

Big Data

INTRODUZIONE

Una definizione per i Big Data

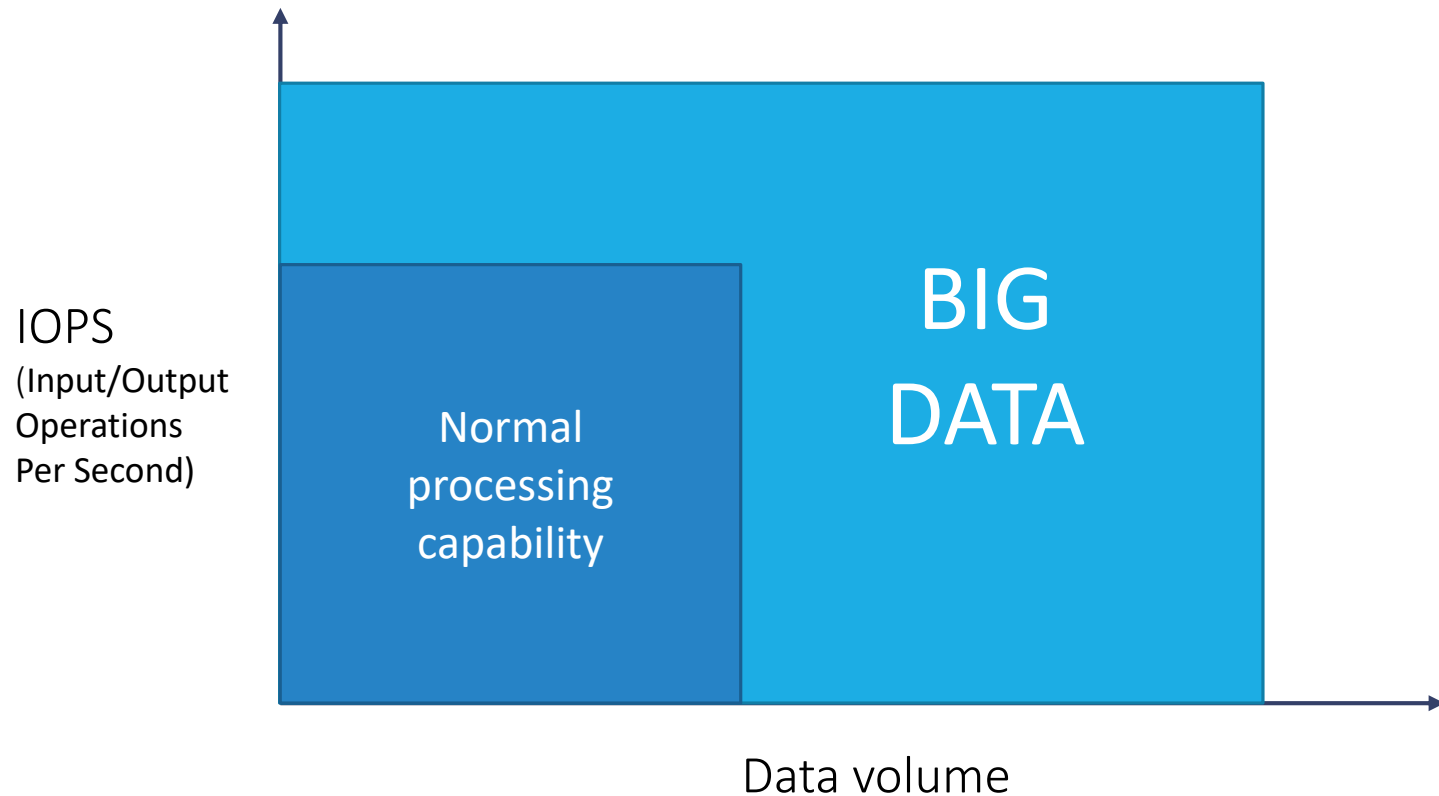
In realtà, ci sono diverse definizioni

"Big data exceeds the reach of commonly used hardware environments and software tools to capture, manage, and process it with in a tolerable elapsed time for its user population." - *Teradata Magazine article, 2011*

"Big data refers to data sets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze." -
The McKinsey Global Institute, 2012

"Big data is data sets that are so voluminous and complex that traditional data processing application software are inadequate to deal with them." -
Wikipedia, 2018

Quando i dati diventano "Big"?



Alcuni numeri

Quanti dati nel mondo?

- 800 Terabytes, 2000
- 160 Exabytes, 2006 ($1\text{EB} = 10^{18}\text{B}$)
- 500 Exabytes, 2009
- 2.7 Zettabytes, 2012 ($1\text{ZB} = 10^{21}\text{B}$)
- 35 Zettabytes nel 2020

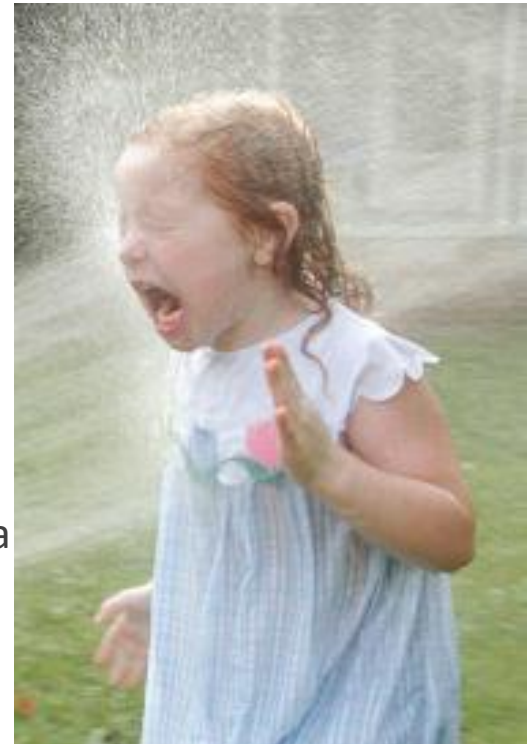
Zettabyte?

- 1,000,000,000,000,000,000,000 bytes
- Una pila di dischi rigidi da 1 TB di 25.400 km di altezza

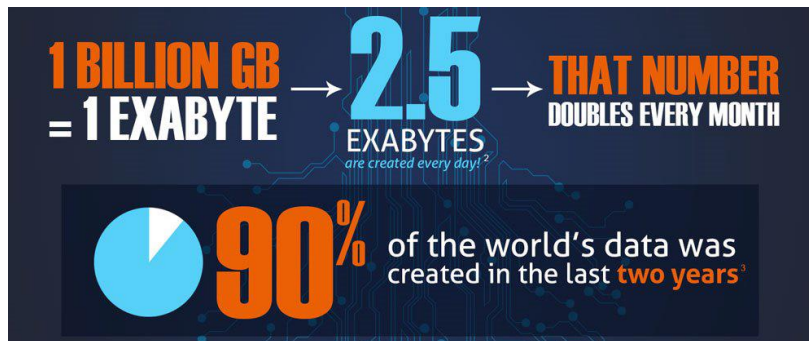
Quanti dati in un giorno?

- 7 TB, Twitter
- 10 TB, Facebook

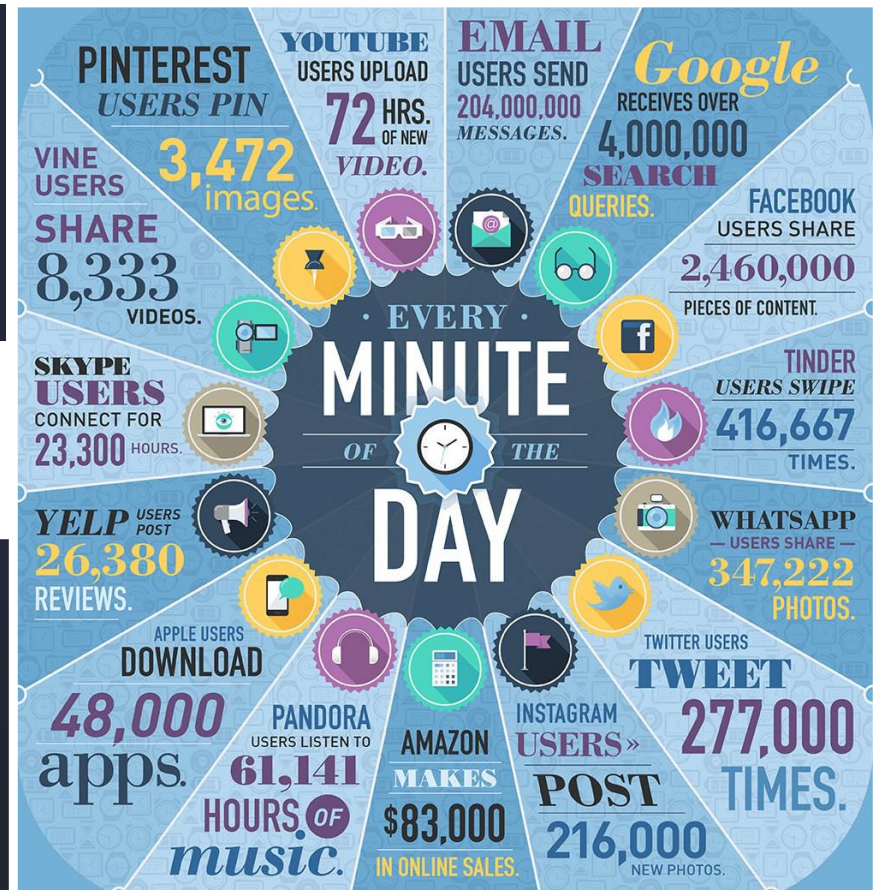
90% dei dati mondiali generati negli ultimi due anni!



