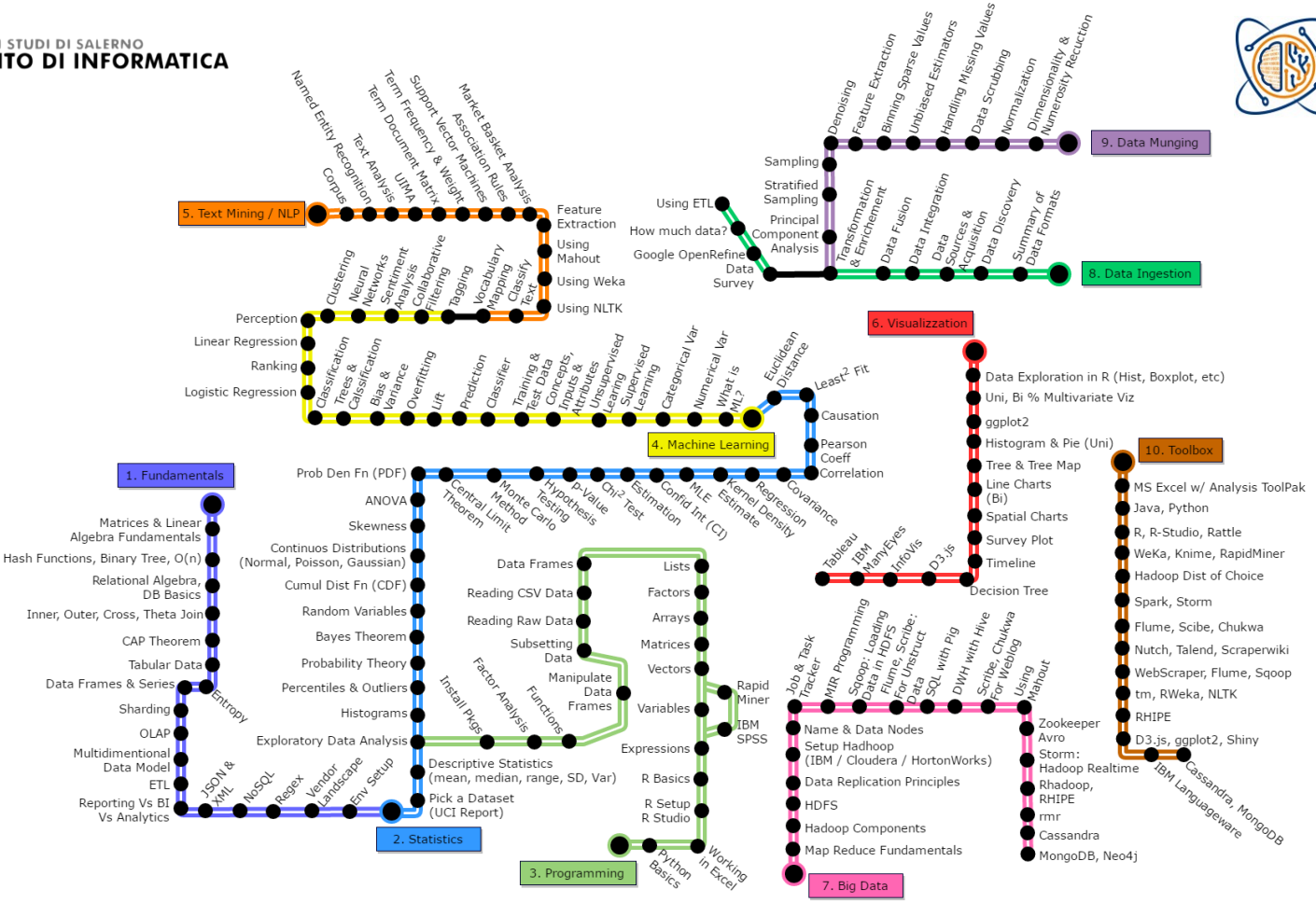
UNIVERSITÀ DEGLI STUDI DI SALERNO
DIPARTIMENTO DI INFORMATICA



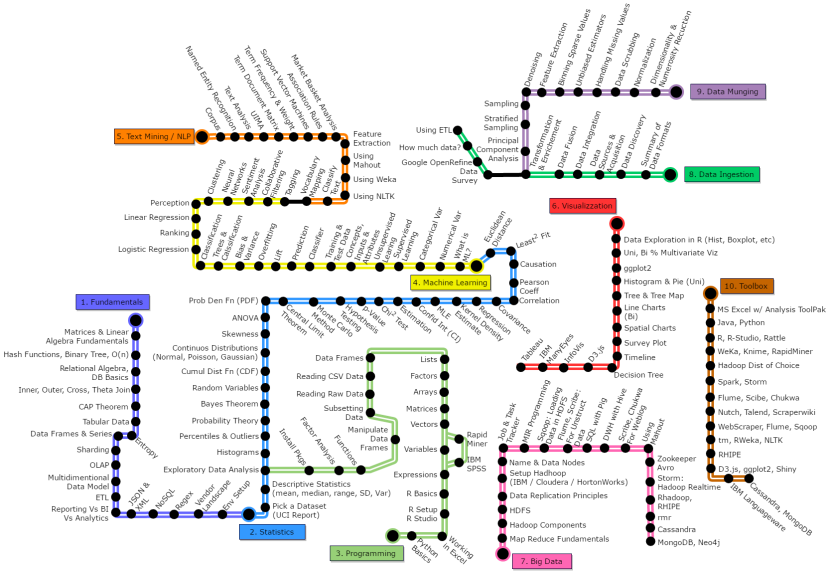
Fondamenti di Data Science & Machine Learning

Panoramica Introduttiva

Prof. Giuseppe Polese, aa 2024-25

Overview

1. **Organizzazione del corso**
2. Principali argomenti trattati
3. Modalità d'esame
4. Che cos'è la Data Science
5. (Big) Data
   - Data Sources
   - Technologies



Ricevimento

- Online su Team *Ricevimento Prof. Giuseppe Polese (Codice Team **n31tfkf**)*
- Comunicazioni online tramite Piattaforma E-learning:
  - elearning.informatica.unisa.it/el-platform/
- Email: gpolese@unisa.it
- Orario di ricevimento (A distanza):
  - Lunedì 15:00-16:30;
  - Mercoledì 15:00-16:30;

Materiale Didattico

- Articoli e dispense forniti dal docente
- Testi:
  - Aurélien Géron, "***Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*** ", O'Reilly ed., 2023.
  - Jure Leskovec, Anand Rajaraman, Jeffrey D. Ullman, "***Mining of Massive Datasets***", 3^ Edizione, Cambridge University Press, 2020.
  - C. Bishop, ***Deep Learning: Foundations and Concepts***, *Springer Nature Switzerland, 2023.*
  - *Chirag Shah,* ***A Hands-On Introduction To Data Science****, Cambridge University Press, 2020.*
  - *P. Deitel, H. Deitel,* ***Introduzione A Python*** *– Per L'informatica E La Data Science, Pearson 2021.*
  - Foster Provost, Tom Fawcett, ***Data Science for Business: What you need to know about data mining and data-analytic thinking,*** O'Reilly ed, 2013.

Obiettivi formativi del corso

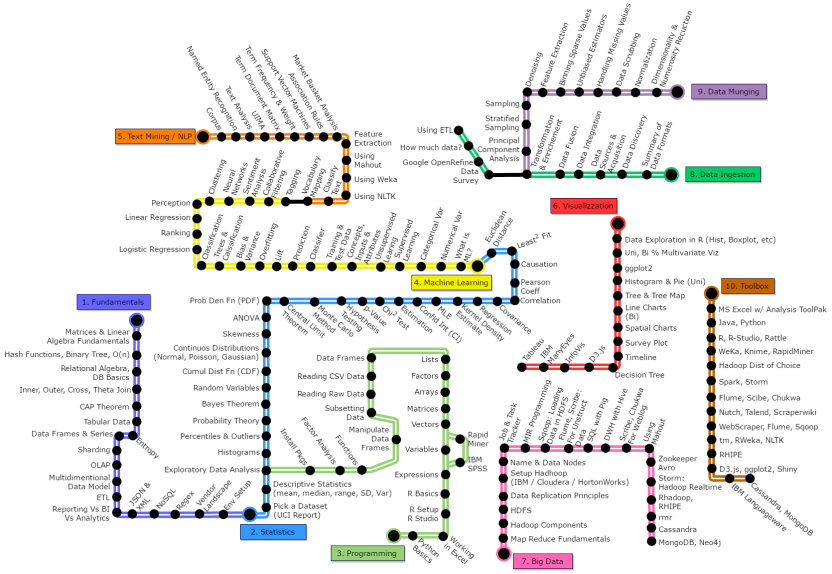
- L'insegnamento mira a fornire le competenze metodologiche e tecnologiche necessarie per estrarre conoscenza da grossi volumi di dati, mediante tecniche di Data Mining e Machine Learning, utilizzando opportune strategie di visualizzazione dei risultati.
- **Prerequisiti:** fondamenti di data management, sistemi distribuiti, paradigma ad oggetti ed un linguaggio di programmazione.

Modalità d'Esame

- L'esame prevede una prova scritta (o una prova in itinere a metà corso) ed una prova orale.
- Opzionalmente gli studenti possono sviluppare un progetto, individualmente o in gruppi di massimo 3 persone.
- Il voto finale scaturisce dalla media dei voti conseguiti alla prova scritta (o a quella in itinere) ed a quella orale, con la possibilità di incrementare il punteggio così ottenuto fino a 3 punti, tramite lo sviluppo del progetto.

Overview

1. Organizzazione del corso
2. **Principali argomenti trattati**
3. Modalità d'esame
4. Che cos'è la Data Science
5. (Big) Data
   - Data Sources
   - Technologies



Principali Argomenti Trattati

- Panoramica sui Big Data
- Data Curation & Data Quality
- Estrazione di Conoscenza da Big Data
- Machine Learning
- Reti neurali
- **Laboratorio**
  - Linguaggio Python
  - Piattaforma Pentaho (Weka)

Data Curation

- Data Profiling
- Dipendenze funzionali approssimate e data quality
- Data Cleaning
- Integrazione dati da sorgenti multiple

Estrazione Conoscenza da Big Data

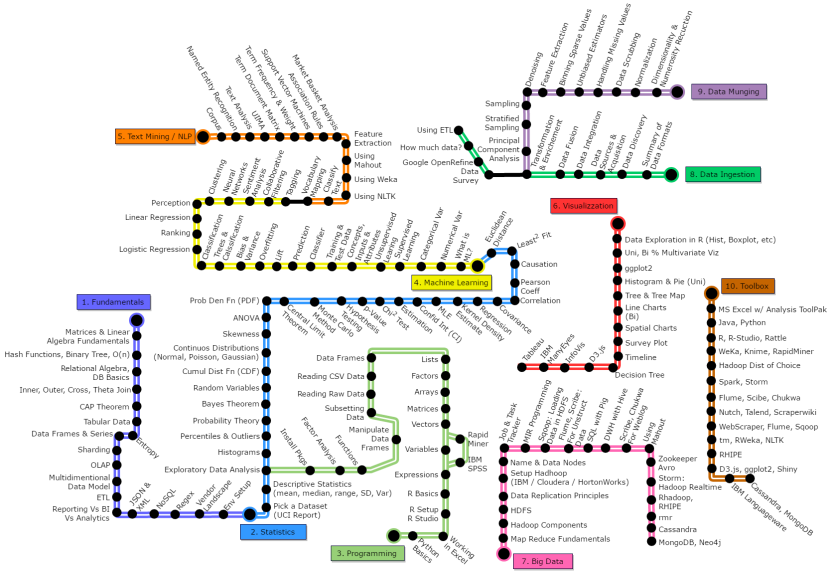
- Data Mining
- Mapreduce
- Funzioni di similarità
- Classificazione
- Clustering

Machine Learning

- Alberi di decisione
- Ensemble learning and Random Forest
- Reti Neurali
- Reti Convoluzionali
- Reti Ricorrenti
- Scikit
- Tensor flow

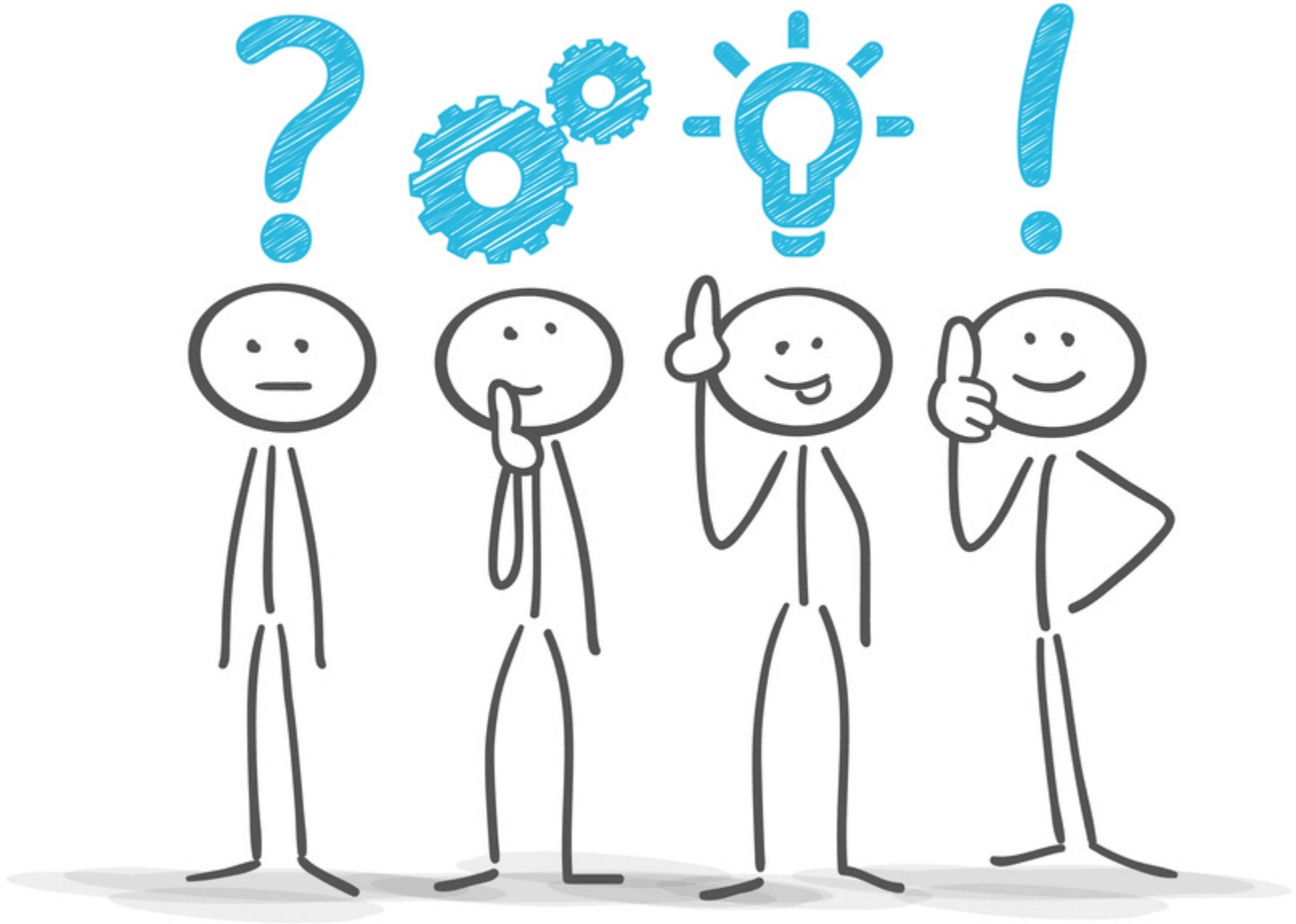
Overview

1. Organizzazione del corso
2. Principali argomenti trattati
3. **Modalità d'esame**
4. Che cos'è la Data Science
5. (Big) Data
   - Data Sources
   - Technologies



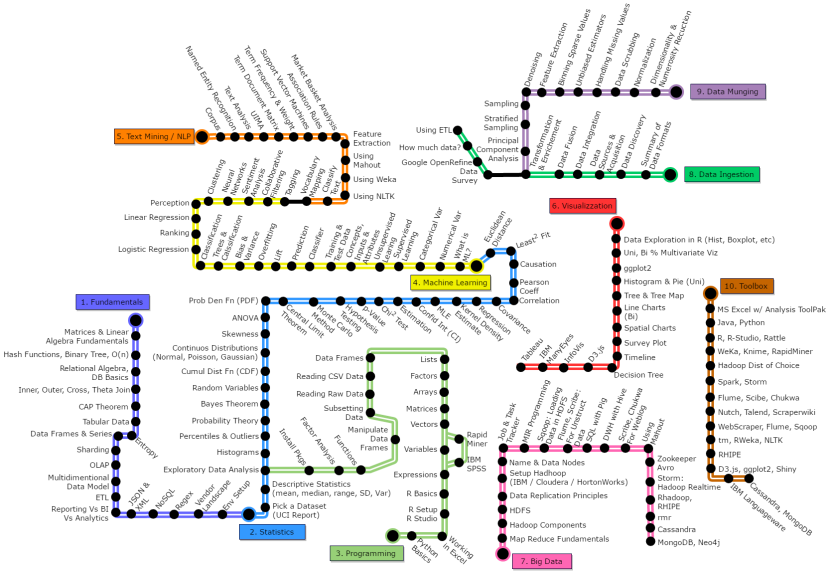
Modalità D'Esame

- **Prova Scritta o prova in itinere (50%)**
- **Progetto Facoltativo (fino a 3 punti)**:
  - Progetto di Corso
  - Implementazione o sperimentazione di un tool
  - Survey
  - Ricerca
  - *Possibilità Progetto-Tesi, Progetto-Tesi Esterna, Progetto combinato con corsi AI, DEL, NLP, SAD*
- **Esame orale (50%)**



