Notes from Thursday, June 7, 2018 conference call

IRS infrastructure

* Large clusters
* Software is 2 years old: Spark, PySpark, Hadoop
* Data is 400,000 rows, INSOL has 3,300 rows, RTF has 1,500, mostly numerical, categorical
* Who is Lec?
* Mike Strudler and Mike Weber prefer the RTF as a starting point.
* C++ programs do some of the initial editing