Welcome to Algorithms for Big Data

Instructor: Hossein Jowhari

Department of Computer Science and Statistics
Faculty of Mathematics
K. N. Toosi University of Technology

Spring 2021

What do we mean by Big Data?

- No precise definition ☹

- " Big data is a field that treats ways to analyze, systematically extract information from, or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application software."

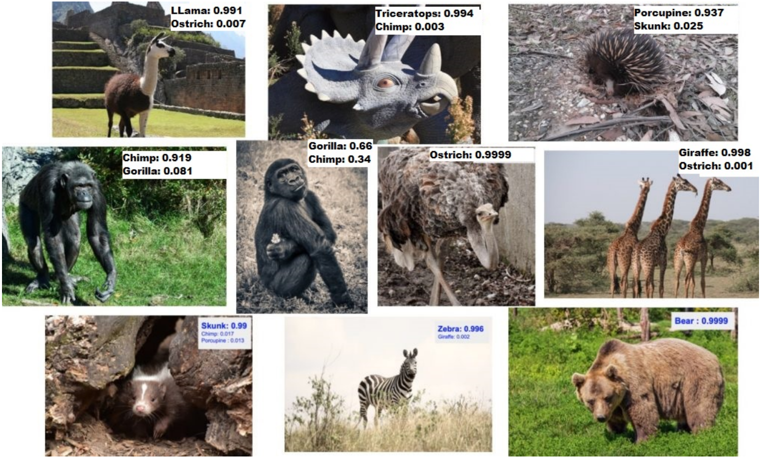
  Wikipedia

- *"Big data refers to things one can do at a large scale that cannot be done at a smaller one."*

  Big Data: A Revolution .. Mayer-Schonberger, Cukier

A Few Motivating Stores

Image Classification



A Few Motivating Stories

AlexNet: Image Classification

- AlexNet: A deep neural network for image classification
- ImageNet contest: 14 million images, 20,000 categories
- In 2012, AlexNet achieved 15% error rate on TOP-5 contest
- Nearly 11% lower than the runner-up
- High intensive computations using GPU (Graphical Processing Units)
- In 2018 the error rate has dropped to 2% using more involved networks and more GPU cards

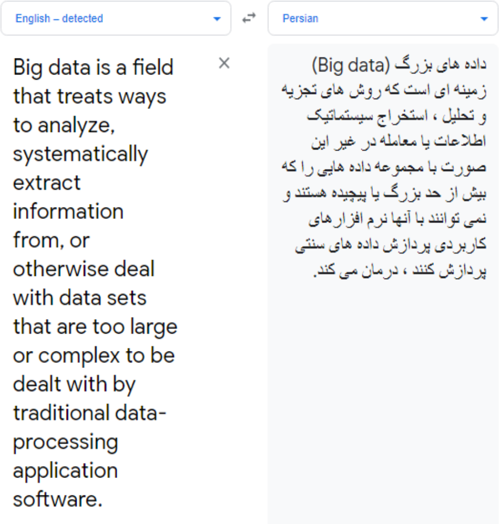
Alex Krizhevsky,

Ilya Sutskever,

Geoffrey E. Hinton

A Few Motivating Stores

Machine Translation



English – detected

Persian

Big data is a field that treats ways to analyze, systematically extract information from, or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application software.

داده های بزرگ (Big data) زمینه ای است که روش های تجزیه و تحلیل ، استخراج سیستماتیک اطلاعات یا معامله در غیر این صورت با مجموعه داده هایی را که بیش از حد بزرگ یا پیچیده هستند و نمی توانند با آنها نرم افزارهای کاربردی پردازش داده های سنتی پردازش کنند ، درمان می کند.

A Few Motivating Stores

Candid and Google Translate: Exact vs Messy

- In 1990, IBM's Candid project used 10 years of parliamentary published transcripts in French and English to create a statistical-based machine translation service. It used around 3 million sentence pairs.

