Machine Learning and Big Data for

LIFE EXPECTANCY

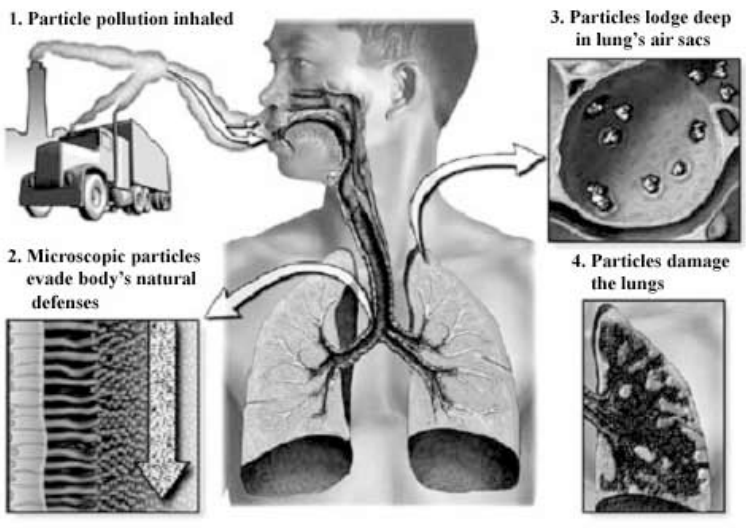
Prof David J. Lary http://davidlary.info

EFFECTIVE PUBLIC HEALTH POLICY

A new age is dawning with actionable insights being provided by a combination of holistic datasets comprehensively describing issues, coupled with unprecedented computing power and machine learning. For the first time in man's history we have the chance to comprehensively characterize problems and objectively focus on the key issues.

Decades of research have also attempted to isolate the specific factors that matter the most, not only as an academic exercise but to help policy makers set priorities in a decision-making environment of limited resources. In a community beset with high rates of chronic diseases, should a city council or board of supervisors give priority to hospital budgets, expanding primary care, passing laws to ban smoking indoors, addressing unemployment, strengthening schools, and so on? All are clearly important, but which matter the most and will give the best "bang for the buck"?

Posed with this question scientists have typically resorted to traditional statistical techniques, such as regression equations, to try to quantify or model the relative importance of different factors. This approach has yielded appreciation of some of the key factors but carries limitations, two of which bear mention here. First, these calculations often examine associations rather than causality: for example, the fact that people who have not graduated from high school have worse health doesn't mean that handing out diplomas will fully erase the disparity, rather, the educational level proves to be a useful proxy. Second, the variables that researchers plug into their formulas are chosen selectively based on the variables for which data are available and those that the researchers "think" are most important to consider. For example, a researcher forced to choose whether to adjust for poverty, voter registration, or social trust will invariably choose poverty because there is more evidence available linking poverty to adverse health outcomes.



The need to be selective and for humans to choose the variables to consider is partly a legacy of the age-old scientific method—pose a hypothesis first and then collect data to support or refute it—but partly a practical necessity because examining all the data has always been an untenable option, especially as the volume of available data has expanded.

The advent of machine learning is removing the second barrier at a time when the availability of "big data" is ascendant in all fields. Machine learning can provide a valuable regression tool for empirically estimating variables of interest when we do not have a complete theoretical description of a process but we do have a useful dataset. Machine learning encompasses a very broad range of algorithms (for example, Neural Networks, Support Vector Machines, Gaussian Processes, Decision Trees, etc.) that can provide multi-variate non-linear non-parametric regression or classification based on a training dataset.