ALY 6110

Data Management and Big Data

Instructor: Valeriy Shevchenko

Northeastern

Lecture 1 (Week 1)

Introduction
to
Big Data

Learning Objectives

This session will explain:

- Introduction to a Big Data
- Big Data vs Traditional Data
- Four Vs Big Data
- Examples of four Vs

Introduction to Big Data

What is Big Data?

Introduction to Big Data

Big Data Definition by Oracle:

… is the **derivation of value**
from traditional relational database-driven business decision making,
augmented with **new sources**
of unstructured data.

,Oracle

Introduction to Big Data

Big Data Definition by Intel:

… **opportunities** emerge in organizations generating a median of 300 terabytes of data a week.

,**Intel**

Introduction to Big Data

Big Data Definition by Intel (cont'd):

… opportunities emerge in organizations generating a median of 300 terabytes of data a week.

The **most common forms** of data analyzed in this way are business transactions stored in relational databases, followed by documents, e-mail, sensor data, blogs, and social media..

,Intel

Introduction to Big Data

Big Data Definition by Microsoft:

… the term increasingly used to describe the **process** of **applying** serious computing power—latest in **machine learning** and artificial intelligence—

to seriously **massive** and often highly complex **sets** of information.”

‚Microsoft

Introduction to Big Data

Big Data Definition by NIST:

… is **data** which “exceed(s) the capacity or capability of current or conventional methods and systems.”

In other words, the notion of “big” is relative to the current standard of computation.

,***National Institute of Standards and Technology***

Introduction to Big Data

- **Big data** is exciting and can change health care policy decisions and the way we do business.

- **But** to harness these benefits, we need to address several challenges first.

Introduction to Big Data

- Organizations have a long **tradition** of capturing transactional data.

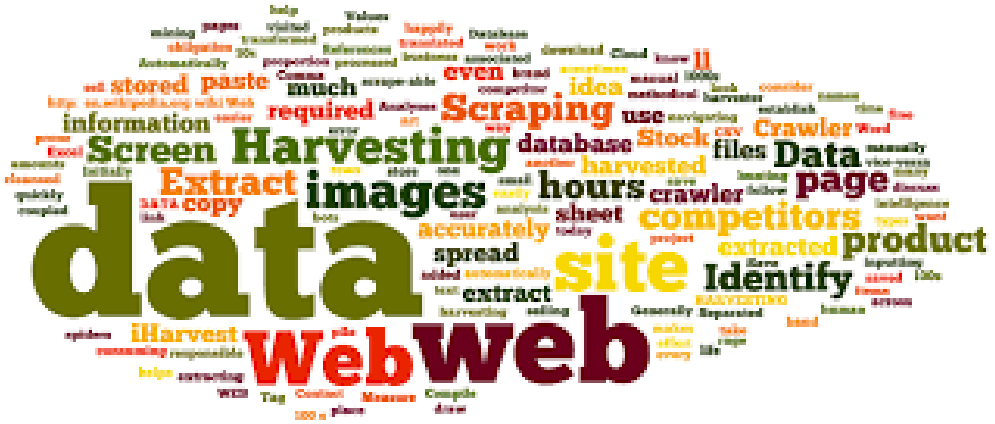
- Apart from that, organizations nowadays are **capturing additional data** from its operational environment at an increasingly fast speed.

Introduction to Big Data

Examples of Big Data

- **Web data**

Customer level web behavior data such as page views, searches, reading reviews, purchasing, can be captured. They can enhance performance in areas such as next best offer, churn modelling, customer segmentation and targeted advertisement.



Introduction to Big Data

Examples of Big Data

- **Text data**

E-mail, news, Facebook feeds, documents, etc) is one of the biggest and most widely applicable types of big data. The focus is typically on extracting key facts from the text and then use the facts as inputs to other analytic process (for example, automatically classify insurance claims as fraudulent or not.)

