Automated Data Collection and Compression System for CO2 Monitoring (DE-SC0019854)

Hamed Soroush, PI

Dr. Julian Yao, Dr Salah Faroughhi, Dr Ali Payani, Dr Kamal Shadi, Dr Mohamad Zamini, Petrolern

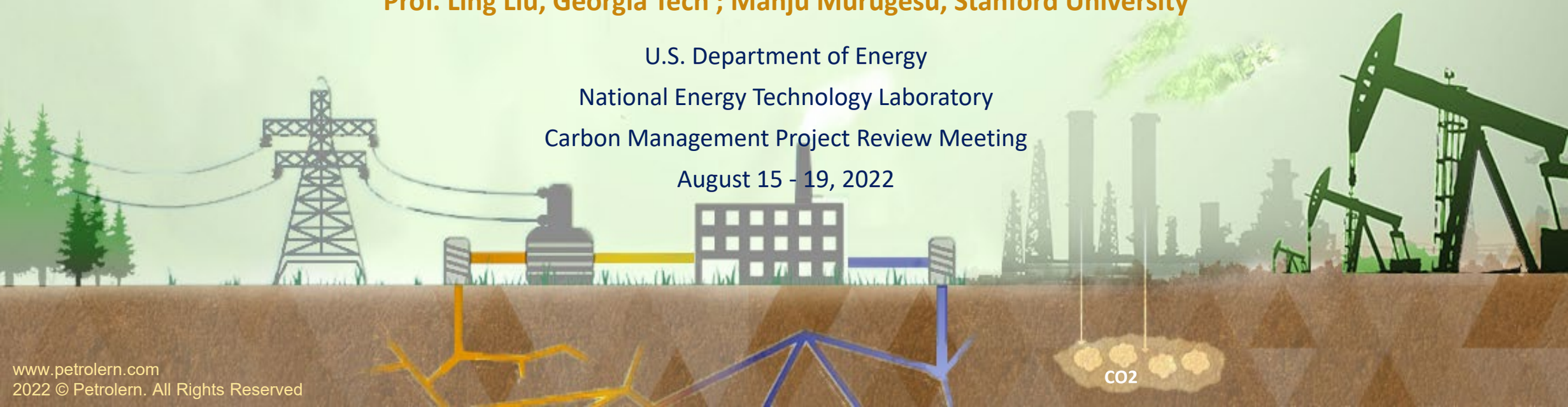
Prof. Ling Liu, Georgia Tech ; Manju Murugesu, Stanford University

U.S. Department of Energy

National Energy Technology Laboratory

Carbon Management Project Review Meeting

August 15 - 19, 2022



Outline

1. Problem Statement
Real-time access to big monitoring datasets

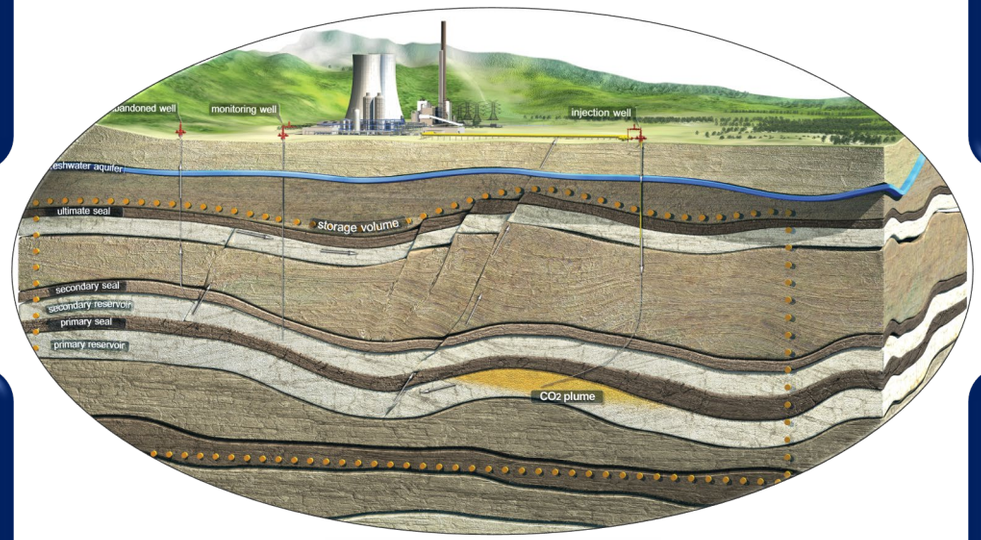
2. Objective
Multi-modal composite data compression

