ANALISI DEI DATI

Ariel Cedola
Data Scientist

Agenda

1. Data science… cos’è?
2. Processo della Data Science
3. Analisi dei dati come parte del processo
4. Tipi di analisi
5. Applicazioni e casi di uso
6. Strumenti
7. Lab 1: Identificare i dati disponibili da cui estrarre valore
8. Lab 2: Analisi basico
9. Lab 3: Analisi predittiva

1. Data Science

Dato

dato <sost.m.> *IT*

Ciascuno degli elementi di cui si dispone per formulare un giudizio o per risolvere un problema.

data <noun> *UK*

Information, especially facts or numbers, collected to be examined and considered and used to help decision-making.

dato <sust.m.> *ES*

Información sobre algo concreto que permite su conocimiento exacto o sirve para deducir las consecuencias derivadas de un hecho.

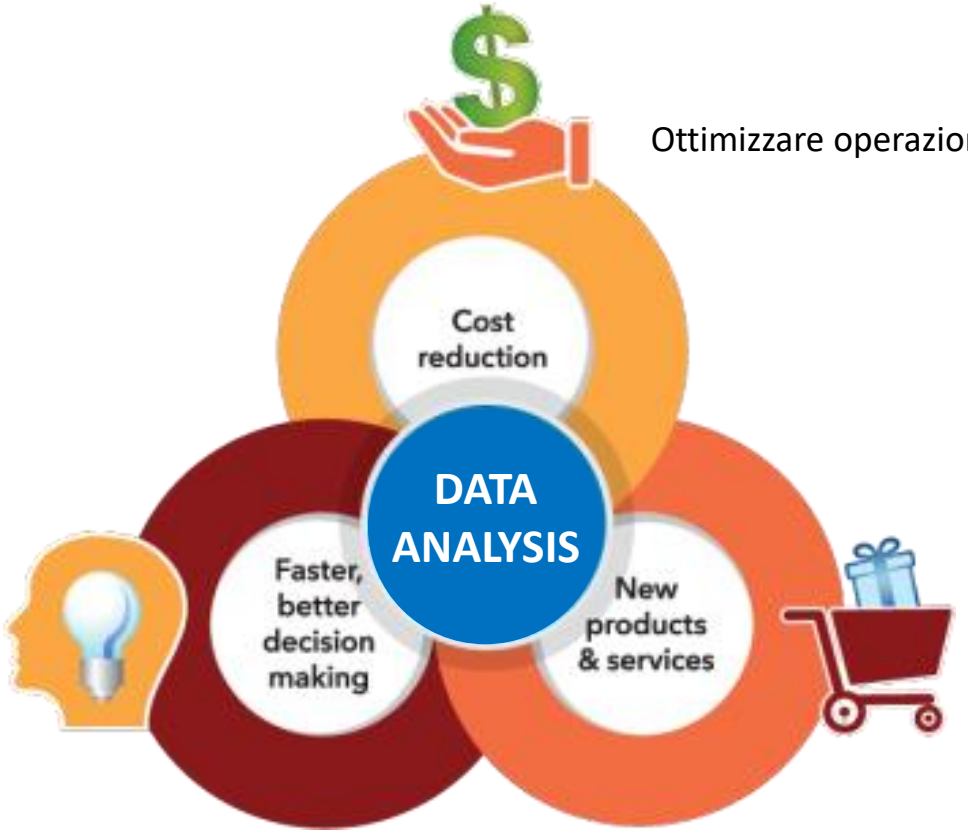
“Il dato è il nuovo petrolio”, Clive Humby (2006)

Dato = nuovo petrolio

“Data is the new oil. It’s valuable, but if unrefined it cannot really be used. It has to be changed into gas, plastic, chemicals, etc to create a valuable entity that drives profitable activity; so must data be broken down, analyzed for it to have value.” Clive Humby (2006)

Perché è così importante l’analisi dei dati?

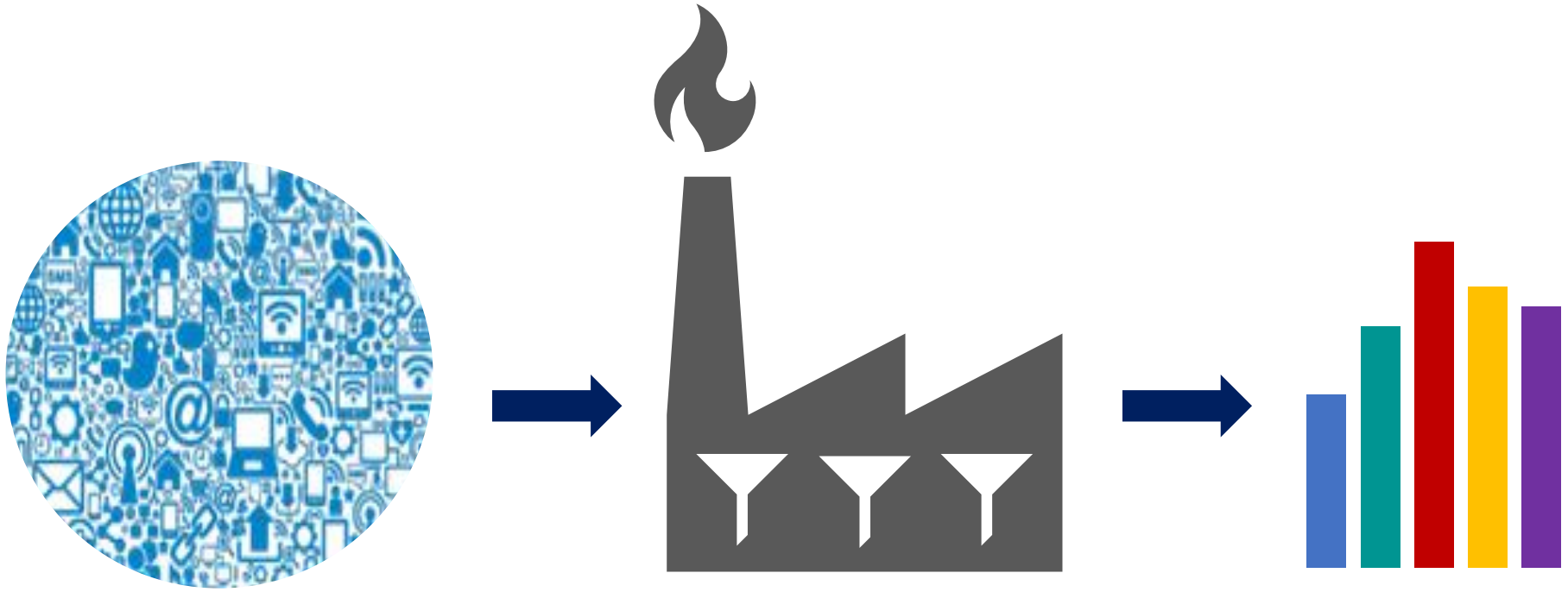
Ottimizzare operazioni e infrastrutture



Adattare prodotti e servizi per soddisfare le esigenze dei clienti

Trovare nuove fonti di reddito

Dati = Informazioni (ma nascosta…)



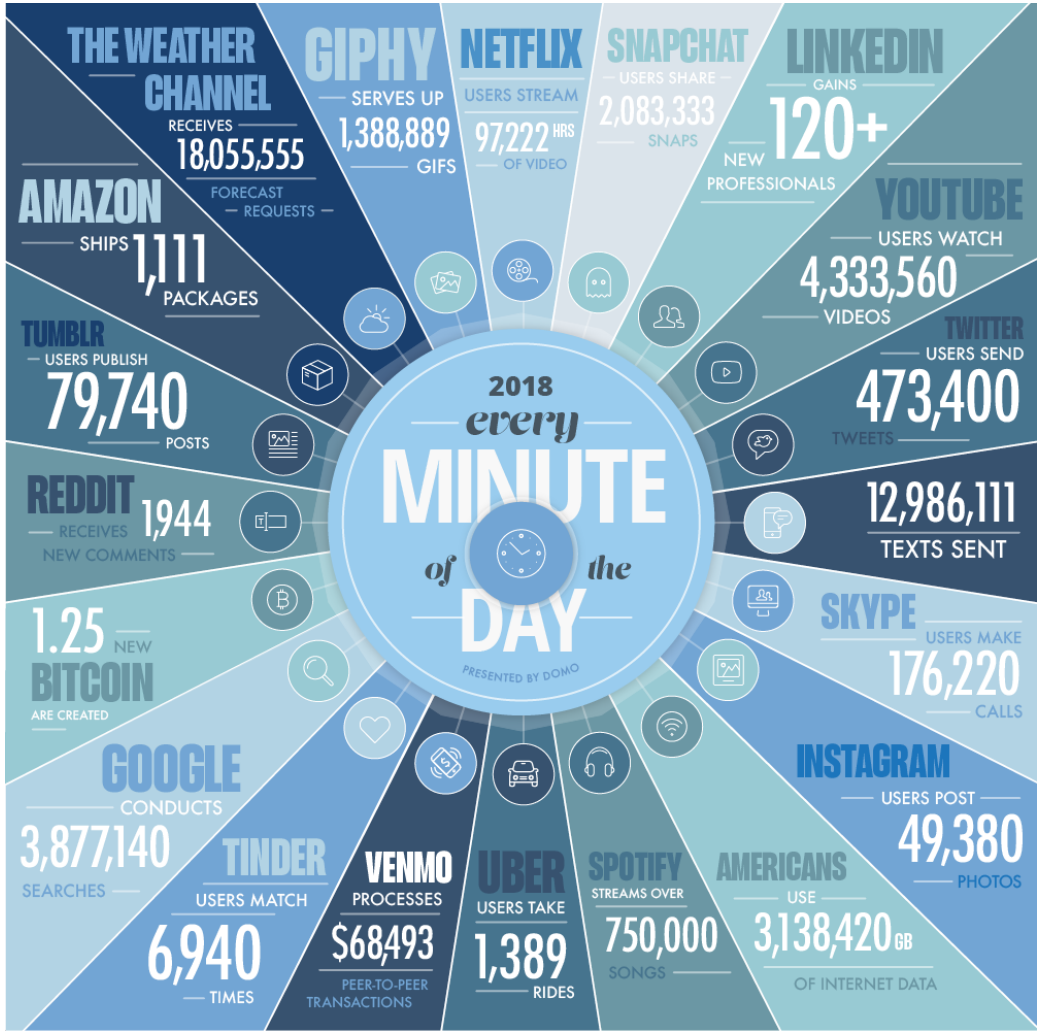
DATI GREZZI **DATA SCIENCE** **VALORE**

sdsadnuasdhskbusduasdsadsandsndishdy478324783278ftacxfascxsgvcdvbfbg,gt´ñ2´3020eekweofjewo9fjwdkod,d
wpkd9i192i239eu8wehisuffdijgogjoerkpei948395738jehfdfkvlkdjkgherutsfishfuieiugUUEEGIYWGIHSUSKSKUHKkh
khfkehfeksnvkdnkiwuru7t3u91284932uri8yfubvbxcvbxcnbfhmtokyhe9gjiuhsytt5wer32rur3tiryosugihdgjkfjbxbvbh
gfyuetri8rwitoljgkndfkg,nflhjrhokrokldfkfnbkdfjglirjou5o9eu9u8iy2i32jvdsvlksnvksdnvidshvihvvb888bv48vksnvsm
x.v,xcvlsddjonwuo8ow8i7wy3dgi2yriyeoferyouryi67p89067940iffpiewojncvxnxc.mñhljyul7oi0riyjrgijdfigheiri8ruty
sfdqe5reytsfhvdhgfvxncvnxcvmcxnv,smlkwpi’23i403i5gueritirjgekrgeljowekr934j5o934uto9ueroifjdknkxcvnlsfkp3i
rguwrorogeoitjteriohto3ury72t47g4yiruhweukfhdkgkhlrtkhp065iy9459385823yu8ryweufhkdgwjfpwkfomo43ufjfn
vo23unnunc2u3nhe84yejr0pcw’’’w39rushfkzjcsfñeñfpkw9r8h284787irsefsdhfkfdsngjerogtnimi03vriv4t9nu4n85v
nnfsjniecp2i9tueihgkknxmnv,xvlnjfvownifvpkñuyjkl{8íp89ouiñk,{jk{‘mo,pm9nron9bibfu7vqe6tv6dastcdqtwcdysg
hbkaCxicwrusaOwrkjmsinvNsnfsmpvrOzmo9nogfjsSnknkpvp02vCronweaxnvEryinbtypruNvofrtvoZmgvmérbioAyv
593u852yb8irhwkf,sndv.dmgp´’0inosdbfihsb7vu732r32prv’tob´n´gklcv,cxnbxcmb.l,m{´mgpbjinye7bu7eb7jvewbiw
ovwpkerpvmnskfslvddlngvkrtv8rvb623t7frweirsnlbkhtyljkñp8i´’87im’jkghnmkvbnciniixjdvgusdtrvc6etvyfstdvwy6
er632yt473y85u4ntidkgn,mb.nggmjñymjp76iuhijdiguhbnugbyusdbgfyfdtyf632te6t32ryv734bvi4ytivsdifhhdkfbkgd
jfgnbl6kyn6050po56´nuklgmkfmnñmul60iidhgu7bbu7yvewibyv84bvyt8i4uvi8rwitoljgkndfkg,nflhjrhokrokldfkfnbk
dfjglirjou5o9eu9u8iy2i32jvdsvlksnvksdnvidshvihvvb888bv48vksnvsmxryosugihdgjkfjbxbvbhgfyuetri8rwitoljgkndf
kgeriohto3ury72t47g4yiruhweukfhdkgkhlrtkhp065iy9459385823y43ufjfnvo23unnunc2u3nhe84yejr0pcwu8ryweuf
hkdgwjfpwkfomojfififhuefhuwehiwdwidjoweo32u823y7e8wqydsghcvjwg62153723489380543oudkvksjhiwjeed82
dbsgycgyafdy32y6t0ppljm,jñm,jpmlphkotykonigjibnfnubuvbdycywtdtwdr532refstcvsgcvdnvbjnkjrotigi4ti0gipgvioj
fknv,dfmlbreonting9jdknksdnfkue8bfrbu3ru0enwiofjdskvnxjvbu7vc2ry94yt9u05iybobgfmb,vmnh,gpuo56udjfhvb
q7ey2e2c1y932r003btigilfbmvnmlkn`60m05bimtodjvknfkjwehrc3y432u9v0nilogxknblbkm54o0iojgnfhwbehc7rt37
8cy32u32cni2nidvxncb.mgñljlmì’omniu9figjxuhfsutdcv6q2te7c32t5843y5o9u460547onmfghkmghlmgmñylu`76º80
560iyn4yu83bytbv7237rtcygfdshvxcbvmnfkgjtryopbyet0v3ni95u23ryeuibrvhsukhfbfivregobdfgjnbolgknbfokgnb5n

Perché si parla così tanto adesso?