Introduction to Big Data

Examples of Big Data

- **Time and location data**

GPS and mobile phone as well as Wi-Fi connection makes time and location information a growing source of data. At an individual level, many organizations come to realize the power of knowing when their customers are at which location.

Introduction to Big Data

Examples of Big Data

- **Time and location data (cont.)**

Equally important is to look at time and location data at an aggregated level. As more individuals open up their time and location data more publicly, lots of interesting applications start to emerge. Time and location data is one of the most privacy-sensitive types of big data and should be treated with great caution.

Introduction to Big Data

Examples of Big Data

- **Smart grid and sensor data**

Sensor data are collected nowadays from cars, oil pipes, windmill turbines, and they are collected in extremely high frequency. Sensor data provides powerful information on the performance of engines and machinery. It enables diagnosis of problems more easily and faster development of mitigation procedures.

Big Data vs Traditional Data

Big Data vs Traditional Data

- **Big data** can be an entirely new source of data;

- ***For instance***, most of us have experience with **online shopping**. The transactions we execute are not fundamentally different transactions from what we would have done traditionally.

- An organization may **capture web transactions**, but they are really just more of the same transactions that have been captured for years (e.g. purchasing records).

- However, actually **capturing browsing behavior** (how do you navigate on the site, for instance) as customers execute a transaction creates fundamentally new data.

Big Data vs Traditional Data

- Sometimes one can argue that the speed of data feed has increase to such an extent that it qualifies as a new data source.

- *For instance,* your power meter has probably been read manually each month for years. Now we have a smart meter that automatically read it every 10 minutes.

- One are argue that it is the **same data**. It can also be argued that the frequency is so high now that it enables a very different, more in-depth level of analytics that such data is really a new data source.

Big Data vs Traditional Data

- Increasingly more semi-structured and unstructured data are coming in.
- Most traditional data sources are in the structured realm.
- Structure data are the ones like the receipts from your grocery store, the data on your salary slip, accounting information on the spreadsheet, and pretty much everything that can fit nicely in a relational database.
- Every piece of information included is known ahead of time, comes in a specified format and occurs in a specified order. This makes it easy to work with.

Four Vs Big Data

Big Data Four Vs

Four Vs refer to:

Volume

Velocity

Veracity

Variety

Big Data Four Vs

4 x Vs

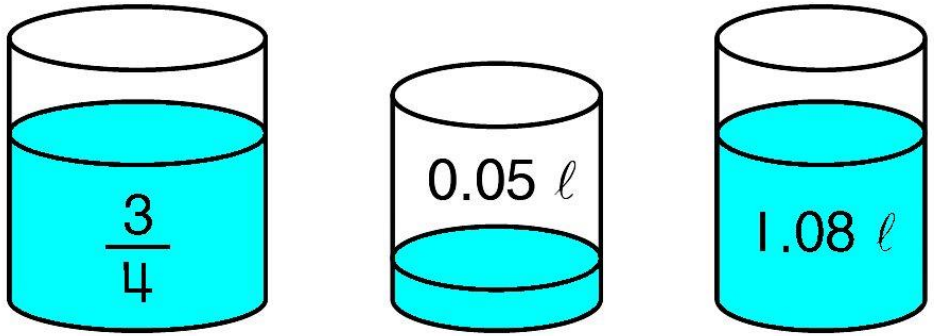
or

V V V V

The four Vs of big data –
captures the challenges
that you will face when dealing with big data

Big Data Four Vs - Volume

- Volume in big data refers to a large amount of data that you have to deal with.
- Nowadays, data is produced in a very large quantity.
- Think about surveillance cameras installed in a major city, such as Boston, LA, or New York.
- The number of these cameras might be in the thousands, and each of them is providing a constant video stream, resulting in massive amounts of data, even within one day.