Overview

1. Organizzazione del corso
2. Principali argomenti trattati
3. Modalità d'esame
4. **Che cos'è la Data Science**
5. (Big) Data
   - Data Sources
   - Technologies

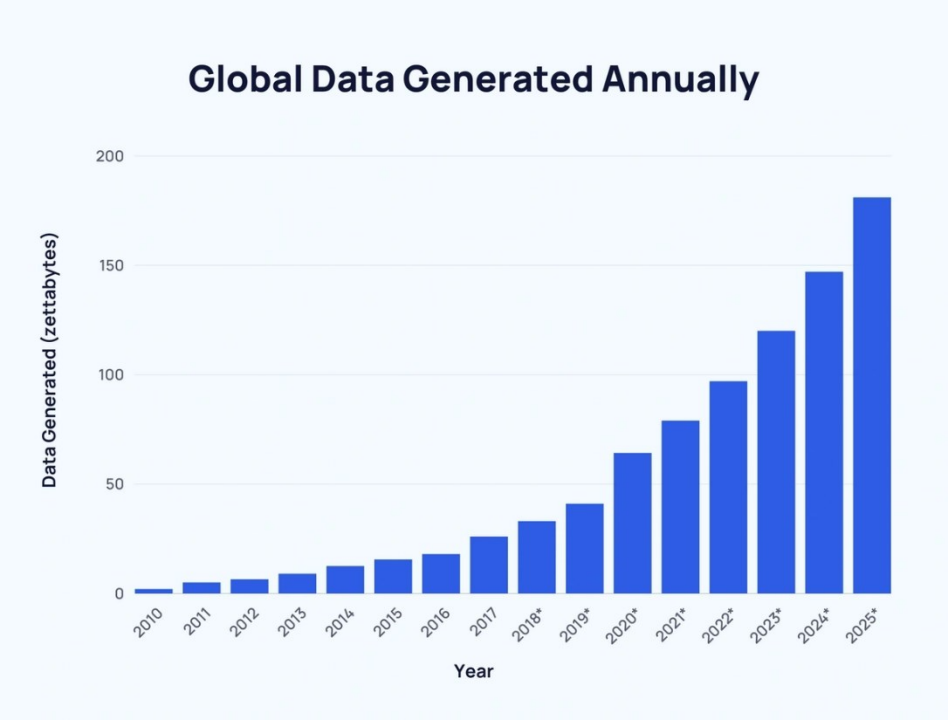


From Wikipedia…..

- Data Science is an interdisciplinary field about scientific methods, processes, and systems to extract knowledge or insights from data in various forms…….
- It employs techniques and theories drawn from many fields within the broad areas of mathematics, statistics, information science, and computer science, in particular from the subdomains of machine learning, classification, cluster analysis, data mining, databases, and visualization.

Data generation: Annual Growth

+4500% ANNUAL DATA GENERATION from 2010 to 2025(ESTIMATED)



- Estimated around 147 zettabytes of data generated[1] in 2024 , 12 zettabytes (ZB) per month, 2.8 ZB per week, or 0.4 ZB every day
- 181 zettabytes of data will be generated in 2025
- Videos account for over half of internet data traffic

[1] Includes data that is newly generated, captured, copied, or consumed

source Statista, Bernard Marr & Co. 2024

Data generation: Annual Growth(2)

- It is estimated that 90% of the world's data was generated in the last two years alone
- Video is responsible for over half (53.72%) of all global data traffic
- Social media is brimming with video content.
- TikTok is entirely based on videos and continues to grow its user base year-over-year
- Facebook’s 51% of content shared on the platform is video-based.

Internet Data Traffic by Category

Category	Proportion of Internet Data Traffic
Video	53.72%
Social	12.69%
Gaming	9.86%
Web browsing	5.67%
Messaging	5.35%
Marketplace	4.54%
File sharing	3.74%
Cloud	2.73%
VPN	1.39%
Audio	0.31%

Internet Data Trafic by Type

Type of Media	Amount per Minute	Amount per Day
Emails sent	231.4 million	333.22 billion
Crypto purchased	90.2 million	129.89 billion
Texts sent	16 million	24.04 billion
Google searches	5.9 million	8.5 billion
Snaps shared on Snapchat	2.43 million	3.5 billion
Content shared on Facebook	1.7 million	2.45 billion
Swipes on Tinder	1.1 million	1.58 billion
Hours streamed	1 million	1.44 billion
USD spent on Amazon	443,000	637.92 million
USD sent on Venmo	437,600	630.14 million
Tweets shared on Twitter	347,200	499.97 million
Hours spent in Zoom meetings	104,600	150.62 million
USD spent on DoorDash	76,400	110.02 million

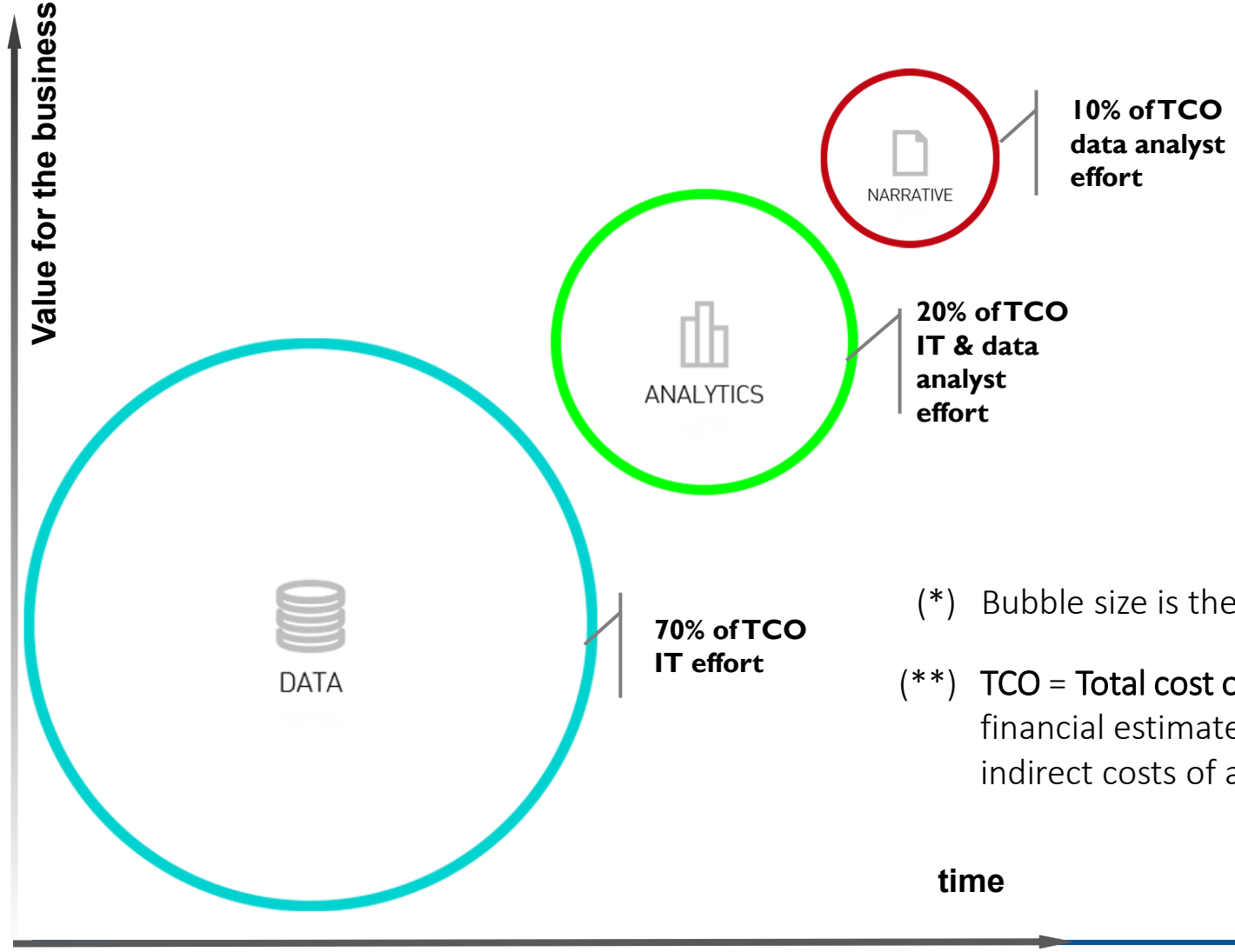
Video Data Traffic

- Watching **YouTube** videos in 480p resolution uses up to 500 MB of data per hour. While 4k YouTube videos use around 30xmore.
- Default **Spotify** settings use 2MB+ per 3-minute song. That's 40MB every hour. Or 960MB per day.
- Each standard definition **Netflix** stream uses 1GB of data per hour (24GB per day). High-definition Netflix streams can use as much as 3GB of data each hour (72GB per day). And ultra HD uses 7GB per hour (168GB per day.)

Data Exploitation

- Almost 90% of data is dark
- Only 12% available for business insights
- 88% is just stored
- 80% recordings, pdfs and texts

Data to Knowledge journey

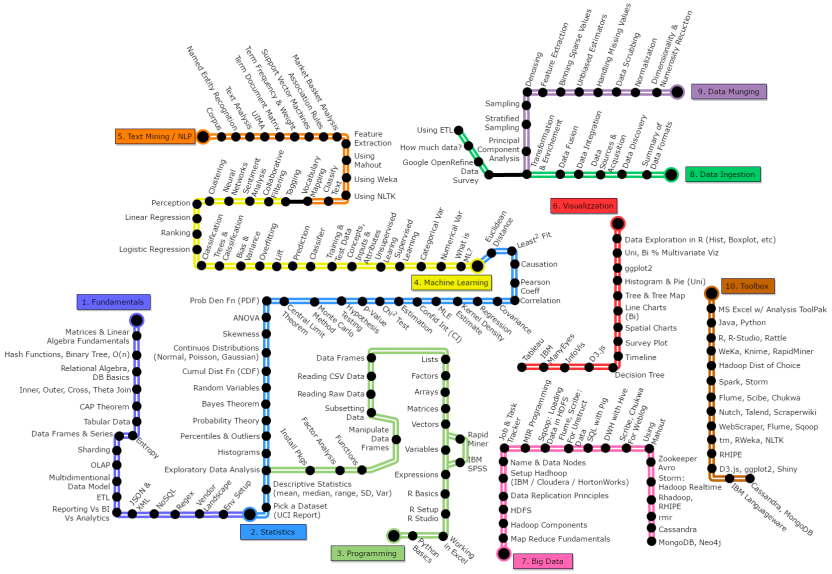


(*) Bubble size is the sub-process cost

(**) TCO = Total cost of ownership is a financial estimate of the direct and indirect costs of a product or system.

Overview

1. Organizzazione del corso
2. Principali argomenti trattati
3. Modalità d'esame
4. Che cos'è la Data Science
5. **(Big) Data**
   - Data Sources
   - Technologies

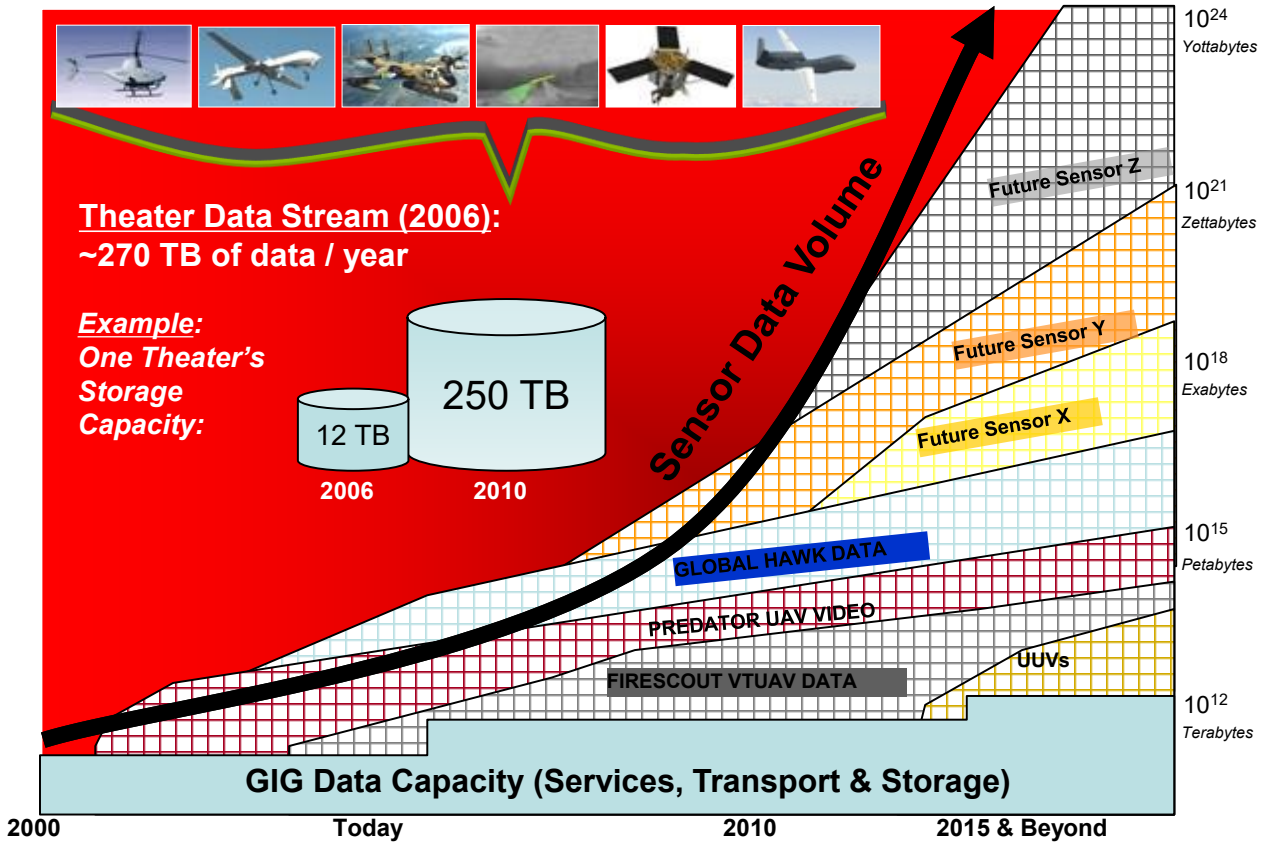


Big Data Keywords

Big Data Motivation

- We're now entering the "Industrial Revolution of Data," where the majority of data will be stamped out by machines: software logs, cameras, microphones, RFID readers, wireless sensor networks, and so on.
- These machines generate data much faster than people can, and their production rates will grow exponentially with Moore's Law.
- Storing this data is cheap, and it can be mined for valuable information.
  - Joe Hellerstein
    http://gigaom.com/2008/11/09/mapreduce-leads-the-way-for-parallel-programming/

Military Projection of Sensor Data Volume



Using 1TB drives, this would require 1 trillion (10^{12}) drives!

Bob Gourley: Thoughts on the future of Information Sharing Technology

FDS: Qualcuno si era sbagliato

- **Oggi si usano**:

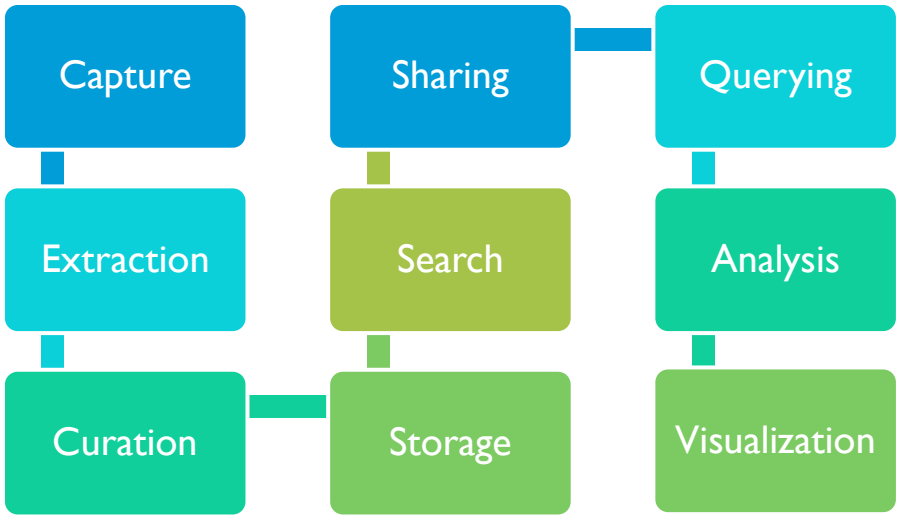
✓ PetaByte (10^{15})
✓ ExaByte (10^{18})
✓ ZettaByte (10^{21})
✓ YottaByte (10^{24})

NOBODY WILL EVER NEED MORE THAN 640K RAM.

