**Overview**

**1 introduce ml**

* **(Before talk about sml, we want to introduce u ml to see the differences. From my prospection, in a few words, ml is…. So, what is sml？ )**

**2 introduce slml**

In many domains, data now arrives faster than we are able to learn from it. To avoid wasting this data, we must switch from the traditional "one-shot" machine learning approach to systems that are able to mine continuous, high-volume, open-ended data streams as they arrive. We have identified a set of desiderata for such systems, and developed an approach to building stream mining algorithms that satisfies all of them. The approach is based on explicitly minimizing the number of examples used in each learning step, while guaranteeing that user-defined targets for predictive performance are met.

* + - **(In the following, we will explain the approach by using examples, but first let’s see why we need sml)**

**3 why we need slml**

* + **（”it is not who has the best algorithm that wins, it is who has the most data”）**
  + **challenges**
  + **1 computation**
  + **when data set growing bigger and bigger, the amounts of examples make training cost more and more time.**
  + **For example, handle 300,000,000 people’s health care information. You need to go throw 3 00million data for every learning step. It’s horrible.**
  + **2 natural language**
  + But what we really want is not just raw data, but computer systems that understand this data
  + For instance: same words in different sentence may lead to different meanings

通常大规模机器学习都是深度学习

Why Deep learning?

* + When it comes to image or natural language processing, normal ML algorithms have bad performance. While going deep into neural network, we can extract the more abstract information contained in the data.
  + **3** **Large-scale unsupervised learning**
* 3. 1 afterwards challenges
* But it also brings problems in storage, I/O, communication, etc.
* Communication happens at least between each node and the scheduler. It is much harder than usual communication, since the individual system is more complex. When talking about communication, we have to choose synch or asynch.
* An additional reason that we need sml is the need of the development of reinforcement learning.
* 4. How
* 4.1 MapReduce
* 4.2 GPU
* What GPU dose is exactly the same as MapReduce but on a single chip. A GPU is divided into several blocks and blocks are divided into threads which do smaller job.
* GPU stands for graphic processing unit. It is designed for processing images using a parallel architecture. CPU is designed for calculating. When processing, the gpu is divided into blocks and blocks are divided into threads. Every thread dose a part of the big job.
* 4.3 Stochastic GD

We can now store and perform computation on large datasets, using things like MapReduce, BigTable, Spanner, Flume, Pregel, or open resource variants like Hadoop, HBase, Cassandra, Giraph..

