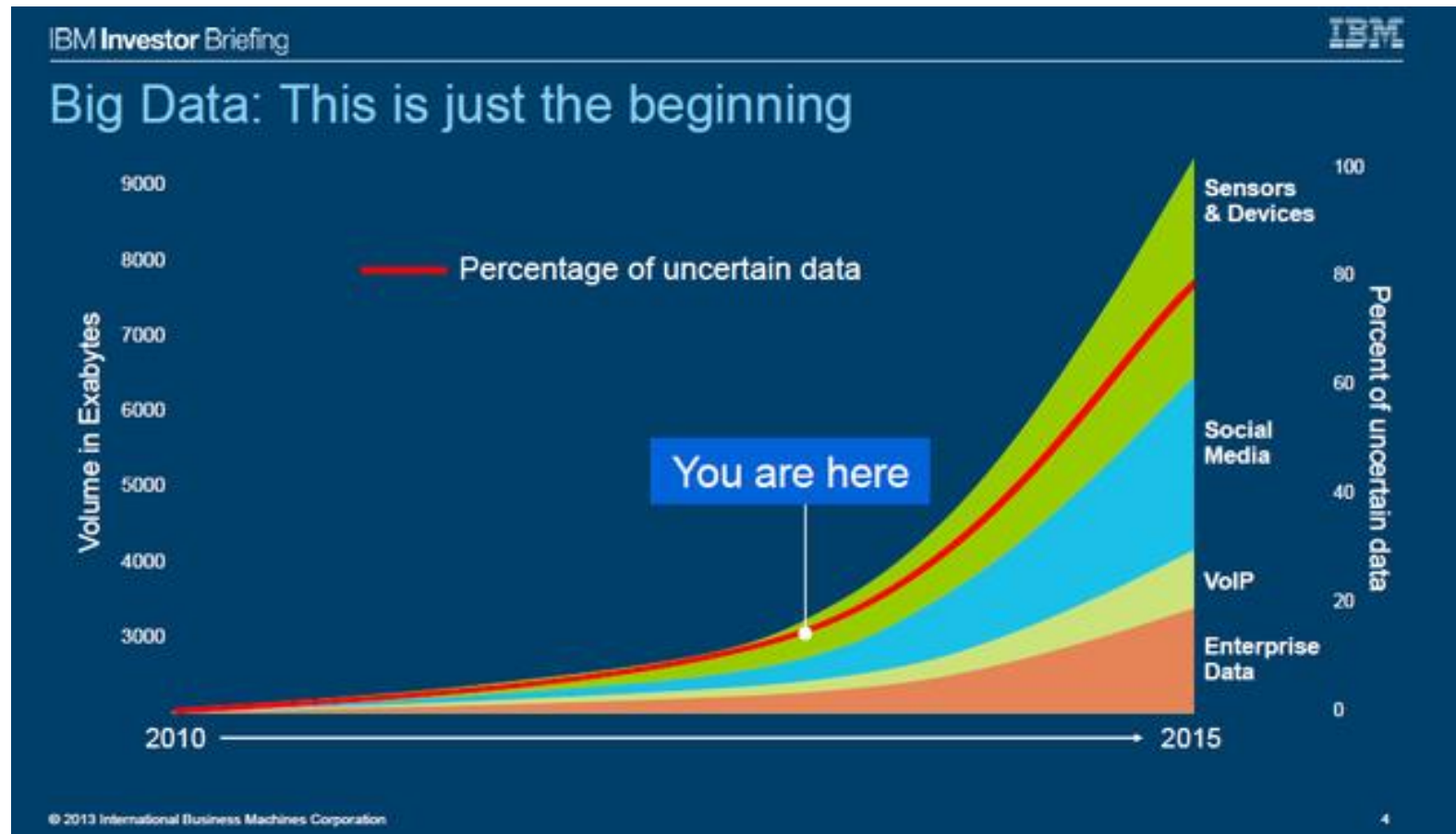
Crescita dei dati



MORE IPHONES
ARE SOLD THAN BABIES BORN

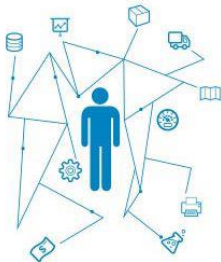


Crescita dei dati

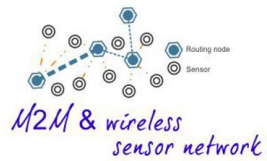
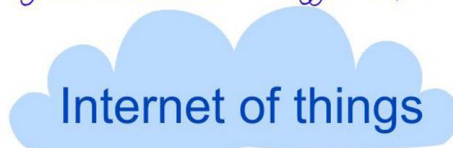


Internet of Things

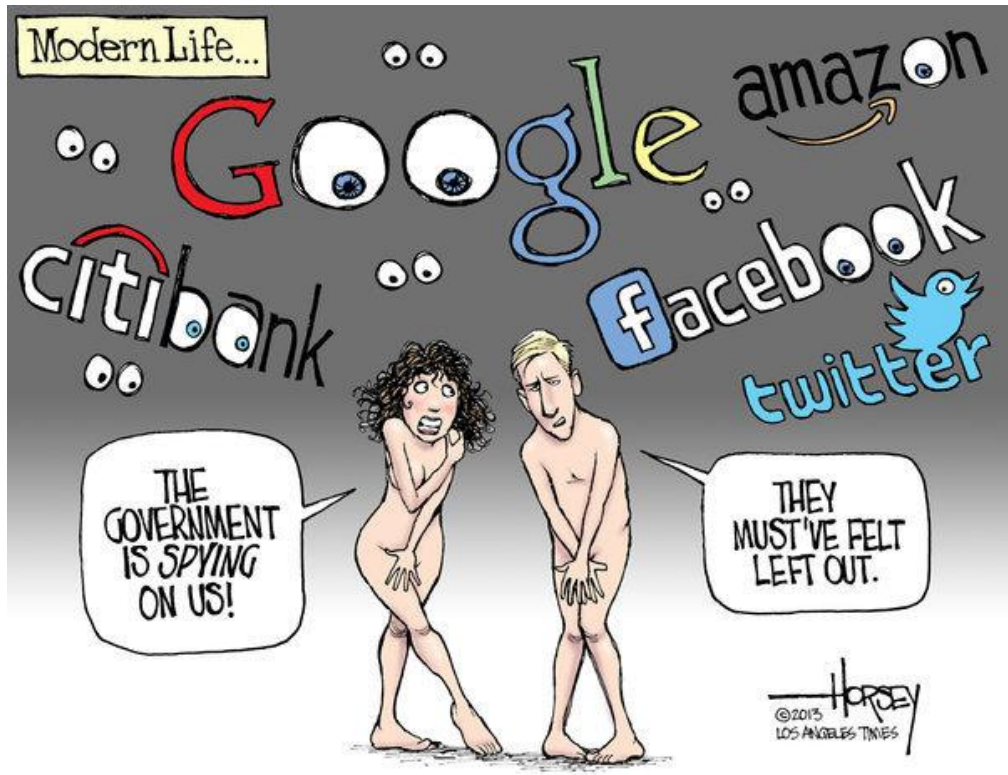
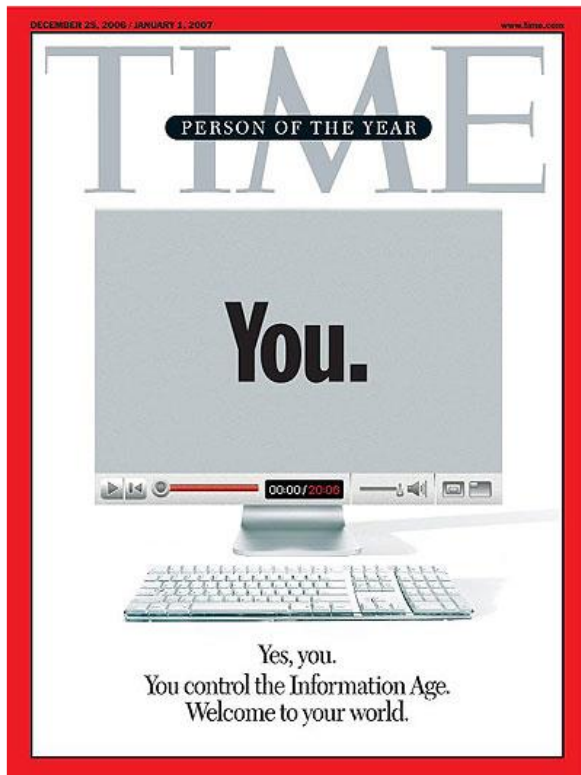
There will be as many as
**40 TO 80
BILLION**
connected objects
by 2020.