Bill Gates, 1981

Defining Big Data

- Big data is a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.



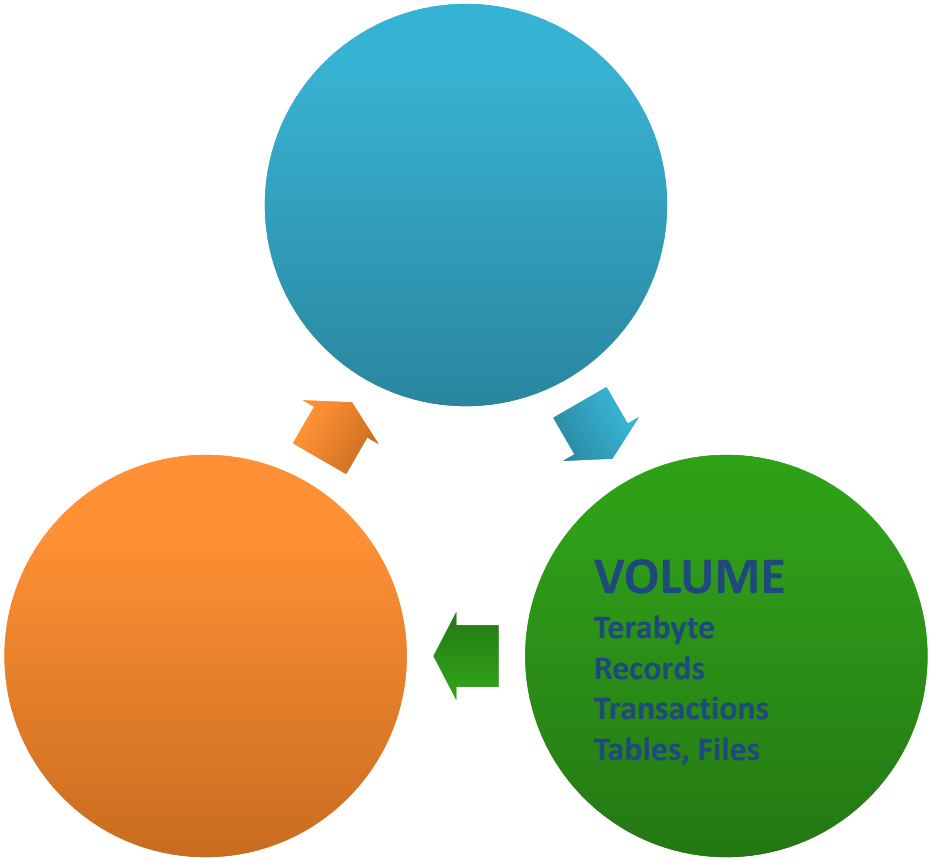
If data is **too big**, **too fast**, or **too hard** for existing tools to process, it is Big Data.

Gartner's 3(+1) V's – Big Data Properties

- **Volume**
  - 12 terabytes of Tweets: product sentiment analysis
  - 350 billion annual meter readings: predict power consumption
- **Velocity**
  - 5 million daily trade events: identify potential fraud
  - 500 million daily call detail records: predict customer churn faster
- **Variety**
  - 100's of live video feeds from surveillance cameras
  - 80% data growth in images, video and documents to improve customer satisfaction
- **Veracity**
  - 1 in 3 business leaders don't trust the information they use to make decisions.

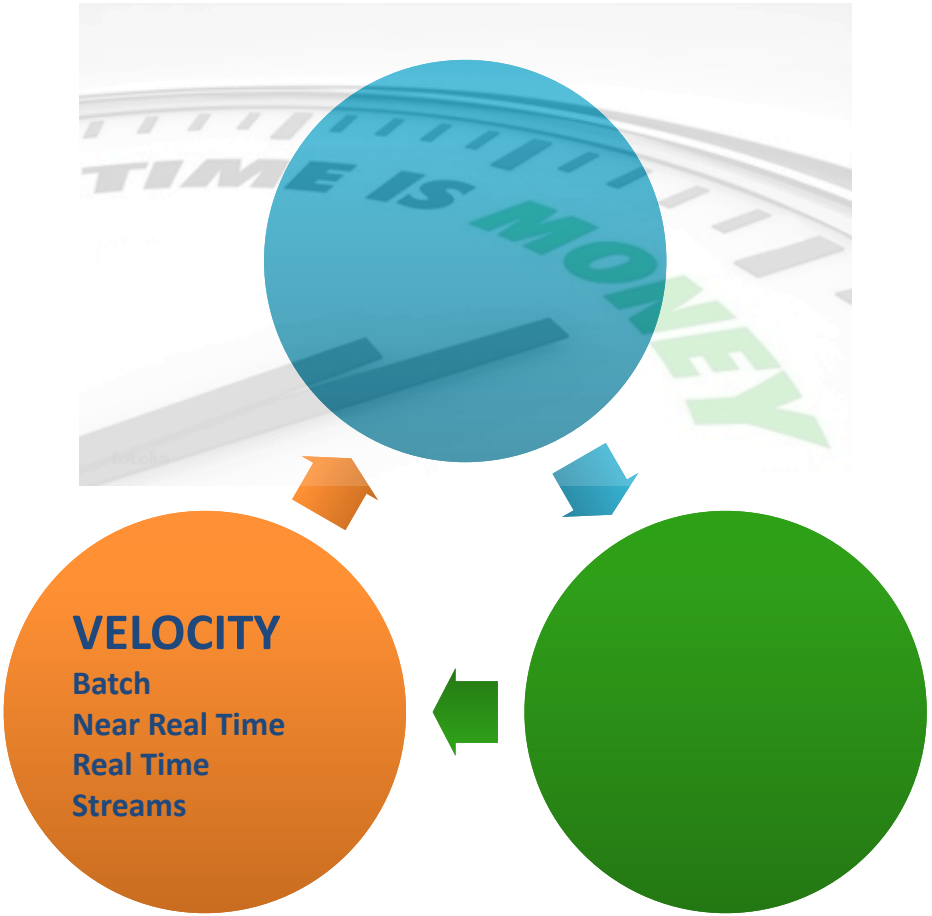
http://www.ibm.com/software/data/bigdata/

The 3V of Big Data



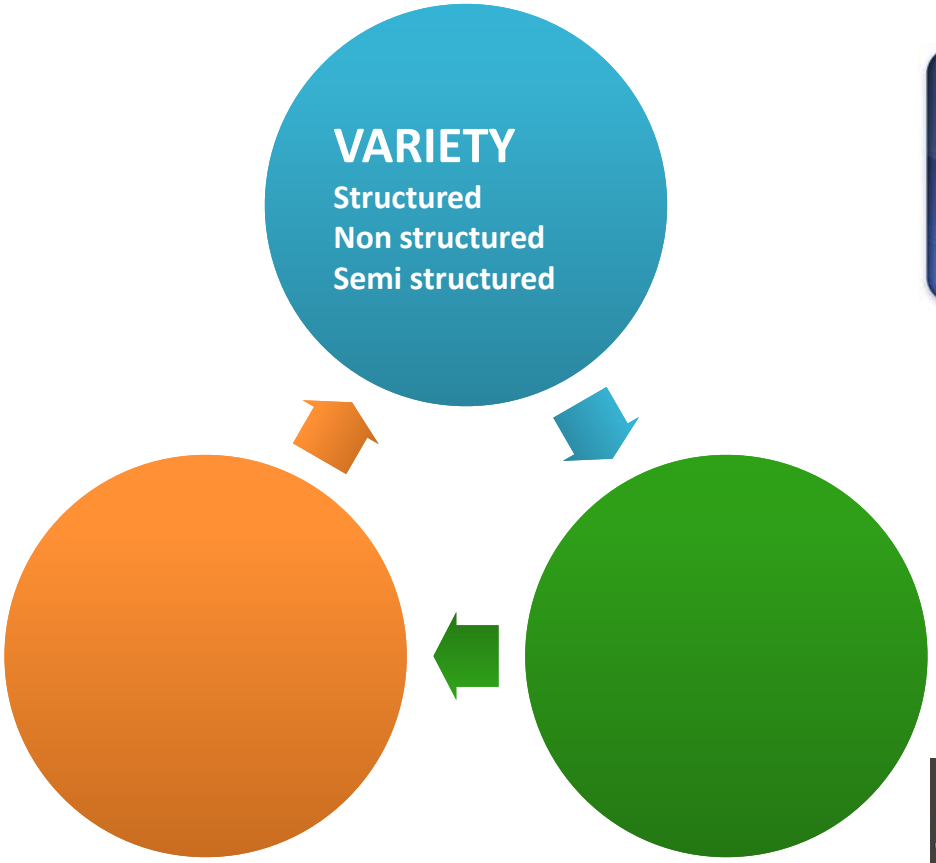
Kilobyte (10^3 Bytes)	**KB**
Megabyte (10^6 Bytes)	**MB**
Gigabyte (10^6 Bytes)	**GB**
Terabyte (10^9 Bytes)	**TB**
Petabyte (10^{15} Bytes)	**PB**
Exabyte (10^{18} Bytes)	**EB**
Zettabyte (10^{21} Bytes)	**ZB**
Yottabyte (10^{24} Bytes)	**YB**

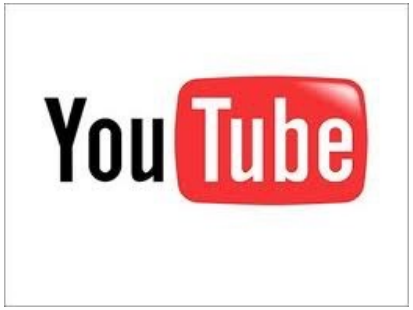
The 3V of Big Data



- Data produced with growing velocity and frequency, forcing organizations to make decisions much more rapidly.
- Think about *sentiment* monitoring, through the analysis of *social networks* of messages mentioning a given *brand*.
- Quick obsolescence of Information.

The 3V of Big Data





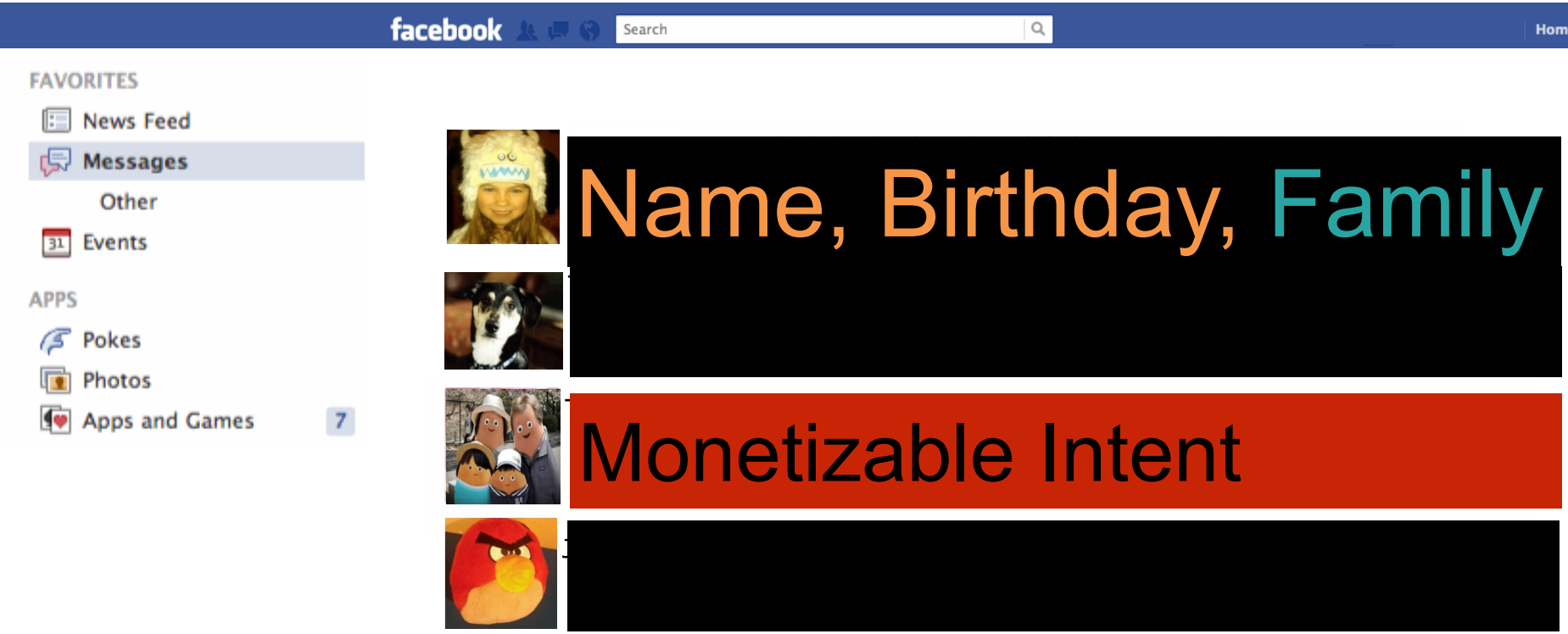
More V's

- **Viscosity**
  - Integration and dataflow friction
- **Venue**
  - Different locations that require different access & extraction methods
- **Vocabulary**
  - Different languages and vocabularies
- **Value**
  - Added-value of data to organization and use-case
- **Virality**
  - Speed of dispersal among community
- **Variability**
  - Data, formats, schema, semantics change. Also, the interpretation of the same data can change depending on the context in which it is collected and analyzed.

Big and Small

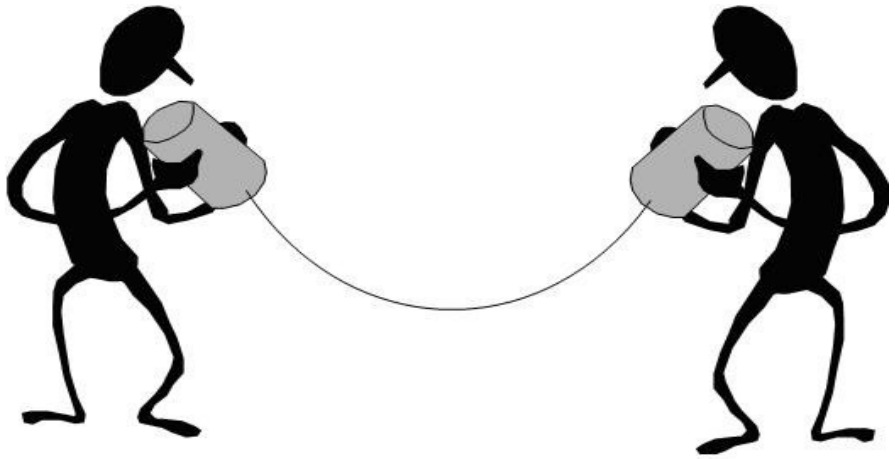
- Big Data can be very small
  - Streaming data from aircraft sensors
  - Hundred thousand sensors on an aircraft is "big data"
  - Each producing an eight byte reading every second
  - Less than 3GB of data in an hour of flying
    - (100,000 sensors x 60 minutes x 60 seconds x 8 bytes).
- Not all large datasets are "big".
  - Video streams plus metadata
  - Telco calls and internet connections
  - Can be parsed extremely quickly if content is well structured.
  - From http://mike2.openmethodology.org/wiki/Big_Data_Definition
- The task at hand makes data "big".

