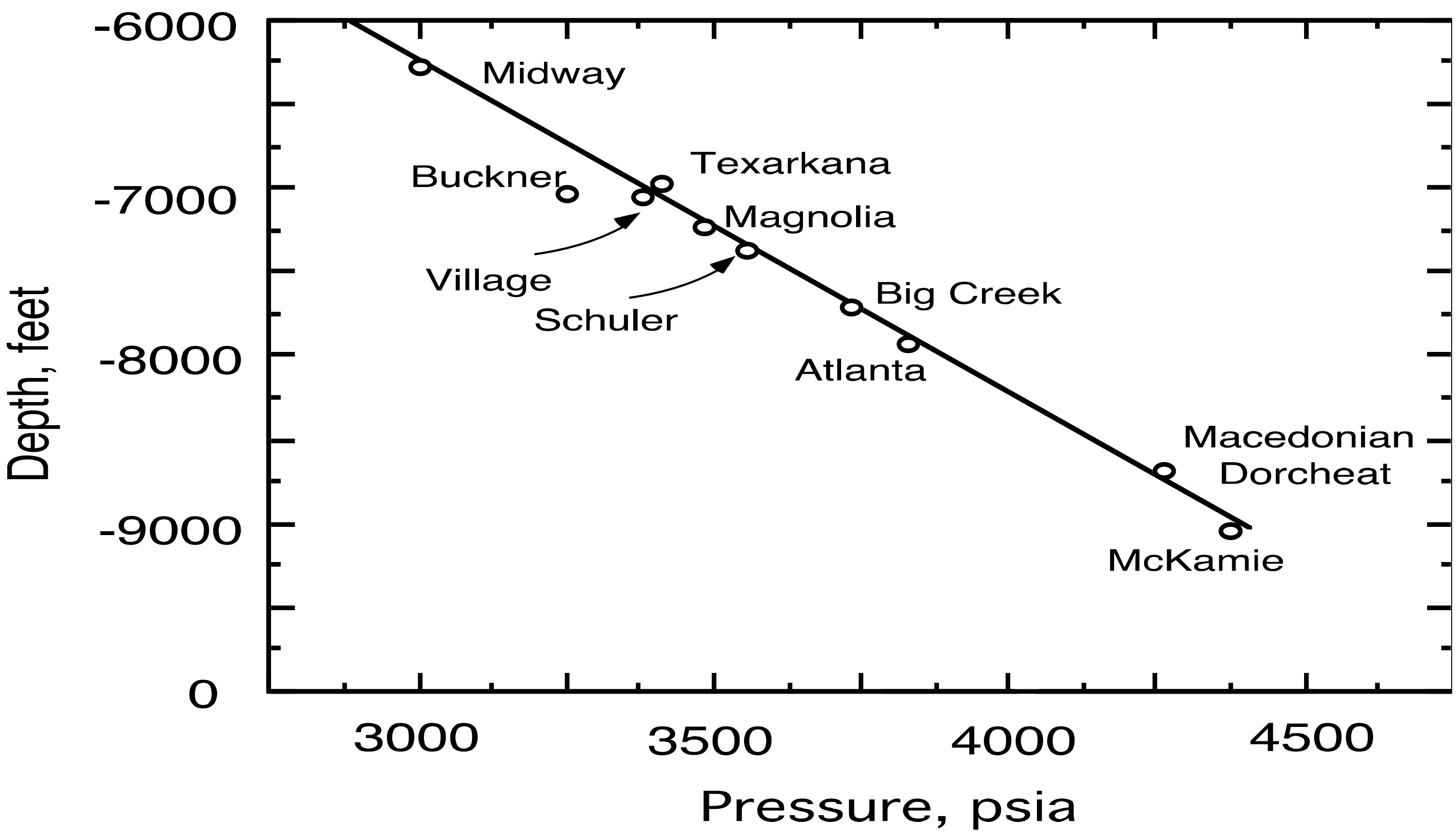
Ref. WL

Reservoir Pressure

z = subsurface depth

α_P = geopressure gradient

(0.433 psi/ft fresh water; 0.465 psi/ft brine)



Ref. WL

Reservoir Temperature

α_T = geothermal gradient (typical 0.01 F/ft)

Field	Geothermal Gradient $^oF/100$ ft
East Texas Woodbine	2.20
Burbank, OK	2.20
North Pettus, Bee County, TX	2.17
Leduc, Alberta, Canada	2.10
Fort St. John, British Columbia	1.81
Deep Lake, La.	1.15
Oklahoma City, OK	1.14
Hugoton, OK	0.84
Panhandle, TX	0.70
Monument, New Mex.	0.60

Ref. WL

Quick overview of fundamental properties in petroleum engineering

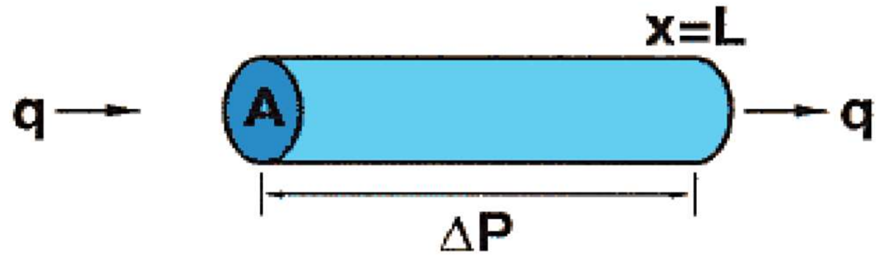
Porosity: A Static Petrophysical Property

$$\phi = \frac{\text{Interconnected pore volume}}{\text{Bulk volume}}$$

- Depends on packing and pore size distribution
- Depends weakly on grain size
- Determined by sedimentation and diagenesis
- Question: what is a common value for porosity?