3. Compression Techniques
- Time series-based
- Model-based

4. Understand Data Type
- Floating points
- Integers
- Number of bytes

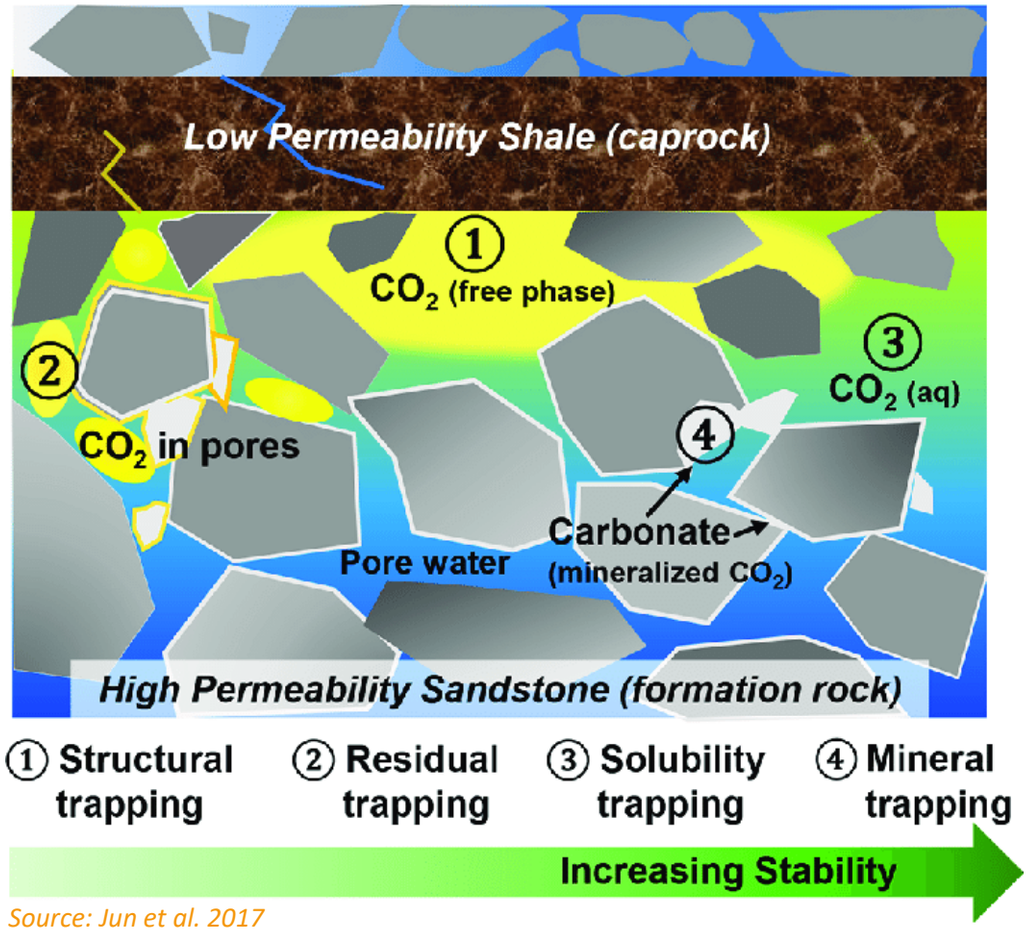
5. Applied Methodologies
- Lossy vs. Lossless
- Tabular data compression