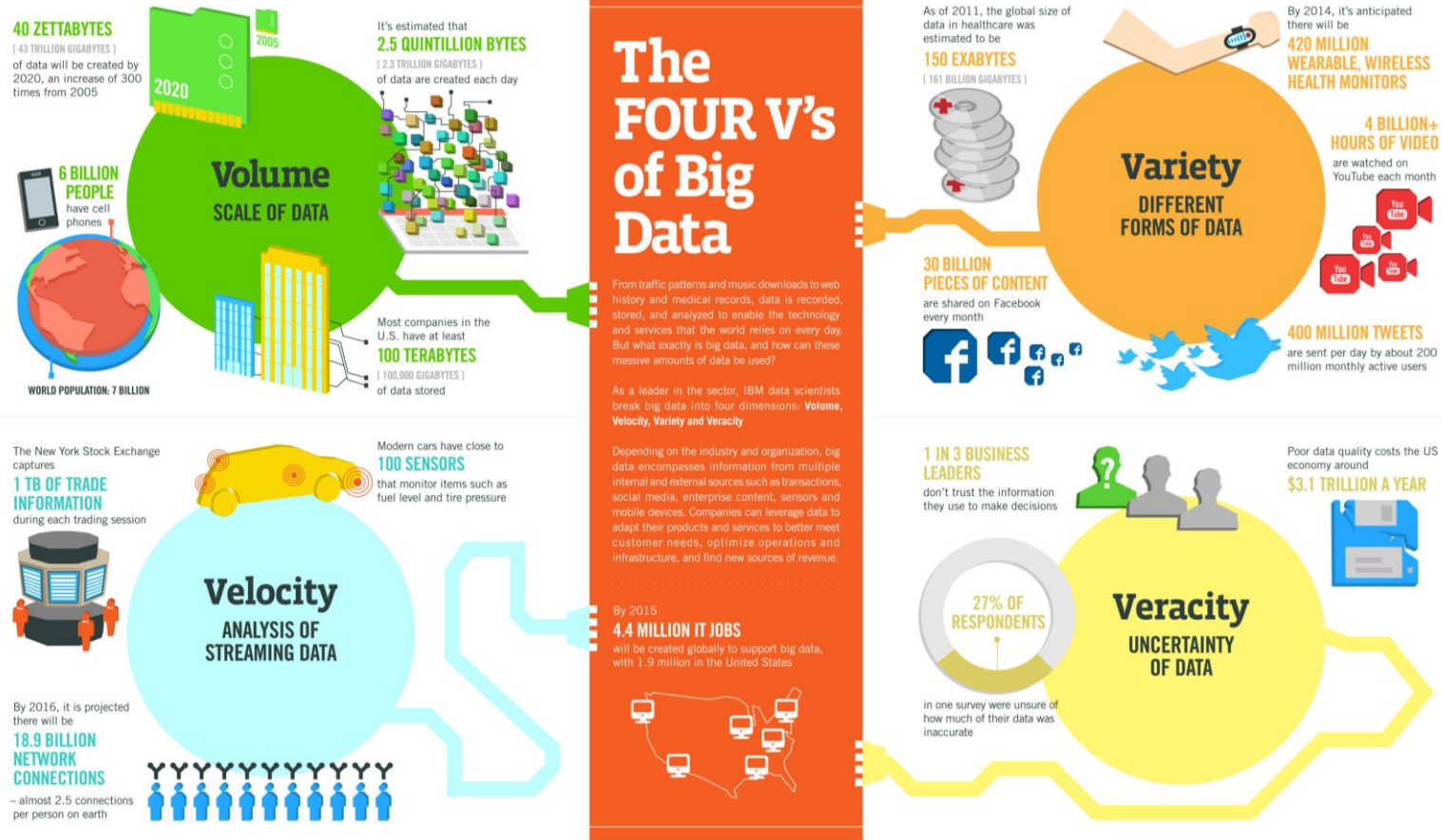
There will be
10 connected
objects
for every man,
woman, and child
on the **PLANET.**



Contenuti generati dagli utenti



Le quattro "V" di Big Data



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

http://www.ibmbigdatahub.com/sites/default/files/infographic_file/4-Vs-of-big-data.jpg

IBM

Big Data: V⁴+VALUE

Volume

- Terabyte(10^{12}), Petabyte(10^{15}), Exabyte(10^{18}), Zettabyte (10^{21})

Variety

- Strutturato, semistrutturato, non strutturato
- Testo, immagine, audio, video, registrazione

Velocity

- Periodico, quasi in tempo reale, in tempo reale

Veracity

- La qualità dei dati può variare notevolmente

Value

- I Big data possono generare enormi vantaggi competitivi

Cosa c'è di nuovo?

L'ampia disponibilità di dati ci permette di applicare modelli più sofisticati e ottenere risultati molto più accurati rispetto al passato!



Anthony
Goldbloom



William
Deming

It is a capital mistake to theorize before one has data

The bigger the data set you have, the more accurate the predictions about the future will be

If you torture the data long enough, it will always confess

In God we trust; all others must bring data



Ronald
Coase

Bigger = Smarter?

Sì!

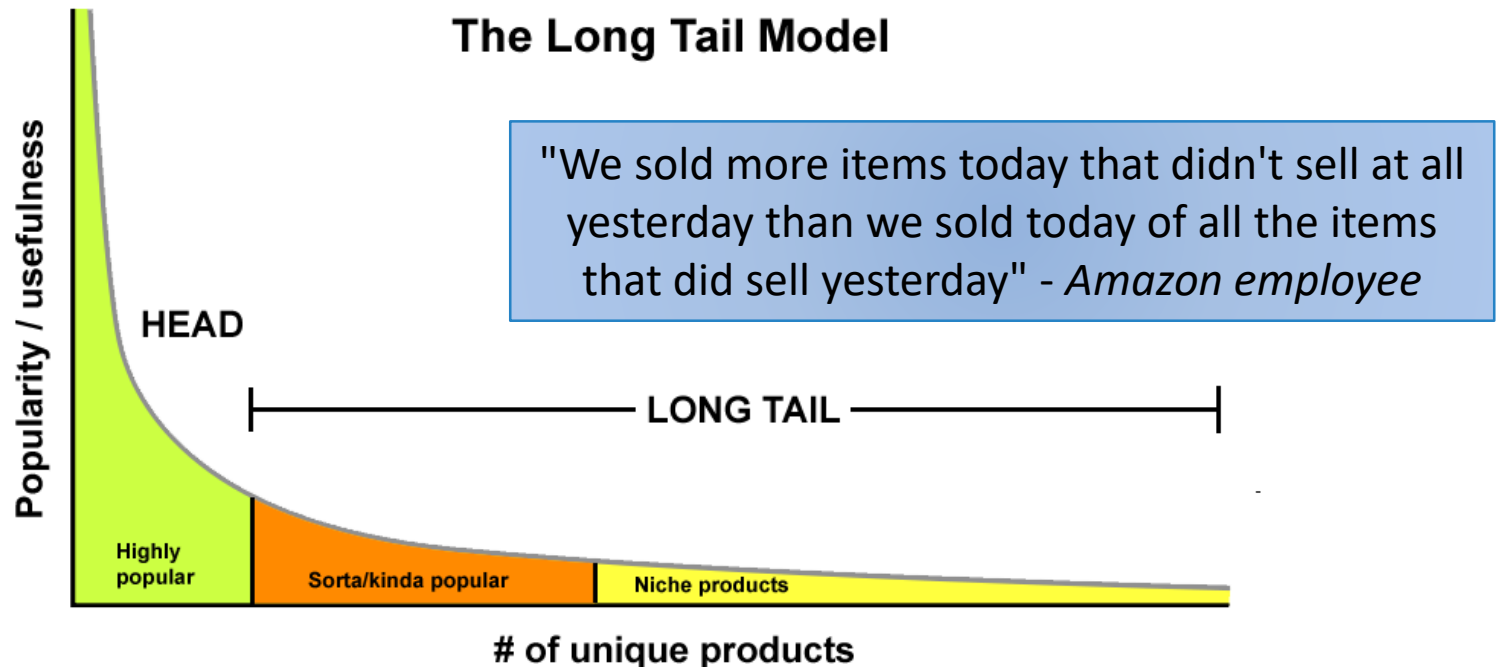
- Gli algoritmi funzionano molto meglio
- Tollerare errori
- Sfruttare la "long tail"

Ma...

- Più eterogeneità; più dati, più errori
- I dati crescono più velocemente rispetto alla capacità dei chip
- Con abbastanza dati si può provare qualsiasi cosa
- Ancora bisogno di esseri umani per porre domande giuste



Long tail



Coda lunga: il valore più alto non proviene da pochi articoli molto popolari (regola 80/20), ma dalla lunga lista di articoli di nicchia

Insieme, i dati insignificanti sono in realtà i più significativi

Perché adesso?

Perché abbiamo dati

- Dati nati già in forma digitale
- 40% di crescita dei dati all'anno



Perché possiamo

- \$400 per un disco in cui memorizzare tutta la musica del mondo
- > 40 anni di legge di Moore e grandi risorse computazionali
- 76% delle organizzazioni hanno investito in grandi dati in 2016
- \$130 miliardi investiti in Big data in 2016



"Perché abbiamo raggiunto la fine della logica"



Un semplice esempio di bigger=smarter

Google translate

- Raccogli frammenti di traduzioni
- È possibile abbinare le frasi a frammenti
- È possibile eseguire il debug continuo del sistema

Perché funziona?