Big Data Four Vs - Velocity

- Velocity refers to the speed at which the data arrives.
- Again, if we consider surveillance cameras, they provide data at constant speed and often at high resolution.
- This is a lot of data at high speeds.
- The internet also provides a vast amount of data at very high speed.
- A company's firewall system has to monitor the high-speed data trying to enter their network.

Big Data Four Vs – Velocity (cont'd)

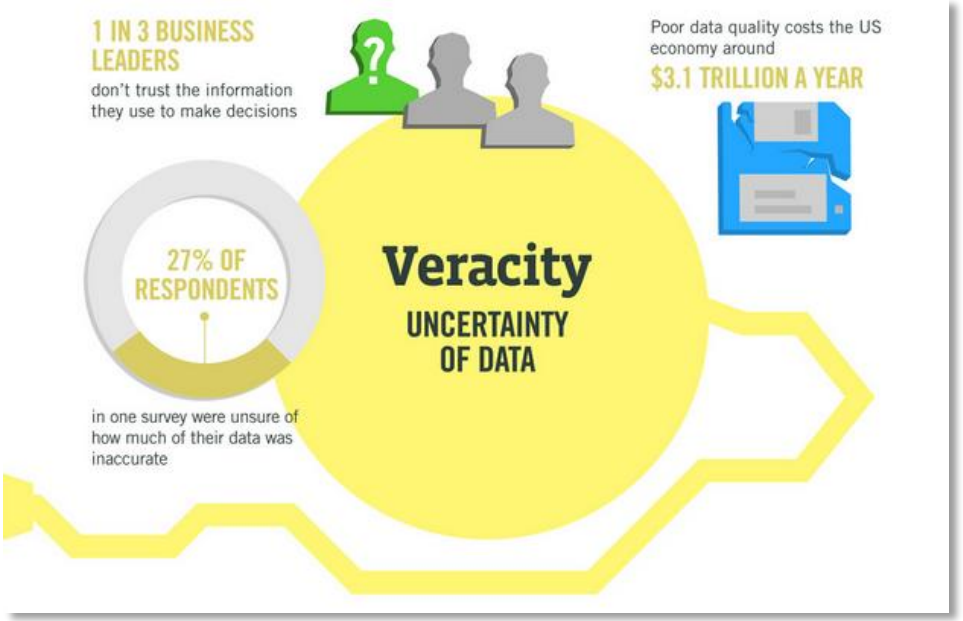
- In the context of cyber security, it's crucial to deal with the data of high velocity and to make sure that it's not a cyber attack.
- Due to the high velocity of data, it might not be feasible to store or check all of the data.
- To deal with this issue, we look at sampling techniques that store a representative fraction of the data.

Big Data Four Vs - Veracity

- Veracity refers to the uncertainty that comes with data.
- Often, data is not complete and can be noisy.
- So you cannot rely completely on all aspects of the data that arrives, and you have to deal with abnormalities of the data.
- Think of location services on phones. If every user provides their location, then this location is usually not precise, but within a range of, let's say, 100 yards.

Big Data Four Vs – Veracity (cont'd)

- The data may not be complete, as the GPS coordinates cannot be obtained at some locations.
- Dealing with this data often requires a data-cleaning process that reduces veracity.
- While this can remove abnormalities, it often still leaves you with incomplete data, and you would need to fill in the blanks when using it for your application.

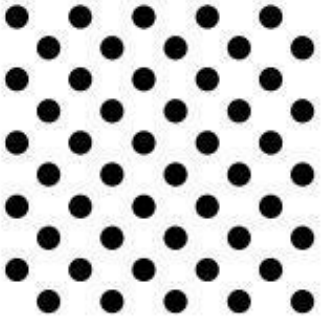


Big Data Four Vs - Variety

- The variety of big data refers to the different sources of data.
- Data can come in various forms - images, videos, audio, sensor data, and so on.
- For a specific application, you might have to integrate data from various sources.
- Often the data provided is unstructured and doesn't arrive in a coordinated way.
- Then, you have to rely on your different data sources.