6. Developed Workflow
- Results
- Next steps



PETROLERN

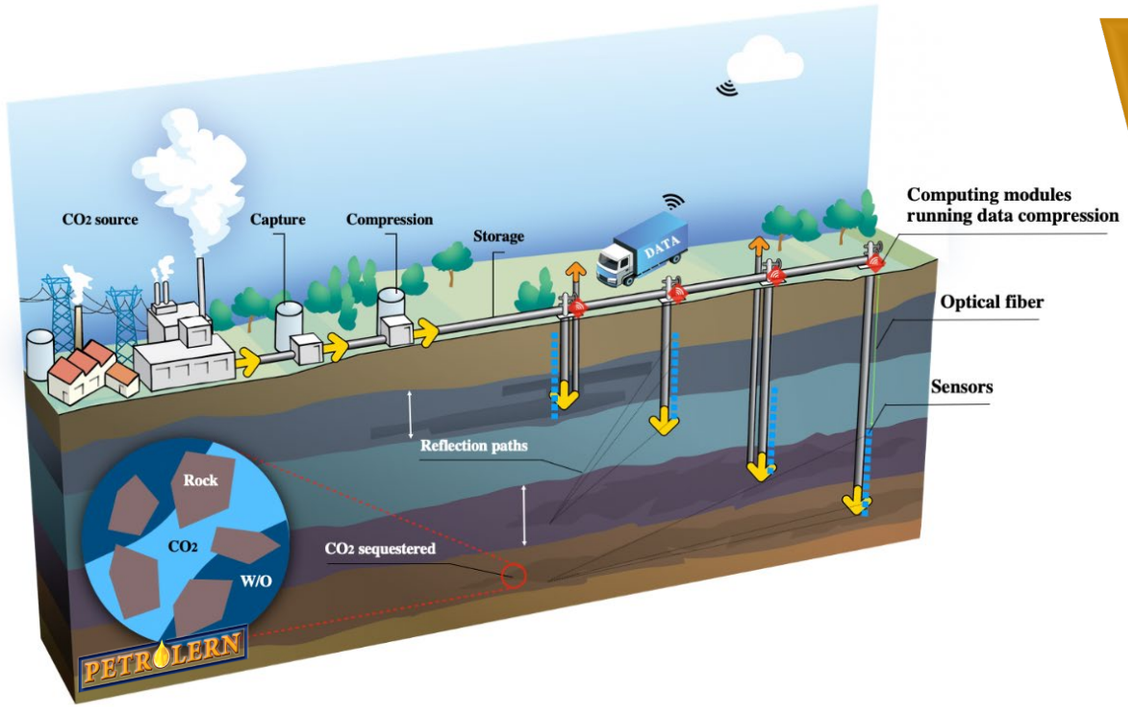
Real-time CO_2 Monitoring is Important



Source: Jun et al. 2017

- To identify CO_2 leakage pathways:
  - Movement to the shale formations
  - Through faults and natural fractures
- To understand kinetics of long-term impact of CO_2 on reservoir
- To improve reservoir stimulation processes in real-time
- To understand long-term impact of geochemical and geomechanical alterations due to CO_2

Data Transmission and Handling is the Bottleneck



100 %

1-6 TB of data per day
≈ 30,000 sensors (pressure, temperature, acoustic)

Data Capture

≈ 40% of all data is never stored; Remainder is stored locally at offshore

Challenge

No interface to handle the large data and enable real-time analytics

Infrastructure:
Data transfer speed limitation

Only ≈ 1% of data is streamed onshore for daily use.

