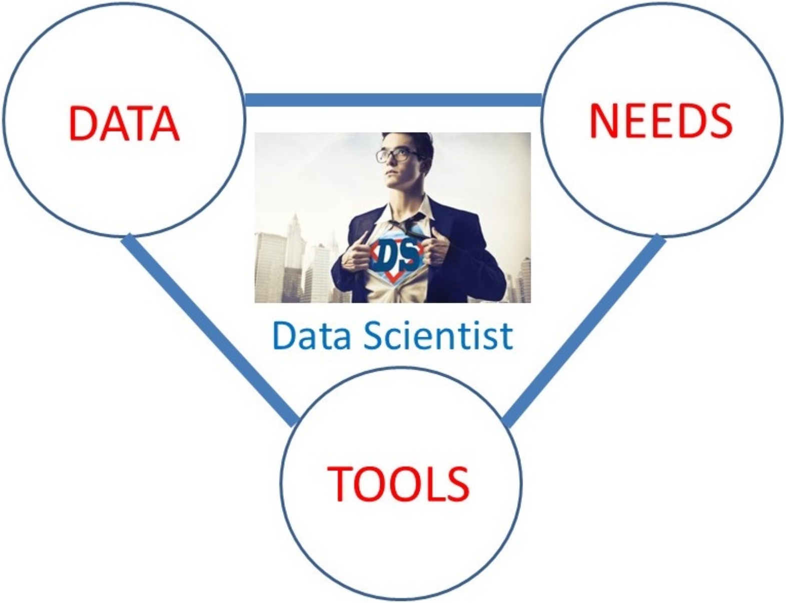
DATA MINING

2016/2017

Will become *Big Data Computing* in 2017/2018



DATA

NEEDS

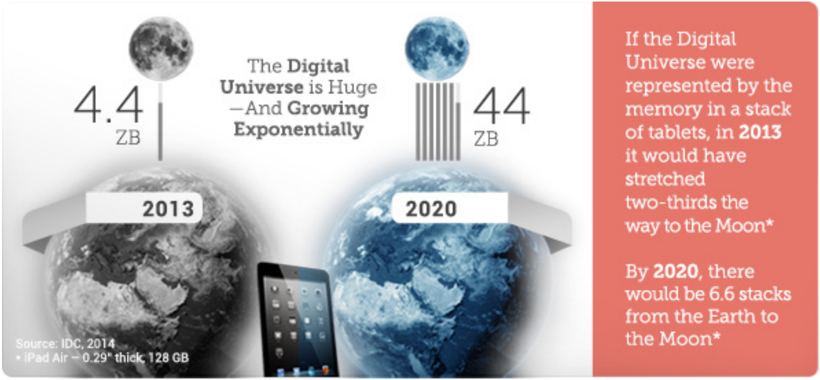
Data Scientist

TOOLS

DATA

1. Technological progress
    - Growth of storage capacity
    - Growth of communication bandwidth
    - Growth of computing power
    - Reduction of ICT costs
2. Digital universe:
    - Integration of digital technologies in *all sorts* of human activities
    - Scientific research: biology (e.g., genomics); physics (e.g., particle physics, LHC); astronomy (e.g., wide-field survey telescopes); climate monitoring; etc.
    - Exponential growth of data (doubles every 2 years)
3. Data can be either unstructured (e.g., textual data) or structured (e.g., database records, networks)

DATA: digital universe



If the Digital Universe were represented by the memory in a stack of tablets, in **2013** it would have stretched two-thirds the way to the Moon*

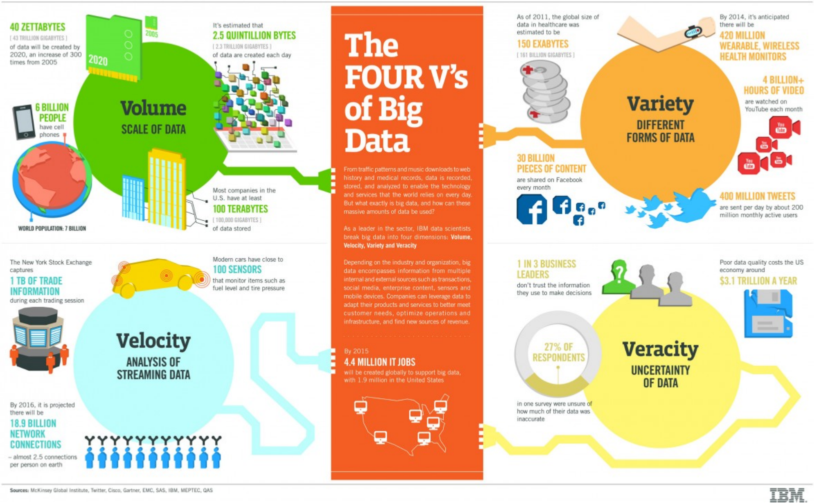
By **2020**, there would be 6.6 stacks from the Earth to the Moon*