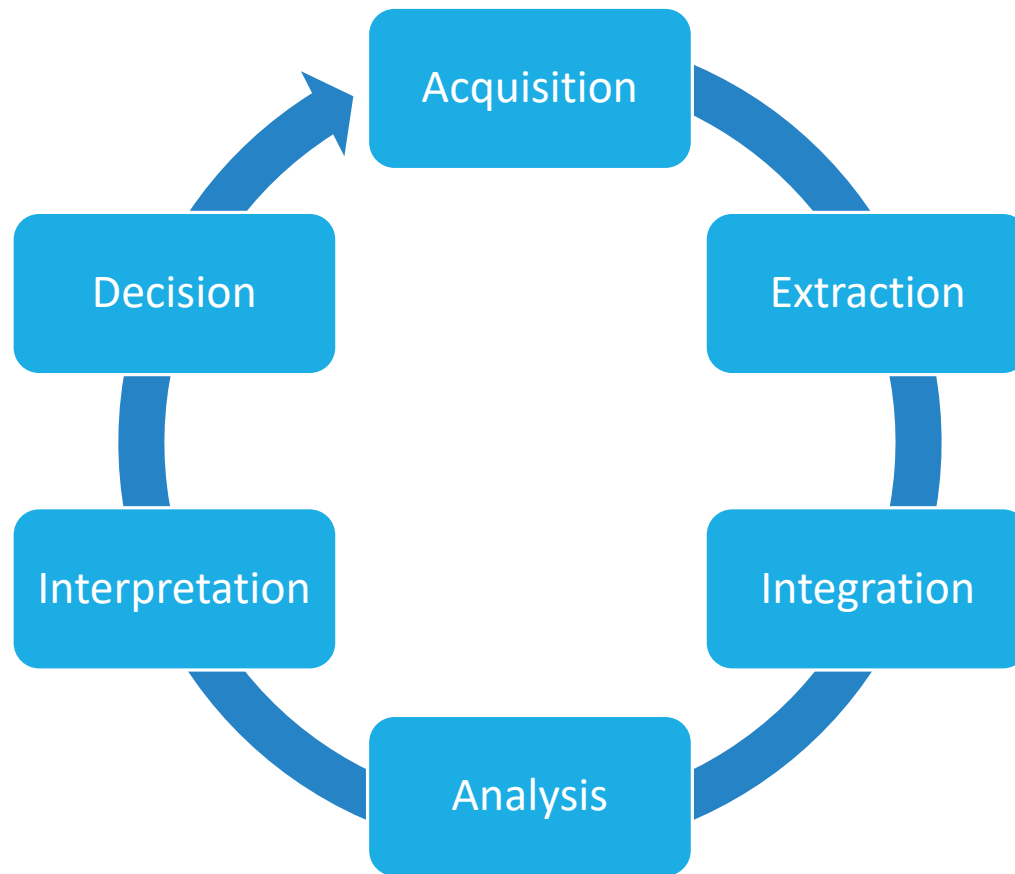
- Ci sono tonnellate di dati sul Web
- La precisione migliora man mano che il training set cresce



Casi di utilizzo

Today's Challenge	New Data	What's Possible
Healthcare Expensive office visits	Remote patient monitoring	Preventive care, reduced hospitalization
Manufacturing In-person support	Product sensors	Automated diagnosis, support
Location-Based Services	Real time location data	Geo-advertising, traffic, local search
Public Sector Standardized services	Citizen surveys	Tailored services, cost reductions
Retail One size fits all marketing	Social media	Sentiment analysis segmentation

