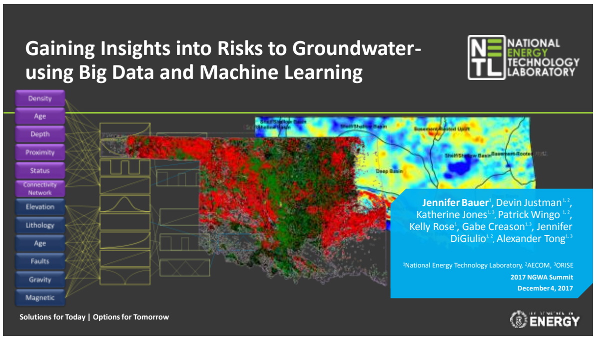


Gaining Insights into Risks to Groundwater- using Big Data and Machine Learning

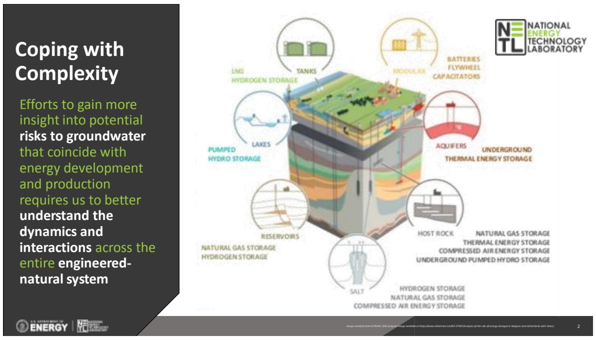
Jennifer Bauer[1], Devin Justman[1, 2], Katherine Jones[1, 3], Patrick Wingo[1, 2], Kelly Rose[1], Gabe Creason[1, 3], Jennifer DiGiulio[1, 2], Alexander Tong[1, 3]

[1]National Energy Technology Laboratory, [2]AECOM, [3]ORISE

2017 NGWA Summit

December 4, 2017

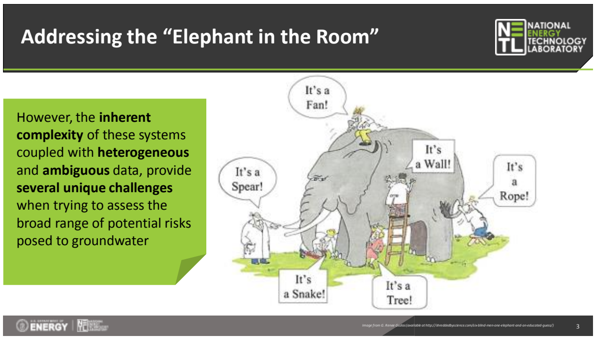
Solutions for Today | Options for Tomorrow



Coping with Complexity

Efforts to gain more insight into potential **risks to groundwater** that coincide with energy development and production requires us to better **understand the dynamics and interactions** across the entire **engineered-natural system**

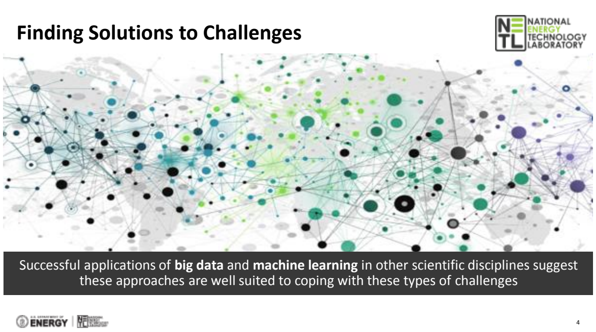
Image modified from ESTMAP, 2016 (original image available at https://www.slideshare.net/IEA-ETSAP/analysis-of-the-role-of-energy-storages-in-belgium-and-netherlands-with-times)



Addressing the "Elephant in the Room"

However, the **inherent complexity** of these systems coupled with **heterogeneous** and **ambiguous** data, provide **several unique challenges** when trying to assess the broad range of potential risks posed to groundwater

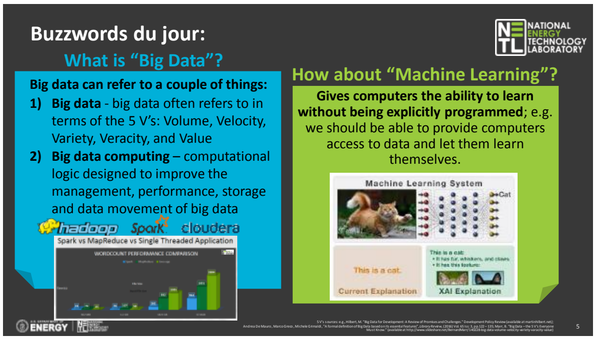
Image from G. Renee Guzlas (available at http://shreddedbyscience.com/six-blind-men-one-elephant-and-an-educated-guess/)



Finding Solutions to Challenges

Successful applications of **big data** and **machine learning** in other scientific disciplines suggest these approaches are well suited to coping with these types of challenges

---



Buzzwords du jour:

What is "Big Data"?