Big Data Four Vs

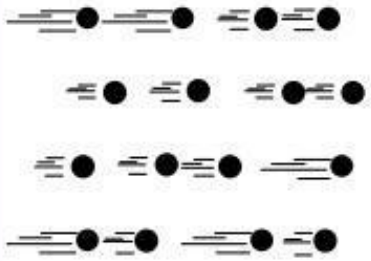
Volume



Data at Rest

Terabytes to exabytes of existing data to process

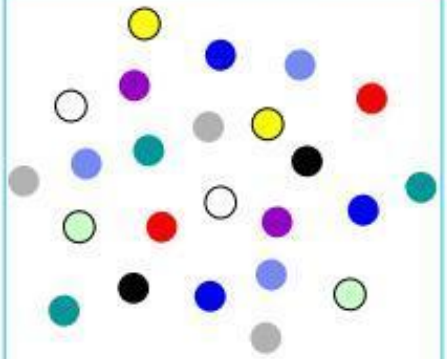
Velocity



Data in Motion

Streaming data, milliseconds to seconds to respond

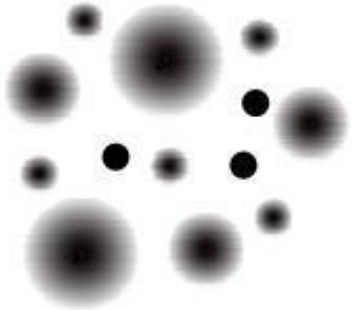
Variety



Data in Many Forms

Structured, unstructured, text, multimedia

Veracity*



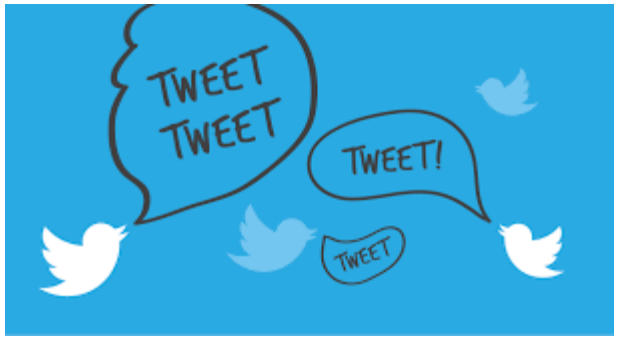
Data in Doubt

Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

Examples of four Vs

Examples of Four Vs – Social Media

- There are millions of people using Facebook and Twitter.
- All the data is produced in an online fashion arriving in the form of a data stream.
- Users post a variety of data online on Facebook, such as text, images, videos. Similarly, Twitter has short text messages.
- The data is high volume and arrives with high velocity at the Facebook/Twitter servers.
- Users may be tagged by their location using GPS coordinates.
- These coordinates are usually imprecise leading to veracity of the data.

