Tackling Big Societal Questions using Big Data Solutions

Lakshminarayanan Subramanian
Associate Professor of Computer Science
Courant Institute of Mathematical Sciences
New York University

Big Data has had a profound impact on human society across a wide spectrum of spheres. In this talk, I will share three brief and diverse journeys in my life where we have both made fundamental scientific advances in big data algorithms as well as successfully deployed big data systems at scale to address important societal challenges with the hope of directly impacting the lives of people.

The first story is at the intersection of artificial intelligence and counterfeit detection. Counterfeit goods represent a $1.7 trillion worldwide problem encompassing nearly 5% of world trade. We describe a new product authentication mechanism that uses high dimensional machine learning algorithms on microscopic images of physical objects to distinguish between genuine and counterfeit versions of the same product. The underlying principle of our system stems from the idea that microscopic characteristics in a genuine product or a class of products (corresponding to the same larger product line), exhibit inherent similarities that can be used to distinguish these products from their corresponding counterfeit versions. Our counterfeit detection system has been operational for luxury goods for the past 6 months and yields more than 99.5% true positive rates with very low false positive rates.

The second story is in the broad area of big-data driven disease surveillance. Accurate disease forecasting at fine-grained location granularities is an extremely challenging problem, due to the lack of reliable disease surveillance data. In developing countries, absence of reliable disease surveillance data makes it extremely challenging to build an early epidemic warning system. I will present results from an operational disease surveillance system that circumvents the unavailability of surveillance data by using changes in calling patterns on a phone hotline as a basis for forecasting an epidemic. Used in Pakistan for over three years to forecast Dengue outbreaks, our system uses patterns of locality-specific call volume data of dengue-related complaints, combined with environmental parameters, to accurately forecast outbreaks (correlation of up to 0.93), 2-3 weeks ahead of time at a fine-grained sub-city level. Our system is the first of its kind that demonstrates that crowd-sourced citizen data from a health hotline can be used to build an accurate, fine-grained, advance warning system in the absence of tradition disease surveillance data.

The third story is in the broad area of event-driven predictive models for socio-economic indicators. In this story, I will describe how one can extract real-world events using NLP algorithms on unstructured news streams and understand the impact of real world events on the volatility of prices or other macro-economic indicators. Our work is built on the hypothesis that factors that trigger sudden fluctuations in commodity prices or other macro-economic indicators can be better characterized with a detailed understanding of real world events and their relationships to specific indicators of interest. We have built an event-driven analytics engine that can automatically learn and categorize location-specific events in real-time from diverse news sources and characterize the relationships between real-world events and fluctuations in prices and macro-economic indicators in a given locality. Specifically, given a specific news corpus pertaining to a given domain, we describe an event class model that can automatically learn different classes of real-world events akin to the concept of classifying documents into topics and automatically learn event-driven predictive models that can be used to infer and potentially predict future fluctuations in specific indicators. We describe specific results of our work on what triggers massive fluctuations in tomato and onion prices in India and why such fluctuations lead to riots!