Data in Conversations

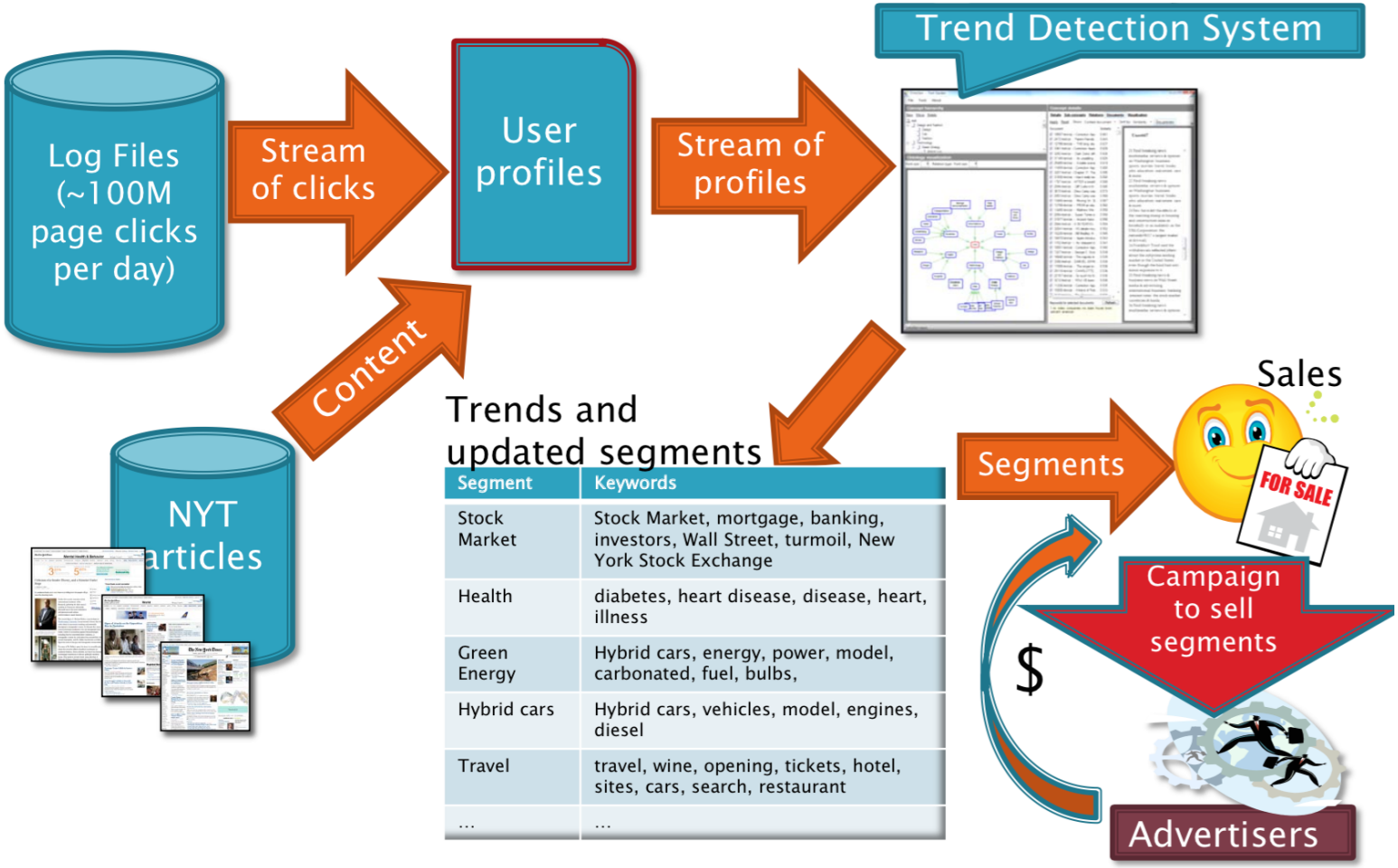


IOT: Internet of things

- IoT connects several devices and sensors.
- For example, street sensors might capture data on trafic flow, which might be sent to your alarm to wake you up earlier if there is trafic jam on the way to your office.



Online Advertising NYTimes



Segment	Keywords
Stock Market	Stock Market, mortgage, banking, investors, Wall Street, turmoil, New York Stock Exchange
Health	diabetes, heart disease, disease, heart, illness
Green Energy	Hybrid cars, energy, power, model, carbonated, fuel, bulbs,
Hybrid cars	Hybrid cars, vehicles, model, engines, diesel
Travel	travel, wine, opening, tickets, hotel, sites, cars, search, restaurant
…	…

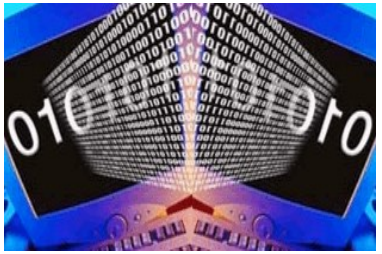
Further applications

Healthcare

Sales

Finance

Log analysis



Sicurity

Trafic monitoring

Telecomunications

Quality control

Manufacturing

Trading Analytics

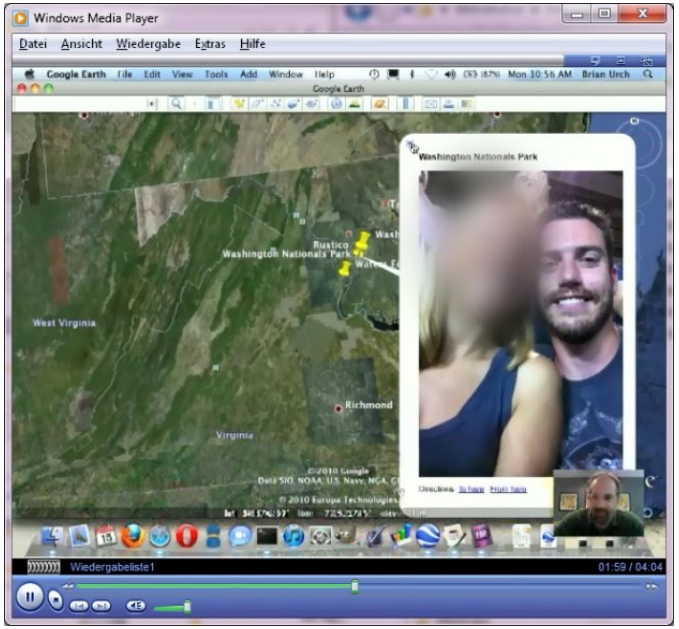
Fraude detection

Retail: Churn, NBO

„Big data“ in business

- Has been used to sell more hardware and software.
- Has become a shallow buzzword (Termine di moda).

- But: The actual big data is there, has added-value, and can be used effectively.
  - Data mining
  - Marketing / advertising
  - Collaborative filtering
  - Raytheon’s RIOT software
  - NSA, etc.
  - Kreditech, Lenddo, Klout, …
  - …

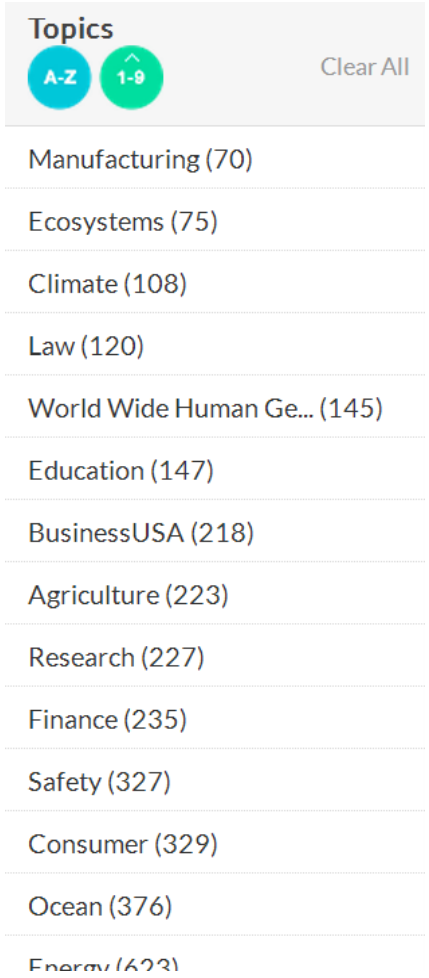


„Big data“ in business

- Amazon.com
  - Millions of back-end operations every day
  - Catalog, searches, clicks, wish lists, shopping carts, third-party sellers, …
- Walmart
  - > 1 million customer transactions per hour
  - 2.5 petabytes (2560 terabytes)
- Facebook
  - 250 PB, 600TB added daily (2013)
  - 1 billion photos on one day (Halloween)
- FICO Credit Card Fraud Detection
  - Protects 2.1 billion active accounts

Big Government Data (USA)

- Big Data Research and Development Initiative
  - Explored how big data addresses important problems facing the government.
  - 84 different big data programs spread across six departments
- Data.gov
  - > 104.000 datasets
- Government owns six of the ten most powerful supercomputers in the world.
- NASA Center for Climate Simulation
  - 32 petabytes of climate observations and simulations



Examples from Wikipedia – Big Science

- Large Hadron Collider
  - 150 million sensors; 40 million deliveries per second
  - 600 million collisions per second
  - Theoretically: 500 exabytes per day (500 quintillion bytes)
  - Filtering: 100 collisions of interest per second
    - Reduction rate of 99.999% of these streams
  - 25 petabytes annual rate before replication (2012)
  - 200 petabytes after replication

Examples from Wikipedia - Science

- Sloan Digital Sky Survey (SDSS)
  - Began collecting astronomical data in 2000
  - Amassed more data in first few weeks than all data collected in the history of astronomy.
  - 200 GB per night
  - Stores 140 terabytes of information
  - Large Synoptic Survey Telescope, successor to SDSS
    - Online in 2016
    - Will acquire that amount of data every five days.
- Human genome
  - Originally took 10 years to process;
  - Now it can be achieved in one day.

Big Data = Science?

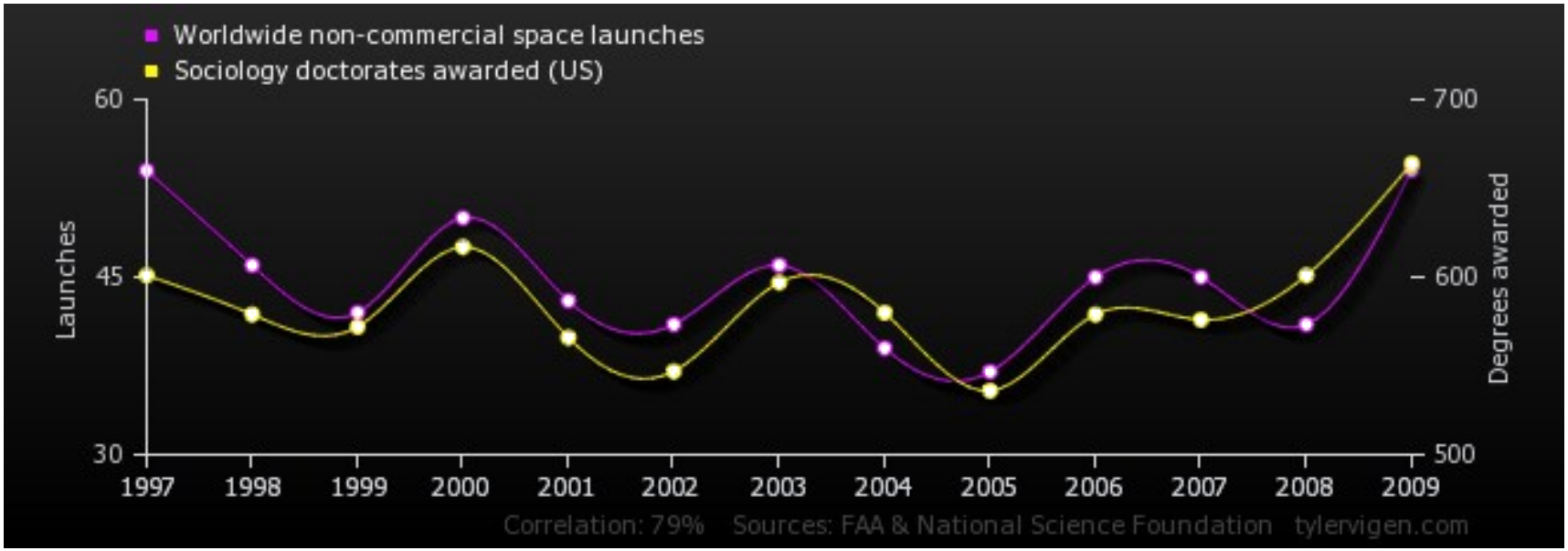
- The End of Theory: The Data Deluge Makes the Scientific Method Obsolete (Chris Anderson, Wired, 2008)
  - All models are wrong, but some are useful. (George Box)
  - All models are wrong, and increasingly you can succeed without them. (Peter Norvig, Google)

- Before Big Data: Correlation is not causation!
- With Big Data: Who cares?
  - Traditional approach to science — hypothesize, model, test — is becoming obsolete.

- Petabytes allow us to say: "**Correlation is enough.**"

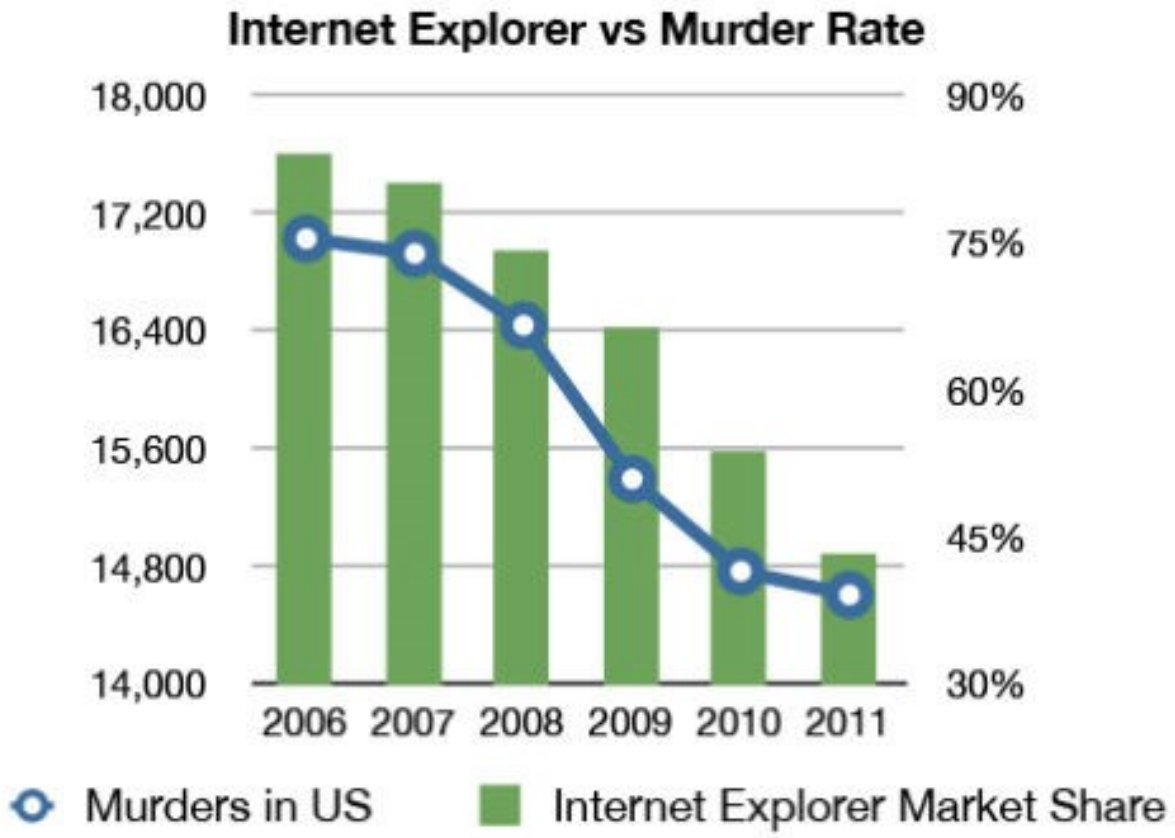
http://www.wired.com/science/discoveries/magazine/16-07/pb_theory

Correlation vs. Causation



Quelle: Spurious correlations (www.tylervigen.com)

Correlation vs. Causation



Addressing Big Data: Parallelization

- Long tradition in databases
- Vertical and horizontal partitioning
- Shared nothing
- Each machine runs same single-machine program

- Other trends
  - Map/Reduce / Hadoop
  - Multicore CPUs
  - GPGPUs

    (general-purpose computing on graphics processing units, meaning the use of a graphics processing unit for purposes different from its traditional use in computer graphics)

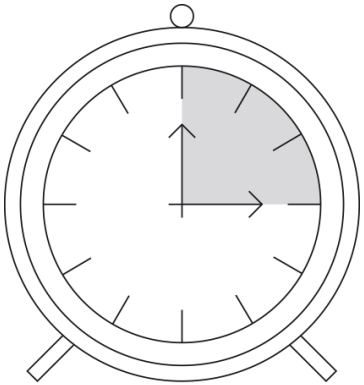
Levels of Parallelism on Hardware

- Instruction-level Parallelism
  - Single instructions are automatically processed in parallel
  - Example: Modern CPUs with multiple pipelines and instruction units.
- Data Parallelism
  - Different data can be processed independently
  - Each processor executes the same operations on its share of the input data.
    - Example: Distributing loop iterations over multiple processors
    - Example: GPU processing
- Task Parallelism
  - Different tasks are distributed among the processors/nodes
  - Each processor executes a different thread/process.
    - Example: Threaded programs.