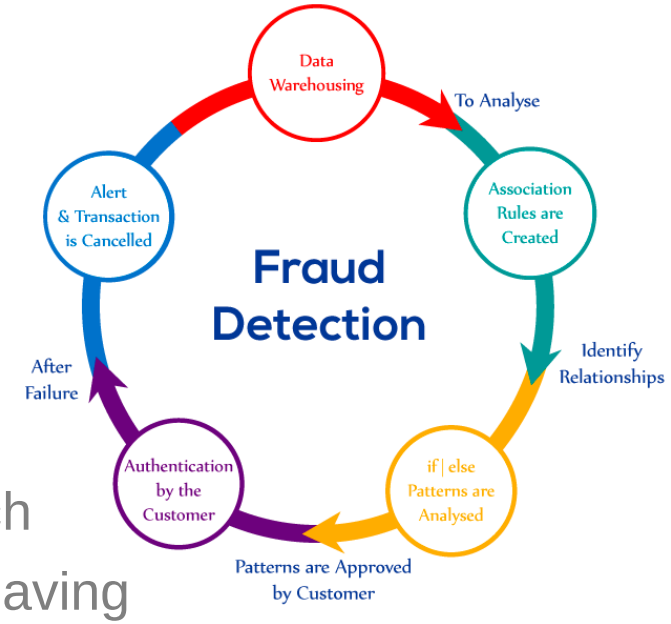


Examples of Four Vs - Banking

Fraud detection in banking transactions

- Banking produces millions of transactions per day. These transactions have to be processed safely and reliably. Thinking about a bank's transactions over a month results in a vast volume of data.
- Fraud detection refers to finding bogus transactions that have been triggered by criminals. This can be by using a stolen credit card or even only its details.

  You see that for fraud detection you would have to deal with large volumes of data, each transaction arriving rapidly, and a decision having to be made as soon as a transaction arrives.



Big Data Four Vs – Banking (cont'd)

Fraud detection in banking transactions

- There are some indicators that can be used to identify fraud, for example a credit card used at an ATM in one country when all other transaction in the previous 2 days have been in another country.

- Finding frauds is hard and the information used to stop a transaction is usually not 100% reliable. You might even have observed this yourself when you tried to use your credit card in a different country and the card was rejected although you were the legitimate user of the card.

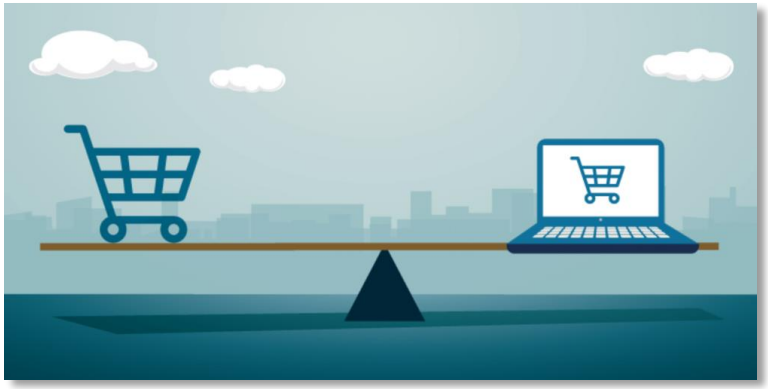
Big Data Four Vs – Online Stores

- Online stores such as Amazon have millions of potential customers that buy a large variety of items from their online servers. These customers produce a very large number of transactions within a short time period.

- Mining these transactions to extract useful information (for example to optimize advertising) has to deal with the large amount of users and the variety of items that they have bought. Making a recommendation to one particular user takes into account what the user has bought so far.

Big Data Four Vs – Online Stores (cont'd)

- The knowledge gathered about a customer is incomplete and a recommendation system has to rely on the imprecise information that it can obtain from the transaction data of customers and their behavior on the online store page.



The main technological components in a Big Data ecosystem

Big Data Ecosystem

- Traditionally, data are stored in relational database (for example a CRM system for customer data, a supply chain management software for vendor related information) and some of these data are extracted periodically from the operational database, transformed and loaded into data warehouse for reporting and further analysis.

- This is typically in the realm of BI (Business Intelligence).

- Such process and tool set fall short when dealing with big data.

Big Data Ecosystem

- For an example, one of the largest publicly discussed Hadoop cluster (Yahoo's) was at 455 petabytes in 2014 and it's grown since then.
- There simply is no parallel relational databases or data warehouse that have come even close to those kinds of numbers.
- Another sweet spot for Hadoop (over relational technology) is when data comes in unstructured format:
  - Audio
  - Video
  - Text

Big Data Ecosystem

- It is worthwhile to mention that there is a general misconception that new technology, such as **Hadoop** is replacing other technologies, such as relational **database**.

- ***It is not the case.*** It is more likely that they are being added ***alongside*** each other.