Big data can refer to a couple of things:

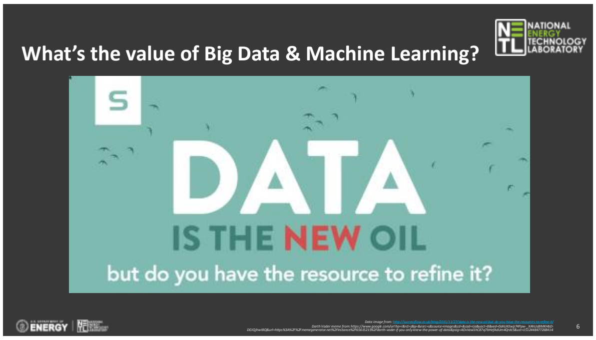
1) **Big data** - big data often refers to in terms of the 5 V's: Volume, Velocity, Variety, Veracity, and Value
2) **Big data computing** – computational logic designed to improve the management, performance, storage and data movement of big data

Spark vs MapReduce vs Single Threaded Application

How about "Machine Learning"?

Gives computers the ability to learn without being explicitly programmed; e.g. we should be able to provide computers access to data and let them learn themselves.

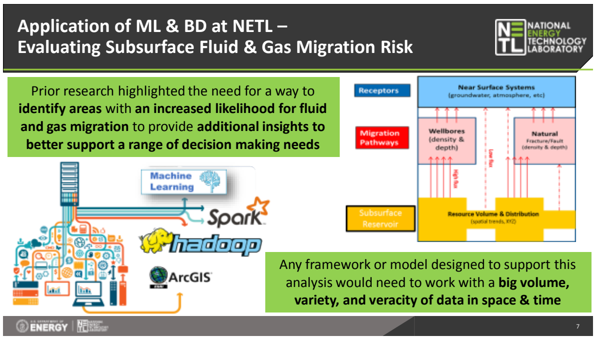
5 V's sources: e.g., Hilbert, M. "Big Data for Development: A Review of Promises and Challenges." Development Policy Review (available at martinhilbert.net); Andrea De Mauro, Marco Greco, Michele Grimaldi, "A formal definition of Big Data based on its essential features", *Library Review*, (2016) Vol. 65 Iss: 3, pp.122 – 135; Marr, B. "Big Data – the 5 V's Everyone Must Know." (available at http://www.slideshare.net/BernardMarr/140228-big-data-volume-velocity-variety-varacity-value)



What's the value of Big Data & Machine Learning?

Data image from: http://successflow.co.uk/blog/2015/11/27/data-is-the-new-oil-but-do-you-have-the-resources-to-refine-it/

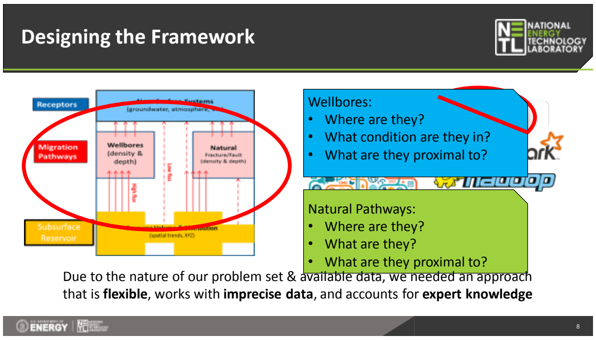
Darth Vader meme from: https://www.google.com/url?sa=i&rct=j&q=&esrc=s&source=images&cd=&cad=rja&uact=8&ved=0ahUKEwjz74Ppw-_XAhUsBlMKHbD-DGIQjhwIBQ&url=https%3A%2F%2Fmemegenerator.net%2Finstance%2F65615215%2Fdarth-vader-if-you-only-knew-the-power-of-data&psig=AOvVaw1hC87vj7bHefAdUm4Qrdc5&ust=1512448477268414



Application of ML & BD at NETL – Evaluating Subsurface Fluid & Gas Migration Risk

Prior research highlighted the need for a way to **identify areas** with **an increased likelihood for fluid and gas migration** to provide **additional insights to better support a range of decision making needs**

Any framework or model designed to support this analysis would need to work with a **big volume, variety, and veracity of data in space & time**

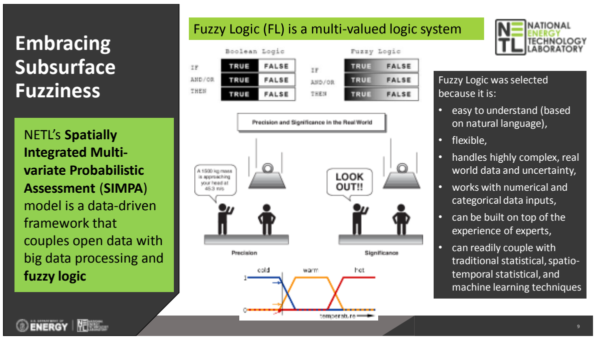


Designing the Framework

Wellbores:
- Where are they?
- What condition are they in?
- What are they proximal to?

Natural Pathways:
- Where are they?
- What are they?
- What are they proximal to?