Pathologies of Big Data

- Store basic demographic information about each person
  - `age`, `sex`, `income`, `ethnicity`, `language`, `religion`, `housing status`, `location`
  - Packed in a 128-bit record
- World population: 6.75 billion rows, 10 columns, 128 bit each
  - About 150 GB
- What is the median age by sex for each country?
  - Algorithmic solution
    - 500$ Desktop: I/O-bound; 15min reading the table
    - 15,000$ Server with RAM: CPUI-bound; <1min
  - Database solution
    - Aborted bulk load to PostgreSQL – disk full (bits vs. integer and DBMS inflation)
  - Small database solution (3 countries, 2% of data)
    - ```
      SELECT country,age,sex,count(*)
      FROM people GROUP BY country,age,sex;
      ```
    - *> 24h, because of poor analysis: Sorting instead of hashing*
    - "PostgreSQL's difficulty here was in **analyzing** [=profiling] the stored data, not in storing it."
      - From http://queue.acm.org/detail.cfm?id=1563874
      ```

"Some" Critical issues with Big Data

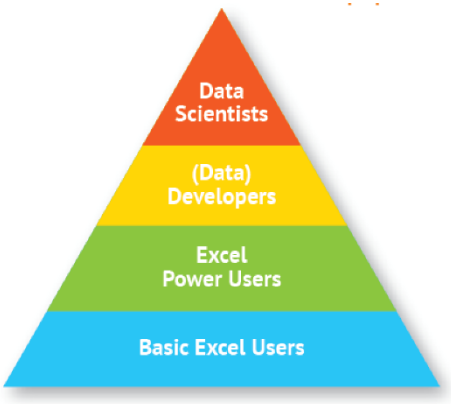


77%

Data Processing

23%

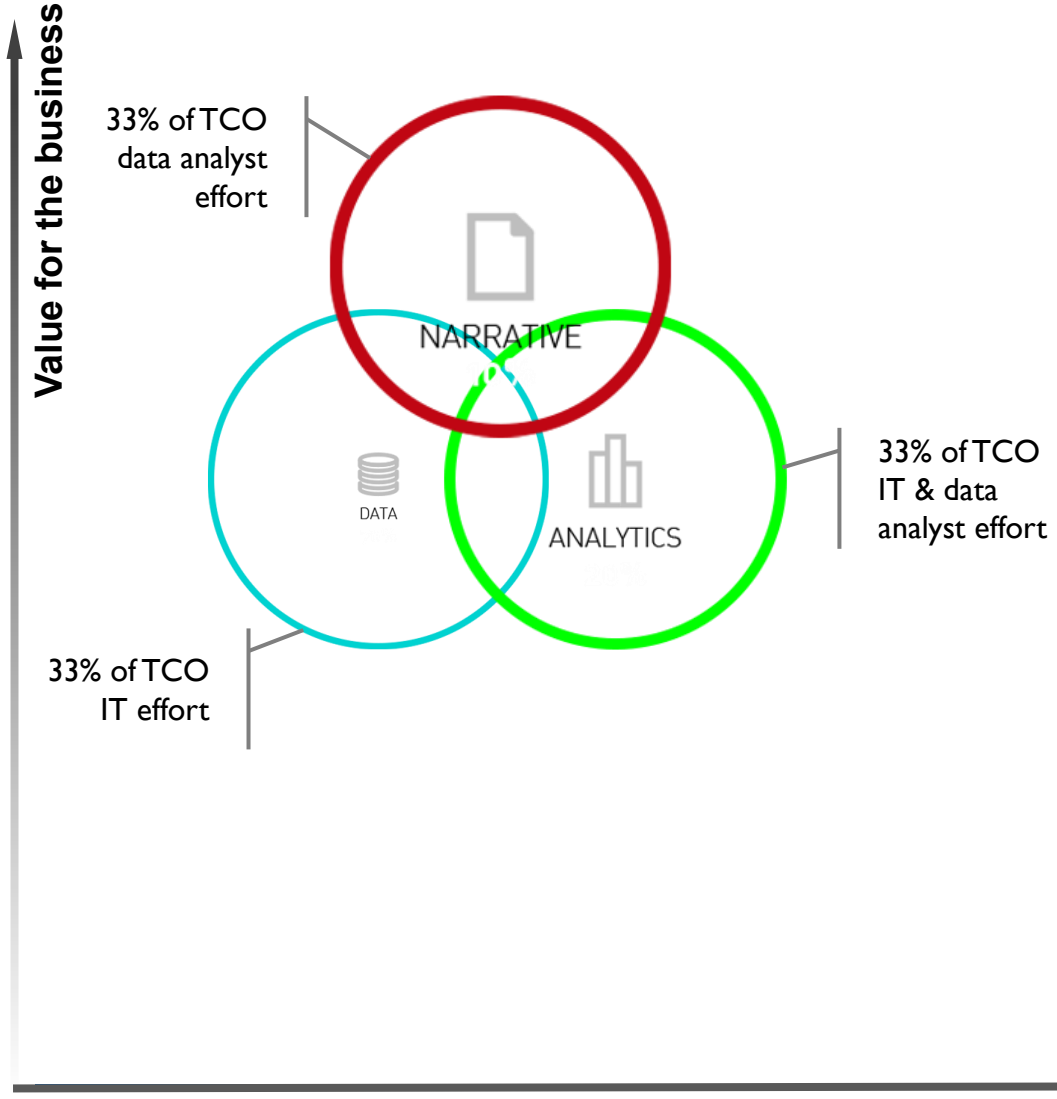
Data Analysis



Source Bloor2016

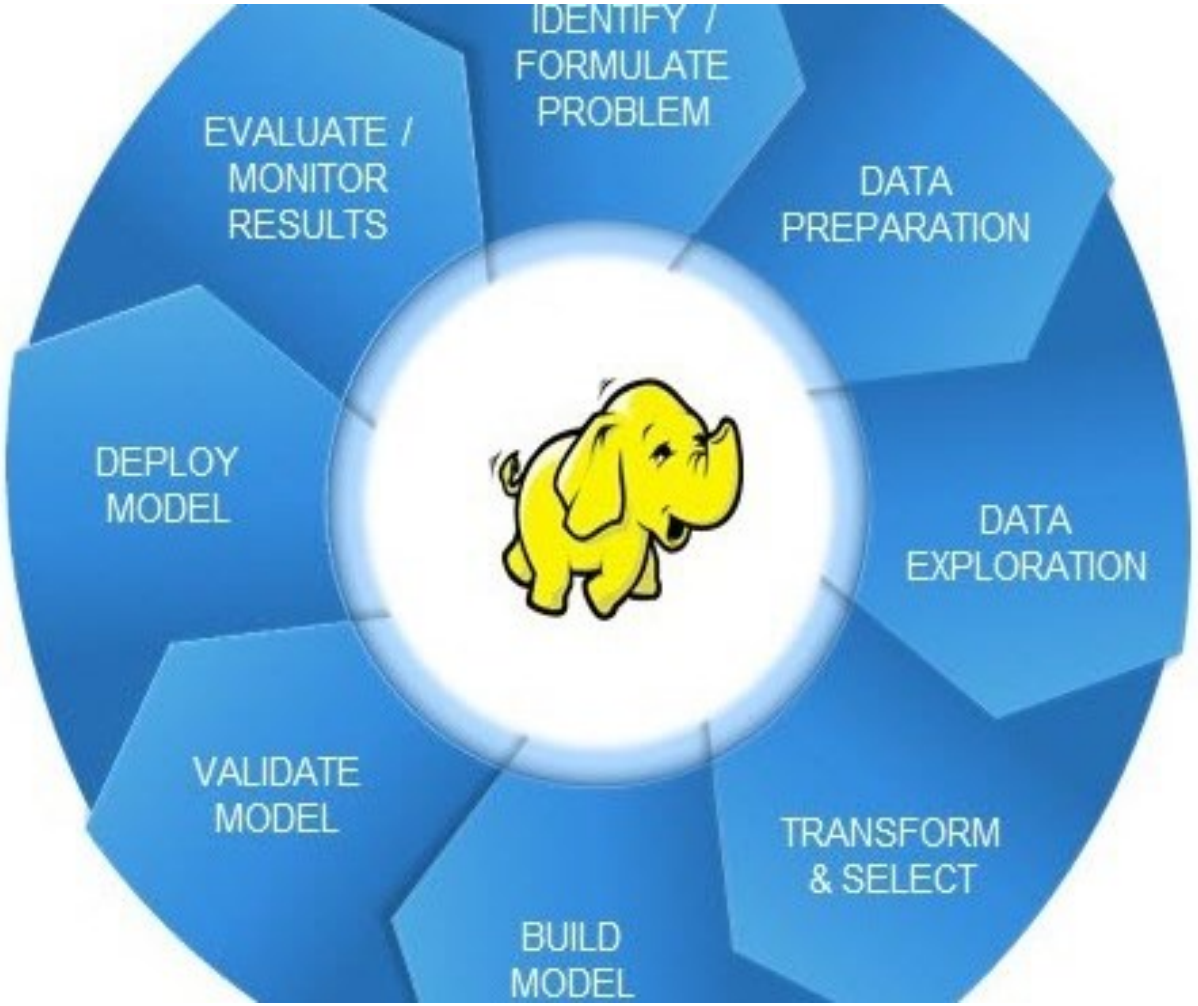
1. OFTEN, THE BUSINESS ANALYST AND THE DATA SCIENTIST DO NOT UNDERSTAND EACH OTHER
2. DARK DATA EXTERNAL TO DATA LAKES CONTINUE TO GROW
3. IT IS REQUIRED LONG TIME FOR MAKING DATA
4. PROBLEMS WITH DATA QUALITY (DATA NEED BE INTEGRATED AND CLEANED FROM MULTIPLE SOURCES)

Data to Knowledge journey - evolutionary



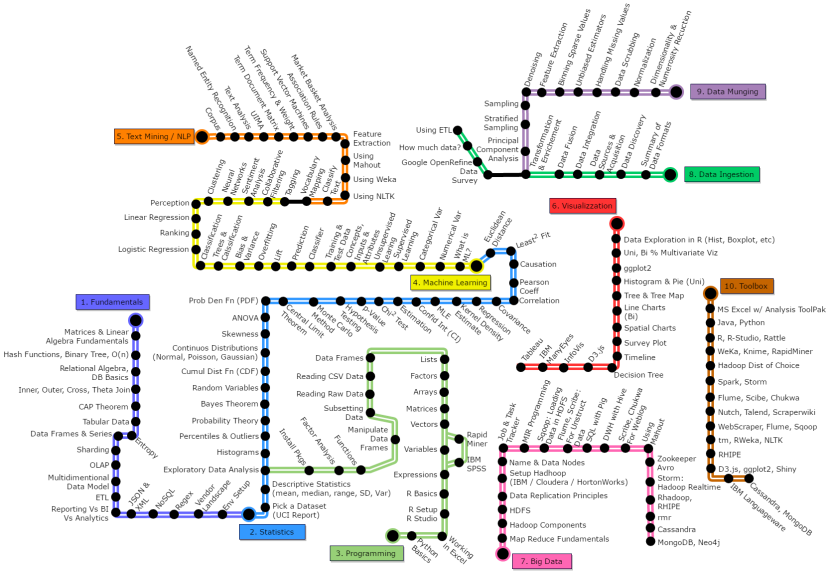
(*) Bubble size is the sub-process cost

The Knowledge Extraction



Overview

1. Organizzazione del corso
2. Principali argomenti trattati
3. Modalità d'esame
4. Che cos'è la Data Science
5. (Big) Data
   - **Data Sources**
   - Technologies



Open vs. closed source

Open

- Linked data
  - http://linkeddata.org/
- Government data
  - data.gov, data.gov.uk
  - Eurostat
- Scientific data
  - Genes, proteins, chemicals
  - Scientific articles
  - Climate
  - Astronomy
- Published data
  - Tweet (limited)
  - Crawls
- Historical data
  - Stock prices

Closed

- Transactional data
  - Music purchases
  - Retail-data
- Social networks
  - Tweets, Facebook data
  - Likes, ratings
- E-Mails
- Web logs
  - Per person
  - Per site
- Sensor data
- Military data

Wikipedia Infoboxes

```
{{Infobox company
|name             = International Business Machines Corporation
|logo             = <br />[[File:IBM logo.svg|200px]]<br />
|caption          = Logo since 1972, designed by [[Paul Rand]]
|type             = [[Public company|Public]]
|traded_as        = {{New York Stock Exchange|IBM}}<br />[[Dow Jones Industrial Average|Dow Jones Compon
Component]]
|industry         = [[Personal computer hardware|Computer hardware]], [[Software|Computer software]], [[
services]], [[Information technology consulting|IT consulting]]
|products         = [[List of IBM products|See IBM products]]
|founder          = [[Charles Ranlett Flint]]
|foundation       = [[Endicott, New York|Endicott]], New York, U.S.<br />{{{Start date|1911|06|16}}}
|location_city    = [[Armonk, New York|Armonk]], New York
|location_country = U.S.
|area_served      = Worldwide
|key_people       = [[Ginni Rometty]]<br />{{{small|Chairman, President, and CEO}}}
|revenue          = {{Increase}} US$ 106.91 [[1000000000 (number)|billion]] <small>(2011)</small><ref na
|url=http://rcpmag.com/articles/2012/01/20/intel-ibm-exceed-earnings-estimates-google-falls-short.aspx|
International Business Machines Corporation |work=United States Securities and Exchange Commission}}</re
|operating_income = {{Increase}} US$ {{0|0}}20.28 billion <small>(2011)</small><ref name=10K/>
|net_income       = {{Increase}} US$ {{0|0}}15.85 billion <small>(2011)</small><ref name=10K/>
|assets           = {{Increase}} US$ 116.43 billion <small>(2011)</small><ref name=10K/>
|equity           = {{Decrease}} US$ {{0|0}}20.13 billion <small>(2011)</small><ref name=10K/>
|num_employees    = 433,362 <small>(2012)</small><ref name="Fortune 500: IBM employees"/>
```

International Business Machines Corporation

Logo since 1972, designed by Paul Rand

Type	Public
Traded as	NYSE: IBM Dow Jones Component S&P 500 Component
Industry	Computer hardware, Computer software, IT services, IT consulting
Founded	Endicott, New York, U.S. (June 16, 1911)
Founder(s)	Charles Ranlett Flint
Headquarters	Armonk, New York, U.S.
Area served	Worldwide
Key people	Ginni Rometty (Chairman, President, and CEO)
Products	See IBM products
Revenue	▲ US$ 106.91 billion (2011)[1]
Operating income	▲ US$ 20.28 billion (2011)[1]
Net income	▲ US$ 15.85 billion (2011)[1]
Total assets	▲ US$ 116.43 billion (2011)[1]
Total equity	▼ US$ 20.13 billion (2011)[1]
Employees	433,362 (2012)[2]
Divisions	Financing, Hardware, Services, Software
Website	IBM.com

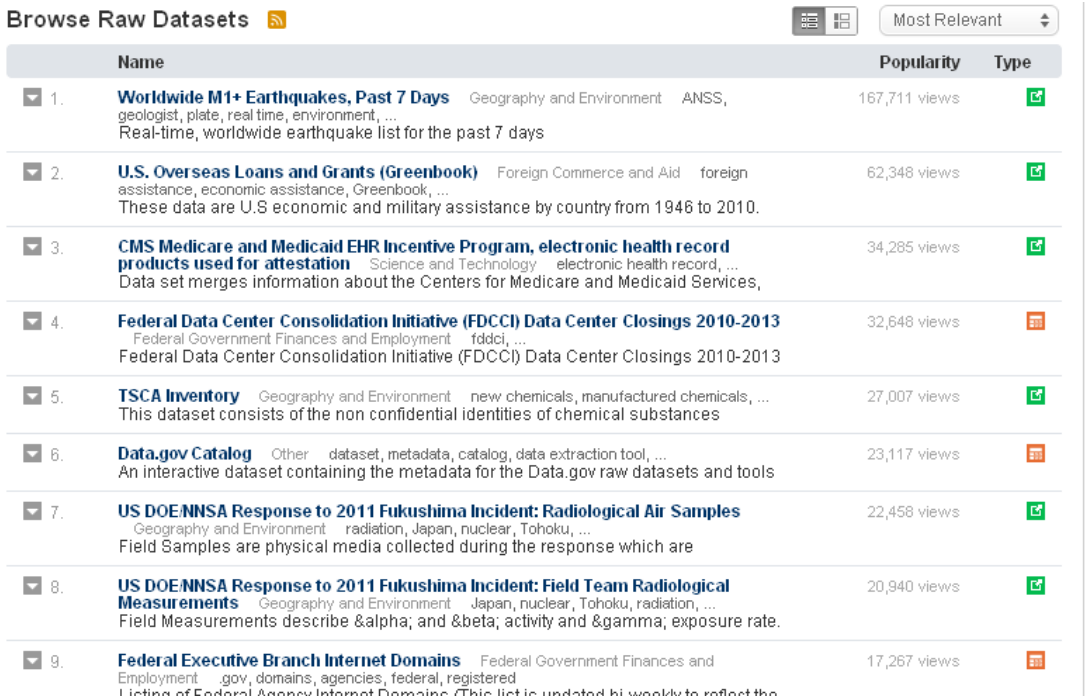
DBpedia statistics

- From 125 languages of Wikipedia
- 3 billion triples
  - 580 million English
- English DBpedia
  - 4.6 million things
  - 1,445,000 persons
  - 735,000 places
  - 411,000 creative works
  - 241,000 organizations
  - 251,000 species
  - …

