

A Statistician's View of Big Data

Max Kuhn, Ph.D (Pfizer Global R&D, Groton, CT)

Kjell Johnson, Ph.D (Arbor Analytics, Ann Arbor MI)

What Does Big Data Mean?


The advantages and issues related to Big Data can be broken down into two areas:

- Big N : more samples or cases
- Big P : more variables or attributes or fields

Mostly, I think the catchphrase is associated more with N than P .

Does Big Data solve my problems?

Maybe¹

¹the basic answer given by every statistician throughout time 

Can You Be More Specific?

It depends on

- what are you using it for?
- does it solve some *unmet need*?
- does it get in the way?

Basically, it comes down to:

Bigger Data \neq Better Data

at least not necessarily.

Big N - Interpolation

One situation where it probably doesn't help when samples are added *within the mainstream of the data*

In effect, we are just filling in the predictor space by increasing the granularity.

After the first 10K observations, the model will not change very much and it becomes a game of nm.

This does pose an interesting interaction within the **variance–bias trade off**.

Big N goes a long way to reducing the model variance. Given this, can high variance/low bias models be improved?

Variance–Bias Trade Off

Maybe.

Many high(ish) variance/low bias models tend to be very complex and computationally demanding.

Adding more data allows these models to more accurately reflect the complexity of the data but would require specialized solutions to be feasible.

At this point, the best approach is supplanted by the available approaches (not good).

Form should still follow function.

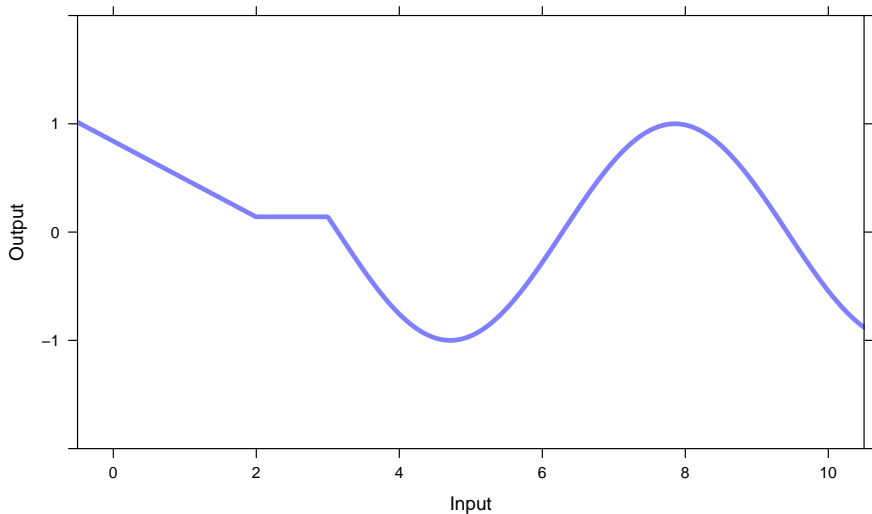
Variance–Bias Trade Off

What about low variance/high bias models?

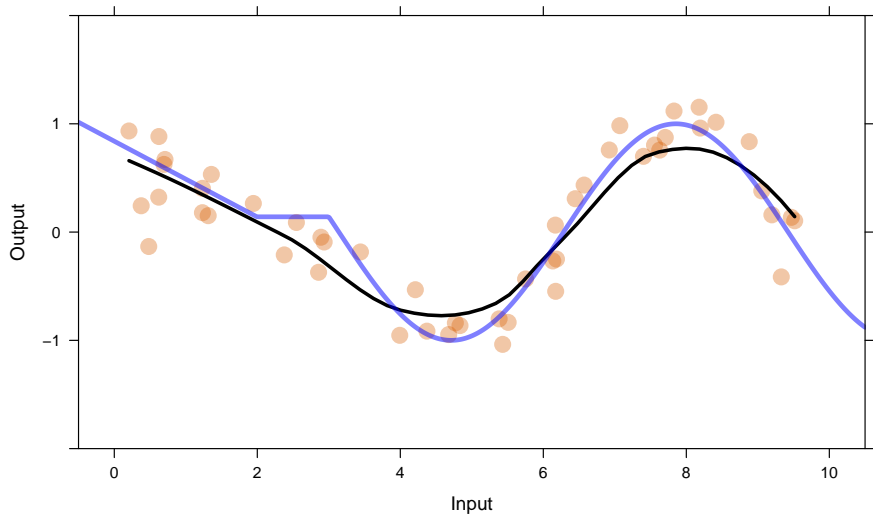
There is some room here for improvement since the abundance of data allows more opportunities for exploratory data analysis to tease apart the functional forms to lower the bias (i.e. improved *feature engineering*) or to select features.