+ 1.000.000.000 GB di dati
si producono ogni giorno

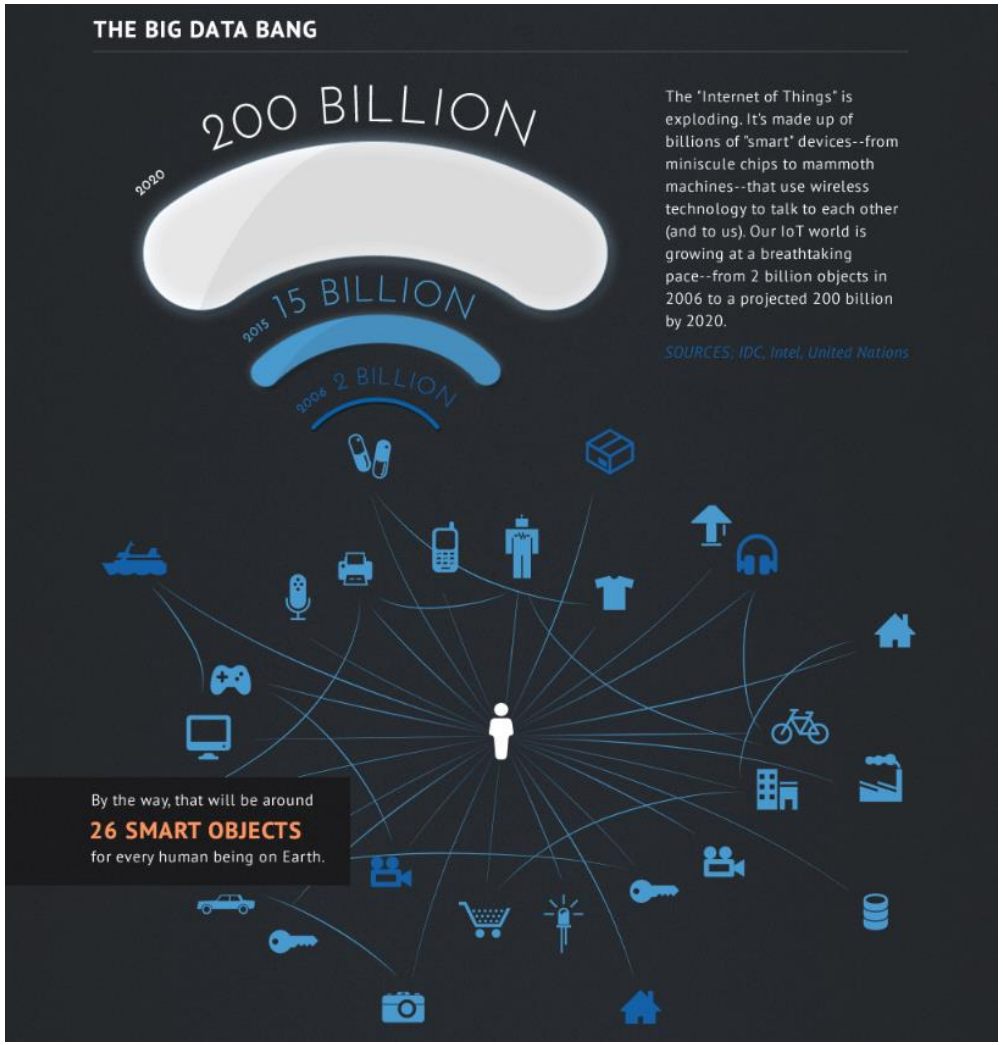
1 EB = 1024 PB
1 PB = 1024 TB
1 TB = 1024 GB



https://www.domo.com/learn/data-never-sleeps-6

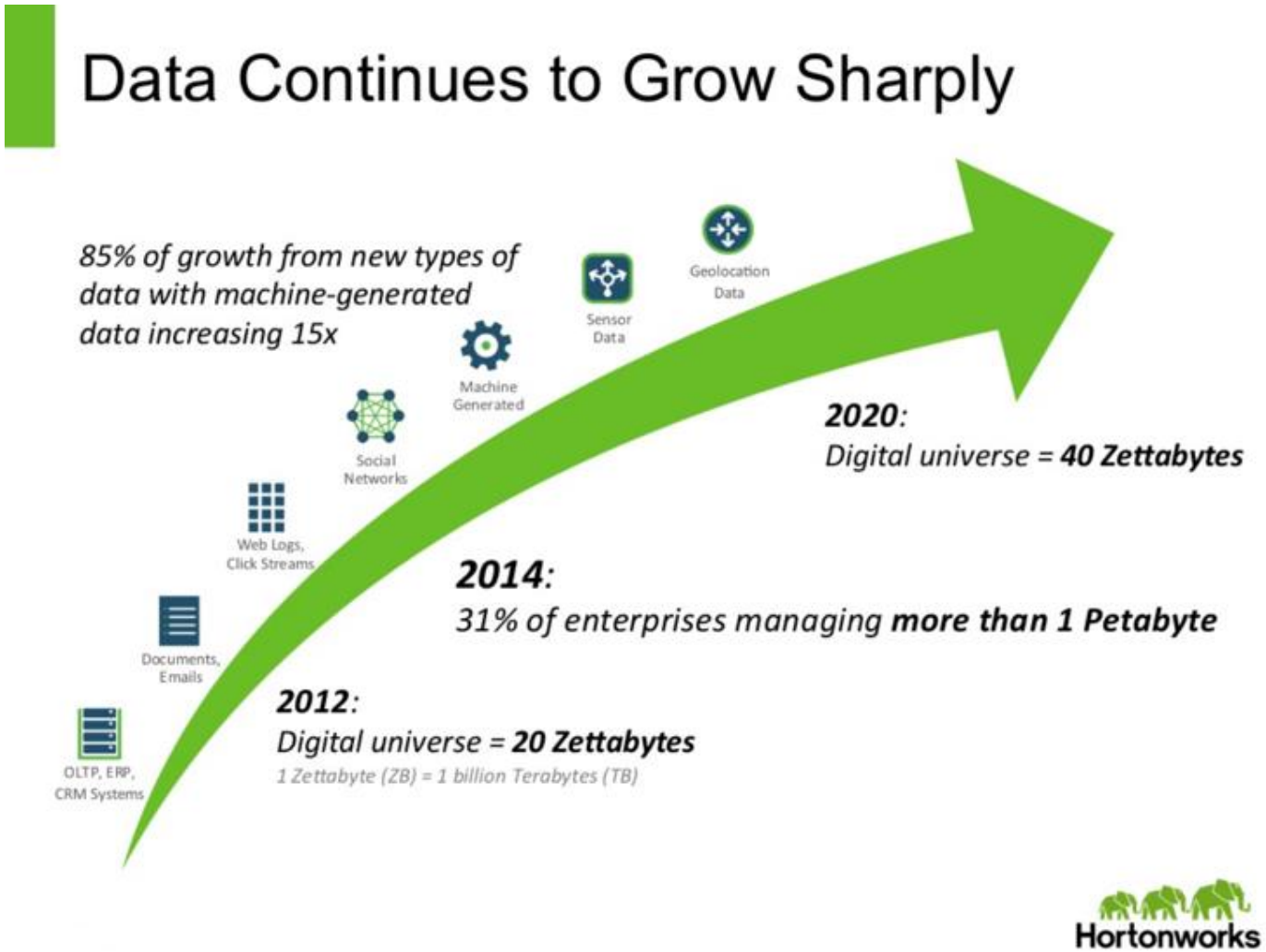
Perché si parla così tanto adesso?

Internet of Things (IoT)



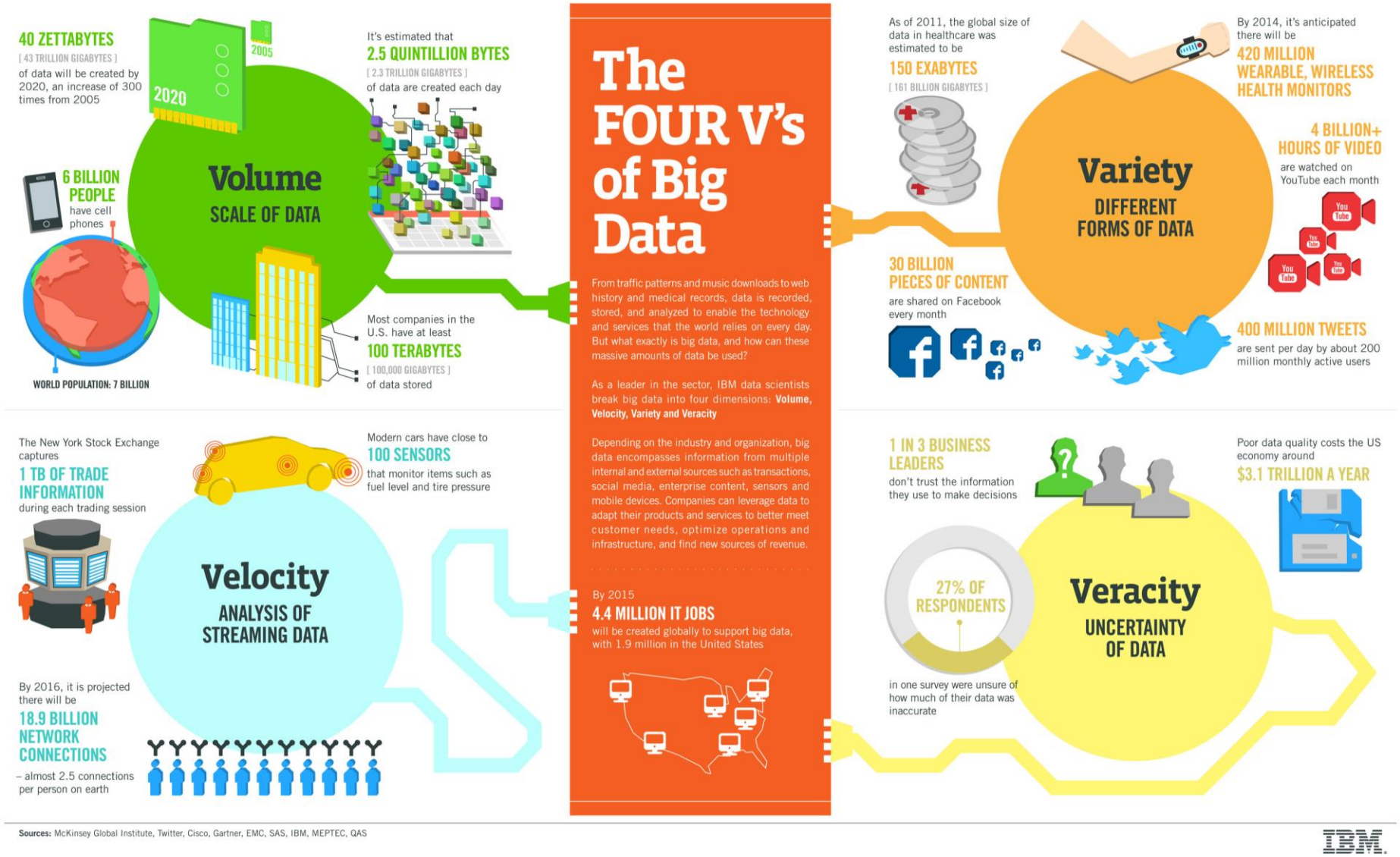
https://www.intel.com/content/dam/www/public/us/en/images/iot/guide-to-iot-infographic.png