- http://wiki.dbpedia.org/about/facts-figures

And more sources

- Government data
  - www.data.gov (380k data sets)
  - data.gov.uk (9k)
  - ec.europa.eu/eurostat
- Finance / business data
- Scientific databases
  - www.uniprot.org
  - skyserver.sdss.org
- The Web
  - HTML tables and lists: billions
  - General sources: Dbpedia (3.7m), freebase (23m), microformats…
  - Domain-specific sources: IMDB, Gracenote, isbndb, …

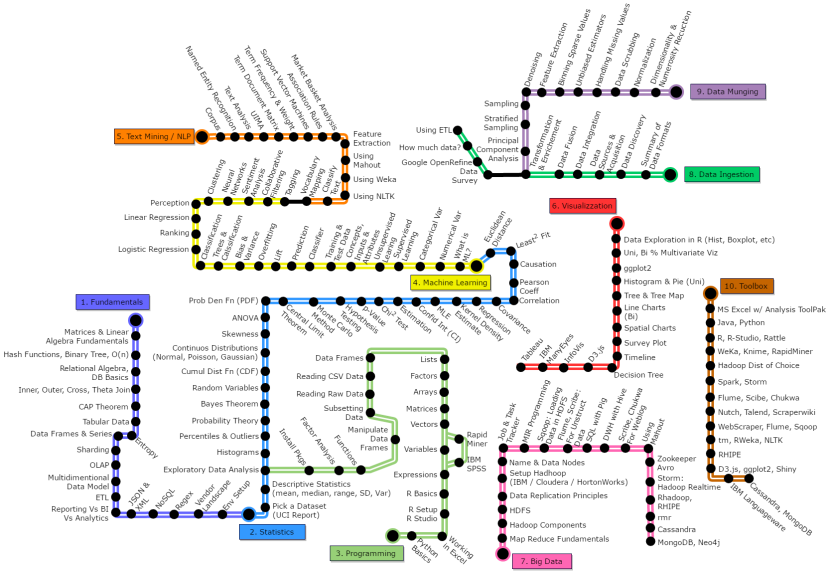


Getting the data

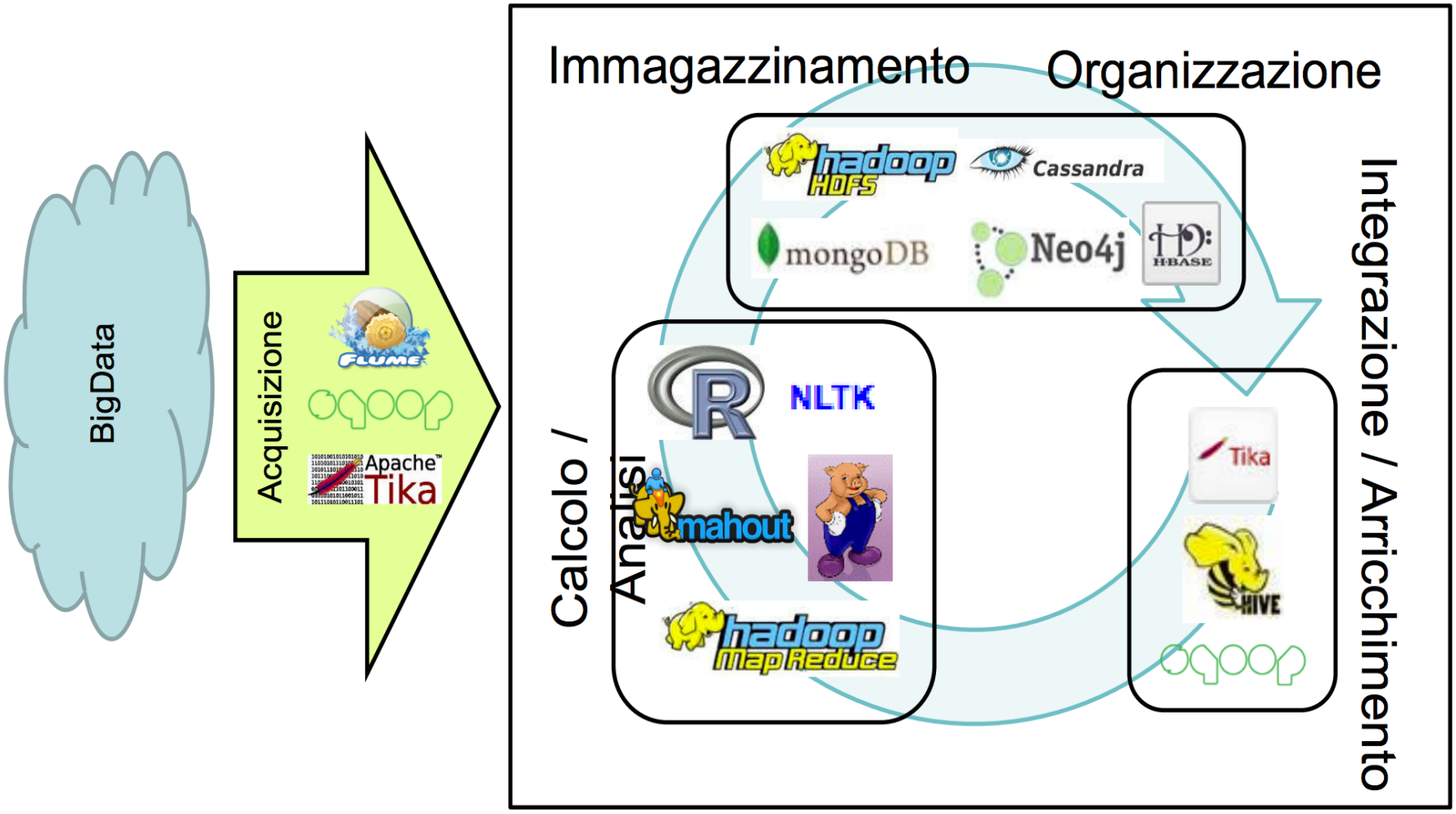
- Download
  - Data volumes make this increasingly infeasible
  - Fedex HDDs
  - Fedex tissue samples instead of sequence data
- Generating big (but synthetic) data
  1. Automatically insert interesting features and properties
  2. Then „magically“ detect them
- Sharing data
  - Repeatability of experiments
  - Not possible for commercial organizations

Overview

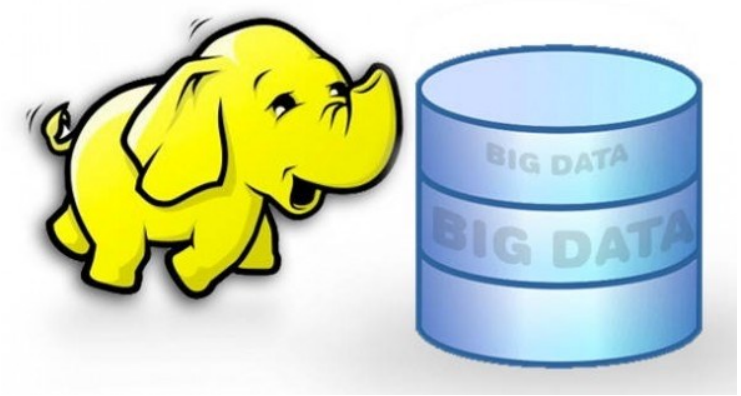
1. Organizzazione del corso
2. Principali argomenti trattati
3. Modalità d'esame
4. Che cos'è la Data Science
5. (Big) Data
   - Data Sources
   - **Technologies**



Some Technologies for Big Data

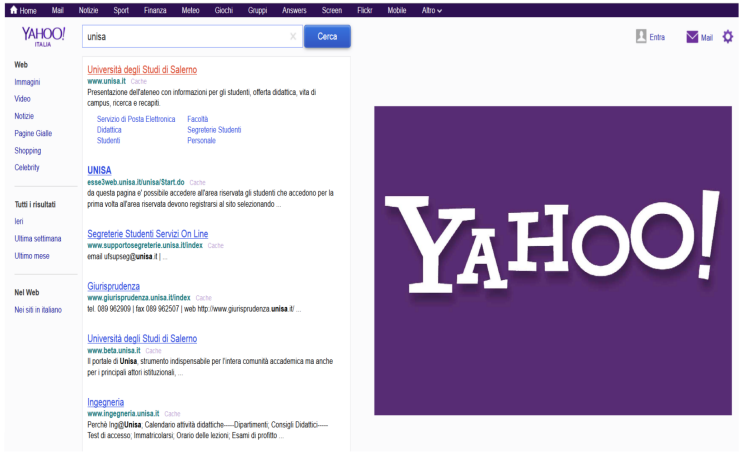


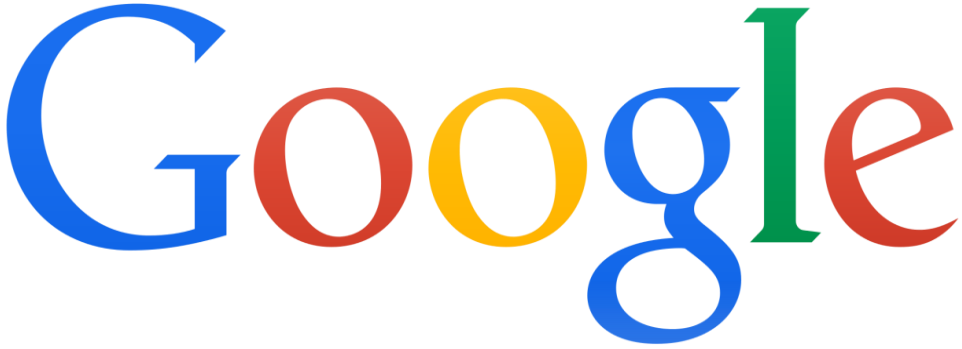
Hadoop



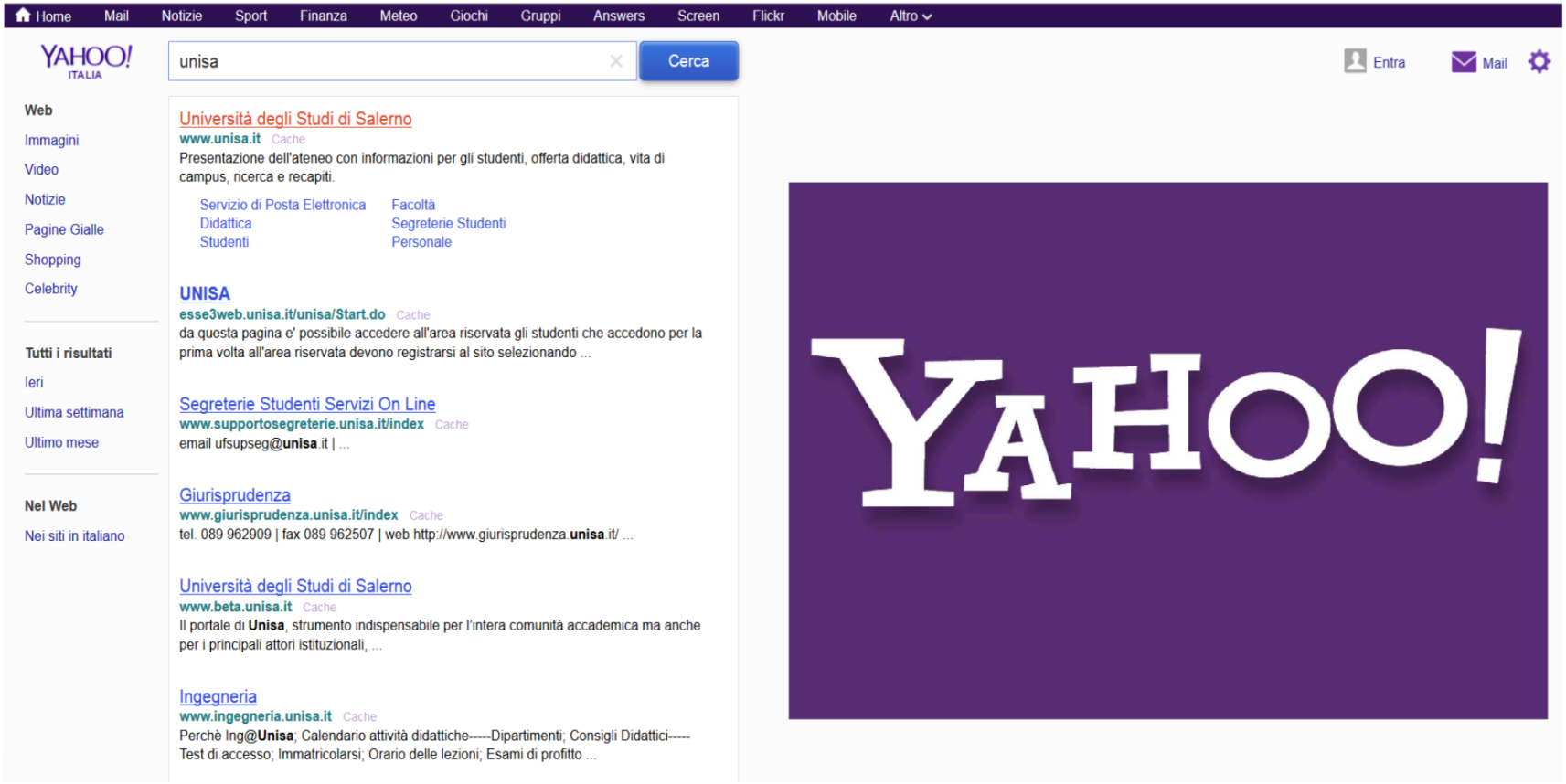
- Released by Apache Software Foundation
- Written in Java
- Big data processing in distributed applications
  - high number of hardware commodities
  - batch processing oriented
  - map/reduce computational model

Who Uses Hadoop ?





Yahoo! uses Hadoop

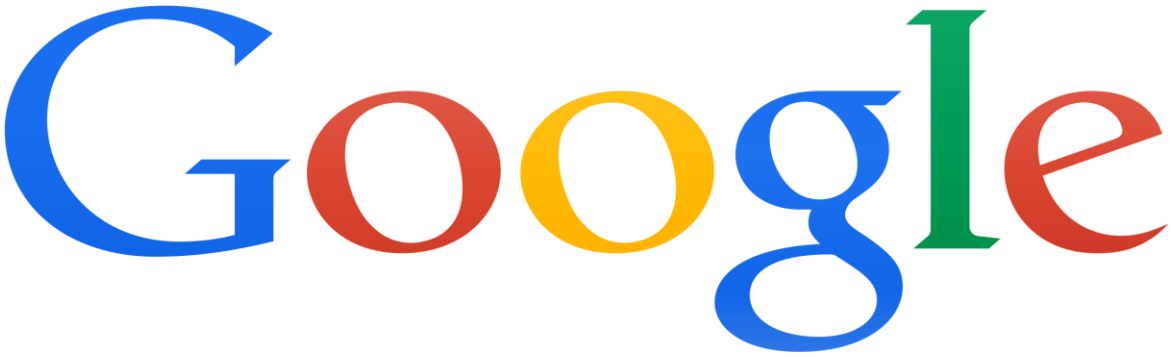


Facebook uses Hadoop

Twitter uses Hadoop

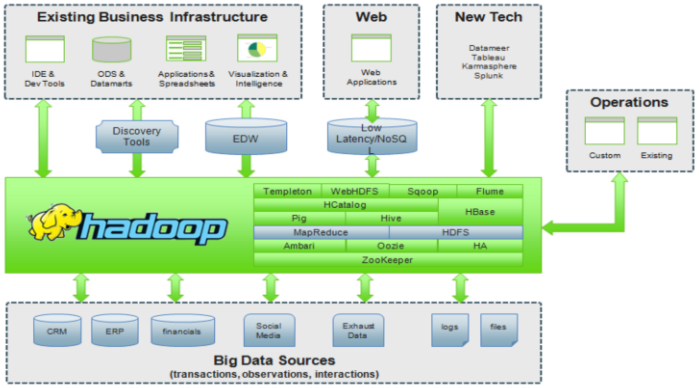
History of Hadoop

- Hadoop history starts with Nutch
  - An open source web search engine and web crawler aiming to capture billions of URLs monthly
  - Scalability problems
- Early 2000 years Google published Google File System (GFS) and MapReduce
  - A solution to the problems arisen with Nutch



Hadoop (2)

- Hadoop is composed of 4 macro projects:
  - **Hadoop Common**: basic libraries for correct Hadoop execution
  - **Hadoop Distributed File System** (HDFS): A distributed file system providing fast access to data
  - **Hadoop MapReduce**: a system for parallel processing of massive data
  - **Hadoop YARN**: to schedule processes and manage cluster resources