For example, non–linear terms for logistic regression models can be parametrically formalized based on the results of spline or loess smoothers.

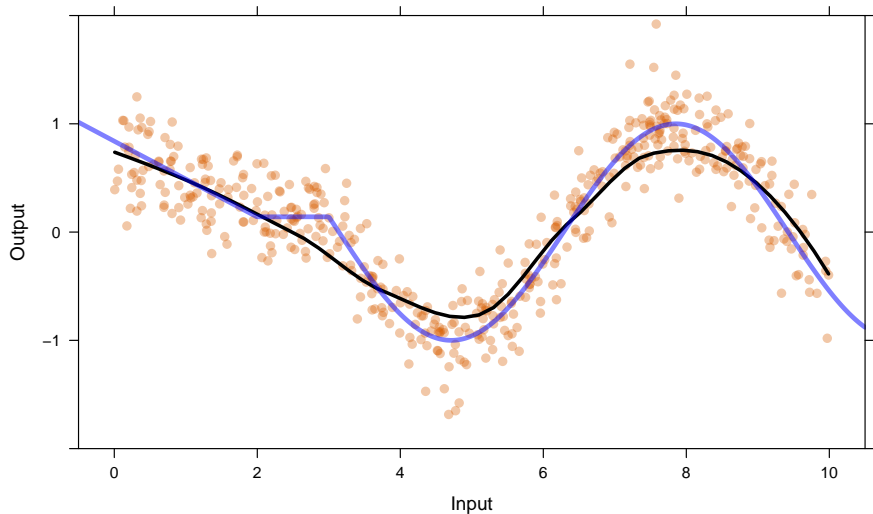
A Simulated Pattern



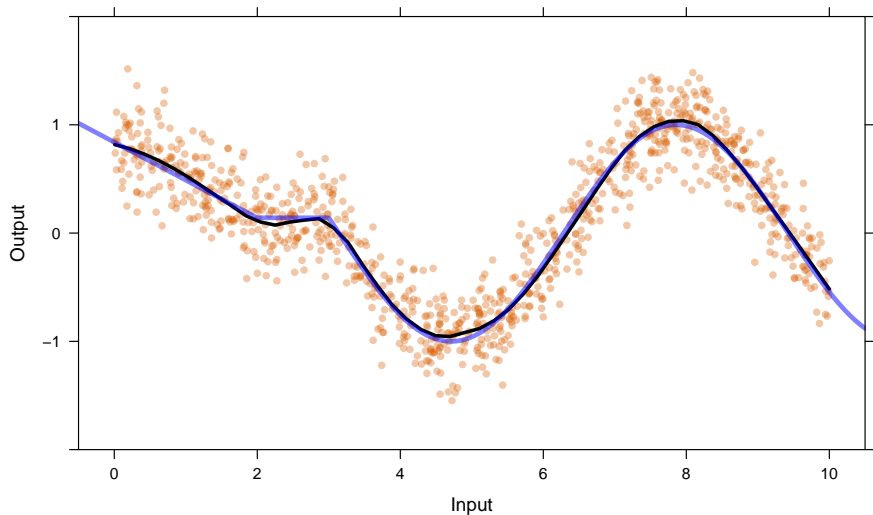
High Bias Model (LOESS) with $n = 50$



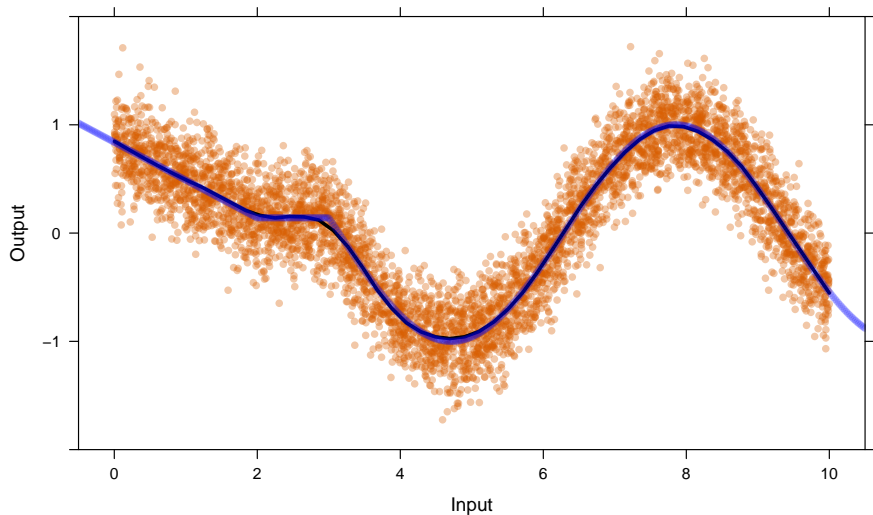
High Bias Model (LOESS) with $n = 500$



Lower Bias Model (LOESS) with $n = 1K$



Lower Bias Model (LOESS) with $n = 5K$



Diminishing Returns on N^*

At some point, adding more ($*$ of the same) data does not do any good.

Performance stabilizes but computational complexity increases.

The modeler becomes hand-cuffed to whatever technology is available to handle large amounts of data.

Determining what data to use is more important than worrying about the model.

Two Examples from Computational Chemistry

Pfizer relies on quantitative structural–activity relationship (QSAR) models for medicinal chemistry.

Assay data for existing compounds are used to create a relationship between the characteristic of interest and molecular descriptors (e.g. molecular weight, the number of carbons, etc)

When new compounds are designed, we can estimate their characteristics prior to synthesis.

We also have existing assay data on large numbers of compounds to build predictive models.

Global Versus Local Models

When developing a compound, once a “hit” is obtained, the chemists begin to tweak the structure to make improvements. This leads to a *chemical series*

We could build QSAR models on the large number of existing compounds (a global model) or on the series of interest (a local model).

Our experience is that local models beat global models the majority of the time.

Here, fewer (of the most relevant) compounds is better.

Our first inclination is to use all the data because our (overall) error rate should get better.

Like politics, *All Problems Are Local*.

Data Quality

One “Tier 1” screen is an assay for $\log P$ (the partition coefficient) which we use as a measure of “greasiness”.

Tier 1 means that $\log P$ is estimated for most compounds (via model and/or assay)

There was an existing, high-throughput assay on a large number of historical compounds. However, the data quality was poor.

Several years ago, a new assay was developed that was lower throughput and higher quality. This became the default assay for $\log P$.

The model was re-built on a small (1K) set chemically diverse compounds.

In the end, fewer compounds were assayed but the model performance was much better and costs were lowered.

Big N and/or P - Reducing Extrapolation

However, Big N might start to sample from rare populations.

For example:

- a customer cluster that is less frequent but have high profitability (via Big N).
- a specific mutation or polymorphism that helps derive a new drug target (via Big N and Big P)
- highly non-linear “activity cliffs” in computational chemistry can be elucidated (via Big N)

Now, we have the ability to solve some unmet need.