Due to the nature of our problem set & available data, we needed an approach that is **flexible**, works with **imprecise data**, and accounts for **expert knowledge**

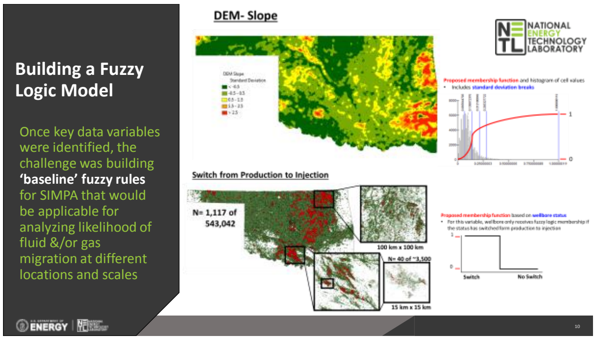


Embracing Subsurface Fuzziness

NETL's **Spatially Integrated Multi-variate Probabilistic Assessment (SIMPA)** model is a data-driven framework that couples open data with big data processing and **fuzzy logic**

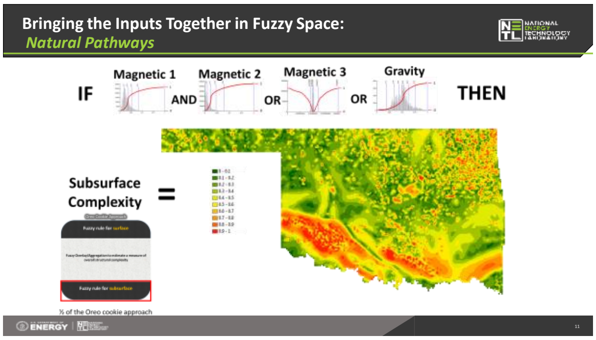
Fuzzy Logic (FL) is a multi-valued logic system

Fuzzy Logic was selected because it is:
- easy to understand (based on natural language),
- flexible,
- handles highly complex, real world data and uncertainty,
- works with numerical and categorical data inputs,
- can be built on top of the experience of experts,
- can readily couple with traditional statistical, spatio-temporal statistical, and machine learning techniques



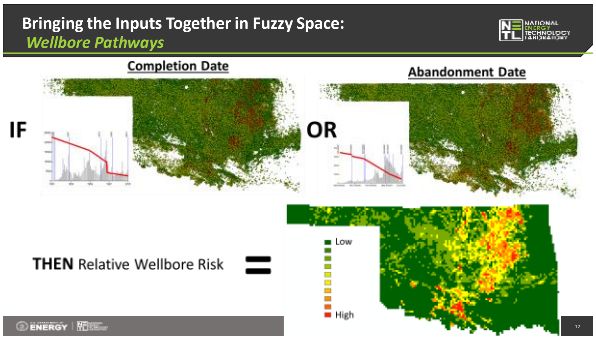
Building a Fuzzy Logic Model

Once key data variables were identified, the challenge was building **'baseline' fuzzy rules** for SIMPA that would be applicable for analyzing likelihood of fluid &/or gas migration at different locations and scales

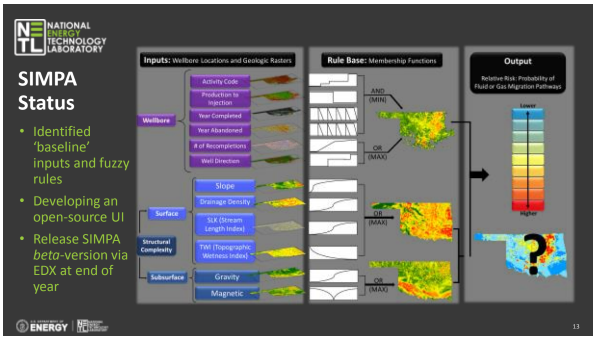


Bringing the Inputs Together in Fuzzy Space: *Natural Pathways*

½ of the Oreo cookie approach

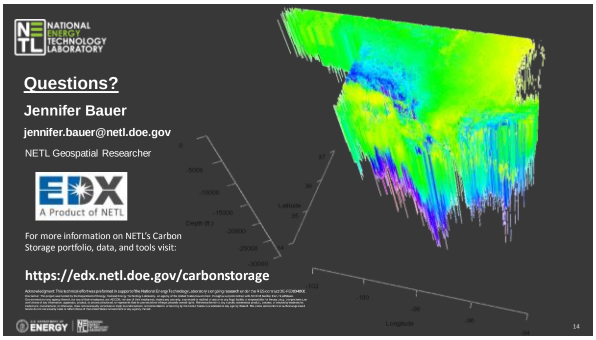


Bringing the Inputs Together in Fuzzy Space: *Wellbore Pathways*



SIMPA Status

- Identified 'baseline' inputs and fuzzy rules
- Developing an open-source UI
- Release SIMPA *beta*-version via EDX at end of year



Questions?

Jennifer Bauer

jennifer.bauer@netl.doe.gov

NETL Geospatial Researcher

For more information on NETL's Carbon Storage portfolio, data, and tools visit:

https://edx.netl.doe.gov/carbonstorage

Acknowledgment: This technical effort was preformed in support of the National Energy Technology Laboratory's ongoing research under the RES contract DE-FE0004000.