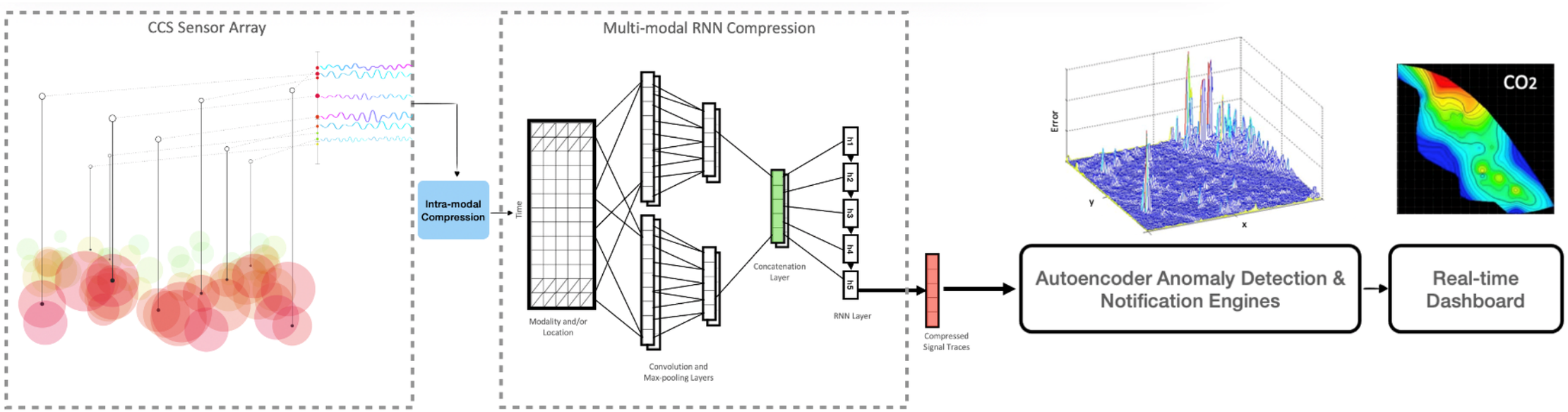
<1%

Of data used for decision making

Adapted and modified from Cisco

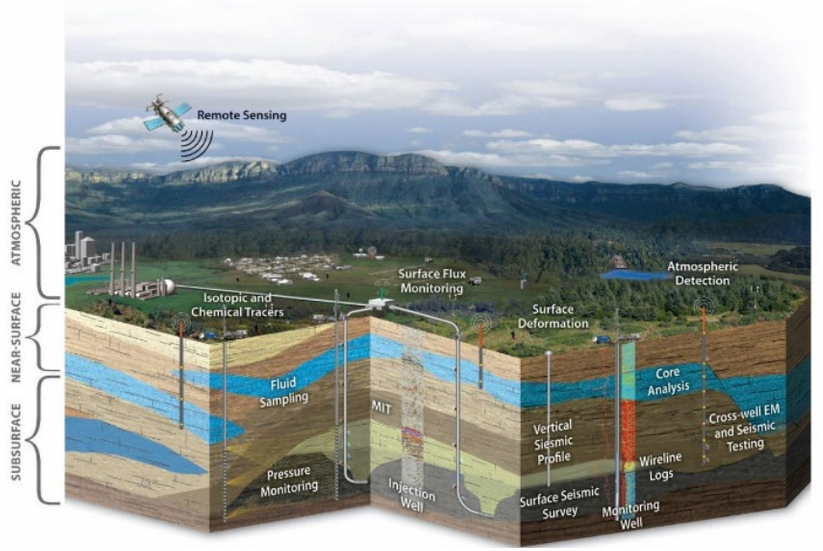
Project Objective

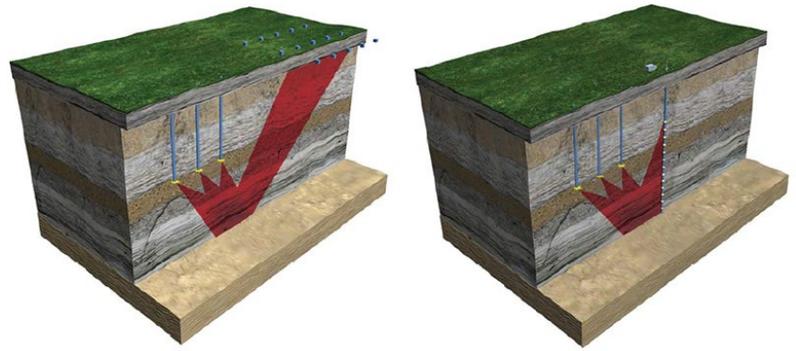
- Develop an **automated data compression system** to address the issues of high latency, inadequate bandwidth, and limited storage
- Account for the interdependency between data using a multimodal compression method



Value Proposition

- Enabling large scale data collection required for CO_2 MVA.
- Enabling efficient and real-time decision making.
- Reducing the cost of data transmission, processing and storage.

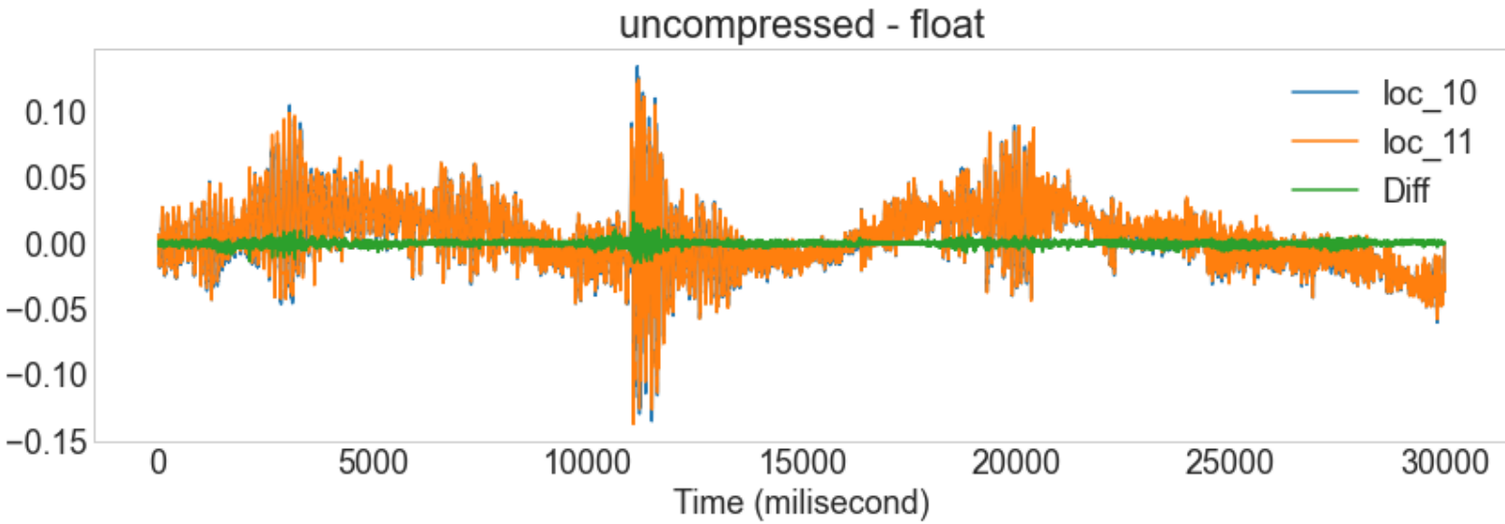




https://www.netl.doe.gov/coal/carbon-storage/advanced-storage-r-d/monitoring-verification-accounting-and-assessment

Compression Technology Gap

- Real-valued time series data is non-stationary and float (over large alphabet usually 32 bits)
- Traditional compression methods are not suitable:
  - Correlations across parallel sensors are not leveraged
  - They usually do not offer a trade-off between compression gain and error rate.