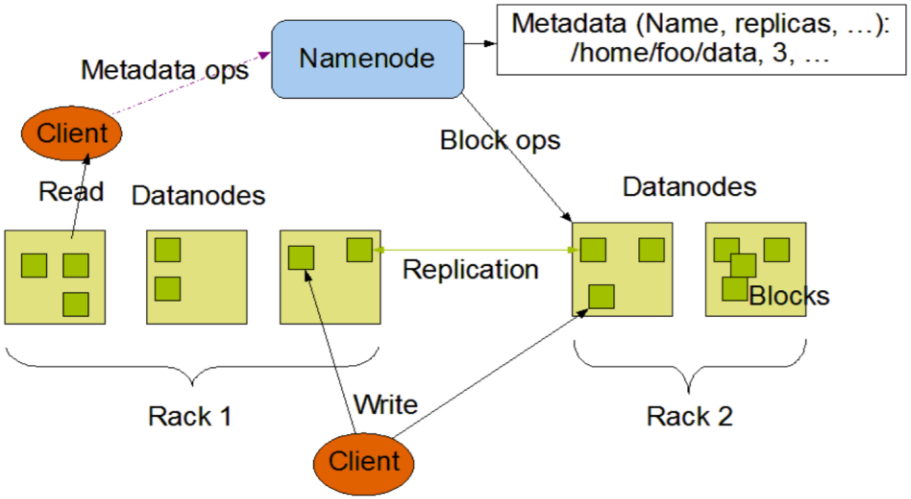


HDFS (1)

- Inspired to GFS
- Block-structured file system
  - Single files are stored as blocks of fixed size
  - The blocks of a file are not necessarily stored on the same machine

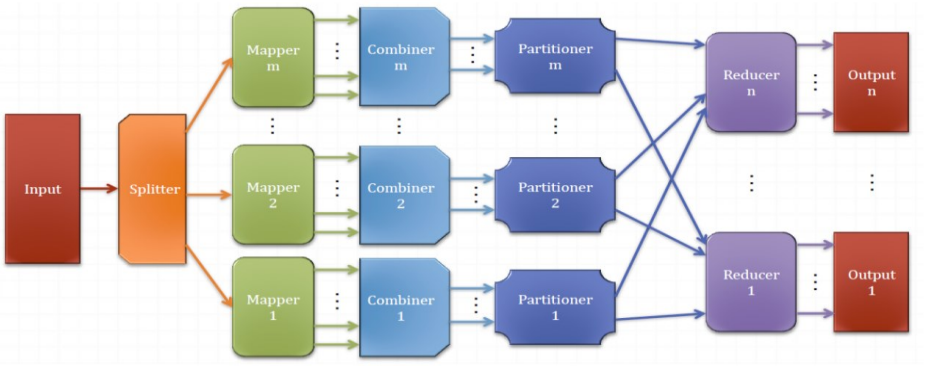
HDFS (2)

- Master/slave architecture
- A HDFS cluster consists of a single NameNode, a master managing the name space and access of clients to files
- There are many DataNodes, typically one for each cluster node

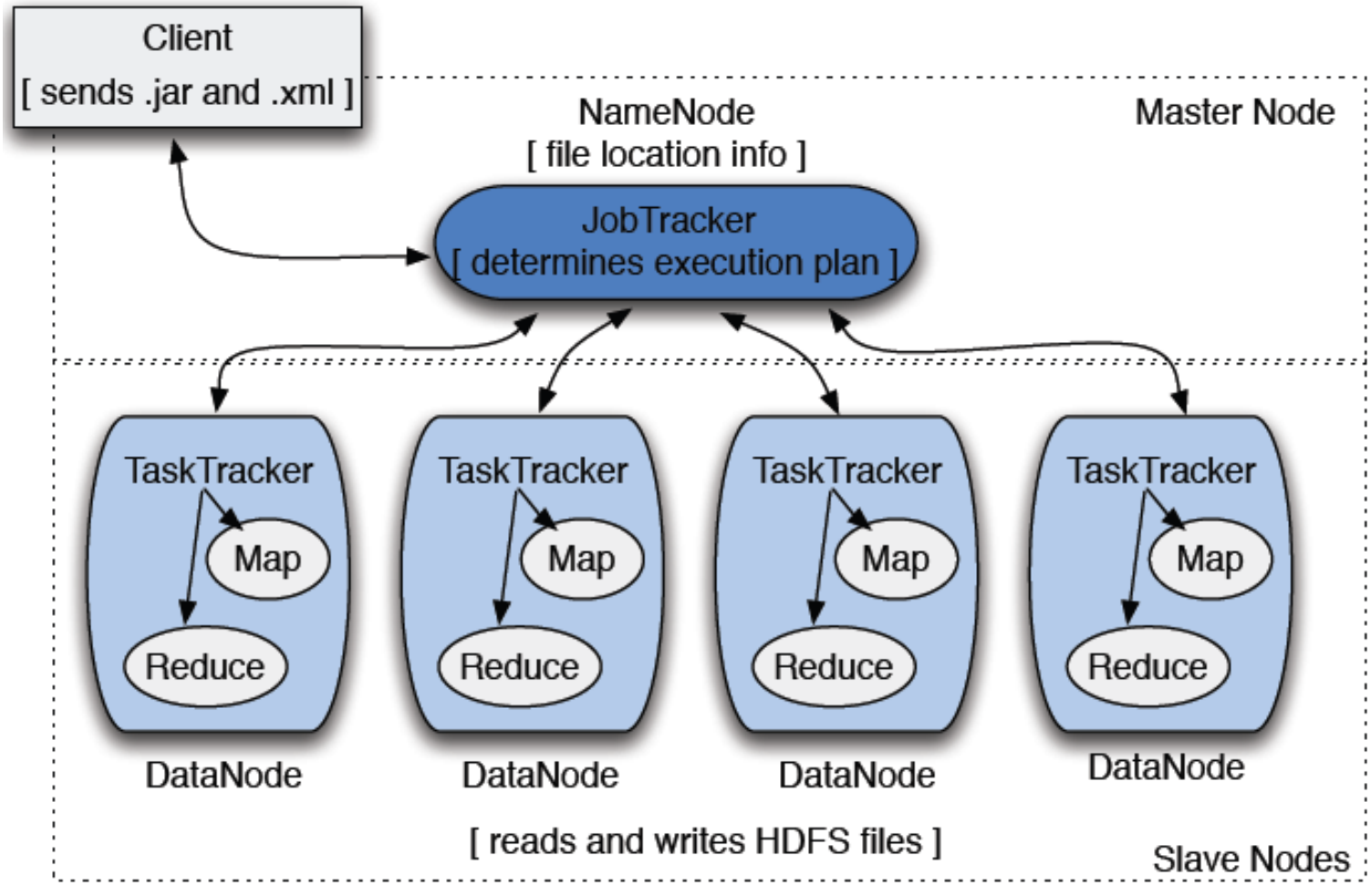


MapReduce (1)

- MapReduce was issued by Google as a programming paradigm for processing “Big Data”
- Map e Reduce primitives can be found in the functional programming paradigms, like in the Lisp language
- Programs written with this paradigm can be easily parallelized on a great number of machines
  - It abstracts from the low level details of parallel programming
  - It automatically handles node failures



MapReduce (2)


Client
[sends .jar and .xml]
NameNode
[file location info]
Master Node
JobTracker
[determines execution plan]
TaskTracker
Map
Reduce
DataNode
TaskTracker
Map
Reduce
DataNode
TaskTracker
Map
Reduce
DataNode
TaskTracker
Map
Reduce
DataNode
[reads and writes HDFS files]
Slave Nodes

Technologies for Big Data acquisition

- API: Twitter API, Facebook API ed API dei motori di ricerca
- Web scraping: cURL, Apache Tika
- ETL: Sqoop
- Stream: Apache Flume

Twitter API

- The launch of API twitter API in 2009 has insipred numerous research projects
- Sentiment analysis [Datumbox], Communication during emergencies [Building a Data Warehouse for Twitter Stream Exploration]

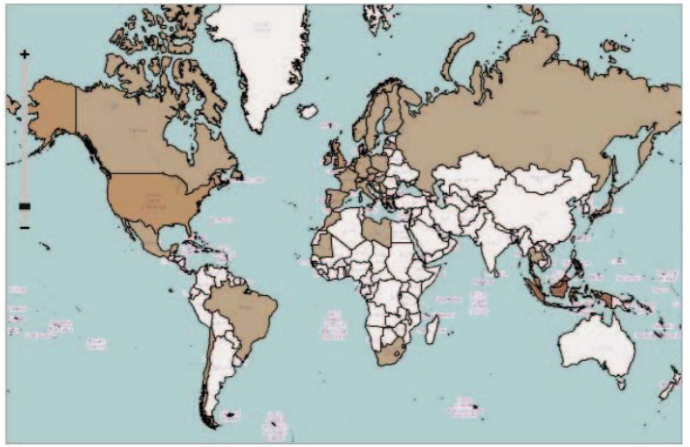
Datumbox Twitter Sentiment Analysis

Inserisci la parola sotto per eseguire l'analisi del sentimento:

Keyword: lollipop android Invia

Results for "lollipop android"

Id	User	Text	Twitter Link	Sentiment
537640724771381248	Robonto	I made a wish and it didn't come true. @Android @googlenexus 4 #lollipop □ http://t.co/ZUkelf4IFO	View	negative
537640397770481664	SJT	Xiaomi's Hugo Barra: We might launch an Android One phone, MIUI with Lollipop to arrive Q1 2015 http://t.co/6aVft6alQp	View	neutral
537640358663159808	Juan Ignacio Yarza	Android 5.0 Lollipop review: the biggest Android update. - http://t.co/sY6Q7qrRrC http://t.co/01w07hmq1y	View	neutral
537640248742645760	Domenico Loiacono	#Android Lollipop arriva su Android Wear - Wired.it http://t.co/La6Khxx2Fk	View	neutral
537640109521526784	Israel Banda	RT @chungu_mwila: 16 Things You Can Do in Android Lollipop That You Couldn't Do in KitKat http://t.co/wHgk7r8xBT	View	positive
537640071772381184	Android Cdma	#AndroidCdma Xiaomi's Hugo Barra: We might launch an Android One phone, MIUI with Lollipop to arrive Q1 2015 http://t.co/dsiErFJ1x5	View	neutral
537640071311011841	Android Pres News	#android Xiaomi's Hugo Barra: We might launch an Android One phone, MIUI with Lollipop to arrive Q1 2015 http://t.co/V8MpbBJnfE	View	neutral
537639952310599680	Redtwit Bot	No more □ hairy heart emoji on Lollipop http://t.co/MqH4YbkkUN	View	neutral
537639408539680768	SIAM_MS	Waiting For Moto G 2nd Gen To Be In Stock Again !□ #Flipkart #Android #Kitkat #Lollipop #Waiting #MotoG2ndGen #Google #Moto @Flipkart Moto	View	positive

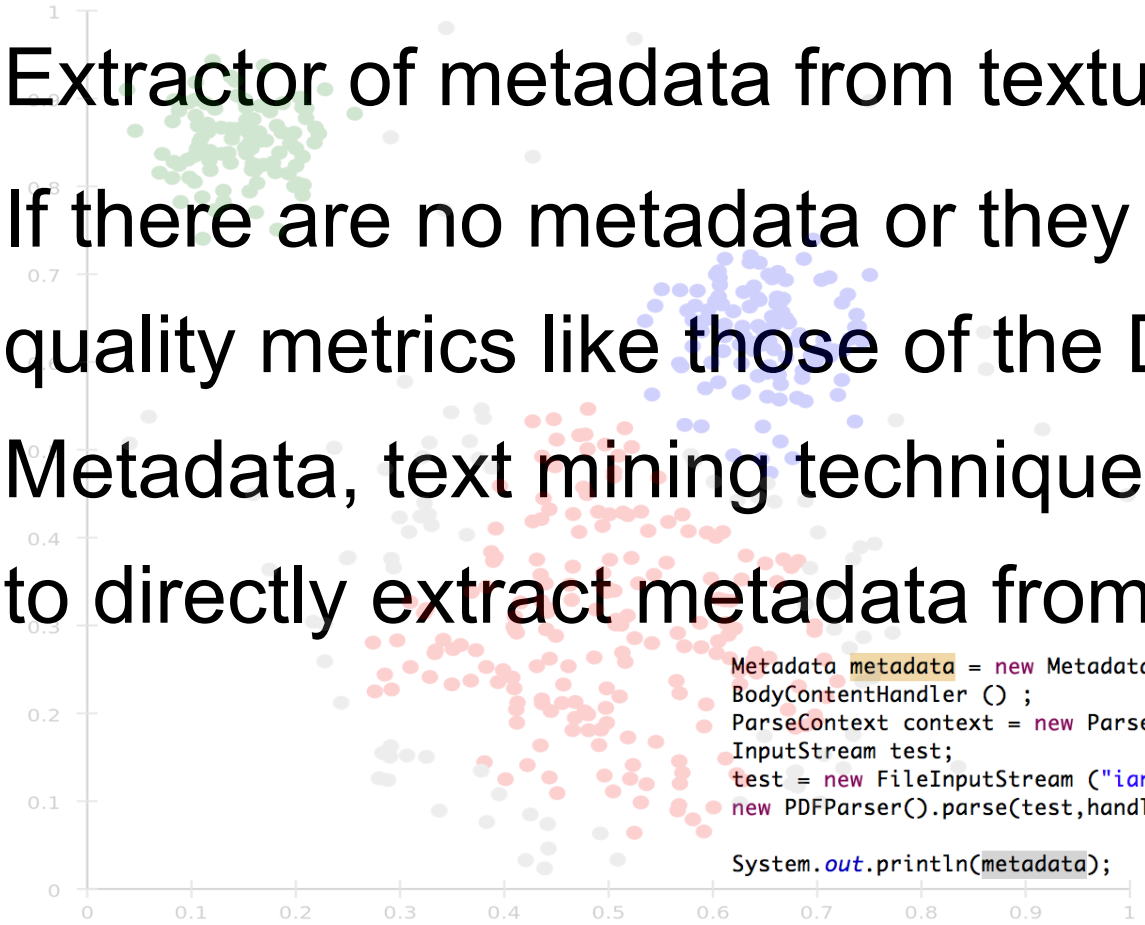


Web scraping tools - cURL

- Simulate human navigation on the Web
- Enable comparisons of online prices, monitoring of meteorological data, and so on….
- Possibility to specify multiple URLs: http://site.{one,two,three}.com
- Sequences: ftp://ftp.numericals.com/file[1-100].txt
- Use of cookies, proxy

Apache Tika

- Knowledge management tool
- Extractor of metadata from textual documents
- If there are no metadata or they do not fulfill quality metrics like those of the Dublin Core Metadata, text mining techniques will be used to directly extract metadata from the document