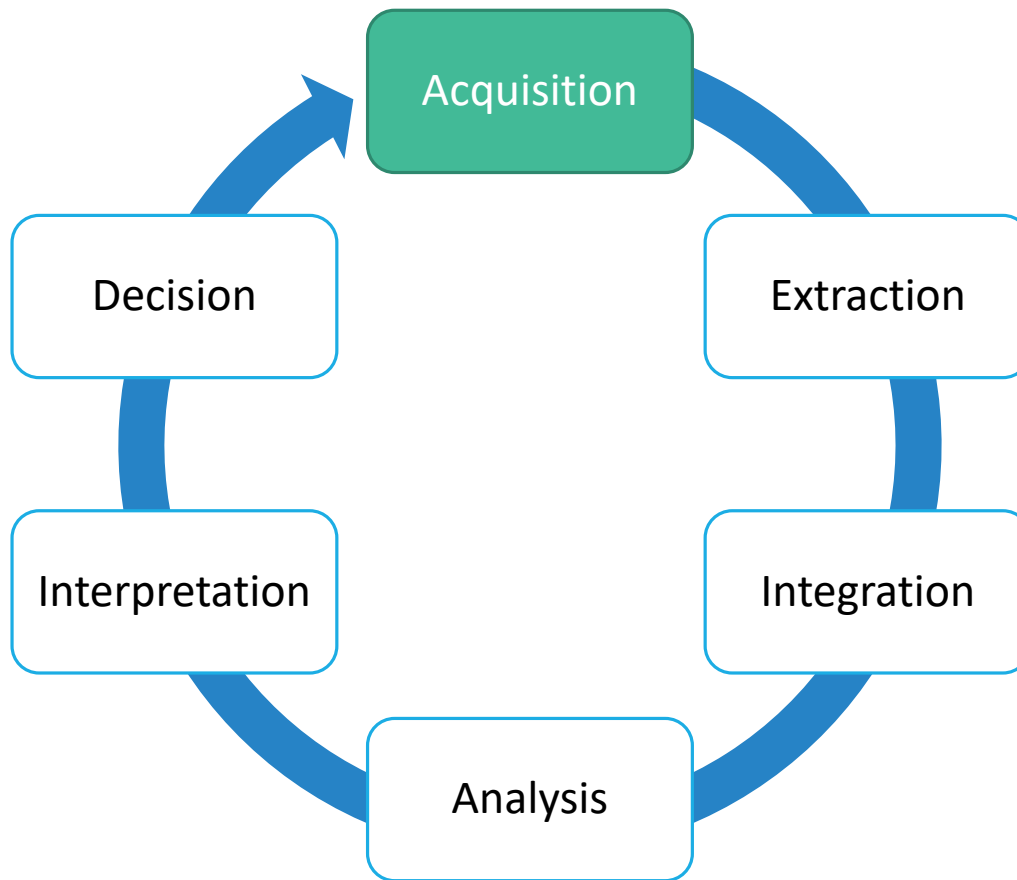
Big Data in action



Obiettivo

Prendere decisioni strategiche efficaci sfruttando la disponibilità di Big Data

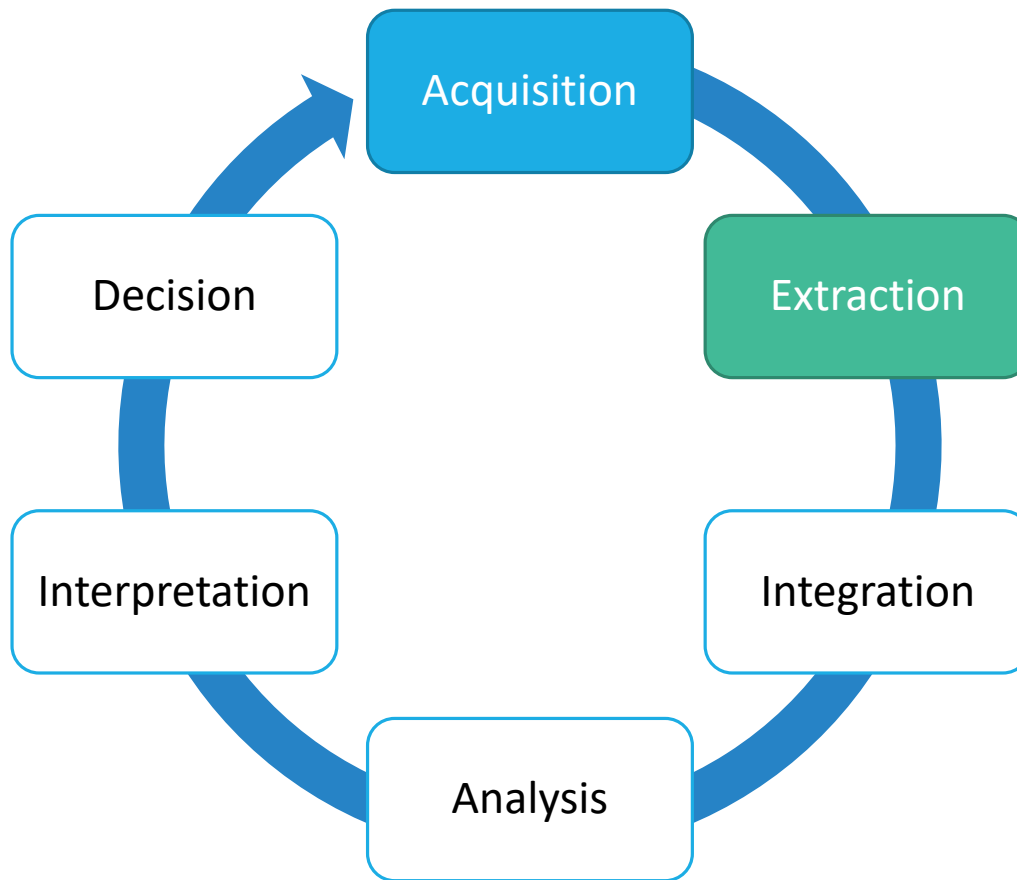
Big Data in action



Richiede

- Selezione
- Filtro
- Generazione di metadati
- Gestione della provenienza

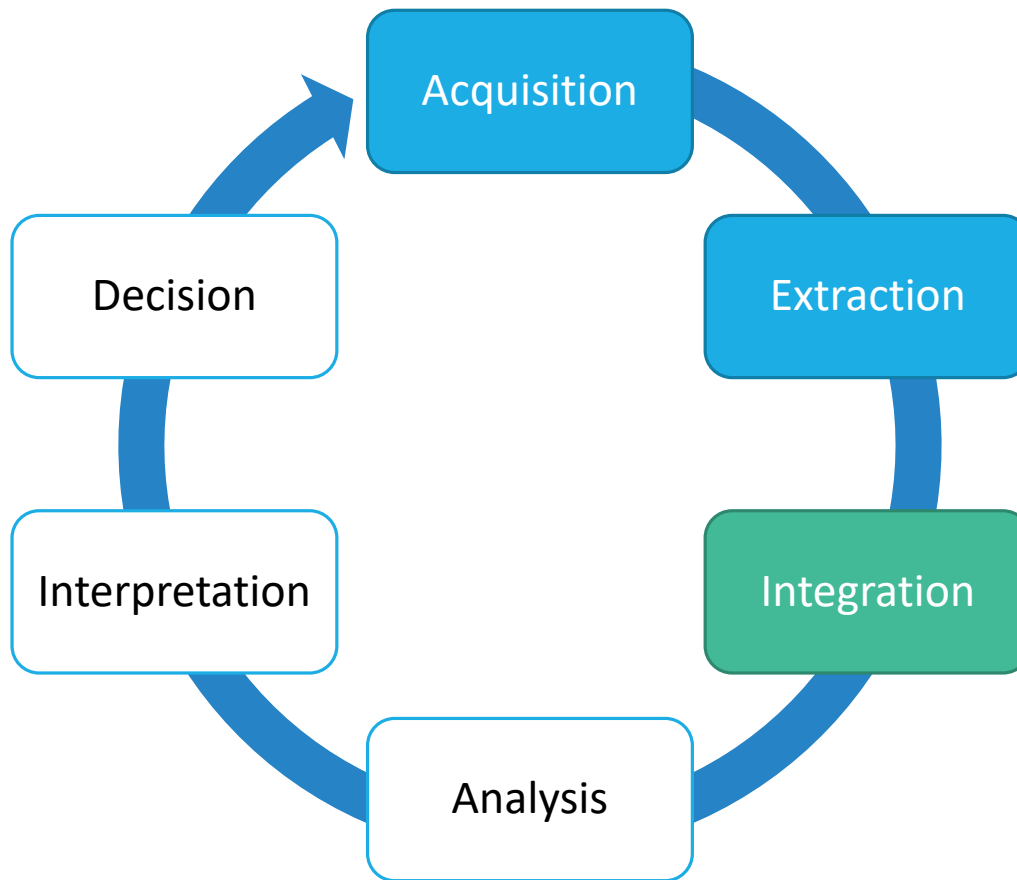
Big Data in action



Richiede

- Trasformazione
- Normalizzazione
- Pulizia
- Aggregazione
- Gestione degli errori

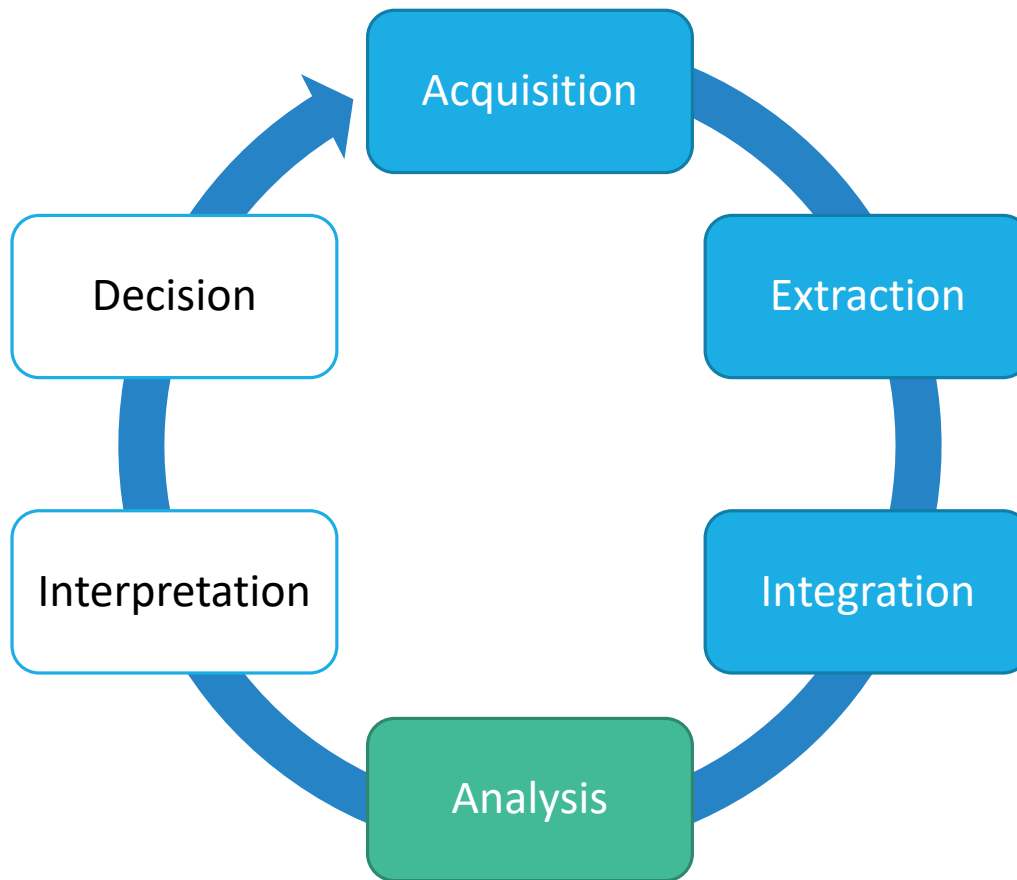
Big Data in action



Richiede

- Standardizzazione
- Gestione dei conflitti
- Riconciliazione
- Definizione di mappatura