Source: IDC, 2014
• iPad Air – 0.29" thick, 128 GB

Source: *The Digital Universe of Opportunities*, by ICD (2014)

1 ZettaByte (ZB) $= 10^{21}$B

DATA: the four V's



The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015 **4.4 MILLION IT JOBS** will be created globally to support big data, with 1.9 million in the United States

Volume
SCALE OF DATA

40 ZETTABYTES [43 TRILLION GIGABYTES] of data will be created by 2020, an increase of 300 times from 2005

6 BILLION PEOPLE have cell phones

WORLD POPULATION: 7 BILLION

It's estimated that **2.5 QUINTILLION BYTES** [2.3 TRILLION GIGABYTES] of data are created each day

Most companies in the U.S. have at least **100 TERABYTES** [100,000 GIGABYTES] of data stored

Velocity
ANALYSIS OF STREAMING DATA

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

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

By 2016, it is projected there will be **18.9 BILLION NETWORK CONNECTIONS** – almost 2.5 connections per person on earth

Variety
DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be **150 EXABYTES** [161 BILLION GIGABYTES]

By 2014, it's anticipated there will be **420 MILLION WEARABLE, WIRELESS HEALTH MONITORS**

4 BILLION+ HOURS OF VIDEO are watched on YouTube each month

30 BILLION PIECES OF CONTENT are shared on Facebook every month

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

Veracity
UNCERTAINTY OF DATA

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

Poor data quality costs the US economy around **$3.1 TRILLION A YEAR**

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

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

Source: IBM Big Data & Analytics Hub

NEEDS

- Either predetermined or induced by avilability of data/tools
- Examples:
    - Retailing: product improvement; recommendation systems
    - Banking/Finance: fraud detection; risk prediction; financial forecast
    - Telecommunications: user profiling; quality-of-service improvement
    - Science: validation of methods/results
    - Medicine: diagnosis/therapy improvement
    - Social impact: analysis of social networks; smart cities

TOOLS

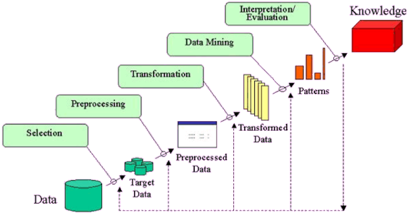
- Statistics
- Machine Learning
- Algorithmics: tradeoffs between accuracy and space-time efficiency; novel computing/programming frameworks
- High-performance computing
- Natural Language Processing
- Visualization

What is Data Mining?

Given a potentially large dataset, discover (through automatic procedues) patterns, models, properties, which are

- *Valid*
- *Useful*
- *Unexpected* (i.e., not explicitly contained in the data) and *previously unknown*
- *Understandable* (to humans)

Process of Knowledge Discovery in Databases



Presentation of the Course

What will we learn?

1. A sample of key primitives for data analysis
   - Rigorous setting
   - Algorithmic solutions with focus on large inputs
2. Novel computing/programming frameworks for (large) data analysis: theory and practice

More specific contents

1. Association Analysis
2. Computational Frameworks: MapReduce/Spark, Streaming
3. Clustering
4. Graph Analytics
5. Mining primitives for data streams
6. Similarity Search

Administative issues

Evaluation

- Written test: 75%
- Project: 25%

Online tools

- **Moodle:** registration (by march 12, password: DM1617), forum, results of written tests
- **Uniweb:** official exam lists, final grades
- **Course web site** (http://www.dei.unipd.it/∼capri/DM): complete info, slides, links, solutions