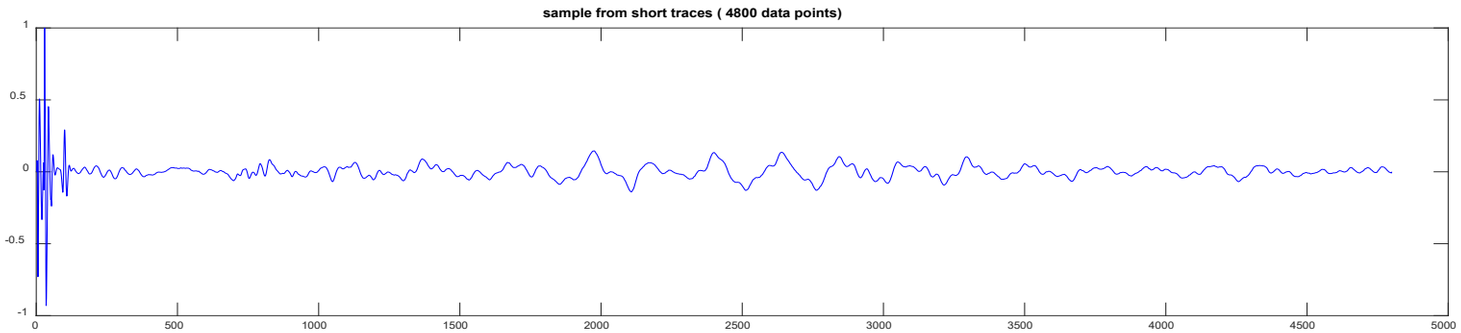
Methodology

- First Level compression methods:
  - **Lossless:** reduces bits by identifying and eliminating statistical redundancy.
  - **Lossy:** reduces bits by removing less important information
- Second level compression methods
  - **Predictive Coding:** use a predictive filter to remove temporal correlation, and then compress the residual using a lossy/lossless universal coding
    - **Pros:** Show good performance for low noise signals and are lightweight
    - **Cons:** Do not learn from the past
  - **Model-based Learning:** represent data using well-established approximation models (e.g., Neural Network)
    - **Pros:** Learn the statistics of signal and usually higher compression gain
    - **Cons:** Memory intensive and have initial learning cost

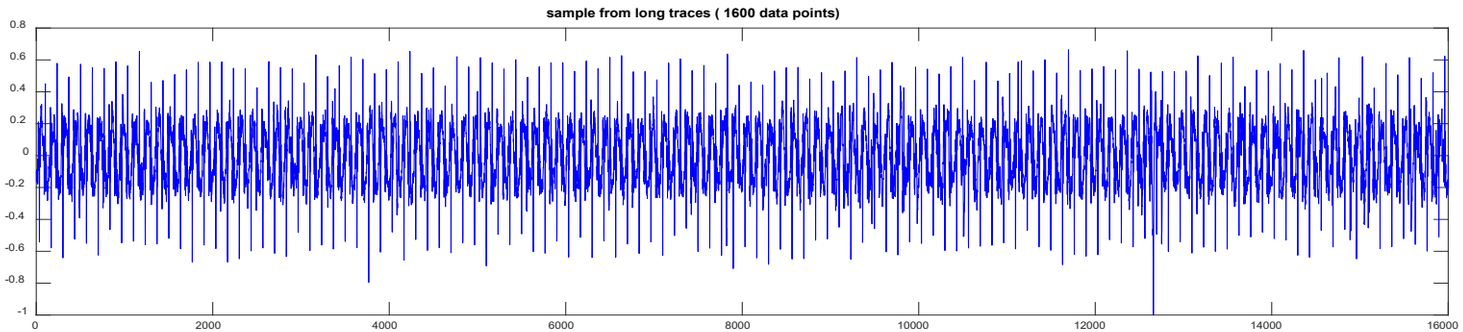
We started with SEGY and moved to the Multimodal Fiber Optic Data

Development on SEGY Data

- SEG-2 seismic datasets acquired in the CASSM experiment (Frio-2 GCS pilot)
- All files correspond to the single source continuous active source seismic monitoring
- Includes:
  - 6953 short traces with 4800 data points per trace

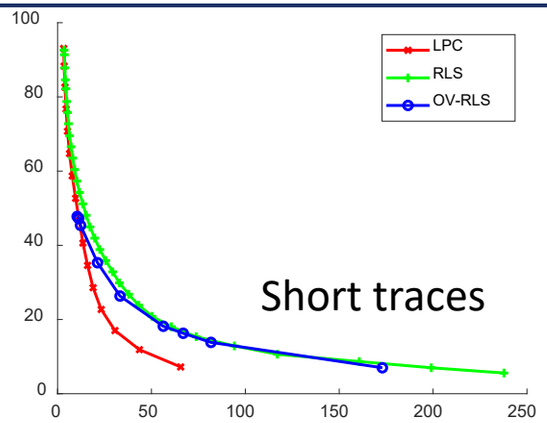


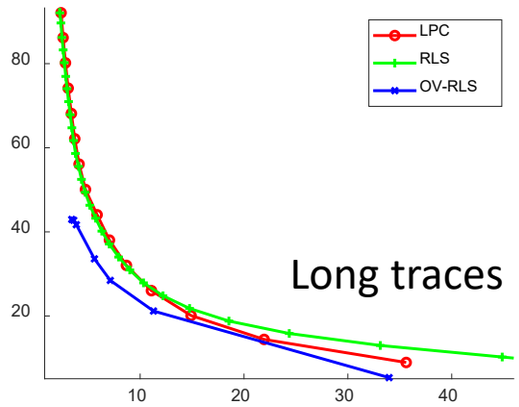
  - 1513 long traces with 16000 data points per trace



PETROLERN

Results on SEGY Data





Compression gain for short traces (ST) and long traces (LT).

