# Computing@PNNL SEMINAR

# *A Network Approach to Topic Models*


## Martin Gerlach, Ph.D.

Postdoctoral Fellow, Department of Chemical & Biological Engineering, Northwestern University


**January 24 | 11 AM | ISB2 Wanapum 155**

One of the main computational and scientific challenges in the modern age is to extract useful information from unstructured texts. Topic modeling is a popular method from machine learning that infers the latent topical structure of a collection of documents and has been successfully applied to a range of problems in sociology, history, linguistics, etc.

In his talk, Dr. Gerlach will present an alternative approach to topic models in the framework of community detection in complex networks, particularly, stochastic block models, by representing text corpora as bipartite networks of documents and words. He will show how this leads to a more principled and versatile formulation of topic modeling, solving many of the intrinsic limitations of state-of-the-art methods, such as Latent Dirichlet allocation. Following these insights, Dr. Gerlach will introduce a new framework to evaluate the performance of topic models based on synthetic benchmark corpora, in analogy to benchmark graphs commonly used for evaluating and comparing community detection algorithms. This approach yields an unbiased and absolute measure of the performance of topic model algorithms, leading to new insights on, for example, principal limitations of topic models to infer topical structures.

Pacific Northwest NATIONAL LABORATORY

*Proudly Operated by Battelle Since 1965*

U.S. DEPARTMENT OF ENERGY

**Host:**
Robert Rallo
Data Sciences Group Lead
ACMD Division
Robert.Rallo@pnnl.gov