- The project did not make much success and got terminated.

- In 2006, Google launched a machine translation project (statistical-based). Google Translate uses a huge text corpus gathered from the Internet (including reliable and unreliable sources). The corpus contains trillions of words.

- Google Translate has been a great success.

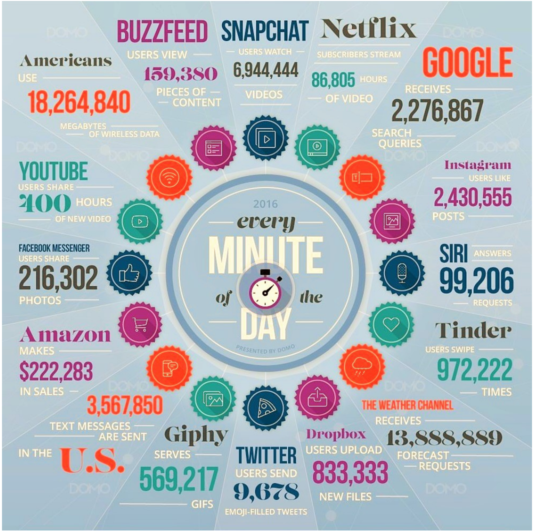
A Few Motivating Stories

Fraud detection: where sampling is not enough

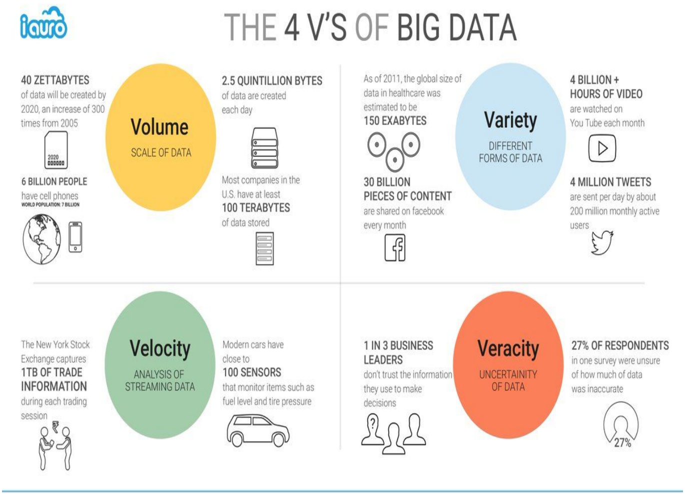
- Xoom is a firm that specializes in international money transfer.
- In 2011, Xoom discovers a slightly higher than average number of Discover Card transactions origination from New Jersey.
- The transactions came from a criminal group.
- To find anomalies one has to crunch all data rather than a sample.

A Few Motivating Stories

Data at High Velocity



The V's of Big Data



THE 4 V'S OF BIG DATA

Volume
SCALE OF DATA

40 ZETTABYTES of data will be created by 2020, an increase of 300 times from 2005

6 BILLION PEOPLE have cell phones
WORLD POPULATION: 7 BILLION

2.5 QUINTILLION BYTES of data are created each day

Most companies in the U.S. have at least **100 TERABYTES** of data stored

Variety
DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be **150 EXABYTES**

30 BILLION PIECES OF CONTENT are shared on facebook every month

4 BILLION + HOURS OF VIDEO are watched on You Tube each month

4 MILLION TWEETS are sent per day by about 200 million monthly active users

Velocity
ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures **1TB OF TRADE INFORMATION** during each trading session

Modern cars have close to **100 SENSORS** that monitor items such as fuel level and tire pressure

