1176-68-129 **David Alvarez-Melis***, daalvare@microsoft.com, MA. *Principled Data Manipulation with Optimal Transport.*

Success stories in data science and machine learning seem to be ubiquitous, but they tend to be concentrated on 'ideal' scenarios where clean, homogenous, labeled data are abundant. But machine learning in practice is rarely so 'pristine'. In most real-life applications, clean data is typically scarce, is collected from multiple heterogeneous sources, and is often only partially labeled. Thus, successful application of ML in practice often requires substantial effort in terms of dataset preprocessing, such as augmenting, merging, mixing, or reducing datasets. In this talk I will present some recent work that seeks to formalize all these forms of dataset 'manipulation' under a unified approach based on the theory of Optimal Transport. Through applications in machine translation, transfer learning, and dataset shaping, I will show that besides enjoying sound theoretical footing, these approaches yield efficient, flexible, and high-performing algorithms. This talk will be based on joint work with Tommi Jaakkola, Stefanie Jegelka, Nicolo Fusi, Youssef Mroueh, and Yair Schiff. (Received January 18, 2022)