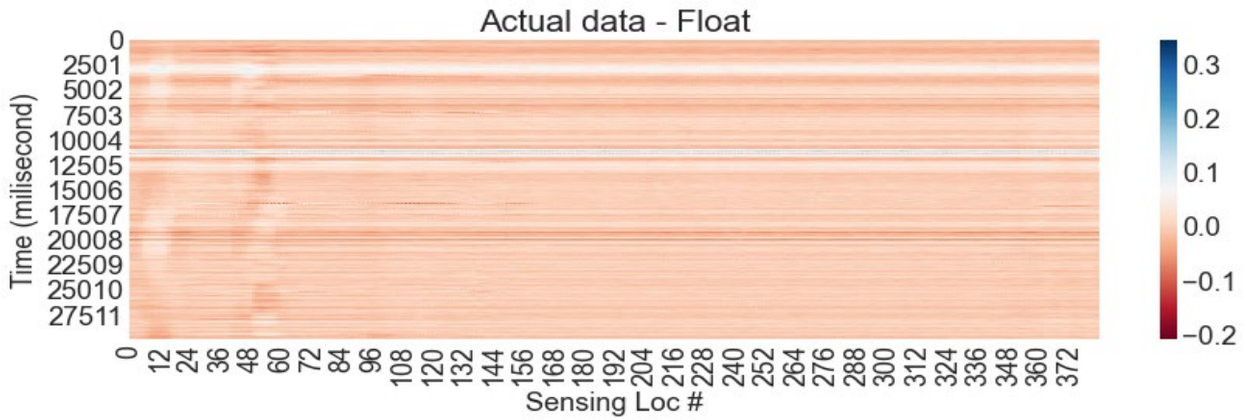
Methods	SNR 20 dB		SNR 30 dB		SNR 40 dB		SNR 50 dB	
	ST	LT	ST	LT	ST	LT	ST	LT
LPC+CTW	27	15	17	9	12	6	15	5
OV+RLS+CTW	53	13	28	7	17	4	8	2
RLS+CTW	54	17	32	9	20	6	15	5
Multi Trace RLS + CTW	55	24	33	13	22	7	18	6

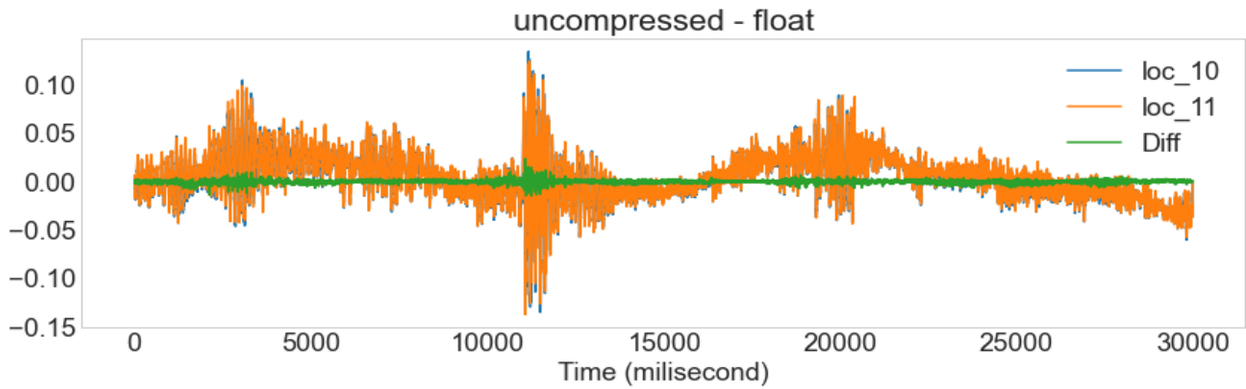
- Predictive Coding via Linear Predictive Filter + CTW
- Predictive Coding via RLS Predictive Filter + CTW
- Multi-Trace Predictive Coding via RLS predictive filter + CTW
- Oversampling + RLS + CTW **(Context Tree Weighting)**

Development on Multimodal Fiber Optic Data (DAS & DTS)

Brady Geothermal Field: (i) Vertical borehole of 384 meters data from 17-26 March 2016

