Are big data always bad data? How to make your best with incomplete data.

I will start by showing that big linguistic enterprises, like data bases, atlases and all sorts of corpora, always contain a certain amount of "noise". They are by definition always incomplete when the hypothesis we want to test is very detailed. Since this is inevitable, I will not present statistical methods to circumvent this problem, but consider some alternative strategies that can help us to find interesting theoretical clues even in data that provide by definition a coarse-grained picture of the linguistic reality. If it is true that big data is never precise enough for a very detailed hypothesis, we can still try to exploit the peculiarity of a blurred image to single out the general outlines of the linguistic panorama, which would remain otherwise uncovered. In this way, using big data mining can nicely complement our introspective type of empirical evidence. In other words, the reason why big data are always "noisy" is that we ask them too much, and the questions we ask are not apt to the type of evidence we have. The solution to the problem I will present is the following: up to now we have only used big corpora to look at the presence versus absence of a given structure in a given language and related it to other structures. An innovative way to think about big data and tailor our questions on the linguistic evidence provided by big data is to consider the type of variation itself as a clue indicating different natural classes of linguistic phenomena. In the talk I will single out three possible distributional types and determine to which type of phenomenon each type of variation is related.

A. I will first take into account the "classic" method of comparing the geographical distribution of different phenomena and consider the theoretical import of different distributional patterns, in particular those in which two phenomena a) completely, b) partially overlap or c) are in complementary distribution on a map. This type of methodology has been used by traditional dialectologists and is still used today in formal frameworks and can be only be adopted when we are comparing two phenomena and trying to establish whether they are intrinsically related or not. A case for a. would then be the one in which the two phenomena depend on the same abstract property. A case for b. would be that the phenomenon that is more largely represented is a necessary but not sufficient condition for the occurrence of the second. A case of c. would be a case of alternative checking of the same property, so that you can only have either the first or the second phenomenon. Still, the distribution we find could only be by chance, but if we have enough languages, the probability that we only have to do with chance reduces the bigger our sample is. Which means, you can use big data, but they have better be really big.

B. I will then present another type of geographical distribution which can provide us with interesting observations that we would not be able to see on the basis of a detailed investigation. This type of distribution is called by traditional dialectologists "leopard spots" because the phenomenon under study occurs precisely with an apparently random distribution. I will show that leopard spot variation is found when we deal with a phenomenon that is only possible when a specific complex constellation of factors is instantiated in the same language. The study of this type of variation can lead us to find out exactly what the complex pre-requisites are that lead to the occurrence of the phenomenon under study.

C. I will show that it is possible to extract syntactic observations from lexical data simply by looking at the type and possible lexical variation for the same functional element starting from a simple but rather strong hypothesis: the index of lexical variation of a functional word within a genetically related set of languages co-varies with the semantic and syntactic complexity of the item itself. I will show that a rather simple count of the possible lexical items used in a set of related languages gives us very precise indications of the featural primitive components the functional element is made up of.
I will conclude by arguing that there are alternative strategies to trying to counterbalance the fact that all our data are noisy, we simply have to use different types of distributional patterns as an indication of different classes of phenomena and use the evidence we have to formulate different types of questions.