Big P is Most Important

I believe that we have reached a plateau in machine learning.

Breakthroughs in predictive performance are not likely to be a result of new algorithms.

Real progress will be made by improving the *data* by enriching the information content available or reducing noise.

Example – RNAseq

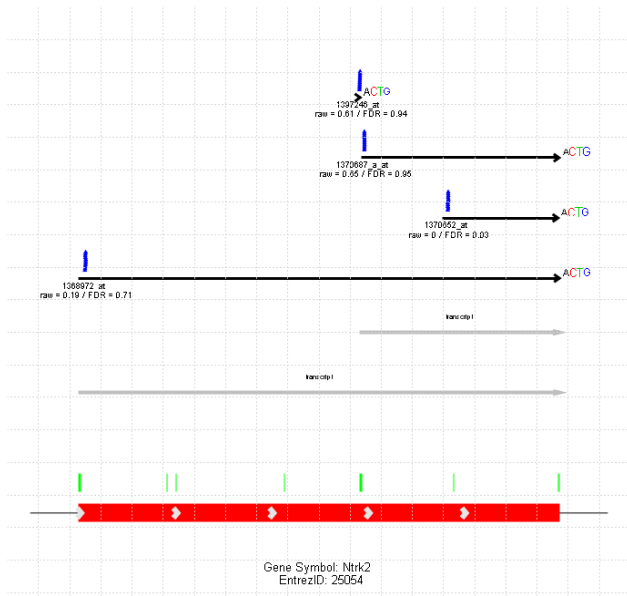
Traditional RNA expression profiling revolves around hybridization arrays (e.g. Affymetrix)

These assays are very noisy, static and bias to the 3' end.

RNAseq applies sequencing technology to RNA transcripts and, as a consequence:

- *digital* estimates of RNA abundance are generated
- the transcript library is no longer static and there is no 3' bias
- the complexity of the data is drastically increased

Affy 3' Bias



Example – RNAseq

The good:

- the sensitivity/resolution is better than previous technologies
- we can now effectively find splice-variants and other genetic components
- the measurement system error can be dramatically improved

The possibly bad:

- the size of the data may let the tool drive the analysis
- many scientists barely have enough time to truly capitalize on the existing, lower-resolution data
- Big N may be required to really make break-throughs due to low event rates

Big N May Enable True Bayesian Machine Learning

$$Pr[Y|X] = \frac{Pr[X|Y]Pr[Y]}{Pr[X]}$$

There are “no Bayesians in foxholes” (Breiman, 1977)

The biggest issue here is $Pr[X]$ (and it's conditional).

We usually make ridiculous assumptions about the multivariate densities (e.g. LDA, QDA, naive Bayes etc) because we have to.

With a good $N:P$ ratio, maybe we can focus more on prediction uncertainty and other issues via Bayesian methods.

Big N May Enable Effective Biological Networks

A good percentage of biological networks are derived from data with $P \gg \gg \gg N$

- Completely over-determined systems lead to severe *ad-hockery*
- Lack of application of good high-dimensional methodologies (e.g. regularization).

This is a good example of where Big N can really offer a contribution.

High Content Screening via Flow Cytometry using cellular data (instead of “wellular data”) dramatically increases N and allows better estimates of networks.

The nature of this technology reduces P significantly (compared to an array) but the unintended side-effect of this is that the researcher usually thinks about what they need to measure.

The problem is that the N here isn't really the N we need (i.e. repeated experimentation) and has the potential to over-fit to the data.

Summary

I think our first inclination is to “go big” because

- It's cool
- I could write a few papers/packages
- More is better, right?

The availability of Big Data should be a trigger to really re-evaluate *what* we are trying to solve and *why* this will help.

Thanks to

- Martin, for the invitation to speak and the stimulating discussions
- Vini Bonato, for the Affy probe visualization