- The sweet spot for a massively parallel relational platform for instance, is dealing with ***high-value transactional data*** that is already structured, that needs to support a large amount of user and applications that ask repeated questions of known data (where a fixed schema and optimization pays off) with enterprise level security and performance guarantee.

Big Data Ecosystem

- It is often called the **Hadoop eco-system** when discussing the various lays of technologies used to deal with big data.

- For a complete list, please refer to https://hadoopecosystemtable.github.io/.

- For instance, stack might look like:
  - **Amazon** web service for infrastructure (in the Cloud and pay as you go).
  - Apache **HDFS** (Hadoop Distributed File System) for distributed file system.
  - **MapReduce** or **Spark** for distributed programming model.
  - **Cassandra** or **HBase** for non-relational distributed database management system.
  - **Hive** for execute SQL on top of Hadoop.
  - **Mahout** for Machine learning library and math library, on top of **MapReduce**.
  - **R** for data analytics and visualization.

Big Data Ecosystem

Analytical techniques

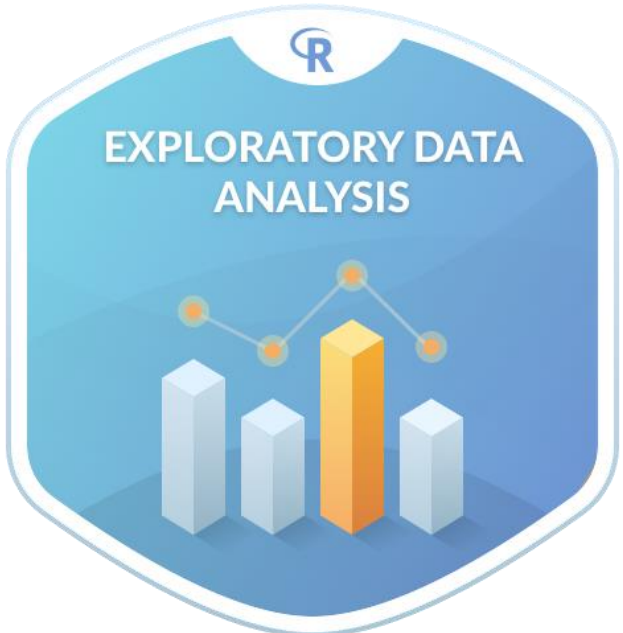
Most of the widely used analytical techniques falls into one of the following categories:

- Statistical methods:
  - Forecasting.
  - Regression analysis.
- Database querying.
- Data warehouse.
- Machine learning and data mining.

Big Data Ecosystem

Analytical techniques (cont'd)

- **Visualization**:
  - When analysis is done, the results need to be **communicated** to various stakeholders.
  - One of the hardest parts of an analysis is producing **quality** supporting **graphics**.
  - Conversely, a good graph is one of the best ways to **present findings**.
  - Graphics are used primarily for two reasons:
    - Exploratory data analysis.
    - Presenting results.



ALY6110

Data Management and Big Data

Q & A

Instructor: Valeriy Shevchenko
v.shevchenko@northeastern.edu

References

https://previews.123rf.com/images/ppbig/ppbig1611/ppbig161100175/67808702-data-analytics-icons.jpg

https://encrypted-tbn0.gstatic.com/images?q=tbn:ANd9GcTc4WtWcH2_dHxmFfJ3dfxNMBzFUQON69KEeKAN7W8_5zqB77B2

https://encrypted-tbn0.gstatic.com/images?q=tbn:ANd9GcS5Db36zmq1D47ksG2zS0sbF8QXGgDB_iXU-wQKBWoxZ3Y6jwPq

https://thumbs.dreamstime.com/z/

http://socialbarrel.com/wp-content/uploads/2018/03

https://securecdn.pymnts.com/wp-content/uploads/2017/01

https://static1.squarespace.com/static

http://www.montblanc-penssale.com

http://wiki.huihoo.com/images/d/d5

https://hadoopecosystemtable.github.io/

https://assets.datacamp.com/production/