1 Introduction

What is (big) data analytics? One can simply define it as the discovery of "models" for data to extract information, draw conclusions and make decisions. A "Model" can be one of several things:

- Statistical model which is the underlying distribution from which the data is drawn.
  Example: *given a set of real numbers, each one independently Gaussian distributed, estimate the mean and variance.* The model for data here is Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ where each data is its independent realization.

- Use the data as a training set for algorithms of machine learning, e.g., Bayes nets, support-vector machines, decision trees, etc.
  Example: ([LRU14]) In "Netflix challenge", the goal was to devise an algorithm that predicts the ranking of movies by users.

- Extract the most prominent features of the data and ignore the rest [LRU14, page 4].
  Example: Feature extraction, similarity, PCA

- Summarization of features
  Example: First example is Page rank (Google's web mining), probability that a random walker on the graph meets that page at any given time. Second example is clustering. Points that are close are summarized, e.g, by their clusters.

One should be careful about the effect of big data analytics. In large random data sets, unusual features occur which are the effect of purely random nature of data. This is called *Bonferroni's principle.*

Example ([LRU14, page. 6])**.** Find evil-doers by looking for people who both were in the same hotel on two different days. Here are the assumptions:

- 10^5 hotels
- Everyone goes to a hotel one day in 100
- 10^9 people
- People pick days and hotels at random independently
- Examine hotel records for 1000 days.

Probability that any two people visit a hotel on any given day is equal to $\frac{1}{100} \times \frac{1}{100}$. Probability that they pick the same hotel is $\frac{1}{10^4} \times \frac{1}{10^5} = 10^{-9}$. Probability that two people visit the same hotel on two different days are $10^{-9} \times 10^{-9} = 10^{-18}$.
Cardinality of the event space is: pairs of people $\binom{10^9}{2}$, pairs of days $\binom{10^3}{2}$. Expected number of evil-doing events, using $\binom{n}{2} \approx \frac{n^2}{2}$, is given by:

$$\binom{10^9}{2} \times \binom{10^3}{2} \times 10^{-18} \approx 5.10^{17}.5.10^5.10^{-18} = 25 \times 10^4 = 250000.$$

Below it is shortly discussed how to carry out computation on large data sets, although it will not be he focus of this lecture.

1.1 MapReduce and Hadoop

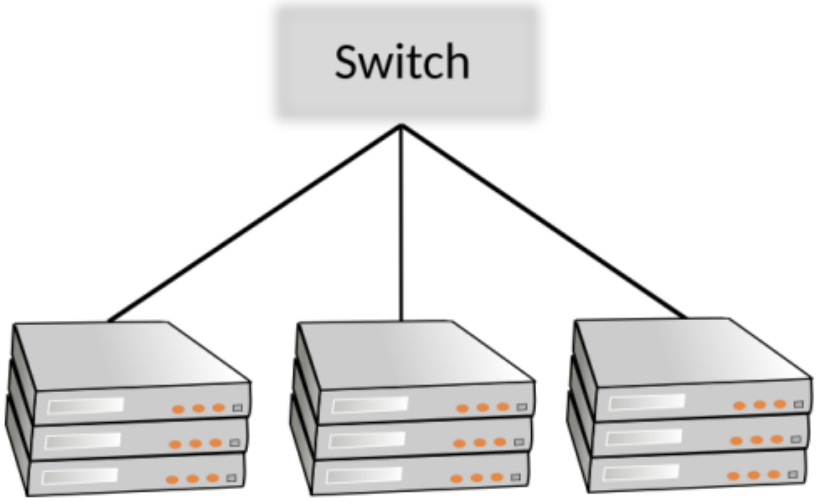

Figure 1.1: Racks of compute nodes

When the computation is to be performed on very large data sets, it is not efficient to fit the whole data in a data-base and perform the computations sequentially. The key idea is to use parallelism from "computing clusters", not a super computer, built of commodity hardware, connected by Ethernet or inexpensive switches.
The software stack consists of distributed file systems (DFS) and MapReduce. In a distributed file system Files are divided into chunks (typically 64 MB) and chunks are replicated, typically 3 times on different racks. There exists a file master mode or name mode with information where to find copies of files. Some of the implementations of DFS are GFS (Google file system), HDFS (Hadoop Distributed File System, Apache) and Cloud Store (open source DFS).
On the other hand MapReduce is the computing paradigm. In MapReduce, the system manages parallel execution and coordination of tasks. Two functions are written by users namely Map and Reduce. The advantage of this system is its robustness to hardware

