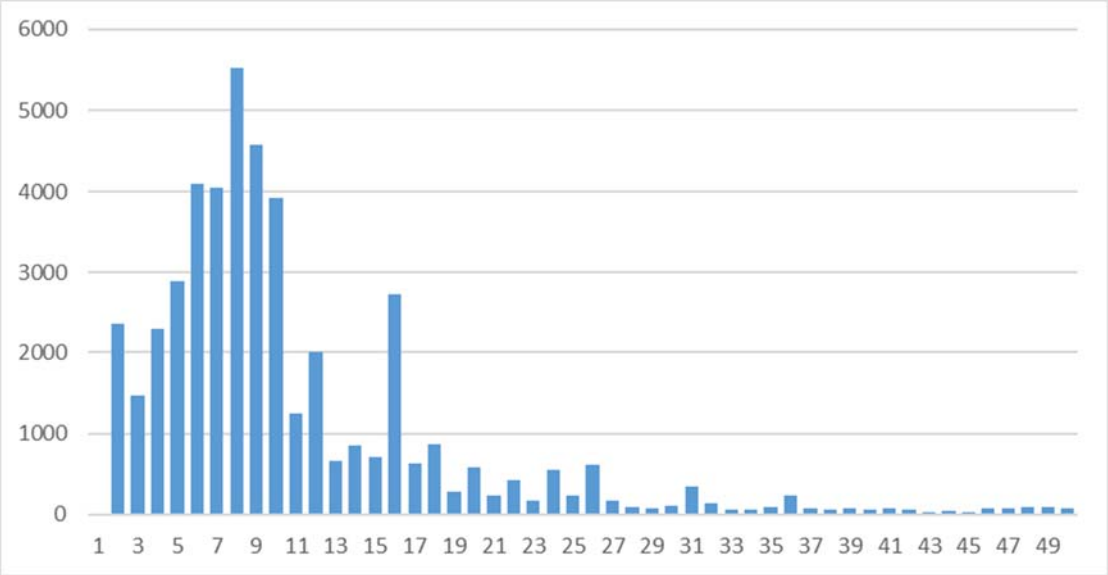
HYPOTHESIS I believe there is a game changing opportunity to build a machine model that analyzes LAS files, using engineer-selected perforation zones (as labels), to build correlations between log curves and the presence of hydrocarbons. Once this service is created, I believe it will identify poor engineer performance and recommend better perforation zones in existing wellbores (yellow/brown fields). Finally, I believe that humans will stop reading and blocking logs for new wellbores (green fields) and turn the entire perforation selection process over to computers.

Note: As you go through this analysis, if you wish to see the supporting data/queries/information, there are many blue hyperlinks (hold control key while clicking the link)

We studied 48,386 distinct LAS log files in North Dakota. These LAS files contain 800,956,027 rows of log data with a total of 6,84,251 curves. We believe ~306,918,771 feet of wellbore are logged in the LAS files (after anomalies are removed). If you want a really big number, there are 12,947,878,106 curve values in these 48,386 LAS files.

The average number of curves per LAS file was 14.14 with a standard deviation of 21.22 (distribution skewed). Can see it better in this chart.



Number of curves is x-axis (capped at 50), count of how many LAS files have X curves is the y-axis. Here is a fun fact. There are 118 LAS files with over 200 curves in the file. In API# 3310503027, there are 220 curves in a single Circumferential Borehole Imaging Log. For grins, try finding the API number or NDIC number (25241) in this LAS file. It's not there. Many servicers do not bother putting any well identifier inside the actual log file. We have to keep an audit trail of what scout tickets contain what LAS files.

200 curves too extreme? 2,163 (4.47%) LAS files have more than 50 curves. 5,033 (1 in 10) LAS files have more than 25 curves.

Depths

The craziness is just getting started…

```
..
 STRT.F              12654:
 STOP.F            -100000:
 STEP.F              -0.5000:
```

Is the 100,000 foot stop depth an obfuscation? We can't ask the service company, as they are not listed in this LAS file.

In this one, we have a start depth of 12,441 and a stop depth of 5 feet.

```
~Well Information Block
STRT.FT             12441.0000: START DEPTH
STOP.FT                 5.0000: STOP DEPTH
STEP.FT                -0.5000: STEP
```

Yet, in this one, we have a start of 11,150 and a stop of 21,596. So, there seems to be two ways to do start depth A) depth closest to the surface and B) depth at the lowest point we logged.

```
STRT       .ft          11150
STOP       .ft          21596
STEP       .ft          1
```

A few appear to measure in meters? Others appear to measure in seconds? Here is a LAS file that only logged 140 feet. Here is an example of where we started logging at -179,000 feet and went to +10,550 feet in increments of 50 feet.

