Group- and Population-level Analysis: Big Data

In the realm of neuroimaging, it becomes increasingly clear that improving the state of current knowledge and the quality of inference requires the joint analysis of large amounts of data, as very large samples sizes are typically required to discover meaningful and reliable features of brain organization. Simultaneously, steady improvements in spatial resolution increase the size of each individual image.

As a consequence, very large datasets are currently assembled and made publicly available, such as the Human Connectome Project (HCP) dataset: 20TB and growing. Even larger cohort will be necessary to decode the impact of genetic variability on brain organization.

In parallel, priceless information can be gained by confronting several such datasets that map human cognition to brain activity patterns. Such a framework is called *meta-analysis*, whereby the joint analysis of large amounts of data yields a consistent and reliable view of brain organization, as it accumulates information across a large variety of experimental settings and cognitive descriptions.

While these ideas are currently gaining interest, the technical and scientific challenges have not been fully acknowledged yet.

In this talk, we will first discuss the importance of sampling large populations to address the modeling of cross-subject variability in neuroimaging and integrate multiple sources of information. In this part, we will mostly review the fundamentals of statistics and machine learning wisdom.

Then, we will focus on the particular case of population comparison in functional connectivity studies, which suffer from limited signal-to-noise ratio. We will describe how the statistical structure of large functional data can be captured to draw meaningful conclusions regarding methods and neuroscience applications.

We will then consider the benefit of large-scale functional neuroimaging studies that consider together a large number of studies to draw sharper conclusions on the cognitive architecture of the brain.

We will conclude the presentation by drawing two consequences from the advent of big data analysis in the domain:

- the need for adapted analysis software that can scale to very large datasets. We will review the current solutions and bottlenecks.
- We will discuss the need of public datasets to enhance the scientific conclusions that can be drawn from neuroimaging studies and increase the reliability of the findings.