failures and it is able to handle large datasets. MapReduce is implemented internally by Google.
The architecture of this system is such that compute nodes are stored on racks, each with its own processor and storage device. Many racks are connected by a switch as presented in Figure 1.1. They are connected by some fast network, interconnection by Gigabit Internet. The principles of this system are as follows. First, files must be stored redundantly to protect against failure of nodes. Second, computations must be divided into independent tasks. If one fails it can be restored without affecting others.
We discuss an example of implementation matrix-vector multiplication using MapReduce [LRU14].

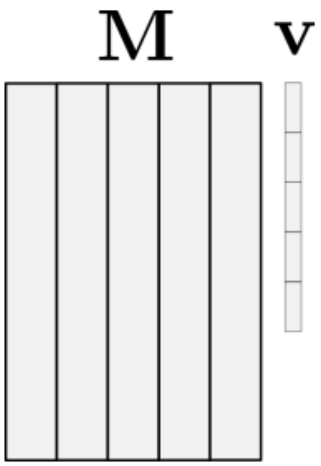

Figure 1.2: Matrix-Vector Multiplication

Example (Matrix-Vector Multiplication by MapReduce)**.** Suppose that the matrix $\mathbf{M} \in \mathbb{R}^{m\times n}$ and the vector $\mathbf{v} \in \mathbb{R}^n$ are given and the goal is to compute their multiplication $\mathbf{x} = \mathbf{M}\mathbf{v}$:

$$x_i = \sum_{j=1}^{n} m_{ij} v_j.$$

When n is large, say 10^7 then the direct computation requires the storage of the whole matrix in the storage which might not be efficient. Particularly in practice the matrix $\mathbf{M}$ can be sparse with say 10 or 15 non-zeros per row.
First the matrix and the vector is stored as the pairs $(i, j, m_{ij}$ and the vector is stored as (i, v_i). MapReduce consists of two main functions, Map function and Reduce function. To implement the multiplication using MapReduce, Map function produces a key-value pair to each entries of the matrix and the vector. To the entry m_{ij} the pair $(i, m_{ij}v_j)$ is associated where i is the key and $m_{ij}v_j$ is the pair. Note that it is assumed here that m is small enough to store the vector $\mathbf{v}$ in its entirety in the memory. The Reduce function receives all the key-value pairs, lists all pairs with key i and sum their values to get $(i, \sum_{j=1}^{n} m_{ij}x_j)$ which gives the ith entry of the product.
If the vector $\mathbf{v}$ cannot fit into the memory then the matrix $\mathbf{M}$ is divided into horizontal strips with certain width and the vector $\mathbf{v}$ is divided into vertical stripes with the same size

as the matrix stripes' width. Accordingly the multiplication can be divided into sub-tasks, each feasible using the MapReduce.

Example (Matrix-Matrix Multiplication by MapReduce)**.** Given two matrices $\mathbf{M} \in \mathbb{R}^{n \times m}$ and $\mathbf{N} \in \mathbb{R}^{m \times r}$, the goal is to compute $\mathbf{MN}$. Map function generates the following key-value pairs:

- For each element m_{ij} of $\mathbf{M}$ produce r key-value pairs $((i,k),(\mathbf{M},j,m_{ij}))$ for $k = 1,\ldots,r$.
- For each element n_{jk} of $\mathbf{N}$ produce n key-value pairs $((i,k),(\mathbf{N},j,n_{jk}))$ for $i = 1,\ldots,n$.