Perché si parla così tanto adesso?



https://it.hortonworks.com/

Big Data

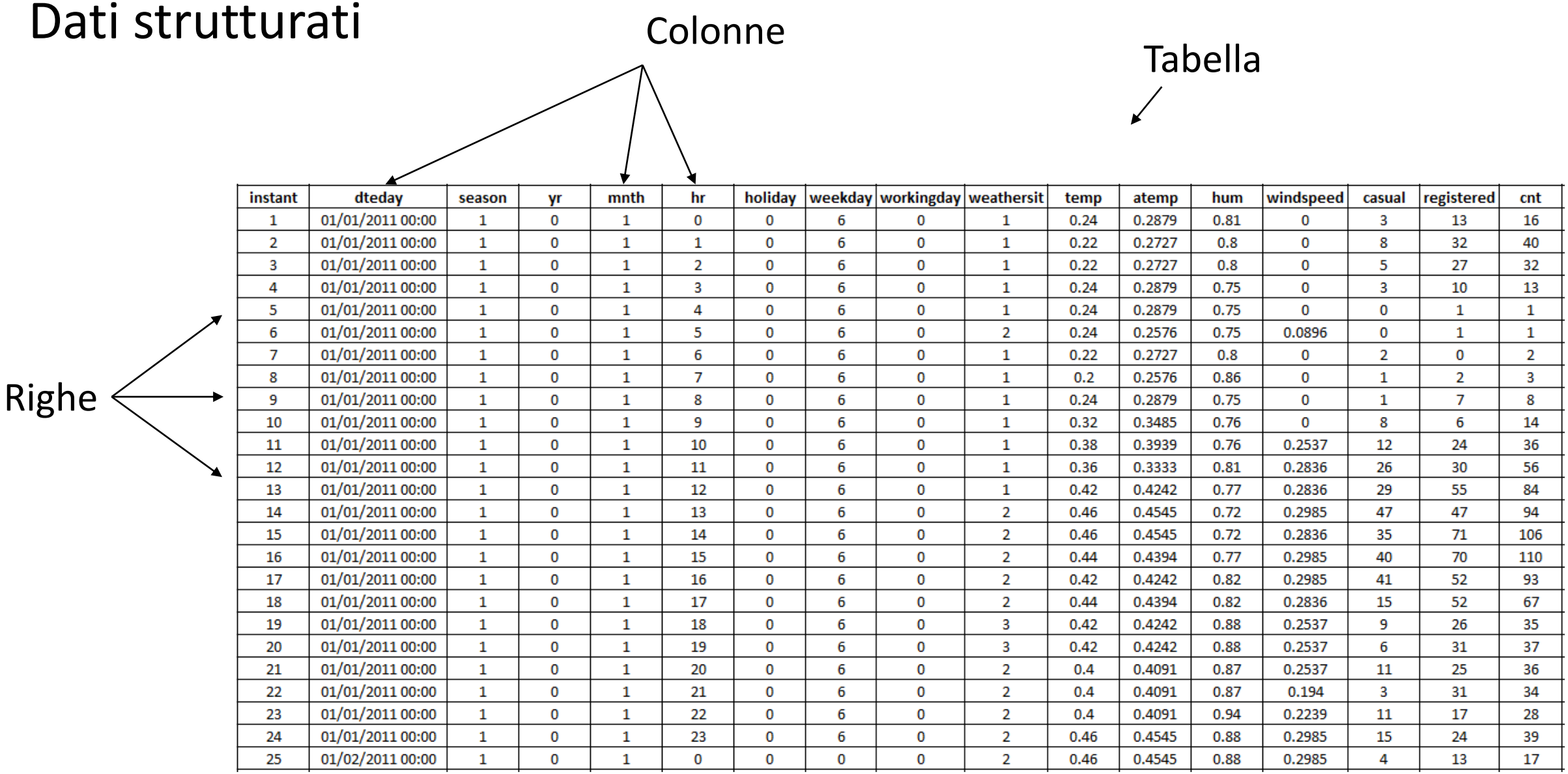


https://www.ibmbigdatahub.com/infographic/four-vs-big-data

Tipi e caratteristiche dei dati

- Dati strutturati
- Dati semi-strutturati
- Testo, documenti, reports
- Linguaggio naturale (NLP)
- Audio, video, immagini
- Telemetria, dati in streaming (time series)
- Network data
- Geolocalizzazione

Dati strutturati



instant	dteday	season	yr	mnth	hr	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
1	01/01/2011 00:00	1	0	1	0	0	6	0	1	0.24	0.2879	0.81	0	3	13	16
2	01/01/2011 00:00	1	0	1	1	0	6	0	1	0.22	0.2727	0.8	0	8	32	40
3	01/01/2011 00:00	1	0	1	2	0	6	0	1	0.22	0.2727	0.8	0	5	27	32
4	01/01/2011 00:00	1	0	1	3	0	6	0	1	0.24	0.2879	0.75	0	3	10	13
5	01/01/2011 00:00	1	0	1	4	0	6	0	1	0.24	0.2879	0.75	0	0	1	1
6	01/01/2011 00:00	1	0	1	5	0	6	0	2	0.24	0.2576	0.75	0.0896	0	1	1
7	01/01/2011 00:00	1	0	1	6	0	6	0	1	0.22	0.2727	0.8	0	2	0	2
8	01/01/2011 00:00	1	0	1	7	0	6	0	1	0.2	0.2576	0.86	0	1	2	3
9	01/01/2011 00:00	1	0	1	8	0	6	0	1	0.24	0.2879	0.75	0	1	7	8
10	01/01/2011 00:00	1	0	1	9	0	6	0	1	0.32	0.3485	0.76	0	8	6	14
11	01/01/2011 00:00	1	0	1	10	0	6	0	1	0.38	0.3939	0.76	0.2537	12	24	36
12	01/01/2011 00:00	1	0	1	11	0	6	0	1	0.36	0.3333	0.81	0.2836	26	30	56
13	01/01/2011 00:00	1	0	1	12	0	6	0	1	0.42	0.4242	0.77	0.2836	29	55	84
14	01/01/2011 00:00	1	0	1	13	0	6	0	2	0.46	0.4545	0.72	0.2985	47	47	94
15	01/01/2011 00:00	1	0	1	14	0	6	0	2	0.46	0.4545	0.72	0.2836	35	71	106
16	01/01/2011 00:00	1	0	1	15	0	6	0	2	0.44	0.4394	0.77	0.2985	40	70	110
17	01/01/2011 00:00	1	0	1	16	0	6	0	2	0.42	0.4242	0.82	0.2985	41	52	93
18	01/01/2011 00:00	1	0	1	17	0	6	0	2	0.44	0.4394	0.82	0.2836	15	52	67
19	01/01/2011 00:00	1	0	1	18	0	6	0	3	0.42	0.4242	0.88	0.2537	9	26	35
20	01/01/2011 00:00	1	0	1	19	0	6	0	3	0.42	0.4242	0.88	0.2537	6	31	37
21	01/01/2011 00:00	1	0	1	20	0	6	0	2	0.4	0.4091	0.87	0.2537	11	25	36
22	01/01/2011 00:00	1	0	1	21	0	6	0	2	0.4	0.4091	0.87	0.194	3	31	34
23	01/01/2011 00:00	1	0	1	22	0	6	0	2	0.4	0.4091	0.94	0.2239	11	17	28
24	01/01/2011 00:00	1	0	1	23	0	6	0	2	0.46	0.4545	0.88	0.2985	15	24	39
25	01/02/2011 00:00	1	0	1	0	0	0	0	2	0.46	0.4545	0.88	0.2985	4	13	17

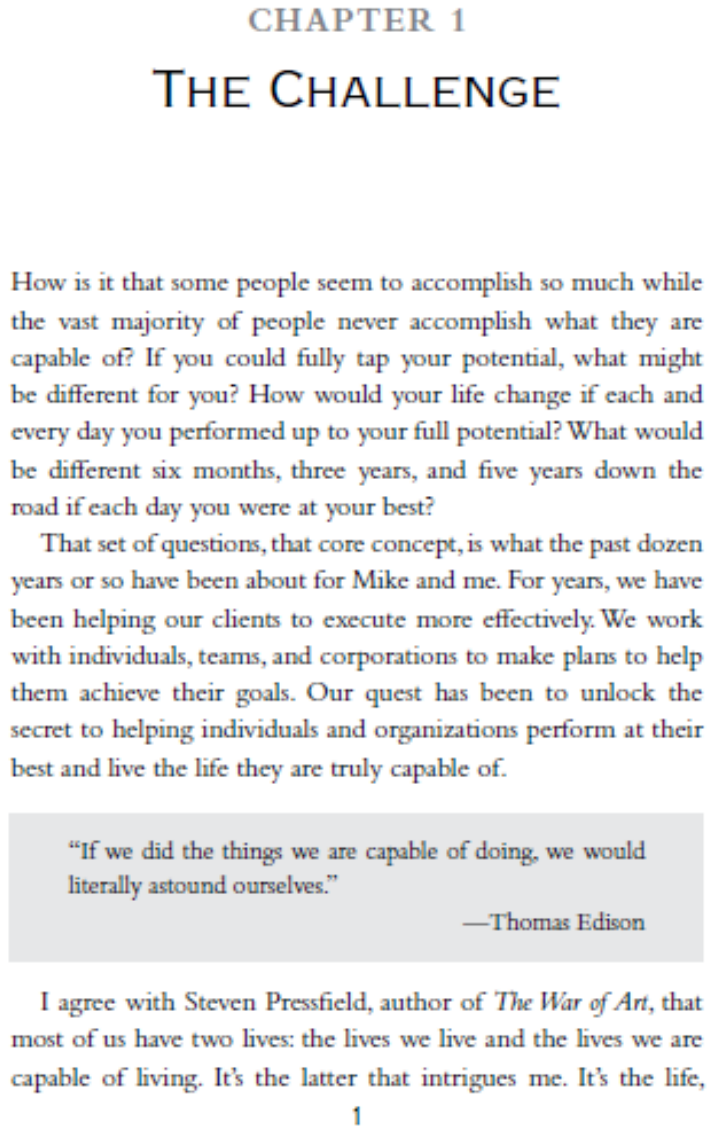
Dati semi-strutturati

- JSON
- XML


```
{
  "locations" : [ {
    "timestampMs" : "1519240942615",
    "latitudeE7" : 450727233,
    "longitudeE7" : 76470873,
    "accuracy" : 13,
    "activity" : [ {
      "timestampMs" : "1519240849799",
      "activity" : [ {
        "type" : "STILL",
        "confidence" : 43
      }, {
        "type" : "ON_FOOT",
        "confidence" : 20
      }, {
        "type" : "WALKING",
        "confidence" : 20
      }, {
        "type" : "UNKNOWN",
        "confidence" : 13
      }, {
        "type" : "ON_BICYCLE",
        "confidence" : 9
      }, {
        "type" : "IN_VEHICLE",
        "confidence" : 7
      }, {
        "type" : "IN_ROAD_VEHICLE",
        "confidence" : 7
      }, {
        "type" : "IN_RAIL_VEHICLE",
        "confidence" : 7
      }, {
        "type" : "RUNNING",
        "confidence" : 3
      } ]
    } ]
  }, {
    "timestampMs" : "1519240670673",
```

Testo, documenti, reports

- pdf
- word, wordx
- txt
- Latex
- ecc.



CHAPTER 1

THE CHALLENGE

How is it that some people seem to accomplish so much while the vast majority of people never accomplish what they are capable of? If you could fully tap your potential, what might be different for you? How would your life change if each and every day you performed up to your full potential? What would be different six months, three years, and five years down the road if each day you were at your best?

That set of questions, that core concept, is what the past dozen years or so have been about for Mike and me. For years, we have been helping our clients to execute more effectively. We work with individuals, teams, and corporations to make plans to help them achieve their goals. Our quest has been to unlock the secret to helping individuals and organizations perform at their best and live the life they are truly capable of.