(ii) Horizontal Trenched cable of 8721 meters data from 11-26 March 2016





DAS

Das / strain	384 x 30000
time	30000
trace	30000
x	384
y	384
z	384

Vertical: 1 TB
Horizontal: 50 TB

DTS

Raman stokes and anti-stokes

Temperature

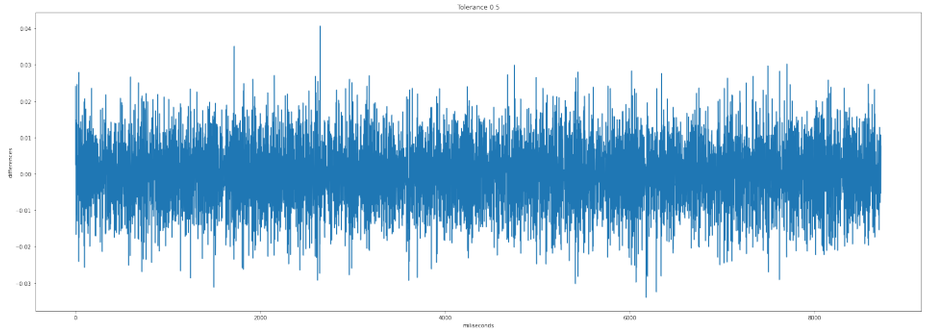
Vertical: 1 GB
Horizontal: 500 MB

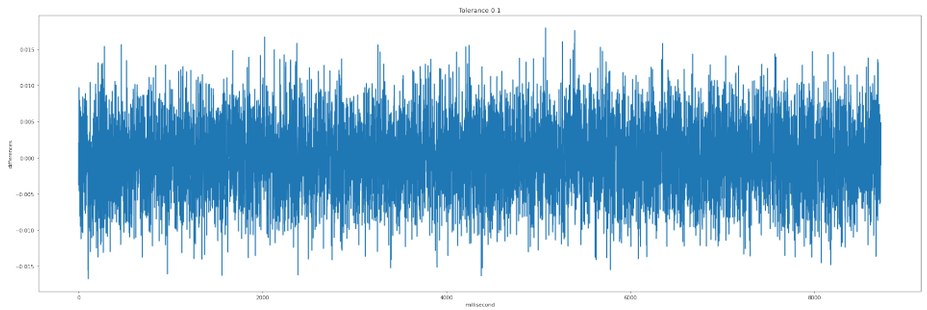
Tested Algorithms

- Lossless algorithms:
  - Lempel–Ziv–Welch (LZW): Adaptive dictionary algorithm – CR = 1.6X
  - Zstandard: Real-time dictionary compression algorithm – CR = 1.1X
  - Brotli: Series of meta blocks – CR = 1.9X
  - G-zip: Arithmetic Encoder, No cross-location dependencies – CR = 2.7X
- Lossy algorithms:
  - Autoencoder: Not suitable
  - Modified Chow Liu Tree: Account for both temporal and spatial dependencies based on probabilistic assumptions – CR = 24X
  - Modified Fixed Rate Near Lossless Compression: Maps small blocks of 4^d values in d dimensions – CG = 53X

Fixed Rate Near Lossless Compression

- Developed by Lindstrom, P. (2014) for floating data.
- Maps small blocks of 4^d values in d dimensions to a fixed, user-specified number of bits per block.
- We modified the algorithm to best suit DAS/DTS data:
  - with tolerance 0.5 ➡ RMS 0.009745346, CR= 53.
  - with tolerance 0.1 ➡ RMS 0.003955638, CR = 21.

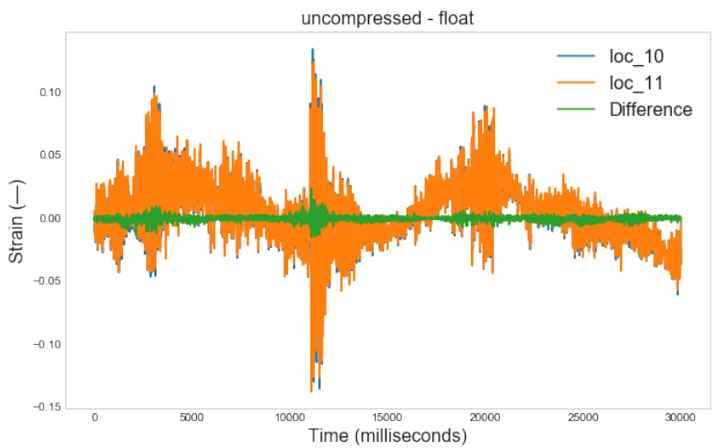


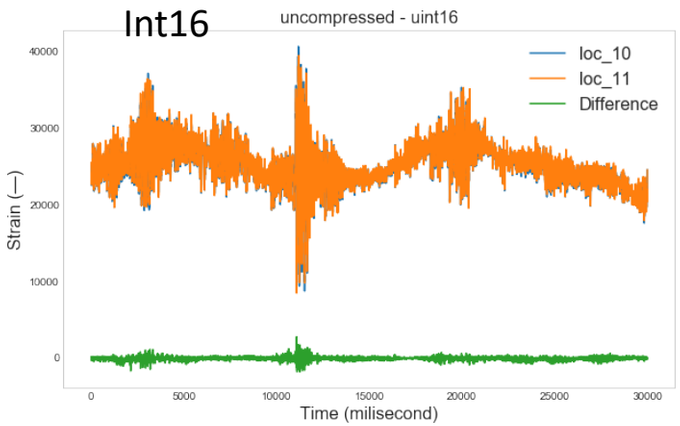


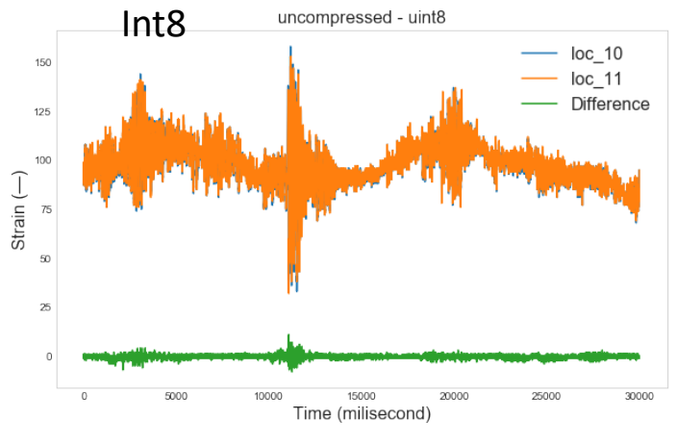
Tolerance	999 GB	Error	Gz+zfpy
0.5	18.92 GB	0.0097	6.81 GB
0.1	48.1 GB	0.0039	39 GB