Darcy's law

$$q = \frac{kA}{\mu}\frac{\Delta P}{L}$$



- q: flow rate
- k: permeability
- A: cross-sectional area
- L: length
- μ: viscosity
- Δpotential: $P_A - P_B$ (horizontal)
- Δpotential: $(P+\rho gh)_A - (P+\rho gh)_B$ (tilted)

Question: what is a common value for permeability?

Different unit systems in petroleum engineering

Table 1-5. Unit systems*

Quantity	Consistent Systems of Units: Darcy Units	mks Units	cgs Units	SI**	Oilfield Units
Area	cm^2	m^2	cm^2	km^2	acres
Compressibility	1/atm	1/pascal	1/μbar	1/kPa	1/psi
Density	g/cm^3	kg/m^3	g/cm^3	kg/m^3	lbm/ft^3
Flow rate (gas)	cm^3/s	m^3/s	cm^3/s	m^3/day	ft^3/day
Flow rate (liquid)	cm^3/s	m^3/s	cm^3/s	m^3/day	bbls/day
Force	N/A	newton (N)	dyne	mN	lb_f
Length	cm	m	cm	m	ft
Mass	g	kg	g	kg	lbm
Molar amount	gmole	kgmole	gmole	kgmole	lbmole
Permeability	darcy	m^2	cm^2	μm^2	md
Pressure	atm	pascal (Pa)	μbar	kPa	psi
Temperature	K	K	K	K, °C	R, °F
Time	s	s	s	days, yrs	days, yrs
Viscosity	cp	Pa-s	poise	Pa-s	cp
Velocity	cm/s	m/s	cm/s	m/s	ft/day
Volume (gas)	cm^3	m^3	cm^3	m^3	ft^3
Volume (liquid)	cm^3	m^3	cm^3	m^3	bbls
Volume (liquid)	cm^3	m^3	cm^3	m^3	bbls

* Notes: A pascal is a newton/m^2; a newton is a kg-m/s^2; a dyne is a g-cm/s^2; a poise is a g/cm-s; a μbar is a dyne/cm^2; a bar is 10^6 dyne/cm^2.

** SPE-approved SI units

Ref: WL

Units

- You should be able to convert common properties from one unit system to another (see the previous Table)

- Sample conversions

$$1 \;\; \mathrm{D} \cong 10^{-12} \mathrm{m}^2 = 1 \;\; \mu\mathrm{m}^2$$

$$1 \;\; \mathrm{ft} = 0.305 \;\; \mathrm{m}$$

$$1 \;\; \mathrm{bbl} = 5.614 \;\; \mathrm{ft}^3$$

$$1 \;\; \mathrm{acre} = 43560 \;\; \mathrm{ft}^2$$

$$14.7 \mathrm{psi} = 0.1 \;\; \mathrm{MPa}$$

$$1 \;\; \mathrm{lb}_{\mathrm{m}} = 0.454 \;\; \mathrm{kg}$$

$$460 \;\; \mathrm{R} = 273 \;\; \mathrm{K}$$

Two examples for unit conversion

α_T = geothermal gradient (0.01 F/ft)

$$\left(0.01\frac{\mathbf{F}}{\mathbf{ft}}\right)\left\{\frac{1\,\mathbf{R}}{1\,\mathbf{F}}\right\}\left\{\frac{273\,\mathbf{K}}{460\,\mathbf{R}}\right\}\left\{\frac{1\,\mathbf{C}}{1\,\mathbf{K}}\right\}\left\{\frac{\mathbf{1\ ft}}{\mathbf{0.305\ m}}\right\} = \mathbf{0.019}\ \frac{\mathbf{C}}{\mathbf{m}}$$

α_P = geopressure gradient (0.433 psi/ft)

$$\left(\mathbf{0.433}\ \frac{\mathbf{psi}}{\mathbf{ft}}\right)\left\{\frac{\mathbf{0.1\ MPa}}{\mathbf{14.7\ psi}}\right\}\left\{\frac{\mathbf{1\ ft}}{\mathbf{0.305\ m}}\right\} = \mathbf{0.0097}\ \frac{\mathbf{MPa}}{\mathbf{m}}$$

$$= \mathbf{9.7}\ \frac{\mathbf{kPa}}{\mathbf{m}}$$

Discussion

- You are tasked to improve the performance of an AI model. What would you do?

Questions on neural nets

- What is a shallow neural network?

Assignment 1

- Read, sign the acknowledgment page of the syllabus, and upload it.
- Go to Google Collab, PyTorch, or any other environment you prefer to access the Jupyter Notebook. Upload a screenshot. We will use Jupyter to practice visualization, and some basic functions needed in this course in the next class.

- Read the Allan Turing paper posted on Canvas (no submission is required)