# INTRODUZIONE AL BIG DATA

## Outline

- What is Big Data?
- How does it look like?
- The Data life-cycle

## WHAT IS BIG DATA?


Big data is defined as

"Extracting insight from an immense volume, variety, and velocity of data, in context, beyond what was previously possible."

## The original need

Digital density is the value data brings to the economy: through the integration of the interactions between single data points (i.e. connections)

## Volume

Organizations today must now handle

- more data
- in more forms
- across a greater number of systems

Storage cost dropping

## Variety

Data is now much more heterogeneous

- → 20 % is structured
- → 80 % is unstructured

Unstructured grows 15x the rate of structured information

## Velocity

It is the rate of generated and consumed information, or the ability to stream or process information flowing at a rapid rate, frequently in real time

## Veracity

Veracity is the truthfulness of data, or data integrity

- → Provenance
- → History of changes
- → Retention
- → Relevance

## Vulnerability

Data is to be kept protected and security is not to be an afterthought

Multilevel security controls are even more important than before

## VALUE

Organizations in the top third of their industry regarding the usage of data-driven decision making were, on average, 5% more profitable than their competitors. (HBR)

## Review Questions

- → What ARE the most important "V"s in Big Data ?
- → Why ?

## HOW DOES BIG DATA LOOK LIKE ?


## Social Data: Twitter

"created_at":"Thu Oct 08 18:43:54 +0000 2015" , "id": 652192712981118977, "id_str":"652192712981118977", "text":"twenty days to find a new #mayor for #rome" , "source":"\u003ca href=\"http:\/\/twitter.com\/download\/android\" rel=\"nofollow\"\u003eTwitter for Android\u003c\/a\u003e", "truncated":false, "in_reply_to_status_id":null ……

## Social Data: Twitter

…,"user":{"id":847963362,"id_str":"847963362", "name":"Angelo Ubaldo" , "screen_name" : "ubaldo_angelo" , "location":null, "url":null, "description":"Il bello dei sistemi umani \u00e8 che progrediscono da soli, senza bisogno dei progressisti" ,"protected":false, "verified":false, "followers_count":793, "friends_count":805 ,"listed_count":4,"favourites_count":1,"statuses_count":…..

## Machine Generated Data

METAR LBBG 041600Z 12003MPS 310V290 1400 R04/P1500 R22/P1500U +S BK 022 OVC050 M04/M07 Q1020 OSIG 9949//91=

## GPS and Spatial Data

loc1,37.786216,-122.409074,500
loc2,37.791134,-122.398774,250
loc3,37.787776,-122.40122,300

## Streaming Data

Streaming takes virtually any kind of data and makes it available in a continuous near real-time feed: Twitter firehose is an example

## Review Questions

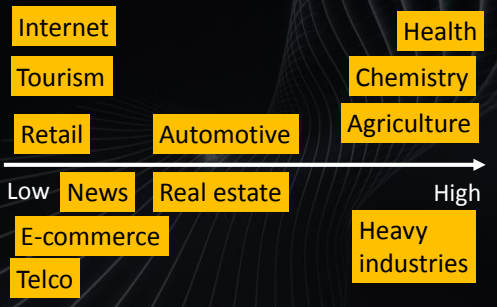
- What are typical formats for "big data" type information?
- Explain streaming data and why it is important

## THE DATA LIFE CYCLE

- Data is worth less as it ages
- Data value decays at different rates (half-life)
- Different factors impact the value decay of various categories

## Industries and Data HL


## Industries and Data HL




## Data Knowledge hierarchy

- Prescriptive
- Predictive
- Descriptive

## Knowledge hierarchy: descriptive analytics

Descriptive analytics, i.e. reporting, aka BI

- "the simplest class of analytics," one that allows you to condense big data into smaller, more useful nuggets of information

- The purpose of descriptive analytics is to summarize what happened. More than 80% of business analytics - including social analytics - are descriptive

Example descriptive analytics

- number of posts, mentions, fans, followers, page views, kudos, +1s, check-ins, pins, etc.
- Revenues per sales rep last month

## Knowledge hierarchy: predictive analytics

- Predictive analytics is the next step up the knowledge path
- It utilizes a variety of statistical, modeling, data mining, and machine learning techniques to study recent and historical data

Example predictive analytics:

- The input to the model is plain text, and the output of that model is a sentiment score, whether it's positive, negative, or something between +1 or -1

## Knowledge hierarchy: prescriptive analytics

The emerging technology of prescriptive analytics goes beyond descriptive and predictive models by recommending one or more courses of action - and showing the likely outcome of each decision

- It's basically when we need to prescribe an action, so the business decision-maker can take this information and act
- It recommends the best course of action for any pre-specified outcome

## Review questions

- Explain why data is never in a "stable state"
- What are the differences between descriptive, predictive and prescriptive analytics ?

## Exercise

- List the key data elements available in Trip Advisor
- Draw them on a chart in terms of their HL
- Identify the missing data elements

## SUMMARY QUESTION

- What ARE the most important "V"'s in Big Data ?
- Why ?
- What are typical formats for "big data" type information?

- Explain streaming data and why it is important
- Explain why data is never in a "stable state"
- What are the differences between descriptive, predictive and prescriptive analytics ?