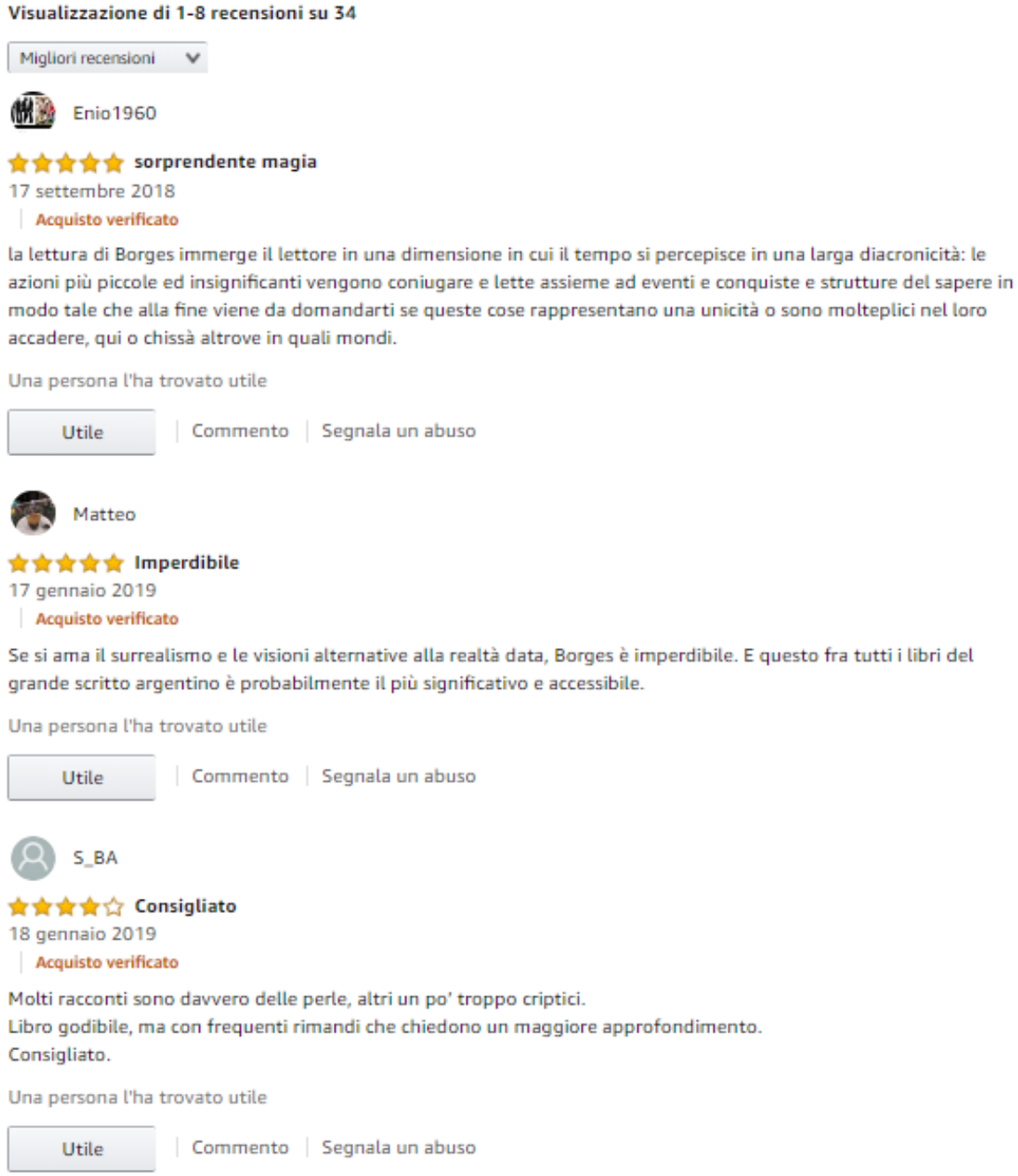
> "If we did the things we are capable of doing, we would literally astound ourselves."
>
> —Thomas Edison

I agree with Steven Pressfield, author of *The War of Art*, that most of us have two lives: the lives we live and the lives we are capable of living. It's the latter that intrigues me. It's the life,

1

Linguaggio naturale (NLP)

- Recensioni di clienti
- Commenti su social network
- Twitters
- Emails
- Google Assistant, Alexa, Cortana
- ecc.



Visualizzazione di 1-8 recensioni su 34

Migliori recensioni

Enio1960

sorprendente magia

17 settembre 2018

Acquisto verificato

la lettura di Borges immerge il lettore in una dimensione in cui il tempo si percepisce in una larga diacronicità: le azioni più piccole ed insignificanti vengono coniugare e lette assieme ad eventi e conquiste e strutture del sapere in modo tale che alla fine viene da domandarti se queste cose rappresentano una unicità o sono molteplici nel loro accadere, qui o chissà altrove in quali mondi.

Una persona l'ha trovato utile

Utile | Commento | Segnala un abuso

Matteo

Imperdibile

17 gennaio 2019

Acquisto verificato

Se si ama il surrealismo e le visioni alternative alla realtà data, Borges è imperdibile. E questo fra tutti i libri del grande scritto argentino è probabilmente il più significativo e accessibile.

Una persona l'ha trovato utile

Utile | Commento | Segnala un abuso

S_BA

Consigliato

18 gennaio 2019

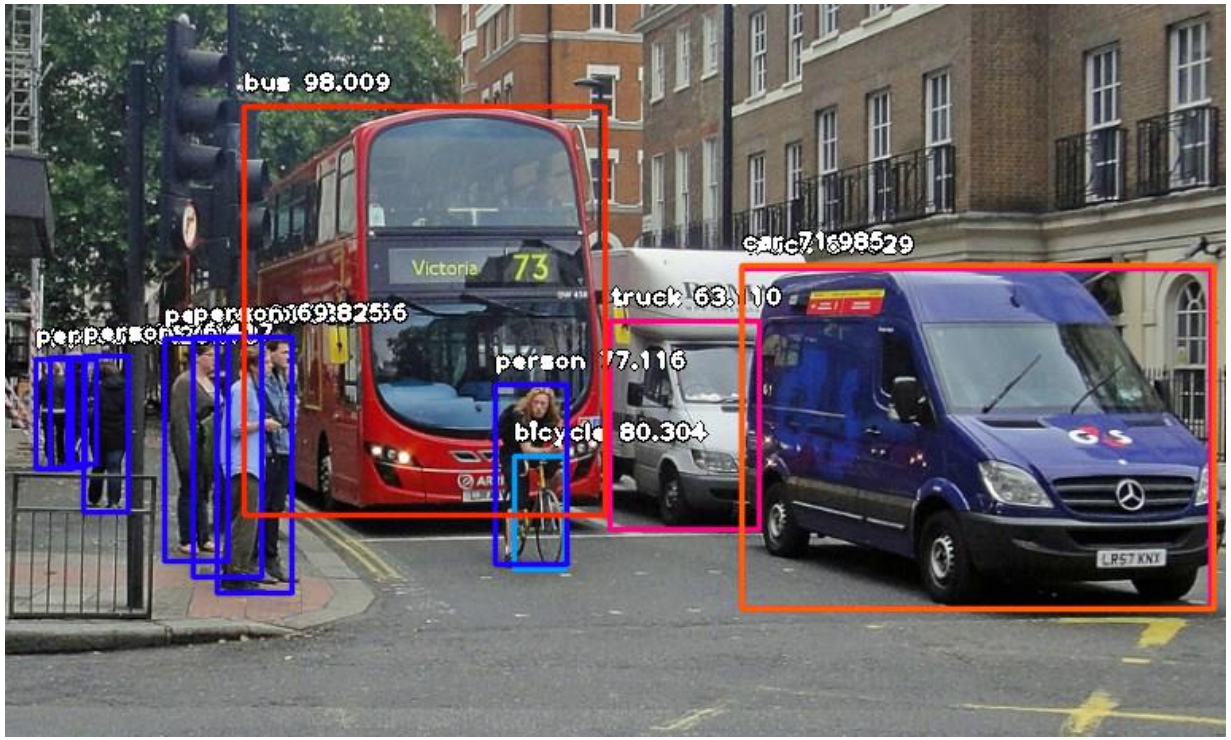
Acquisto verificato

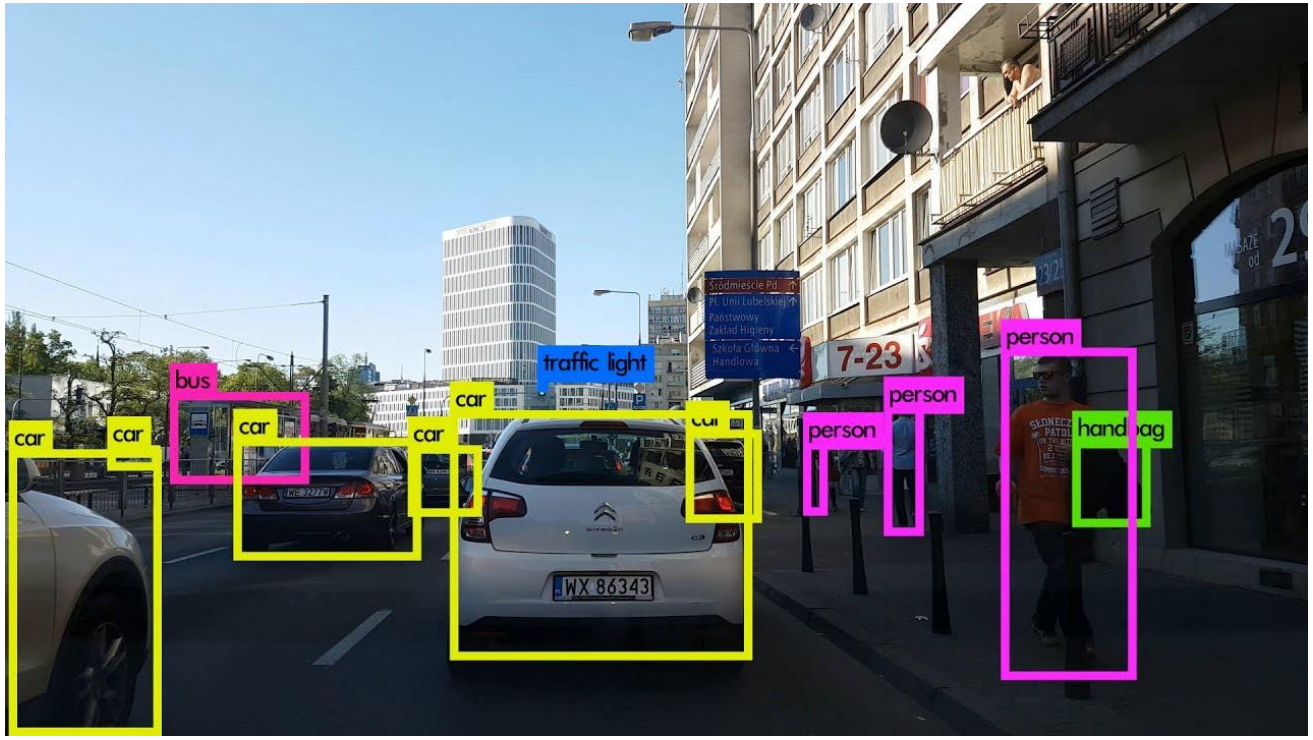
Molti racconti sono davvero delle perle, altri un po' troppo criptici.
Libro godibile, ma con frequenti rimandi che chiedono un maggiore approfondimento.
Consigliato.

