Here we present a more detailed analysis of the problem of exploring high dimensional spaces to discover interesting patterns within data. This subject was touched upon in our Splash event video, but we'd like to further expound upon it here to clarify more explicitly the size of the problem and why it is quite simply intractable via brute force methods.

The **brute force** approach to finding patterns within a dataset would be to explore all patterns within a space of *n* variables and order *r* — the number of variables that are potentially related to a particular outcome.

Generally, the total number of patterns in *n* variables of order *r* is:

$$C(n,r) = \binom{n}{r} = \frac{n!}{r!(n-r)!}$$

This assumes patterns can be found without regard to the orientation, or permutation of variables (A before B before C etc.) If the orientation of variables matters, the problem size is then:

$$P(n,r) = \frac{n!}{(n-r)!}$$

Which is considerably larger.

Let's consider a very small problem where we want to analyze human gene data. In this dataset, *n=20,000* genes and we'd like to find patterns of *r = {3,4,5,6}* order relationships wherein the orientation of variables in an interaction does not matter.

We assume at the extreme lowest end, it would take at minimum 1 floating point operation (FLOP) per example. Let's say that there are $1,000$ patients. In this case, it would take $1,000$ FLOPs to evaluate each pattern. This is of course a very low estimate and in reality it is most likely 1-3 orders of magnitude larger, but let's continue with this assumption.

Next, we would like to determine how long it would take to explore this space. To do this we must consider the total time it takes to explore the space and how fast you can explore it. If we were to utilize a computer that could operate at a specific number of operations per second, we can calculate how much time it would take to explore the entire space of patterns that exist. We used 10^{17} FLOP/s as this rate of exploration, which is faster than the current fastest supercomputer *Sunway Taihu Light* which runs at a reported $9.3 * 10^{16}$ FLOP/s as of November 2017 (https://www.top500.org/lists/2017/11/). We recognize that this will scale further in the coming years.

The whole computation is then as follows:

For the lowest *r=3*:

$$\frac{\binom{20,000}{3} \times 1000\ FLOPs}{10^{17}\ FLOP/s}$$

which is around $1.3 * 10^{-2}$ seconds.

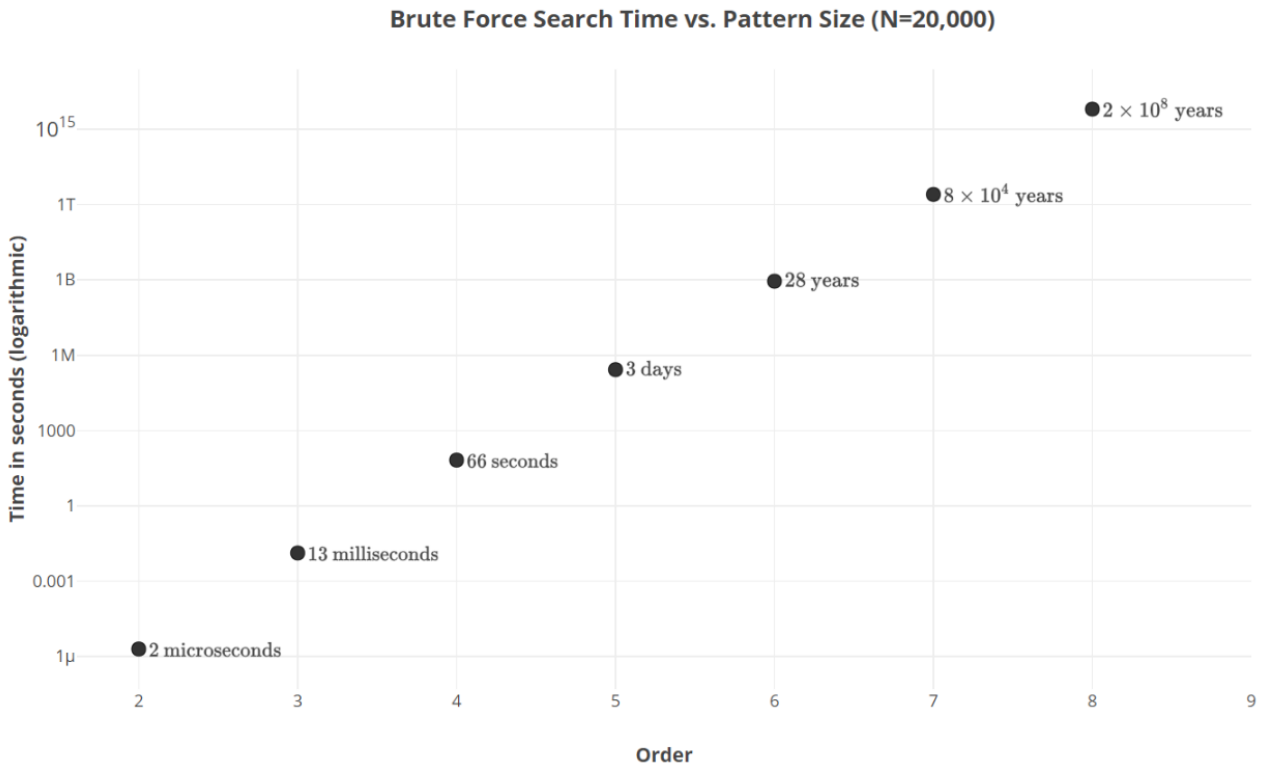
For the highest *r=6*:

$$\frac{\binom{20,000}{6} \times 1000\ FLOPs}{10^{17}\ FLOP/s}$$

which is around $8.8 * 10^{8}$ seconds, or ~28 years.

The true intrinsic patterns within a dataset, however may not just be limited to 3, 4, 5, or 6-order patterns and may be even considerably higher, in which case it rapidly becomes even more impossible to explore by brute force.

The following graph shows the time that it would take to explore all possible patterns of a given order with the above assumptions (note, the y-axis is logarithmic):



As one can see, the problem space very quickly becomes intractable to approach as the search for more interacting variables widens. Therefore, an efficient, elegant algorithmic approach to this is absolutely necessary to explore higher order patterns since it simply cannot currently be explored by brute force in reasonable human timescales.