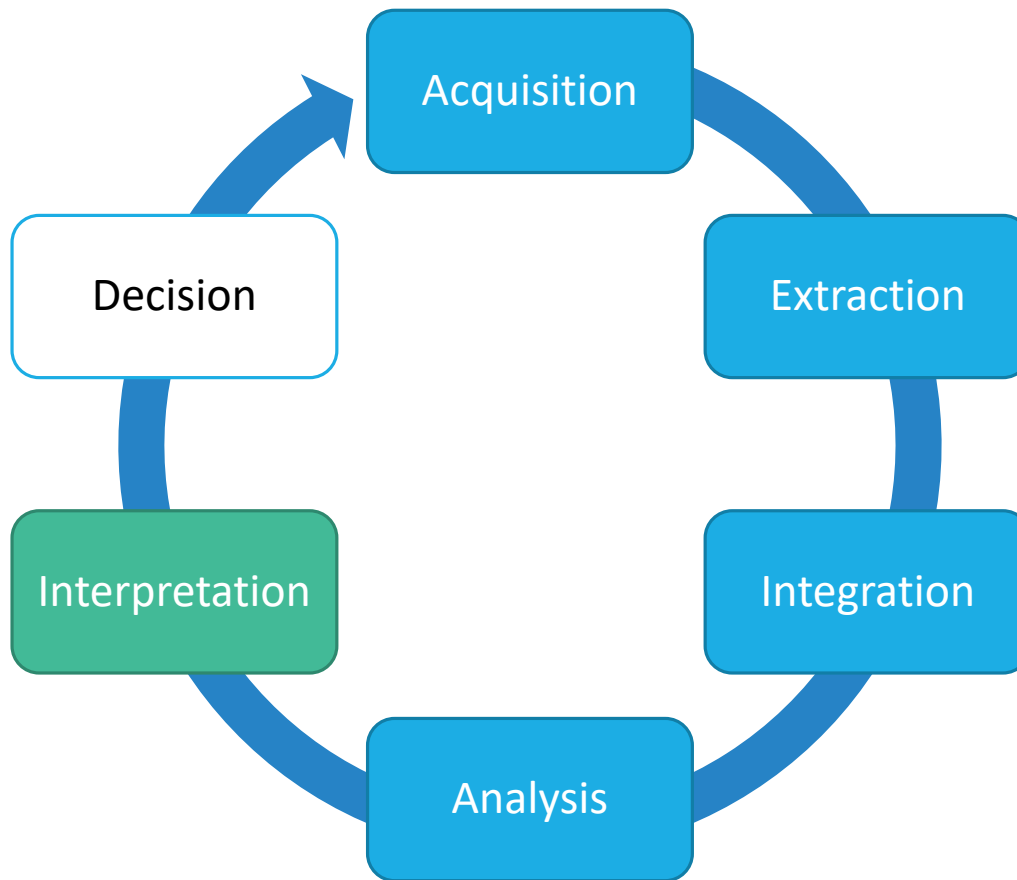
Big Data in action



Richiede

- Esplorazione
- Data mining
- Apprendimento automatico
- Visualizzazione

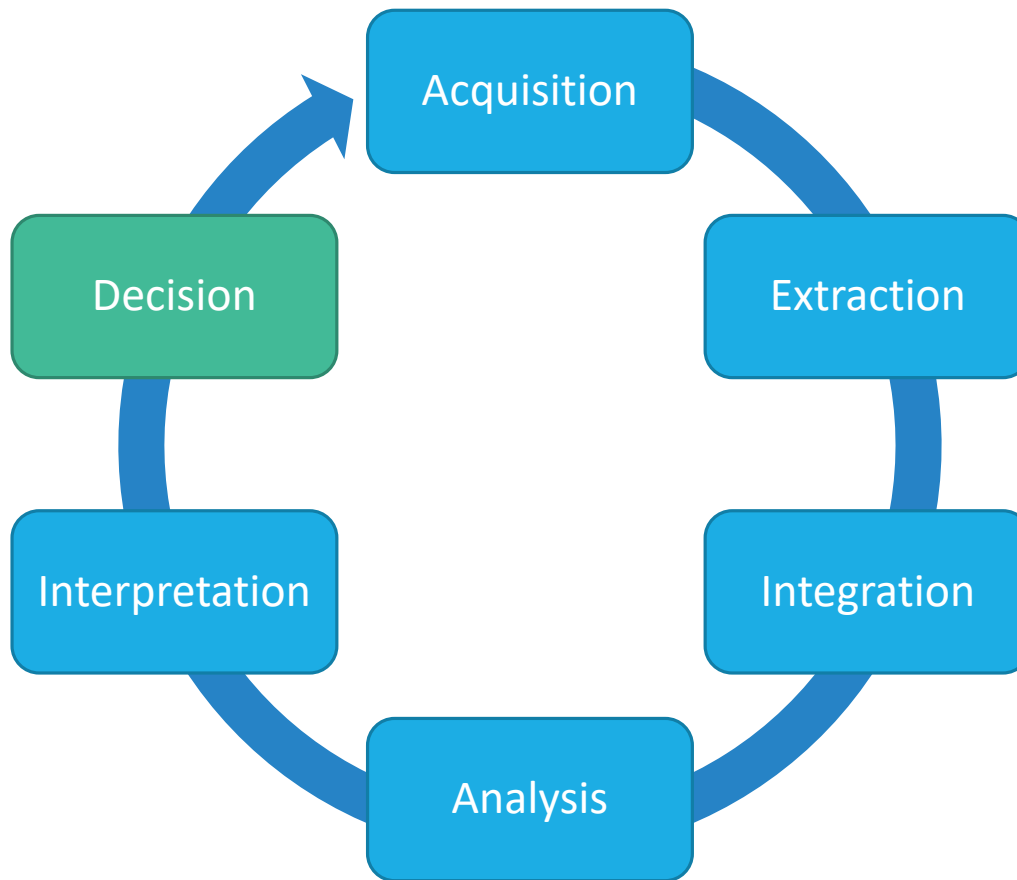
Big Data in action



Richiede

- Conoscenza del dominio
- Conoscenza della provenienza
- Identificazione dei modelli di interesse
- Flessibilità del processo

Big Data in action



Richiede

- Competenze manageriali
- Miglioramento continuo del processo

Un semplice esempio

Problema: La vendita di leccalecca sta scendendo!

Acquisition:

- Vendite per cliente, regione e tempo
- Sondaggi degli utenti
- I social network

Extraction:

- Caricamento dei dati dalle ricevute
- Lettura automatica dei questionari
- Estrazione dati da Twitter

Integration:

- Sulla base dei tipi di utenti

Analysis:

- Leccalecca acquistati da persone più vecchie di 25
- Leccalecca preferito da persone più giovani di 10



Interpretation:

- Le mamme credono: leccalecca = carie

Decision:

- Facciamo leccalecca senza zucchero
- Chiediamo ai dentisti di pubblicizzare i nostri lecci

Sfide

Prestazioni, prestazioni, prestazioni!

- I dati crescono più velocemente rispetto alla capacità di gestirli
- Efficienza
- Scalabilità

Eterogeneità

Efficacia

Flessibilità

Privacy

Costi



BIG DATA LANDSCAPE 2017

INFRASTRUCTURE

HADOOP ON-PREMISE

- cloudera
- Hortonworks
- MAPR
- Pivotal
- IBM InfoSphere
- bluedata
- jethro

HADOOP IN THE CLOUD

- amazon web services
- Microsoft Azure
- Google Cloud Platform
- IBM InfoSphere BigInsights
- treasure data
- aletris
- altriscale
- CAZENA
- CenturyLink

STREAMING / IN-MEMORY

- amazon web services
- confluent
- databricks
- stream
- GridGain
- metamarkets
- DATATORRENT
- dataArtisans
- ORACLE
- hazelcast
- TEPRACOTTA

NOSQL DATABASES

- Google Cloud Platform
- ORACLE
- amazon dynamodb
- Microsoft Azure
- MarkLogic
- mongoDB
- DATASTAX
- redislabs
- Couchbase
- influxdata

NEWSQL DATABASES

- SAP
- Clustrix
- Pivotal
- nuodb
- Cockroach Labs
- memsql
- splice
- VolDB
- citusdata
- Triforce
- deepdb
- paradigm4

GRAPH DBS

- neo4j
- IBM
- ORACLE
- orientDB
- Objectivity

MPP DBS

- TERADATA
- VERTICA
- NETEEZA
- Qcution
- kognitio
- dremio

CLOUD EDW

- amazon web services
- Google Cloud Platform
- Microsoft Azure
- Pivotal
- snowflake
- InfoWorks

DATA TRANSFORMATION

- talend
- pentaho
- alteryx
- TRIFACTA
- tamr
- Paxata
- StreamSets
- UNIFI

DATA INTEGRATION

- informatica
- TEALUM
- enigma
- alooma
- ZALONI
- splicely
- Import
- Stitch

DATA GOVERNANCE

- informatica
- IBM
- skyhigh
- collibra
- Waterline
- Alation

MGMT / MONITORING

- amazon web services
- New Relic
- AppDynamics
- octrino
- WAVEFRONT
- unravel
- splunk
- Tracoda
- pagerduty
- Numerify

STORAGE

- amazon web services
- Google Cloud Platform
- Microsoft Azure
- ALLUXIO
- nimblescale
- Cumulo
- COHO
- panasas

CLUSTER SERVICES

- amazon web services
- Google Cloud Platform
- Microsoft Azure
- ALLUXIO
- nimblescale
- Cumulo
- COHO
- panasas

APP DEV

- Lightbend
- rainforest
- CRSK

CROWDSOURCING

- amazon web services
- upwork
- WorkFusion

HARDWARE

- Google TPU
- ARM
- Graphcore
- MYTHIC
- INVIDIA
- Movidius
- SCORTEX

ANALYTICS

DATA ANALYST PLATFORMS

- Microsoft
- pentaho
- alteryx
- Digital Reasoning
- GUAVUS
- AYASDI
- WATTIVO
- Datameer
- Quid
- ClearStory
- OrigamiLogic
- interlana
- Boffnouse
- ARIMO
- ENDOR
- MODE

DATA SCIENCE PLATFORMS

- IBM
- KNIME
- dataiku
- DOMINO
- yhat
- CONTINUUM
- Alpine
- ALGORITHMIA
- Angoss

BI PLATFORMS

- Microsoft
- amazon web services
- Google Cloud Platform
- looker
- Wave Analytics
- arc42
- GoodData

VISUALIZATION

- tableau
- Google Cloud Platform
- Qlik
- CELONIS
- Periscope
- CHARTIO
- plenty

VERTICAL ANALYTICS

- PREDIX
- CAPE
- UPTAKE
- Optical Insights
- TACHYUS
- Aluminum
- datorama

STATISTICAL COMPUTING

- SAS
- SPSS

DATA SERVICES

- Palantir
- OPERA
- DATA SCIENCE
- kaggle
- DATAKIND

MACHINE LEARNING

- Google Cloud Platform
- H2O
- DataRobot
- occontext relevant
- VIZENZE
- bonsai
- DATAFORMA
- Ionian

HORIZONTAL AI

- IBM Watson
- Cortana
- Face++
- sentient
- Voyager
- clarifai
- Affective
- SENSETIME
- Arconom
- PETUUM
- OSASRO
- BLUE
- VISION

SPEECH & NLP

- Google Cloud Platform
- Amazon
- Microsoft
- IBM
- Facebook
- Twitter
- LinkedIn
- Slack
- WhatsApp
- Telegram
- Skype
- WhatsApp
- Telegram
- Skype

SEARCH

- elastic
- Autodesk
- ORACLE
- THOUGHTSPOT
- Lucidworks
- swiftype
- alphasense
- Searchkick
- SINEQUA

LOG ANALYTICS

- splunk
- sumologic
- loggly
- lwbano
- Logz.io

SOCIAL ANALYTICS

- Hootsuite
- NETBASE
- DATA SIFT
- tracx
- synthesio
- simple reach
- bitly
- predata

WEB / MOBILE / COMMERCE ANALYTICS

- Google Analytics
- mixpanel
- sumall
- retention
- SIGOPT
- granify
- custora

APPLICATIONS - ENTERPRISE

SALES

- VERSTEIN
- CHORUS
- INSIDESALES.COM
- conversica
- clari
- AVISO
- TACT
- fuse|machines
- TROOPS

MARKETING - B2B

- RADIUS
- EVERSTRING
- App Annie
- Lattice
- infer
- MINTIGO
- sense
- tubular
- Datafix
- ENGAGIO

MARKETING - B2C

- bloomreach
- PERSADO
- blueyonder
- kahuna
- SAULTHRU
- QUANTIFIND
- mparticle
- Amplero

CUSTOMER SERVICE

- MEDALLIA
- zendesk
- CLARABRIDGE
- CLICKFOX
- DigitalGenius
- automat
- mscdi
- INTERCOM

HUMAN CAPITAL

- HowVia
- entelo
- hiQ
- textio
- Wade & Wendy
- Stella
- pymetrics

LEGAL

- RAVEL
- Seal
- Everlaw
- Lexipol
- Brevia
- RSS
- casetext

FINANCE

- anaplan
- uora
- tidemark
- SAHANA
- TRADESHIFT

ENTERPRISE PRODUCTIVITY

- slack
- facebook
- ORACLE
- lumiata
- Clara
- talia
- butter.ai
- KASIST

BACK OFFICE AUTOMATION

- HyperScale
- opricity
- AppZen

SECURITY

- TANUIM
- cyberwatch
- SONIFY
- BlueTalon
- Secure
- ARCA
- Secure
- ARCA
- Secure
- ARCA
- Secure