Una persona l'ha trovato utile

Utile | Commento | Segnala un abuso

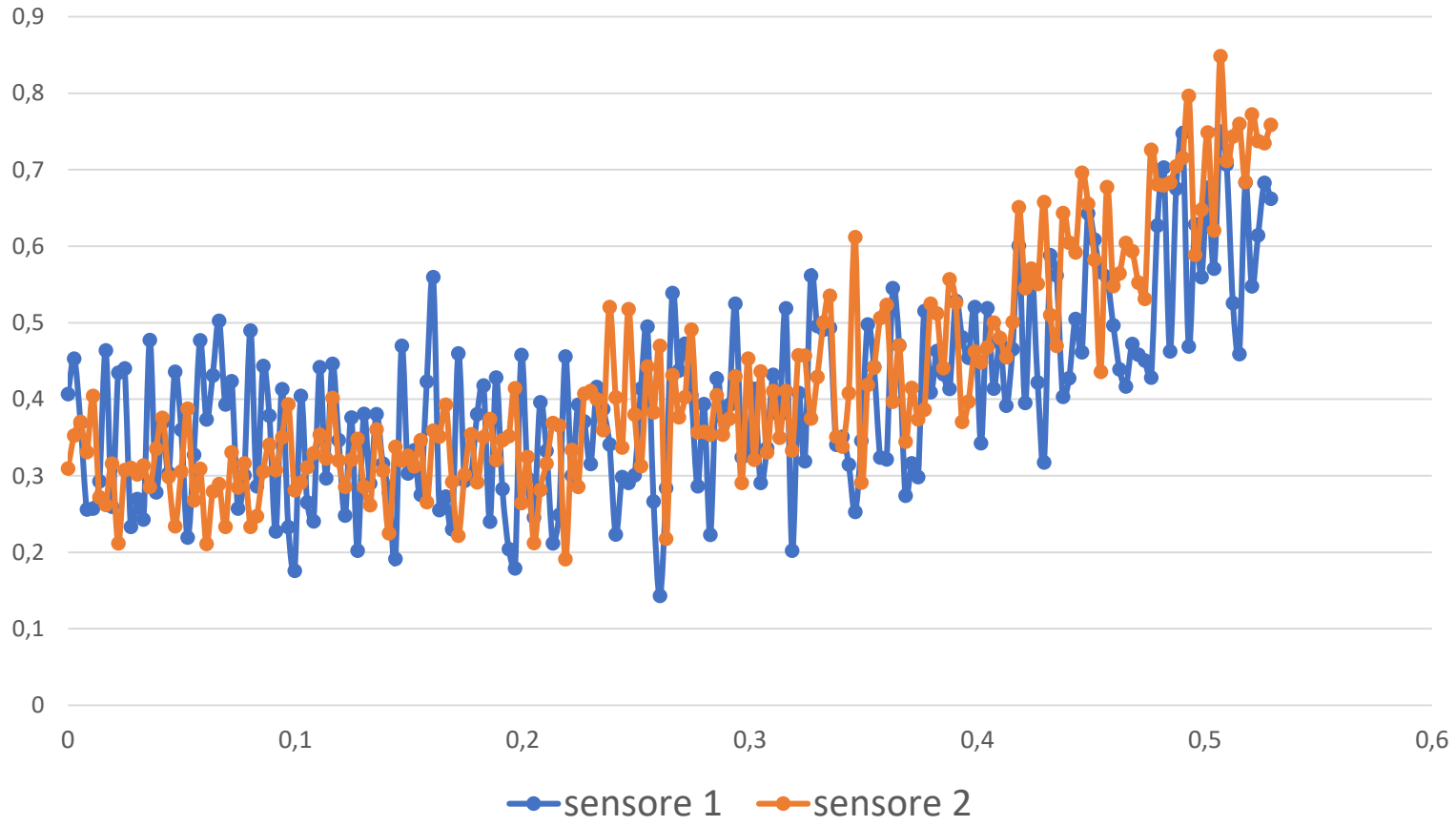
Audio, video, immagini



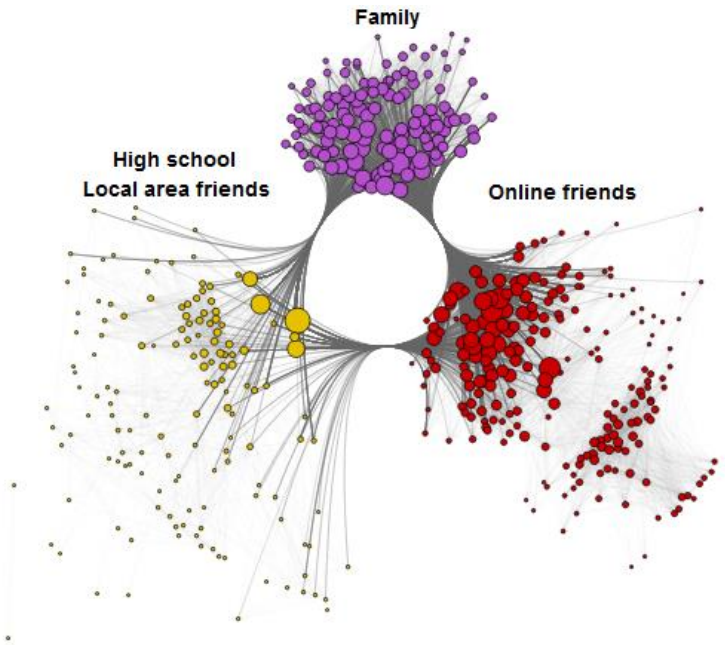


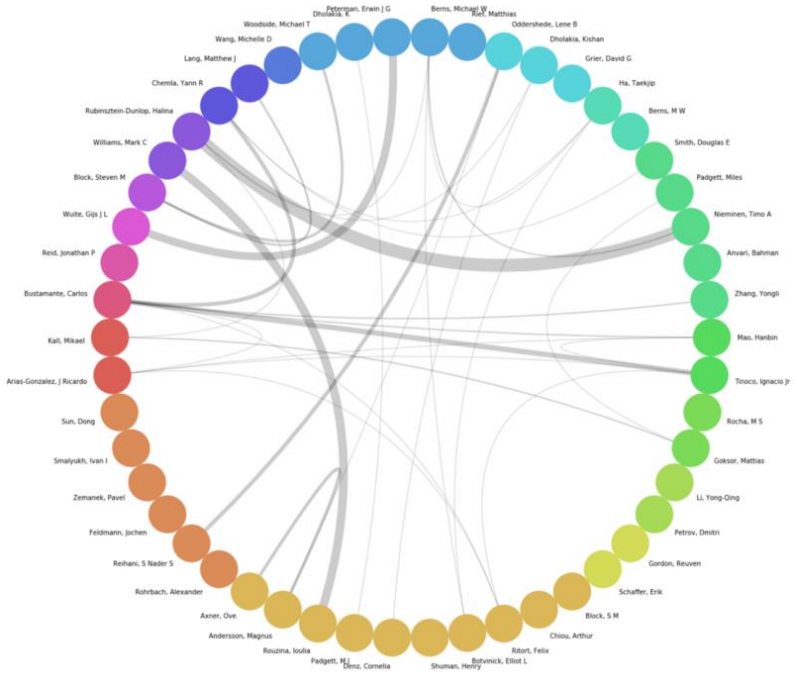
Telemetria

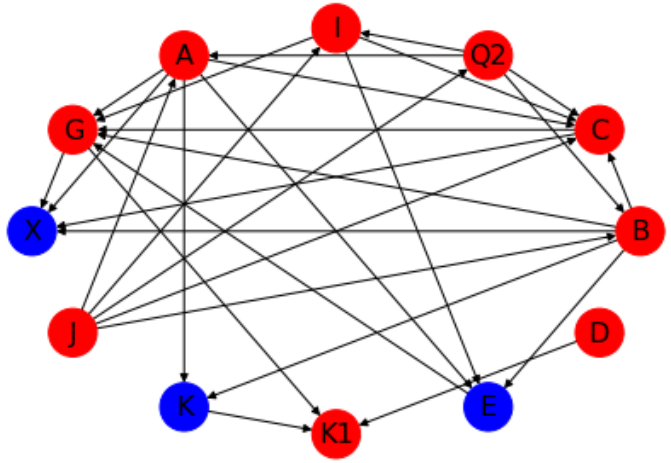
- Dati offline
- Dati in streaming



Network data







Geolocalizzazione

63 lugares

Más visitados

Visitados

90 sin confirmar

1 45° 4' 22.5102" N, 7° 38' 49... 1 de febrero de 2019 + 362 ...	4 45° 2' 1.3812" N, 7° 40' 19.... 6 de enero de 2019 + 10 días	7 45° 3' 54.7488" N, 7° 40' 30... 17 de diciembre de 2018 + ...	10 45° 3' 55.6513" N, 7° 37' 49... 17 de agosto de 2018 + 4 d...
2 45° 3' 43.7184" N, 7° 40' 27... 4 de febrero de 2019 + 60 d...	5 45° 4' 38.7253" N, 7° 38' 18... 30 de septiembre de 2018 ...	8 45° 3' 20.2752" N, 7° 36' 48... 20 de enero de 2019 + 4 días	11 45° 4' 24.528" N, 7° 38' 50.... 4 de febrero de 2019 + 3 dí...
3 45° 3' 44.9561" N, 7° 39' 44... 29 de enero de 2019 + 17 d...	6 45° 4' 9.9408" N, 7° 41' 20.... 10 de octubre de 2018 + 6 ...	9 45° 3' 56.1744" N, 7° 38' 57... 5 de enero de 2019 + 4 días	12 45° 3' 46.3777" N, 7° 39' 30... 31 de mayo de 2018 + 3 días

Fonti dei dati

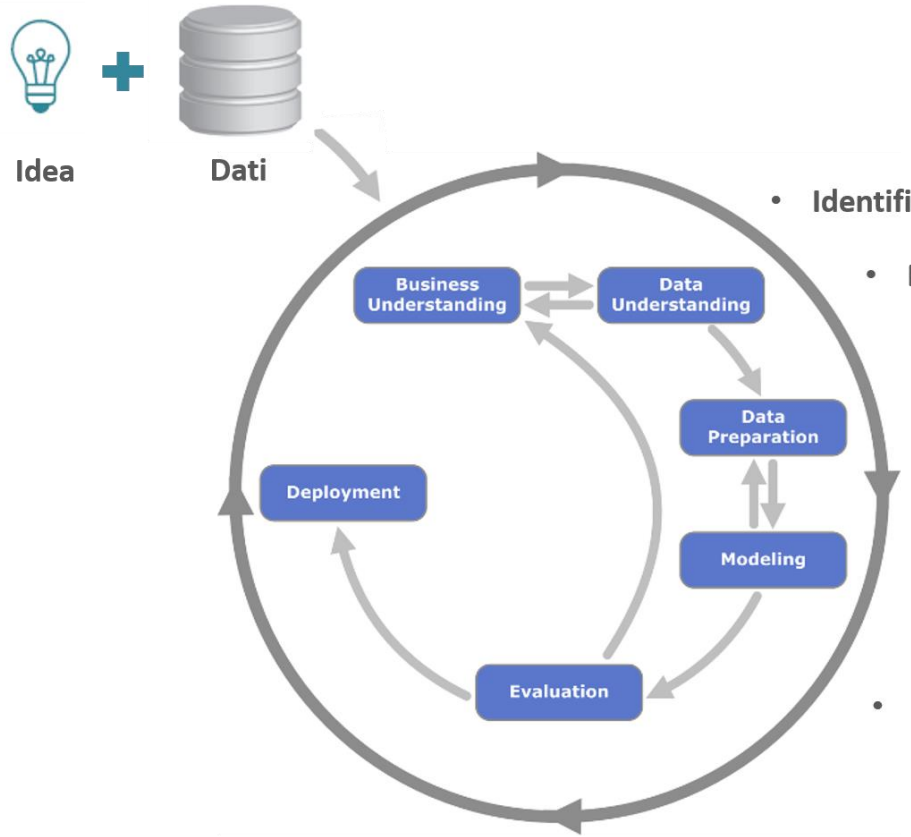
- Files di testo (.txt) o comma separated values (.csv)
- Foglio Excel
- SQL databases (relational databases: SQL Server, MySQL, SQLite)
- NoSQL databases (non-relational databases: MongoDB, CosmosDB, Cassandra)
- Websites
- Dispositivi IoT e sensori inviando dati in streaming
- Mobile data (messaggi, sms, calls, audio, video, foto)
- Banca immagini o altri file multimedia
- Videocamere, immagini via satellite, diagnostica per immagini (medicina)
- …

Esempi di dati

- **COMERCIO**, tabella clienti: Nome, customer ID, azienda, domicilio, telefono, email, dati fatturazione, acquisti, ecc
- **BANCA**, tabella clienti: customer ID, nome, data nascita, codice fiscale, partita IVA, conto corrente, carte di credito, ecc
- **SALUTE**, tabella pazienti ricoverati: Nome, ID, età, ingresso, storia clinica, camera, diagnosi, esami, trattamento, ecc
- **EDUCAZIONE**, tabella studenti: Nome, matricola, carriera, data nascita, domicilio, email, corsi, punteggio medio, ecc
- **CLIMA**, serie temporale di variabili: regione, provincia, città, data e ora, temperatura, umidità, pressione, vento, ecc
- **INDUSTRIA**, telemetria: macchina ID, settore, data e ora, sensore 1, …, sensore N, parametro 1, …, parametro M, ecc
- **INDUSTRIA**, tabella produzione: operatore, ID macchine, settore, data, fascia oraria, prodotto, totale pezzi, ecc
- **ENERGIA**, tabella consumo: regione, provincia, data, domanda di energia, produzione di energia solare, eolica, ecc
- **SORVEGLIANZA**, video: camera 1, camera 2, …, camera N
- **SOCIAL**, rete contatti utente: grafico di network
- **SOCIAL**, recensioni su prodotto o servizio: data, utente, recensione
- **GEOLOCALIZZAZIONE**, movimenti utenti: device ID, data e ora, latitudine, longitudine

Data Science, cosa si tratta?

La scienza dei dati (data science) è un campo interdisciplinare, che utilizza metodi scientifici, processi, algoritmi e sistemi per generare modelli complessi dai dati e trasformarli in informazioni e conoscenza.



- Identificare gli obbiettivi del business
- Raccogliere i dati (files, databases, sensori)
- Pulire i dati e svolgere feature engineering
- Costruire il modello -> approccio predittivo
- Applicare il modello -> ottimizzazione del business
- Valutare la qualità dei risultati e iterare
- Distribuire la soluzione

CRISP-DM modello analitico

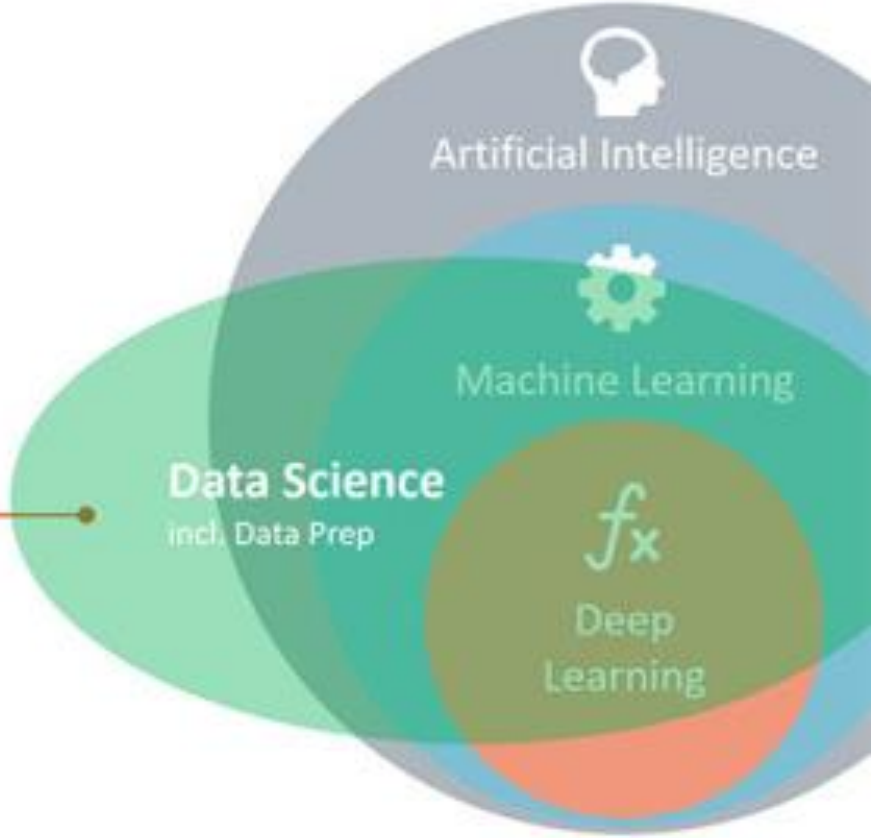
Data Science, cosa si tratta?

