“The End of Theory: The Data Deluge Makes the Scientific Method Obsolete”

Published August 3, 2008

Citation

V2lsbGlnaGFnZW4sIEUuICgyMDA4LCBBdWd1c3QgMykuICJUaGUgRW5kIG9mIFRoZW9yeTogVGhl
IERhdGEgRGVsdWdlIE1ha2VzIHRoZSBTY2llbnRpZmljIE1ldGhvZCBPYnNvbGV0ZSIuIDxpPkNo
ZW0tYmxhLWljczwvaT4uIGh0dHBzOi8vZG9pLm9yZy8xMC41OTM1MC9qNGQxMC0wanQwNQ==

Keywords

Cheminf

Abstract

The thought triggering editorial “The End of Theory: The Data Deluge Makes the Scientific Method Obsolete” by Chris Anderson can’t have escaped your attention. I was shocked when I read the title and the comments made on the blogosphere and on FriendFeed.

The thought triggering editorial "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete" by Chris Anderson can't have escaped your attention. I was shocked when I read the title and the comments made on the blogosphere and on FriendFeed.

How can he say that?! There is no analysis of data anymore?!? Don't we need to understand why X correlated with Y?!? Etc etc.

So, when I read yet another comment, by my respected opensource chemoinformatician Joerg, I just had to read the piece myself. Joerg disagrees with the statement from Chris' editorial that

> *[c]orrelation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all.*

At first, I would agree with Joerg. It's nonsense; any QSAR modeler can explain in details the dangers of overfitting, extrapolation, etc, etc. Not to mention that basically zero mathematical modeling methods can create a statistical signification non-zero regression model with less than 50-100 chemical structures (chemical diversity dependent, etc).

Ok, back to the editorial. There are some arguments on Google, tons of data. Number of incoming links as measure of page importance (brilliant choice, but actually a model, IMHO, which Chris seems to step over). Tons of data. Oh, mentioned that already.

Mmmmm... but wait. Tons of data? The editorial actually refers to petabytes: *Petabytes are stored in the cloud*. (Whatever the cloud is... just another buzzword, trademarketed too, it seems).

Eureka! Chris is right, Joerg is wrong!

Yes! Then it hit me, Chris is actually correct in his statement, and I was wrong (and Joerg too). If we move away from 50-100 molecules in our QSAR training, but use 10k of chemically alike molecules, then our modeling approaches (if capable of handling the matrices) would have a much, much smaller chance for overfitting, extrapolation (there is much, much more interpolation now), etc. The chances of getting random correlation become insignificant! Actually, Chris is making the argument QSAR modelists have been making for decades: we do not know the mode of action in detail, as we can make, given enough training data, a reasonable regression model to predict the action! Joerg and I have been making the same argument as Chris in our PhD theses! We do not need theory; our QSAR regressions make theory obsolete! (Well, surely, we'd still prefer the theory behind the action, but we lack the measuring techniques to see what actually is happening. Joerg, still agreeing with you, so to say ;)

Except for one thing. Joerg and I suggested 'enough' molecules are required for statistical sound regression. Chris, on the other hand, even makes the point that regression is no longer needed at all at the petabyte scene: we just look up what is happening. Does this hold for chemistry? For QSAR? Petabyte data equals about, say 10kB data per structure, maybe less if we use InChI and neglect conformer info, 100.000.000.000 structures. About 5000 times ChemSpider, if not

miscounting the zeros (we don't care about a ten-fold at this scale anymore). Maybe, maybe not. Maybe chemical space is too diverse for that, considering a petabyte of chemical structures is enormously insignificant to the full drugable space (was about 10^{60}, not?)

But not at all? This lookup approach is actually commonly used in chemoinformatics! Even at a way-below-pentybyte scale: HOSE-code-based NMR prediction is a nice example of this! We do not theorize on the chemical carbon NMR shift, we just look it up!

Certainly worth reading, this Wired editorial!

PS. One last remark on the title... I'd say the the *scientific method* is more than just making theories... I feel a bit left out as data analyst... :(I guess the title should have said 'one of the Scientific Methods'...