Removing the meters, seconds, 100K stop depths… actually, let's just assume that 13K is a limit on log length (just for this analysis) and use absolute value to account for start being top or bottom of hole.

This gives us 47,187 log files that match our 13K filter. The average length logged is 6,504 feet. The average step value is 1.03 feet.

Even in this 13K length example, there are 54 different step values. But .5 feet and 1 foot are clearly the most popular step values (chart here).

GUTCHECK Logs make the Frac Focus database look like a Mona Lisa. I suspect logging is a free-for-all… truck-by-truck, no standards, no consistency even within a service company. Only standard is the way the LAS has to be presented (e.g. LAS 3.0). I also believe that no one has ever tackled an aggregate analysis of LAS files. And I don't want to even think about whether service companies are all (consistently) using subsea depths in these LAS files!

Service Company

There are 311 (recorded) service companies that performed log services within the 48K LAS files. Schlumberger is clearly the leader with 11,635 LAS files generated. Looks like Schlumberger did most of their work for Continental (no surprise). This is a good time to point out that all LAS metadata needs to be cleaned and curated… company names, MNEMs, log descriptions, etc. For example, Schlumberger has 16 different versions of Continental Resources.

26,709 (55.02%) LAS files leave the service company name blank.

Let's pick on Weatherford with 2,226 LAS files (accounts for different spellings of Weatherford). There are 60,854 total curves.

Weatherford is nice enough to provide us with a chart book, although the latest I could find is 2009. Weatherford appears to have 922 unique descriptions for log curves. I have to assume many of these are spelling variations? When we analyze by MNEM, it looks like a free-for-all, with the number jumping to 3,231 unique mnemonics. A reason for this delta is the extra data Weatherford will take into a mnemonic. What's more interesting is how this single service company has 22 ways to describe a gamma ray. Maybe it's a truck-by-truck decision on whether to use GRGC or GR as the MNEM.

You might be thinking that I am splitting hairs, being anal… but stay with me. Even though XY Signature 5FT is the single most popular curve in this Weatherford analysis, it only appears in 10 distinct LAS files.

2,152 of the 2,226 have a gamma curve. More interesting is the 10,959 porosity curves recorded by Weatherford. 1,456 (65%) LAS files have at least one porosity curve. There are 24 different porosity MNEMs. There are 30 different porosity descriptions. If log A has a Limestone Neutron Porosity curve and log B has a Sandstone Neutron Porosity curve, can we analyze A against B?

GUTCHECK If a well logger has read this far, the reaction is likely to be "who cares." None of this matters. But remember an engineer is reading (eyeballing) one log at a time or comparing a (very) small set of rasters laid out on a table/screen. We are attempting to get the data into a "state" where we can analyze hundreds of thousands of LAS files side-by-side (and look for correlations). Data cleansing and curation matter.

Curves

One would think that even though a service company can free-type curve descriptions, there would be some standardization when it comes to MNEMs. Sadly, there's not. 48K LAS files have 9,942 unique mnemonics. There will need to be a lot of cleanup to do aggregate analysis. This is a bit frightening. Even within a single service company (Schlumberger) there are 42 ways to describe a RHOZ log. We can conclude there is little-to-no consistency even inside a service company. A truck going north might do a Standard Resolution Formation Density curve and a truck heading south might call it a Density using Gardner Equation and DT-Compressional. This is just crazy!

Let's flip it on its head. There are 4 different MNEMs to describe "Gamma Ray {13.4}." This feels very subtle, but we are starting to conclude that Schlumberger (the company) does not believe in drop down menus. They leave it to each truck to free-type both MNEMs and curve descriptions. Therefore, it's a challenge to study/analyze

Schlumberger LAS files, much less files across 311 different service companies. After working with BHI, I would not be surprised if different trucks are using different logging software altogether.

And don't count on Schlumberger making it easy on us. First, you have to be a client/partner to get their chart book. I found a 2009 version online and, of course, there is no mention of RHOZ, Gardner equation or Standard Resolution Formation Density. I was forced to go here to find anything.

Let's switch gears a bit. Realize that this entire email is only an analysis of metadata (top of the LAS files) so far. There's 90MB of CSV files representing actual curve data, with some LAS files over 200MB each. Because there are almost 10K different curve types, does this mean we will have 10K columns in our table (and 800MM rows)? Will we snip the curves we want to use for machine modeling and combine different MNEMs that are measuring the same event types, but may have a different MNEM?

Then we need a process to pour the data into the table… for example, here we have a 4 curve log. The fourth curve is IHV. But in this LAS file, IHV is column/curve 11. In our table, IHV may be column 986 and we need to make sure all IHV data (where present) drops into the right spot in the table.