Intelligenza Artificiale consente ai computer di pensare, prendere decisioni come un essere umano

Machine Learning è la tecnica che permette ai computer di apprendere a partire dai dati

Deep Learning è un sottogruppo del Machine Learning basato nell'utillizo di reti neuronali

Data Science è l'applicazione pratica di
Intelligenza Artificiale
Machine Learning
Deep Learning

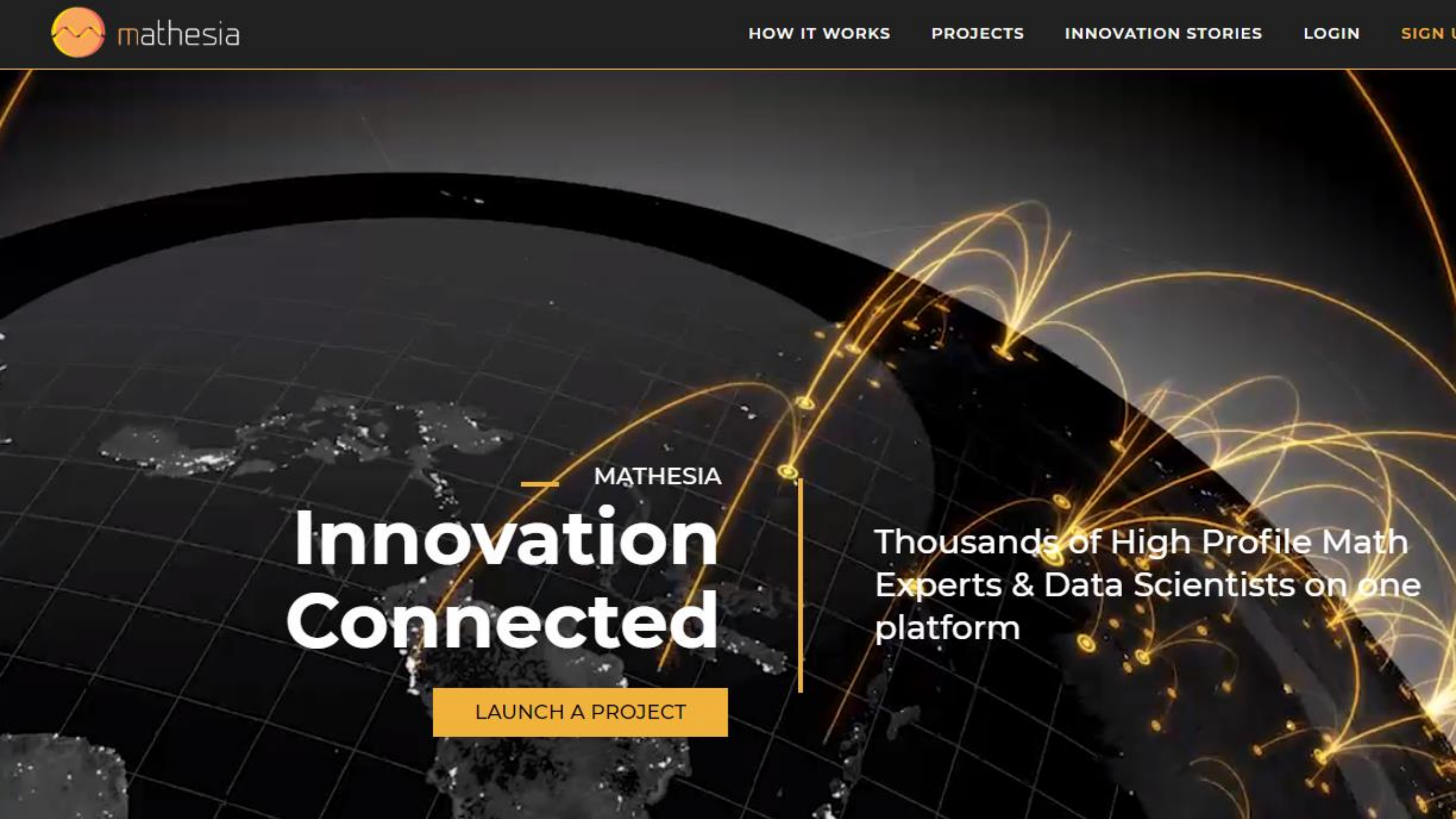


Data Science, a cosa serve?

Secondo la ricerca degli Osservatori Digital Innovation - School of Management (PoliMi) :

Gli obiettivi aziendali che è possibile raggiungere sono:

- migliorare l'engagement con il cliente
- incrementare le vendite
- ridurre il time to market
- ampliare l'offerta di nuovi prodotti e servizi
- ottimizzare l'offerta attuale al fine di aumentare i margini
- ridurre i costi
- identificare nuovi mercati


mathesia
HOW IT WORKS
PROJECTS
INNOVATION STORIES
LOGIN
SIGN U
MATHESIA
Innovation Connected
LAUNCH A PROJECT
Thousands of High Profile Math Experts & Data Scientists on one platform

Business

Data Science Utilization

48%

One in two of the businesses surveyed declare to make use of data science tools.

Adopted technologies

Machine Learning

The most widely used technologies include machine learning, advanced analytics, big data, databases and cloud computing.

Business Area

Marketing

Data science tools are mainly used in marketing & sales, R&D, and production and operation.

Investments

67%

67% plan to invest in the sector, especially in machine learning and predictive/ advanced analytics.

Areas with positive expectations

Process Optimization

As for their expectations, many indicate that they are looking closely at the business intelligence field and the optimization of the process (process optimization, predictive maintenance, automation, supply chain, KPI/quality analysis).

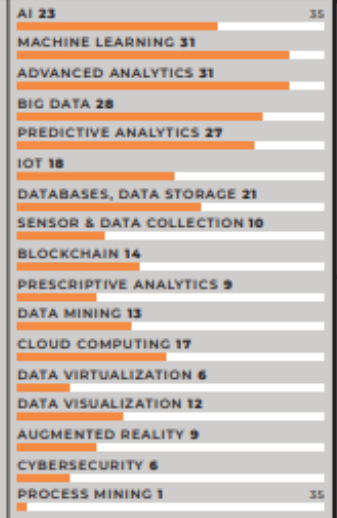
Risks

Privacy

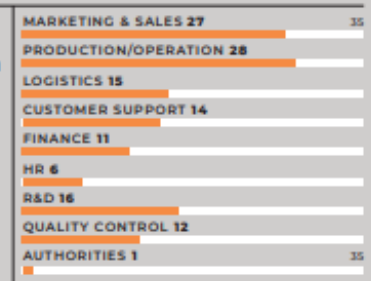
The most critical issues emerge in the areas of privacy, data quality and complexity in adoption of the new technologies, due to the lack of expertise in these areas.

Are you planning to invest in data science technologies in 2019?

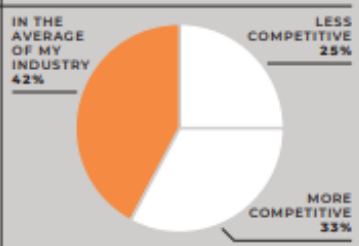
Which technologies do you plan to adopt in 2019?



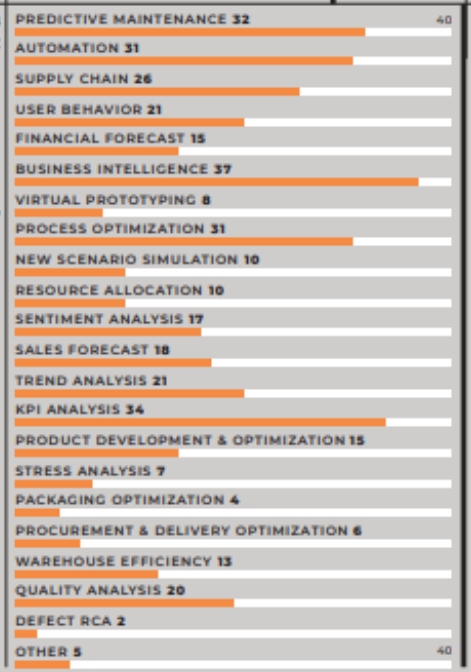
In which business units will you adopt these technologies in 2019?



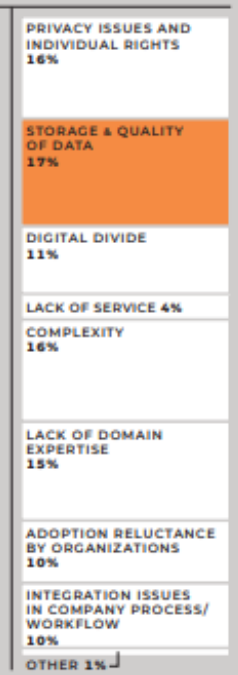
How do you feel about your competitive position since adopting data science technologies?



In which areas do you expect data science technologies to create benefits for your company's strategic development?



What do you expect will be the critical issues and risks data science will have to cope with in the near future?



Experts

Current work areas

Process Optimization

Today experts are working mainly in the fields of process optimization, business intelligence and trend analysis.

Technologies that will make a difference

Artificial Intelligence

In the opinion of the experts, machine learning and artificial intelligence are the most promising technologies, while they place less expectation on IoT, cybersecurity and blockchain.

Fields of application with the greatest impact

Biotech

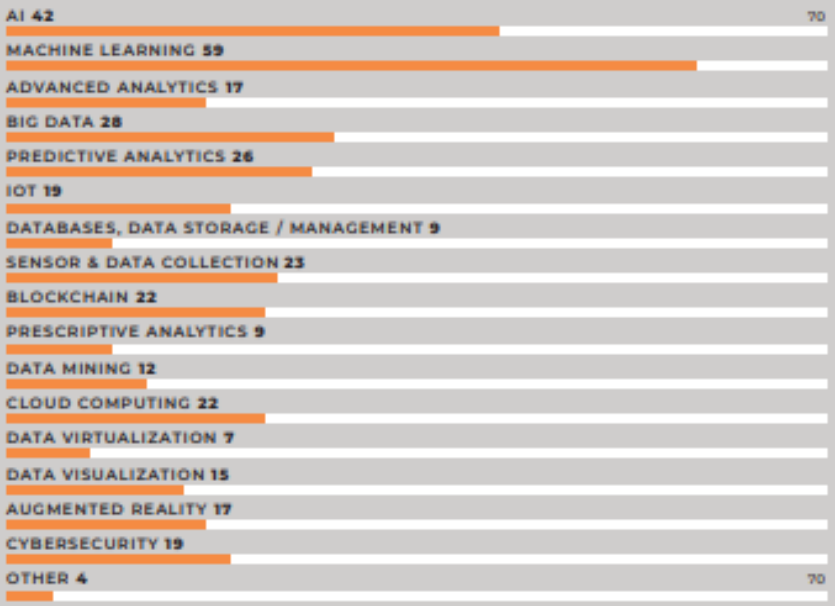
The greatest effects are expected in the medical, biotech and healthcare sector and automation and robotics.

Risks

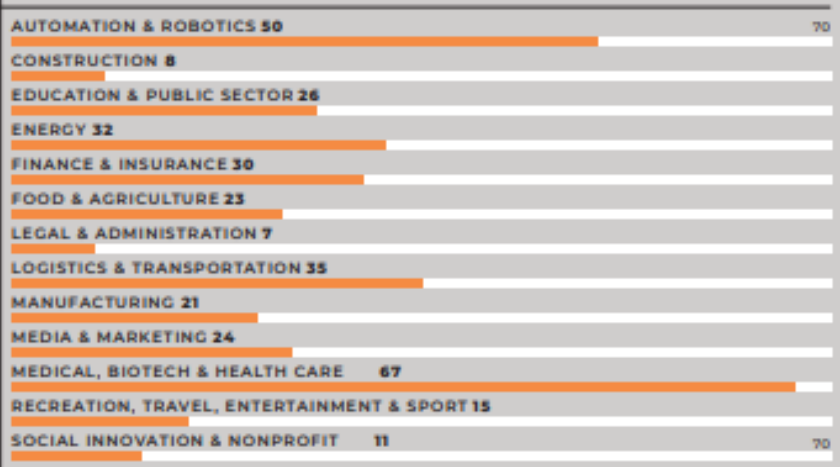
Privacy

The greatest risks are perceived in the area of privacy and individual rights. Other critical issues are storage of data, obtaining information quality from data, and the lack of domain experts.