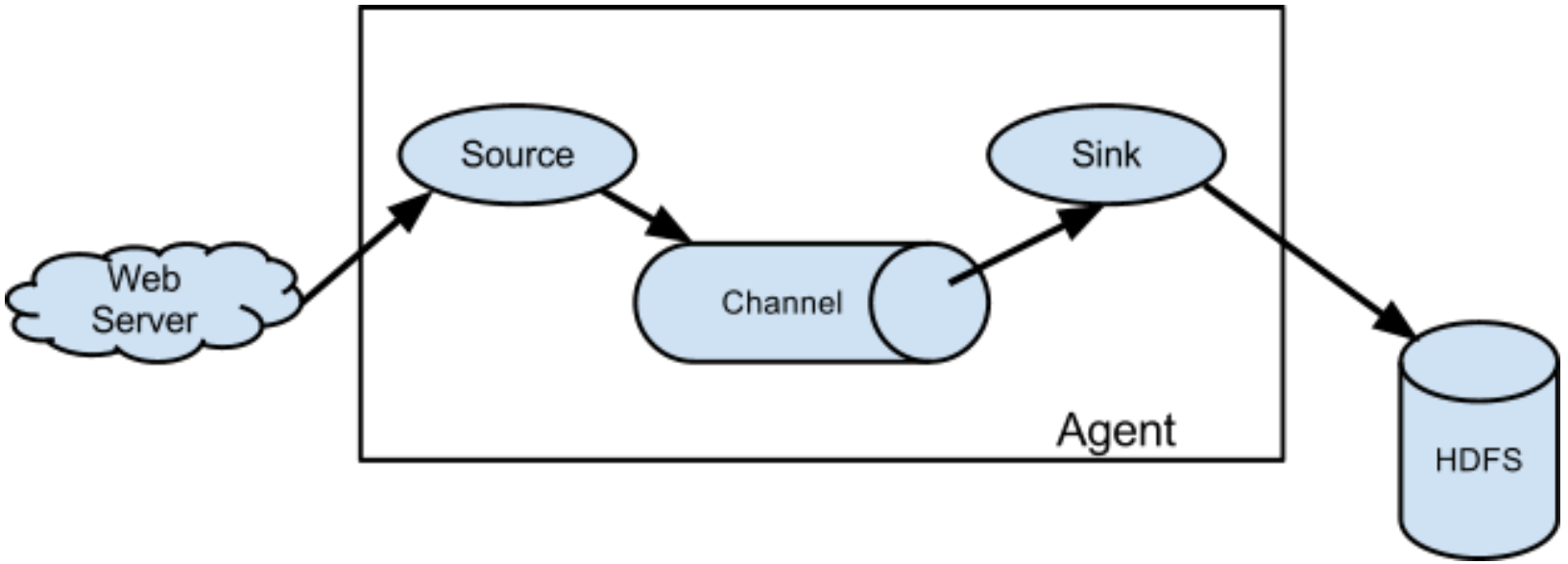


```
Metadata metadata = new Metadata () ; ContentHandler handler = new
BodyContentHandler () ;
ParseContext context = new ParseContext();
InputStream test;
test = new FileInputStream ("ianno-m-(cur-ic).pdf");
new PDFParser().parse(test,handler,metadata,context);

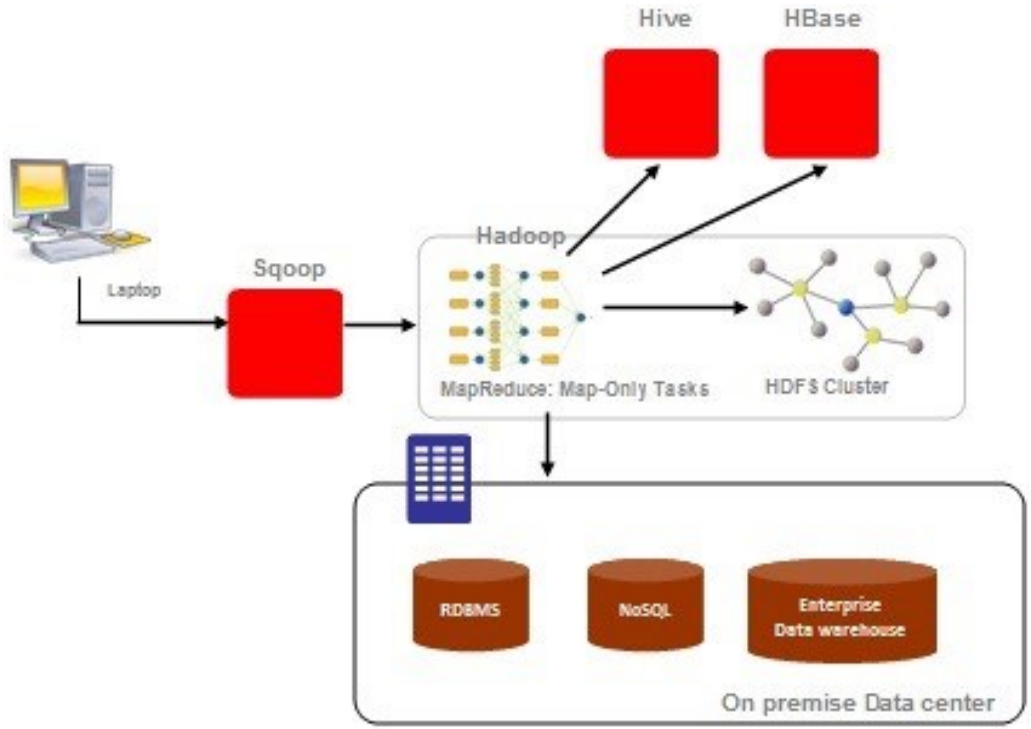
System.out.println(metadata);
```

Flume

- Moves big data streams in the Hadoop's distributed file system
  - Designed to collect web logs in real time



Squoop



- "unlock" of data stored in an enterprise's RDBMSs their trasfer in Hadoop

```
$ sqoop import --connect jdbc:mysql://localhost/test_db --table user_data -m 2 --target-dir /tables/userdata/
```

Hive and Pig

- Tecnology to query Big Data as relational tables
- HiveQL
- Produces MapReduce software in a transparent way
- Initially developed by Facebook
- Data presentation layer

- Like Hive, buy it uses Pig Latin as an high level language
- Data preparation layer

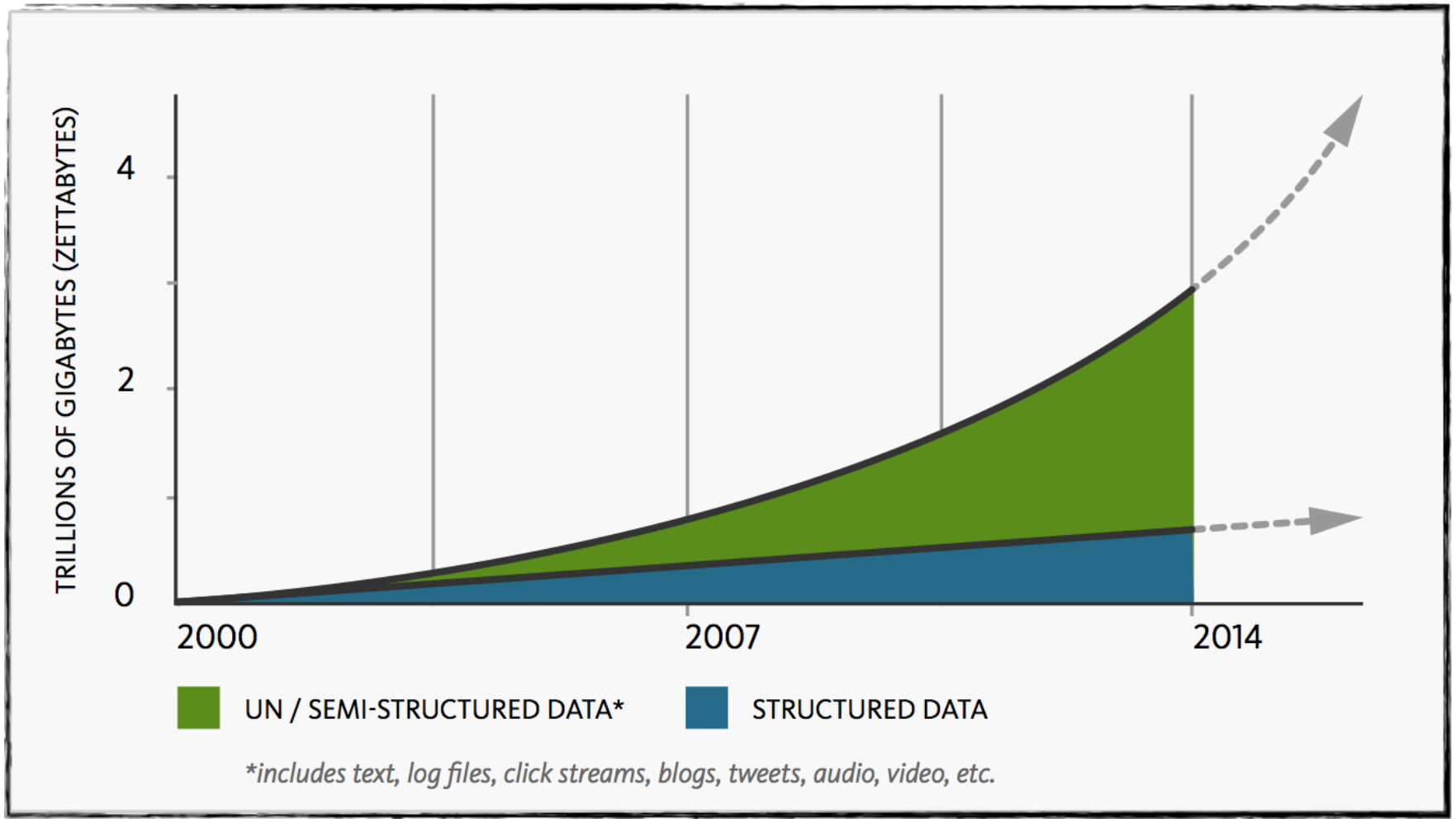

NoSQL
matters

Why SQL?

- Predifined schema for storing structured data
- Already familiar BCNF structure
- Strong consistency
- Transactions
- Mature and thoroughly tested
- Based on ACID properties

- Data Retrieval: Structured Query Language (SQL) - versatile e powerful
- Vertical Scalability : to make an SQL DB scalable, the only possibility is to upgrade the hardware on which the DBMS is installed

Why NoSQL? (1)

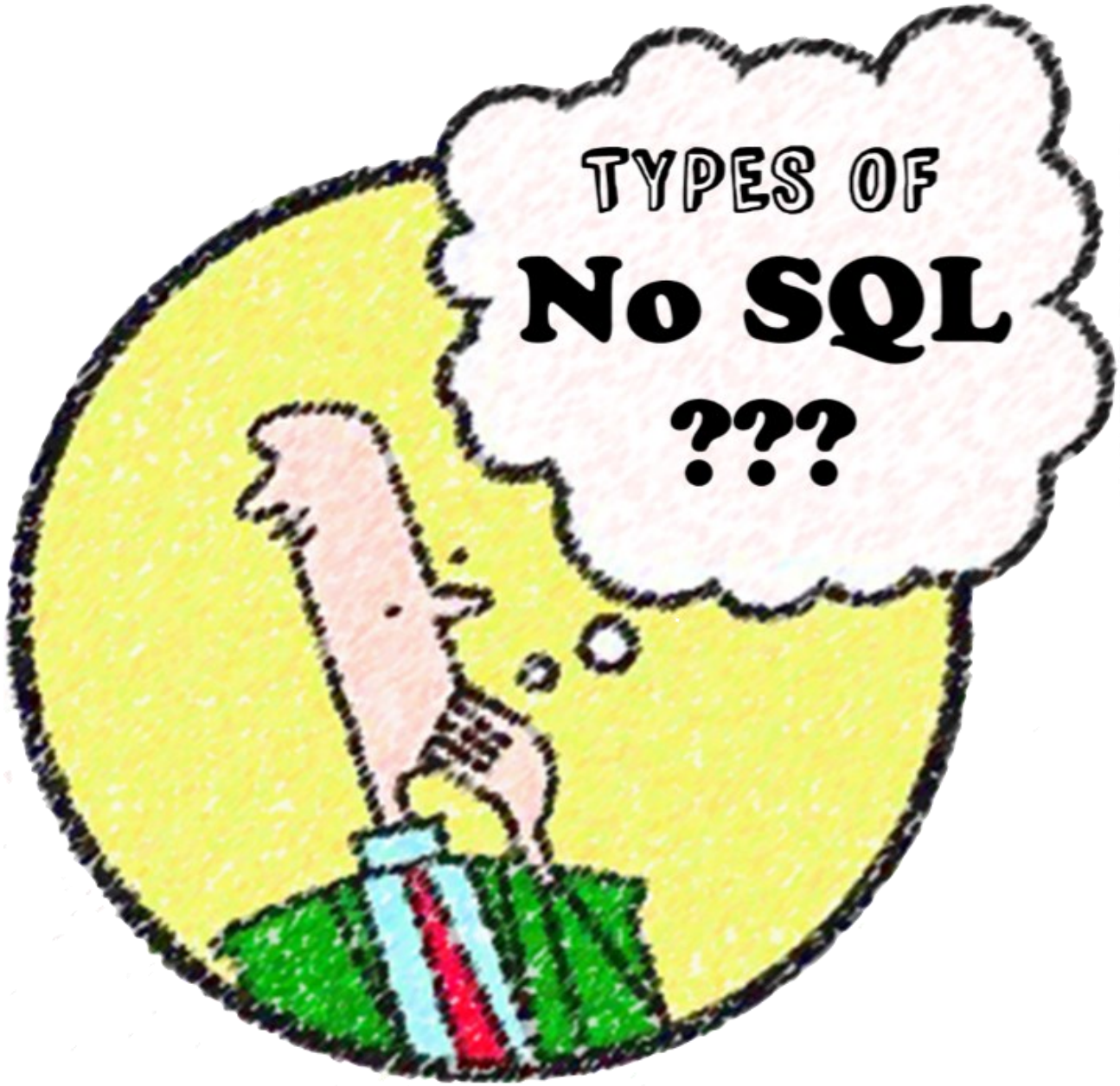

TRILLIONS OF GIGABYTES (ZETTABYTES)
4
2
0
2000
2007
2014
UN / SEMI-STRUCTURED DATA*
STRUCTURED DATA
*includes text, log files, click streams, blogs, tweets, audio, video, etc.

Why NoSQL? (2)

- **Non relational**: the rigid approach of relational DBs does not allow to store strongly dynamic data. NoSQL DBs are "schemaless" and enable the storage of attributes "on the fly", even if they have not been previously defined.
- **Distributed**: the flexibility in data clusterization and replication enables the distribution of the storage on multiple nodes, so as to realize powerful "fault tollerance" systems.
- **Horizontally scalable**: opposedly to vertical scalability, here we have considerably scalable architectures, which enable the storage and management of Big Data.
- **Open-source**: phylosophy underlying the NoSQL community.

Basic Logic

- To be scalable, NoSQL DBs must waive some properties, hence they cannot strictly abide by the ACID model.
- Basic operational Logic:
  - Basically Available: always guarantee data availability.
  - Soft-state: system can change state even if no reads or writes occur
  - Eventual consistency: consistency can be reached later.


TYPES OF
No SQL
???

NoSQL Categories

KEY-VALUE DATA STORE

- Uses an associative array (key-value) as its storage model
- Key-based storage, update, and search
- Primitive data types familiar to programmers
- Simple
- Fast data retrieval
- Big data sizes

COLUMN-ORIENTED DATA STORE

- Data are organized in **columns** instead of rows
- A group of columns is named *family*, a concept similar to relational table
- Columns can be easily **distributed**
- Scalable
- Efficient
- **Fault-tolerant**

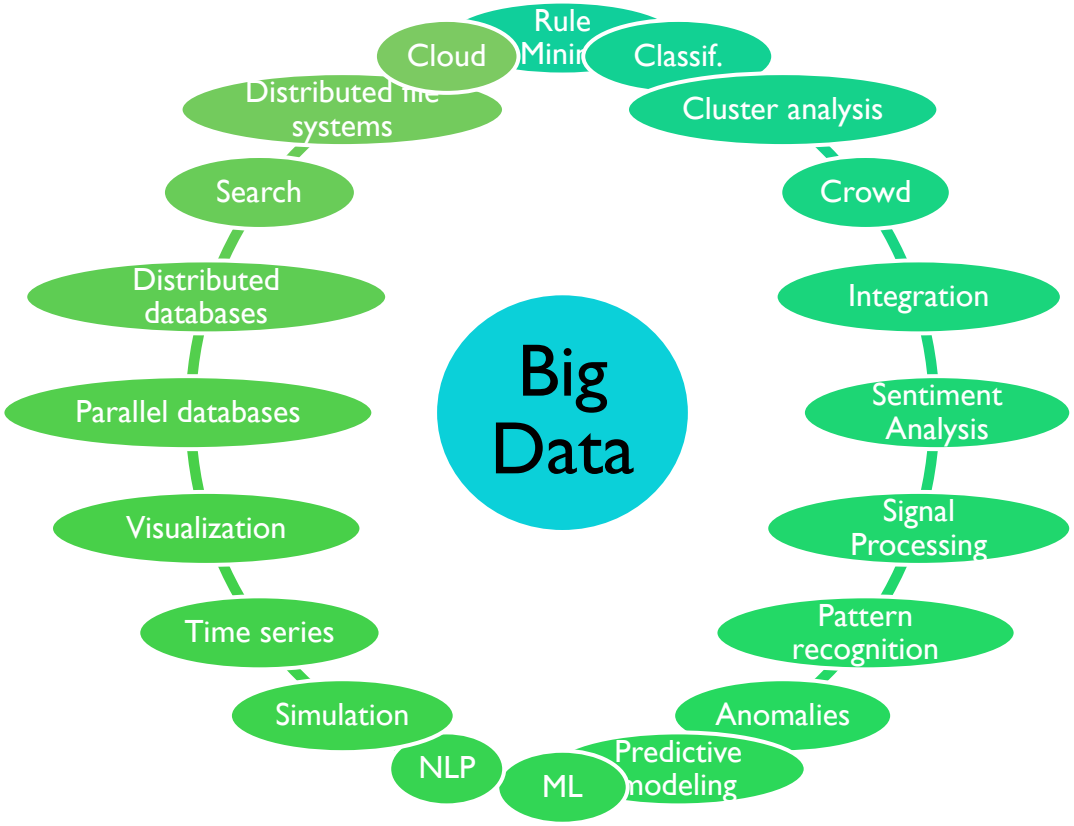
GRAPH-BASED DATA STORE

- Uses nodes (entities), properties (attributes), and edges (relations)
- Simple and intuitive Logical model
- Each element contains a pointer to its adjacent element
- Graph navigation to retrieve data
- Efficient for representing social networks or sparse data
- Relationships among central data

DOCUMENT DATA STORE

- Supports several types of documents
- A document is identified through a primary key
- Schema-less
- Horizontal scalability

Other technologies to approach big data/data science



Data profiling and data cleansing are prerequisites for all of these!