GUTCHECK If you're freaking out at this point, I get it. But show this to any computer scientist and s/he will say it can be done. Show it to a machine learner and s/he will say finding hydrocarbons is probable, but the data has to be clean before submitting to the models. Companies are reporting massive problems with their data quality (page 4, 2nd column). But companies don't have a data problem, they have a human problem… and on both ends. Truck A and truck B are free typing away and no one back at corporate knows how to map, much less analyze millions of rows of log data in one sitting.

Now it's my turn to freak out a little. The most popular curves appear to be a little boring? Where's all the resistivity and porosity and acoustic MNEMs? The only saving grace is there appears to be 294 different MNEMs that describe resistivity, 369 MNEM that refer to porosity, but only 55 MNEMs for permeability. This is a pretty big concern for me. We not only need curve coverage for a large percentage of APIs, but we need combinations of curves consistently across those APIs. We may have more than one LAS per API (let's hope so!). But we need to get a handle on how many APIs have gamma + porosity + resistivity + permeability + anything else critical, even if they appear in different LAS files (under different MNEMs) for the same API.

MNEM	Curve Count	Typical Description
DEPT	28,548	Depth
GR	21,605	Gamma Ray
Depth	18,054	Depth
ROP	17,639	Rate of Penetration
C1	13,860	*No description*
C2	13,829	*No description*
C3	13,813	*No description*
Gamma	12,927	Gamma Ray
TG	12,486	Total Gas
C4	10,999	*No description*
CCL	9,538	Collar Locator
Gas	7,490	*No description*
TENS	6,770	Cable Tension

LTEN	6,040	Surface Line Tension
AMP3FT	5,488	AMP3FT Amplitude
TT3FT	5,304	3FT Travel Time
THMN	4,994	Thickness Minimum Value
THAV	4,934	Thickness Average Value
THMX	4,932	Thickness Maximum Value
IRAV	4,847	Internal Radius Averaged Value
IRMX	4,843	Internal Radius Maximum Value
IRMN	4,840	Internal Radius Minimum Value
LC02	4,271	LC02 Local Comp
C5	4,145	*No description*
NPOR	3,714	Base Neutron Porosity
CTEM	3,463	Cartridge Temperature
GTEM	3,354	Generalized Borehole Temperature
HCAL	3,315	HRCC Cal. Caliper
AMP5FT	3,163	AMP5FT Amplitude
BONDIX	3,134	Bond Index
TT3	3,115	CBL 3' Travel Time
ATT3	3,050	Attenuation from 3'Amplitude
ERAV	2,991	External Radii Average
ERMX	2,988	External Radii Maximum
ERMN	2,988	External Radii Minimum
AMP3	2,920	CBL 3' Amplitude
STIT	2,845	Stuck Tool Indicator
HDRA	2,662	HRDD Density Correction
AMPAVG	2,652	Average Sector Amplitude
LSPD	2,651	Line Speed

We will need to build a master map that combines MNEMs that tell the same story (e.g. resistivity).

PROPOSED STEPS

1. Get our hands on as many LAS files as possible. Scrape them into well and curve metadata + curve details. Technology is already built and was used for this early analysis.
2. Create a map to combine similar MNEMs for analysis. Do this for other metadata (e.g. service company and operator)
3. Import all the curve details into the right columns in a master table.
4. Import perforation intervals for all LAS APIs and set those curve values to a 1. Select an equal area outside the interval and set those values to 0. Zero represents the absence of hydrocarbons.
5. Develop calculations that look across curves and within the same curve to see how values change as depths change through the interval.
6. Build a machine model that correlates hydrocarbon production with curve values. Note: production levels are not needed at this time. Just zone produced (1) or did not produce (0).

7. Create a web service so new LAS files can be uploaded and analyzed by the model(s) to select best perforation intervals.

With this service, any wellbore (green/yellow/brown) can be analyzed for a second opinion on where to perforate to maximize hydrocarbon recovery. This opinion is based on probability and statistics and is like having a single engineer (with a double major in applied math) analyzing 100K+ LAS files and generating a universal solution. The owner of this service can sell access to clients for green field analysis or use the technology to buy up underperforming wellbores and recomplete for passed-over hydrocarbons.

Additional Notes:

1. Check out page 81-82 of the Colorado School of Mines course catalog (link here). To get a degree concentration in "Petroleum Exploration Engineering," one does not have to take a single probability or statistics course. And they stop math at Differential Equations. Petroleum Engineering degree same thing (page 121-122). Yet finding hydrocarbons is 100% probability and statistical risk.
2. Even when presented with digital log data, engineers will still rasterize the data and block it visually. For example, Baker Hughes went nuts over this > https://www.youtube.com/watch?v=TWpz-48z9M4
3. The key is to digitize old-timers' visual and intuitive expertise to make this service work. We need engineers that have actually been to the wellbore, sat in the truck, understand the process, cradle-to-grave. This is a good time to seek out retirees and the unemployed.