Veracity
UNCERTAINITY OF DATA

1 IN 3 BUSINESS LEADERS don't trust the information they use to make decisions

27% OF RESPONDENTS in one survey were unsure of how much of data was inaccurate

Reference : http://www.ibmbigdatahub.com/infographic/four-vs-big-data

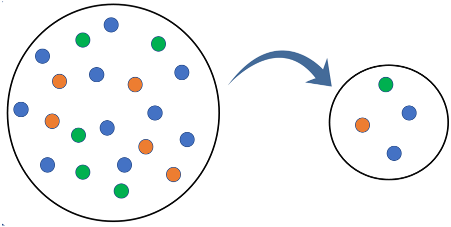
Big Data: A new paradigm for computation and businesses

- Traditional algorithms are not suitable for big data
- We need new computational models and algorithms to cope with big data.
- Computational Power + Big data ⇒ Novel things
- An analogy: A movie is fundamentally different from a frozen photograph.

Major Computation Models for Big Data

- Sampling: Sublinear Time Algorithms
- Parallel Processing: Parallel Algorithms
- Data Stream: Streaming Algorithms, Sketching

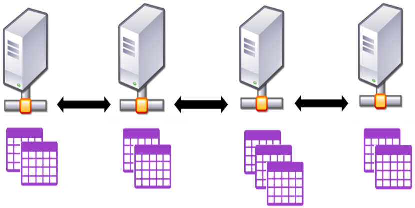
Sampling



- Sampling is always a great tool
- Not always applicable
- Small error margin requires large sample size

Computational Models for Big Data

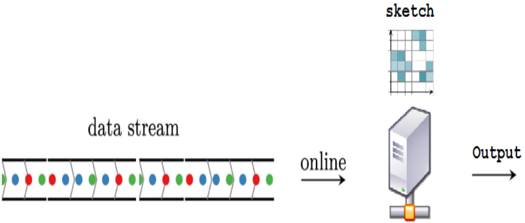
Parallel Processing



- Various Parallel Models: PRAM, MPC, Shared Memory, ...
- Suitable for stored data, offline computation
- Can produce exact answers

Computational Models for Big Data

Data Stream



- Computing over rapid online data
- Not enough memory to store the entire stream
- Fast per-item processing time needed
- Approximate answers, randomized algorithms

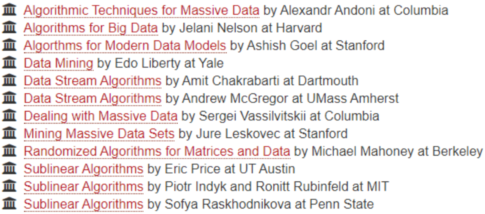
Mathematical Tools for Big Data

- Basic Probability Theory: Deviation Bounds (Markov, Chebyshev, Chernoff bounds, etc)
- Analysis of Algorithms (Time complexity, Space complexity)
- Dimensionality Reduction: JL lemma
- Lower bound techniques: Communication Complexity
- Linear Algebra

Course Information

- Reference: No textbook. Various papers.
- The course is mostly theoretical. Might introduce some software platforms (MapReduce,etc)
- Course Material: Slides, lecture notes
- Evaluation: Midterm, Final, Paper reading and Presentation

Similar Courses for Big Data



- Algorithmic Techniques for Massive Data by Alexandr Andoni at Columbia
- Algorithms for Big Data by Jelani Nelson at Harvard
- Algorithms for Modern Data Models by Ashish Goel at Stanford
- Data Mining by Edo Liberty at Yale
- Data Stream Algorithms by Amit Chakrabarti at Dartmouth
- Data Stream Algorithms by Andrew McGregor at UMass Amherst
- Dealing with Massive Data by Sergei Vassilvitskii at Columbia
- Mining Massive Data Sets by Jure Leskovec at Stanford
- Randomized Algorithms for Matrices and Data by Michael Mahoney at Berkeley
- Sublinear Algorithms by Eric Price at UT Austin
- Sublinear Algorithms by Piotr Indyk and Ronitt Rubinfeld at MIT
- Sublinear Algorithms by Sofya Raskhodnikova at Penn State

source: http://grigory.us/big-data-class.html