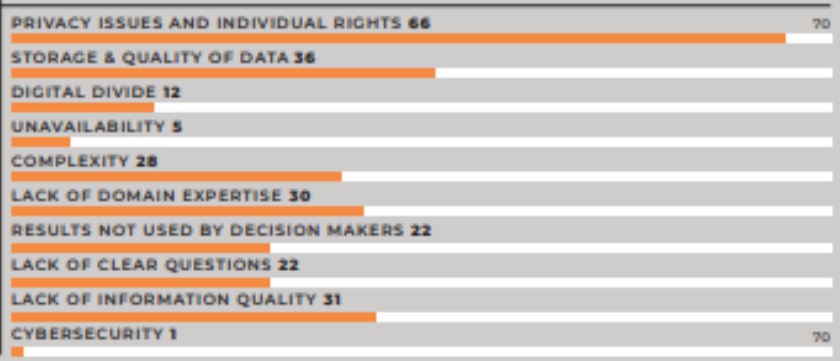
Which technology will truly make a difference in the coming 4-5 years?



In what application field would you say data science/ applied mathematics will truly make a difference in the coming 4-5 years?



What do you expect will be the critical issues and risks data science will have to cope with in the near future?

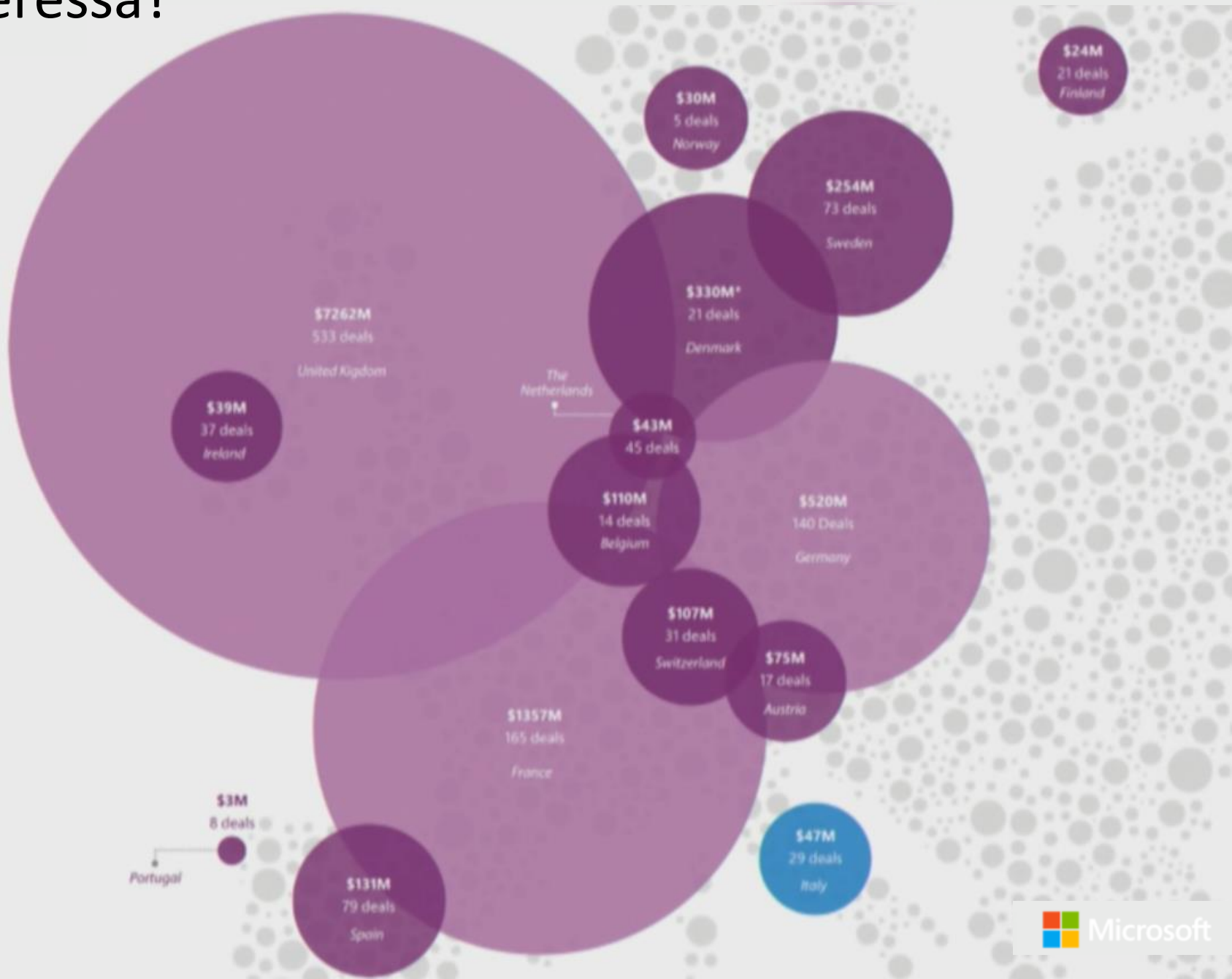


Data Science, a chi interessa?

Business Investments in AI

A few, big AI deals influencing the overall picture

AI companies invested into per country, mUSD (accumulated 2008-2018)