Further Compression by Changing Integers

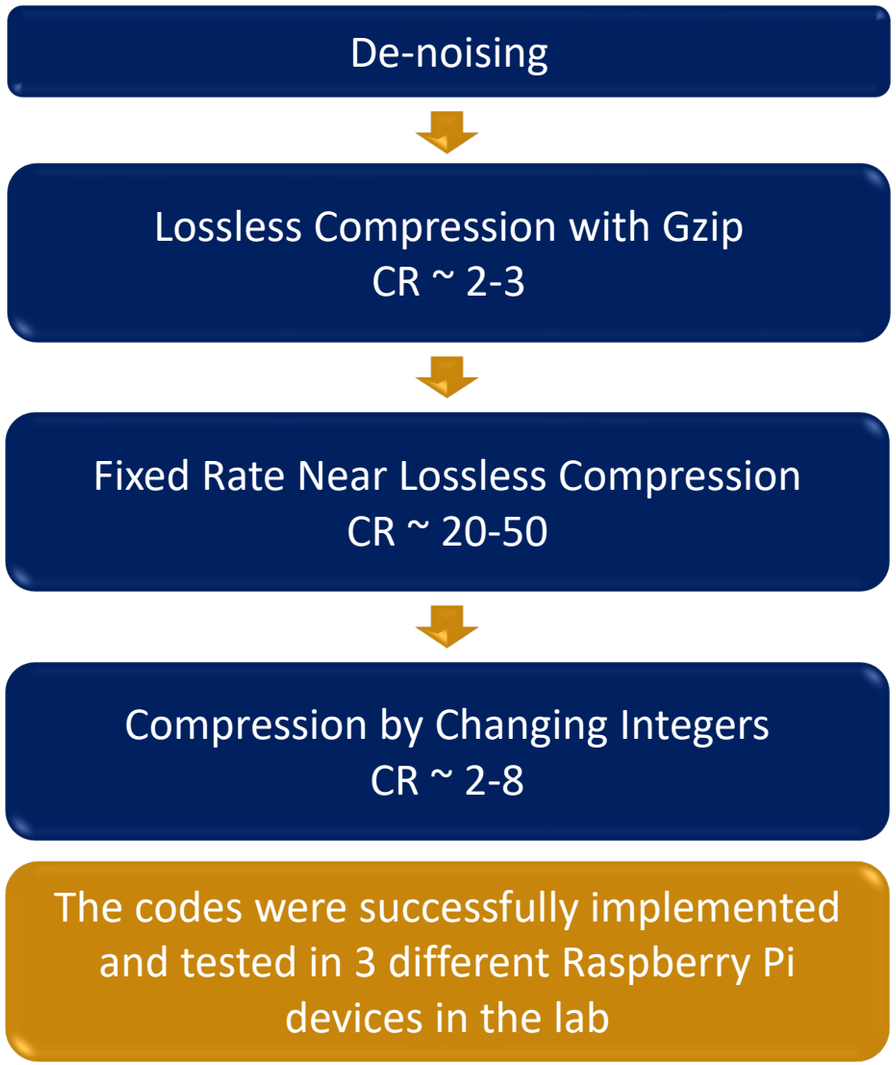
DAS Data (Float64)	999 GB	Error
Compressed with 0.1 Tolerance	48 GB	3.9e-3
Int32	25 GB	7.4e-11
Int16	13 GB	4.8e-06
Int8	6 GB	1.25e-3







Developed Workflow



Next Steps

- Test the workflow with a few more field data to increase generality
- Test the quality of the monitoring results before and after processing
- Develop a software package to automate data collection, compression and transmission including our visualization software, GeoDeck^{TM}
- Real-time field testing
  - Partnered with New Mexico Tech
  - Field testing planned for two CCS projects:
    - Southwest Partnership (SWP) CCUS-EOR project - Farnsworth Unit (FWU) in the Anadarko Basin
    - San Juan Basin CarbonSAFE CCS project - San Juan County, New Mexico

PETROLERN

Summary and conclusion

- Different compression algorithms were tested.
- Most available compression models are developed for string data not floating point data.
- Mapping floating point values to smaller blocks of user-specified numbers lead to high compression rates:
  - Trade off between accuracy and compression rate.
- A Modified Fixed-Rate Compressed Floating-Point Arrays was developed specifically for multi-model fiber optic data that can gain over 50x compression rate.
- A combination of Fixed-Rate Compression with Arithmetic Encoder (Gzip) and changed integer can achieve even more compression rate.
- Algorithms were implemented and tested on edge devices.
- The workflow and codes are ready for field testing.

Innovation is our Passion

Questions?

info@petrolern.com