The Reduce function computes the multiplication as follows:

- For each key (i,k), find the values with the same j.
- Multiply m_{ij} and n_{jk} to get $m_{ij}n_{jk}$.
- Sum up all $m_{ij}n_{jk}$ over j to get $\sum_{j=1}^{m} m_{ij}n_{jk}$.

What exactly is Big Data?

Perhaps nothing will have as large an impact on advanced analytics in the coming years as the ongoing explosion of new and powerful data sources. When analyzing customers, for example, the day s of relying exclusively on demographics and sales history are past. Virtually every industry has at least one completely new data source coming online soon, if it isn't here already. Some of the data sources apply widely across industries; others are primarily relevant to a very small number of industries or niches. Many of these data sources fall under a new term that is receiving a lot of buzz: big data. Big data is sprouting up everywhere and using it appropriately will drive competitive advantage. Ignoring big data will put an organization at risk and cause it to fall behind the competition. To stay competitive, it is imperative that organizations aggressively pursue capturing and analyzing these new data sources to gain the insights that they offer. Analytic professionals have a lot of work to do! It won't be easy to incorporate big data alongside all the other data that has been used for analysis for years.

At first glance, the term seems rather vague, referring to something that is large and full of information. That description does indeed fit the bill, yet it provides no information on what Big Data really is. Big Data is often described as extremely large data sets that have grown beyond the ability to manage and analyze them with traditional data processing tools. Searching the Web for clues reveals an almost universal definition, shared by the majority of those promoting the ideology of Big Data, that can be condensed into something like this: Big Data defines a situation in which data sets have grown to such enormous sizes that conventional information technologies can no longer effectively handle either the size of the data set or the scale and growth of the data set. In other words, the data set has grown so large that it is difficult to manage and even harder to garner value out of it. The primary difficulties are the acquisition,

storage, searching, sharing, analytics, and visualization of data. There is much more to be said about what Big Data actually is. The concept has evolved to include not only the size of the data set but also the processes involved in leveraging the data. Big Data has even become synonymous with other business concepts, such as business intelligence, analytics, and data mining. Paradoxically, Big Data is not that new. Although massive data sets have been created in just the last two years, Big Data has its roots in the scientific and medical communities, where the complex analysis of massive amounts of data has been done for drug development, physics modeling, and other forms of research, all of which involve large data sets. Yet it is these very roots of the concept that have changed what Big Data has come to be.

THE ARRIVAL OF ANALYTICS

As analytics and research were applied to large data sets, scientists came to the conclusion that more is better—in this case, more data, more analysis, and more results. Researchers started to incorporate related data sets, unstructured data, archival data, and real-time data into the process, which in turn gave birth to what we now call Big Data. In the business world, Big Data is all about opportunity. According to IBM, every day we create 2.5 quintillion (2.5 3 1018) bytes of data, so much that 90 percent of the data in the world today has been created in the last two years. These data come from everywhere:

sensors used to gather climate information, posts to social media sites, digital pictures and videos posted online, transaction records of online purchases, and cell phone GPS signals, to name just a few. That is catalyst for Big Data, along with the more important fact that all of these data have intrinsic value that can be extrapolated using analytics, algorithms, and other techniques.

Big Data has already proved its importance and value in several areas. Organizations such as the National Oceanic and Atmospheric Administration (NOAA), the National Aeronautics and Space Administration (NASA), several pharmaceutical companies, and numerous energy companies have amassed huge amounts of data and now leverage Big Data technologies on a daily basis to extract value from them.

NOAA uses Big Data approaches to aid in climate, ecosystem, weather, and commercial research, while NASA uses Big Data for aeronautical and other research. Pharmaceutical companies and energy companies have leveraged Big Data for more tangible results, such as drug testing and geophysical analysis. The New York Times has used Big Data tools for text analysis and Web mining, while the Walt Disney Company uses them to correlate and understand customer behavior in all of its stores, theme parks, and Web properties.