APPLICATIONS - INDUSTRY

ADVERTISING

- AppNexus
- critico
- rocketfuel
- OpenX
- MOAT
- drawbridge
- gumgum
- Yieldmo

EDUCATION

- KNOWTON
- Clever
- Geclara
- kidaptive
- PANDORA
- knowte

GOVERNMENT

- Socrata
- OPENGOV
- mark43
- TrueAccord
- OpenDataSoft

FINANCE - LENDING

- OnDeck
- Kreditech
- AVANT
- INSIKT
- MoneyLion
- TrueAccord
- ignifi
- aire

FINANCE - INVESTING

- Dataminr
- KENSCH
- Quantopian
- NUMERA
- ISENTUM
- claritymoney
- ALGORIC
- ADIVA
- FlowerPack

REAL ESTATE

- Opendoor
- VTS
- CREDICI
- reonomy
- COMPSTAK

INSURANCE

- Metromile
- Lemonade
- CYENCE
- Shift Technology
- Tractable

HEALTHCARE

- FLATIRON
- Healthio
- Gingerio
- COTA
- Acure
- imagic
- Qventus
- Qventus
- Qventus

LIFE SCIENCES

- color
- ocean
- BenvenotAI
- ZEPHYR
- Clear Labs
- Phosphor
- Atomwise
- twoAR
- nap
- panomics

TRANSPORTATION

- UBER
- TESLA
- clearpath
- drive.ai
- OTO
- OTO
- OTO
- OTO
- OTO

AGRICULTURE

- FARMERS
- FarmersEdge
- FarmLogs
- BLUE DRIVER
- mavix
- prosperra

COMMERCE

- instacart
- STITCH FIX
- RetailNext
- HowGood

OTHER

- eHarmony
- stem
- rethink
- robotics
- Boxever
- VERIDICRIS
- Quanta
- Quanta

CROSS-INFRASTRUCTURE/ANALYTICS

amazon web services, Google Cloud Platform, Microsoft, IBM, SAP, Hewlett Packard Enterprise, SAS, data, vmware, TIBCO, TERADATA, ORACLE, NetApp

OPEN SOURCE

FRAMEWORK

- hadoop
- hadoop
- Flink
- YARN
- TEZ
- MESOS
- CDAP

QUERY / DATA FLOW

- Spark
- SQL
- presto
- SLAMDATA
- ARACHNE DRILL
- Google Cloud Dataflow

DATA ACCESS

- cassandra
- nifi
- mongoDB
- COUCHDB
- CHICKEN
- riak
- HERASE
- Spanner
- ACCUTULO

COORDINATION

- talend
- Apache Zookeeper
- Apache Ambari

STREAMING

- Spark
- Flink
- beam
- kafka
- druidd
- STORM

STAT TOOLS

- python
- ScalaLab
- Nucliy
- SciPy

AI / MACHINE LEARNING / DEEP LEARNING

- theano
- Caffe
- TensorFlow
- Apache SINGA
- OpenAI
- neon
- FeatureFu
- DL4J
- DSSTNE
- milib
- Chainer
- DIMSUS
- Aerosolve

SEARCH

- elasticsearch
- Solr

LOG ANALYSIS

- elasticsearch
- kibana
- logstash

VISUALIZATION

- BEAKER
- Rodeo

COLLABORATION

- jupyter
- ANACONDA

SECURITY

- Apache Ranger
- KNOX
- Sentry

DATA SOURCES & APIS

HEALTH

- JAWBONE
- VALIDIC
- practice fusion
- fitbit
- GARMIN
- Human API
- kinsa

IOT

- GE Digital
- UPTAKE
- ThingWorx
- helium
- samsara

FINANCIAL & ECONOMIC DATA

- Bloomberg
- THOMSON REUTERS
- DOW JONES
- S&P CAPITAL IQ
- CBINSIGHTS
- xignite
- quandl
- YODLEE
- PREMISE
- estimote
- Eagle Alpha
- StockTwits
- PLAID
- mattermark

AIR / SPACE / SEA

- PLANET
- Airware
- spire
- SKYYATCH
- AIRBOTICS
- UNDERSTORY
- WINDOWWARD
- DroneDeploy

PEOPLE / ENTITIES

- axiom
- Experian
- EPISILON
- InsideView
- Crimson Hexagon
- BASIS
- quantcast
- SAFEGRAPH

LOCATION INTELLIGENCE

- FOURSQUARE
- Sense
- PlaceIQ
- esri
- factual
- CARTO
- Mapillary
- STREETLINE

OTHER

- qualtrics
- DATA.GOV
- data.world
- panjiva
- enigma

DATA RESOURCES

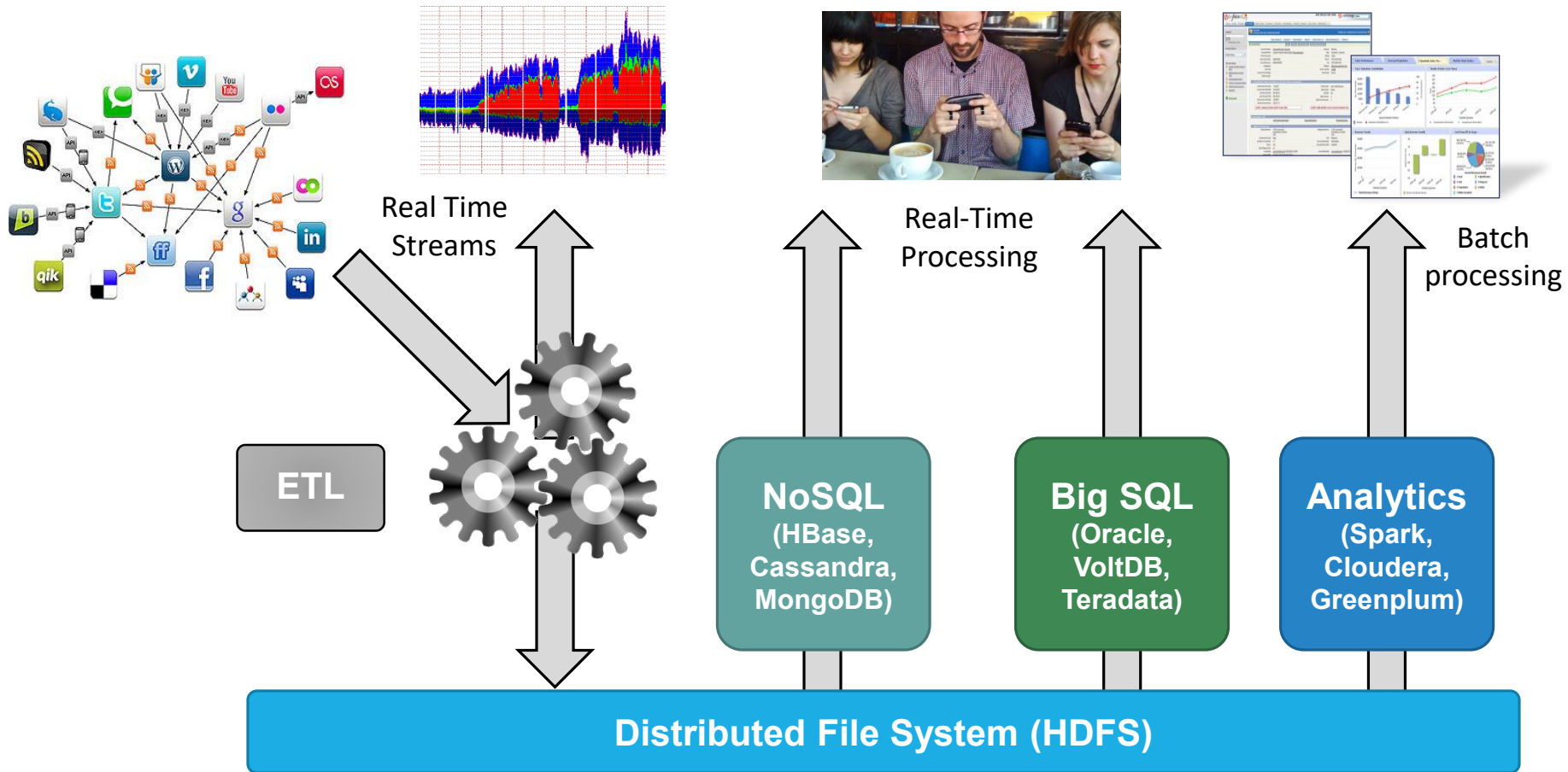
INCUBATORS & SCHOOLS

- PLURALSIGHT
- DATA CAMP
- INSIGHT
- The Data Incubator
- NETIS

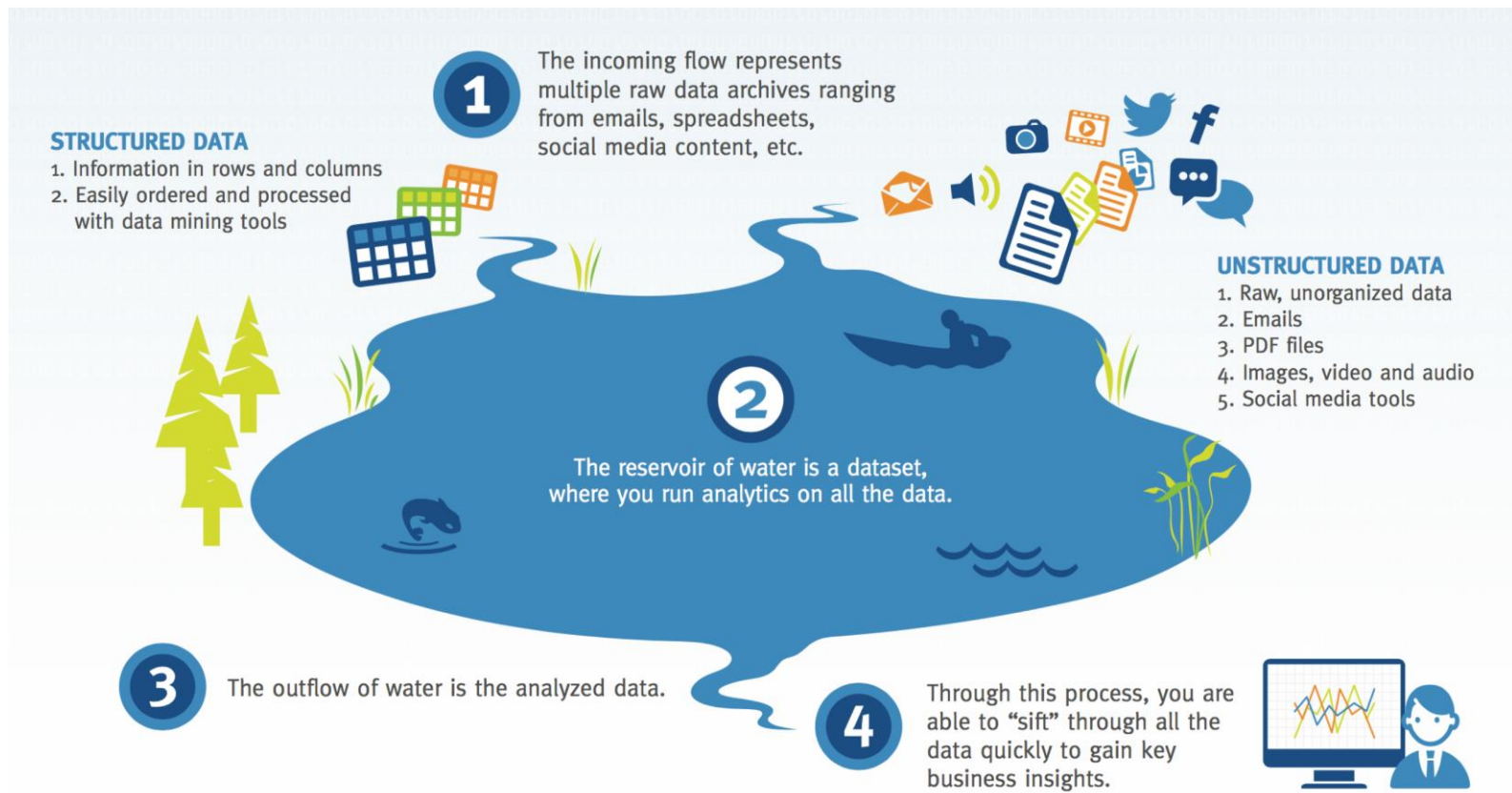
RESEARCH

- facebook research
- OpenAI
- MIRI
- CEAIL
- AI2
- ALLEN INSTITUTE FOR ARTIFICIAL INTELLIGENCE

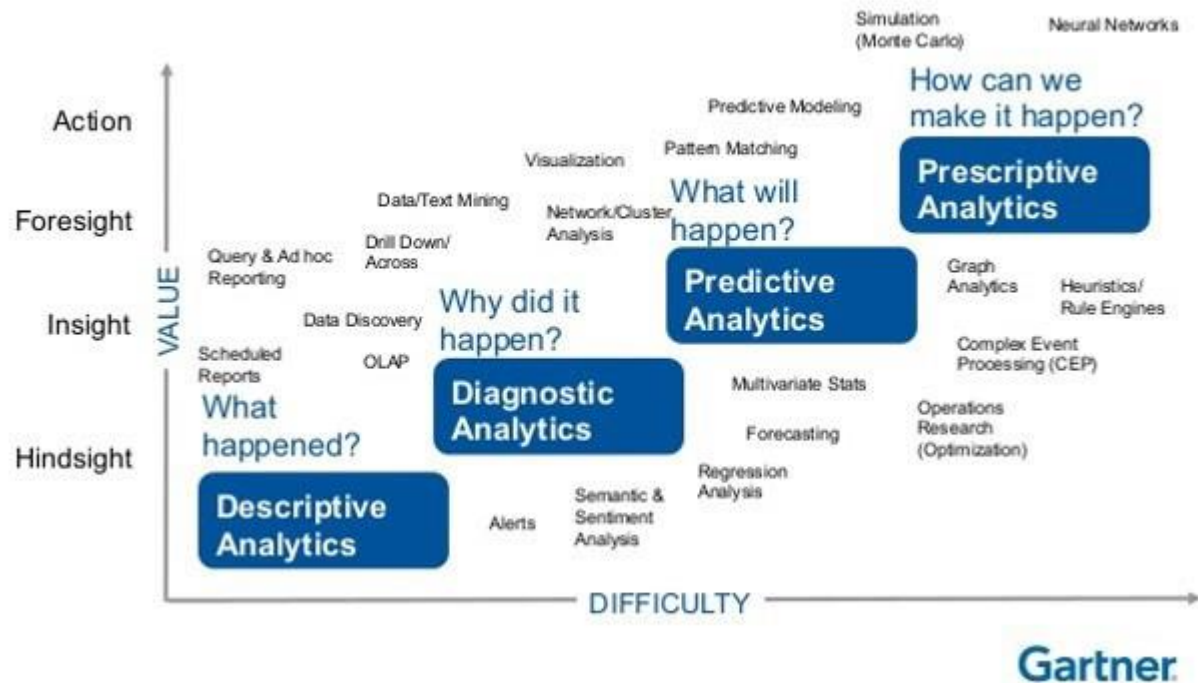
The Big Data flow



Data lake



Obiettivi dell'analisi



Data Scientist: una professione nuova di zecca

Data Scientist: The Sexiest Job of the 21st Century [Harvard Business Review 2013]

Data scientist? A guide to 2015's hottest profession [Mashable 2015]

"It's official – data scientist is the best job in America" [Forbes, 2016]



Competenze data scientist

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of the 21st century, requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- ☆ Machine learning
- ☆ Statistical modeling
- ☆ Experiment design
- ☆ Bayesian inference
- ☆ Supervised learning: decision trees, random forests, logistic regression
- ☆ Unsupervised learning: clustering, dimensionality reduction
- ☆ Optimization: gradient descent and variants

DOMAIN KNOWLEDGE & SOFT SKILLS

- ☆ Passionate about the business
- ☆ Curious about data
- ☆ Influence without authority
- ☆ Hacker mindset
- ☆ Problem solver
- ☆ Strategic, proactive, creative, innovative and collaborative



PROGRAMMING & DATABASE

- ☆ Computer science fundamentals
- ☆ Scripting language e.g. Python
- ☆ Statistical computing packages, e.g., R
- ☆ Databases: SQL and NoSQL
- ☆ Relational algebra
- ☆ Parallel databases and parallel query processing
- ☆ MapReduce concepts
- ☆ Hadoop and Hive/Pig
- ☆ Custom reducers
- ☆ Experience with xaaS like AWS

COMMUNICATION & VISUALIZATION

- ☆ Able to engage with senior management
- ☆ Story telling skills
- ☆ Translate data-driven insights into decisions and actions
- ☆ Visual art design
- ☆ R packages like ggplot or lattice
- ☆ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21st century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- ☆ Machine learning
- ☆ Statistical modeling
- ☆ Experiment design
- ☆ Bayesian inference
- ☆ Supervised learning: decision trees, random forests, logistic regression
- ☆ Unsupervised learning: clustering, dimensionality reduction
- ☆ Optimization: gradient descent and variants

DOMAIN KNOWLEDGE & SOFT SKILLS

- ☆ Passionate about the business
- ☆ Curious about data
- ☆ Influence without authority
- ☆ Hacker mindset
- ☆ Problem solver
- ☆ Strategic, proactive, creative, innovative and collaborative



PROGRAMMING & DATABASE

- ☆ Computer science fundamentals
- ☆ Scripting language e.g. Python
- ☆ Statistical computing package e.g. R
- ☆ Databases SQL and NoSQL
- ☆ Relational algebra
- ☆ Parallel databases and parallel query processing
- ☆ MapReduce concepts
- ☆ Hadoop and Hive/Pig
- ☆ Custom reducers
- ☆ Experience with xaaS like AWS

COMMUNICATION & VISUALIZATION

- ☆ Able to engage with senior management
- ☆ Story telling skills
- ☆ Translate data-driven insights into decisions and actions
- ☆ Visual art design
- ☆ R packages like ggplot or lattice
- ☆ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

Conclusioni

Viviamo nell'era dei Big Data

Ampia gamma di disponibilità in diverse aree

Grandi opportunità per risolvere grandi problemi

Possono creare valore

La sfida è come gestirle e utilizzarle

Occorrono nuove tecnologie

Gli aspetti metodologici sono importanti

Un'area in rapida evoluzione

Data Scientist: l'attuale professione più calda in IT