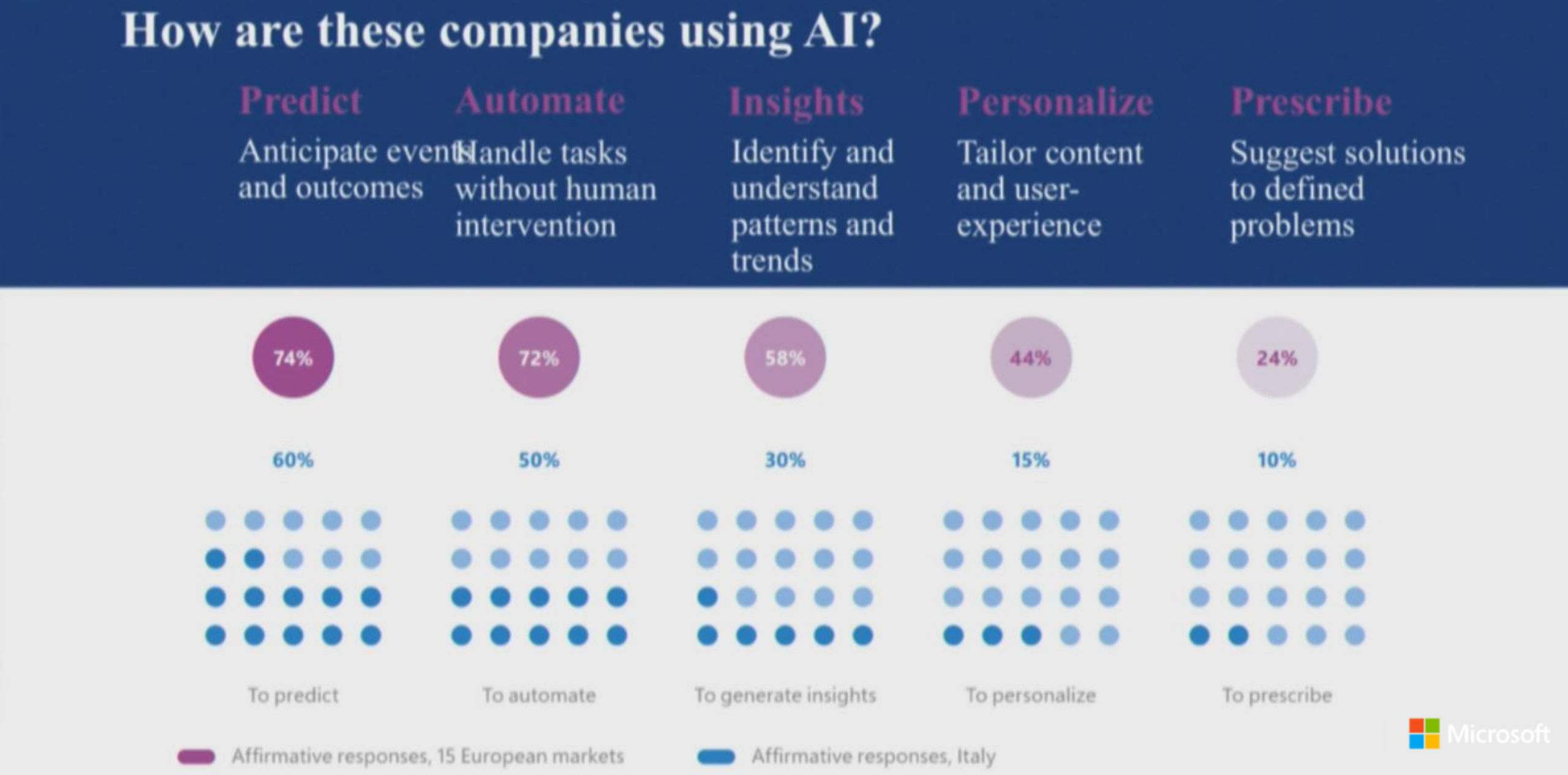


Italy

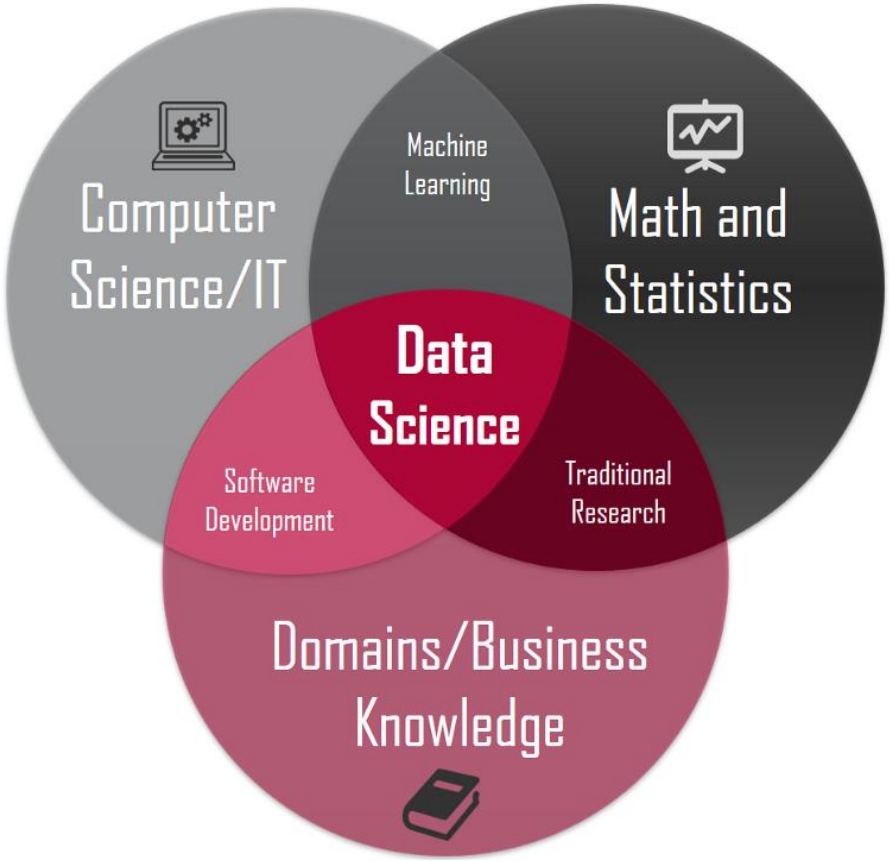
European markets

Microsoft

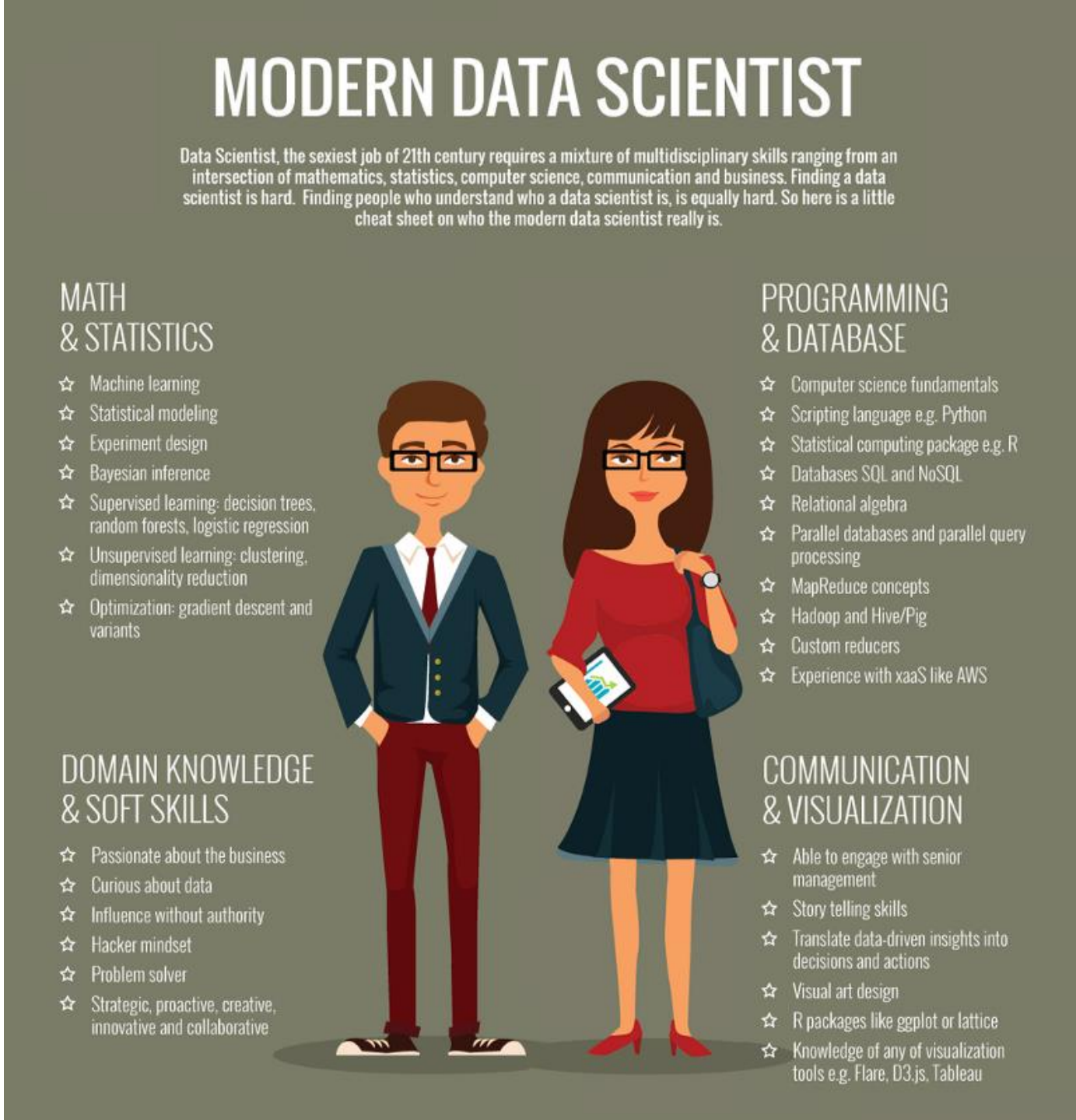
Data Science, a chi interessa?



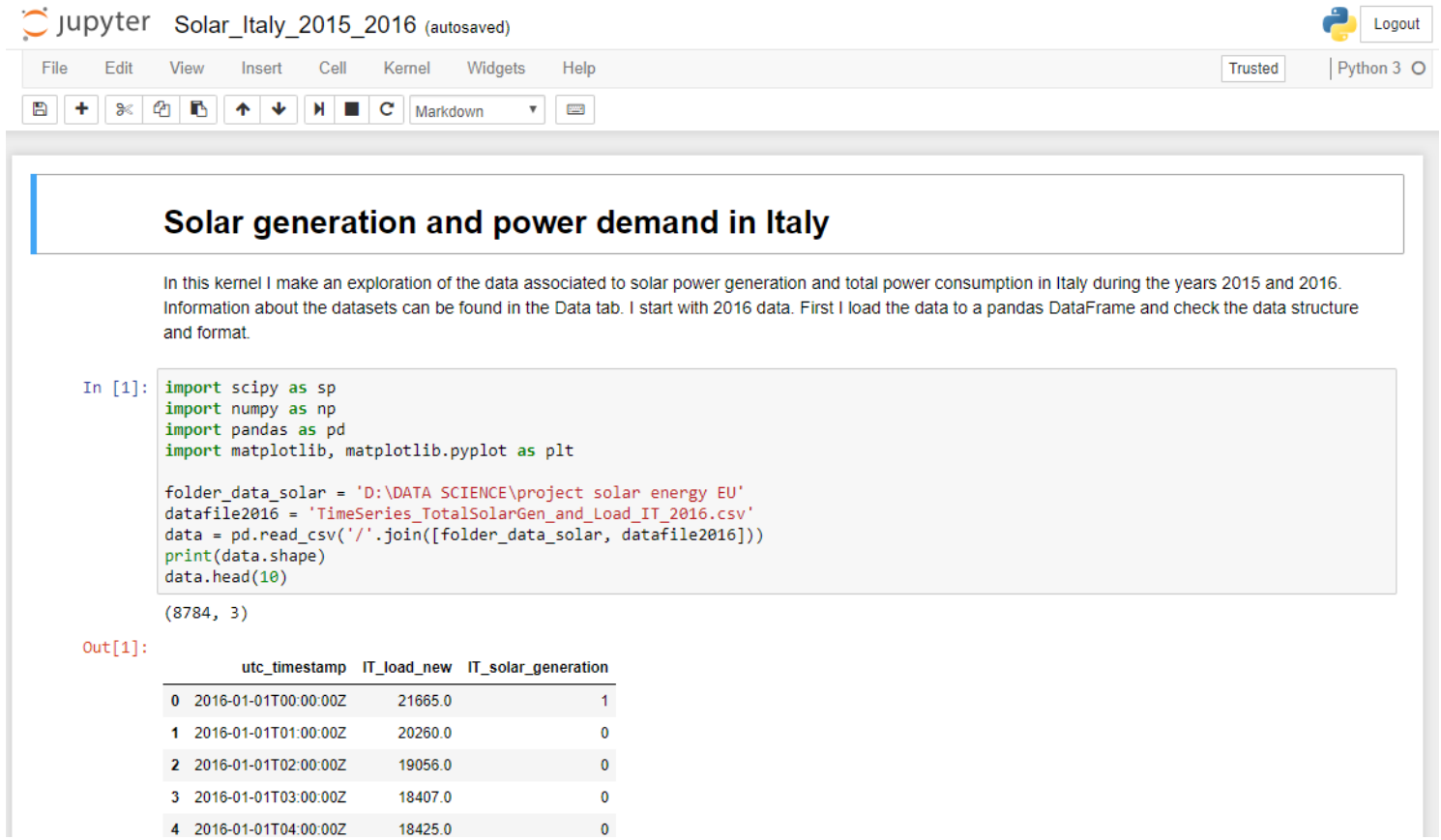
Data Science, chi lo fa?

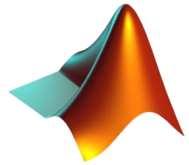


https://towardsdatascience.com/introduction-to-statistics-e9d72d818745



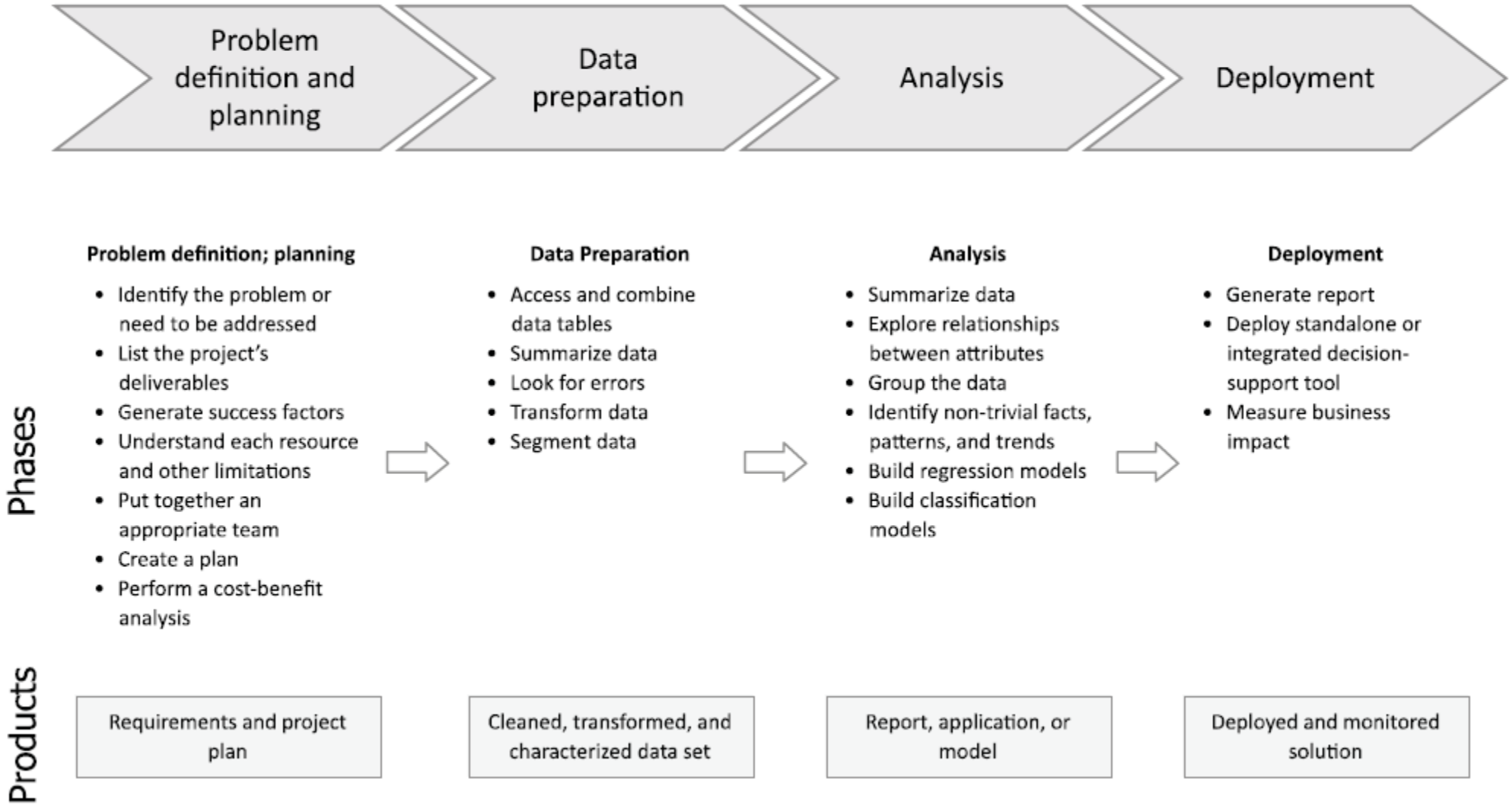
Data Science, come si implementa?





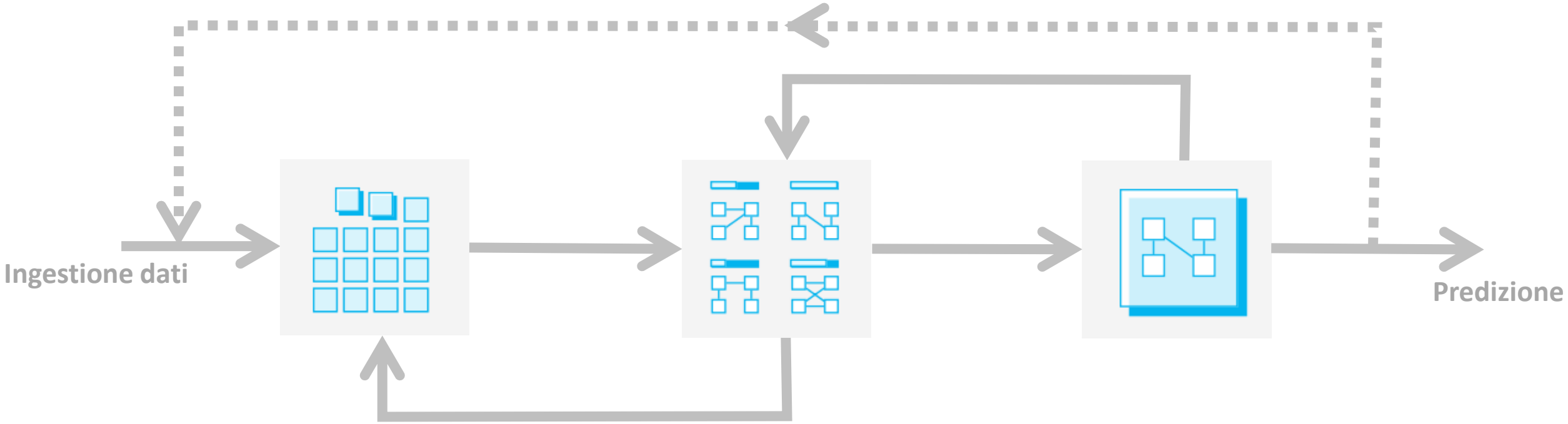
2. Processo della Data Science

Data Science, un processo iterativo



MAKING SENSE OF DATA I - A Practical Guide to Exploratory Data Analysis and Data Mining, G. Myatt & W. Johnson, Wiley 2014

Data Science, un processo iterativo

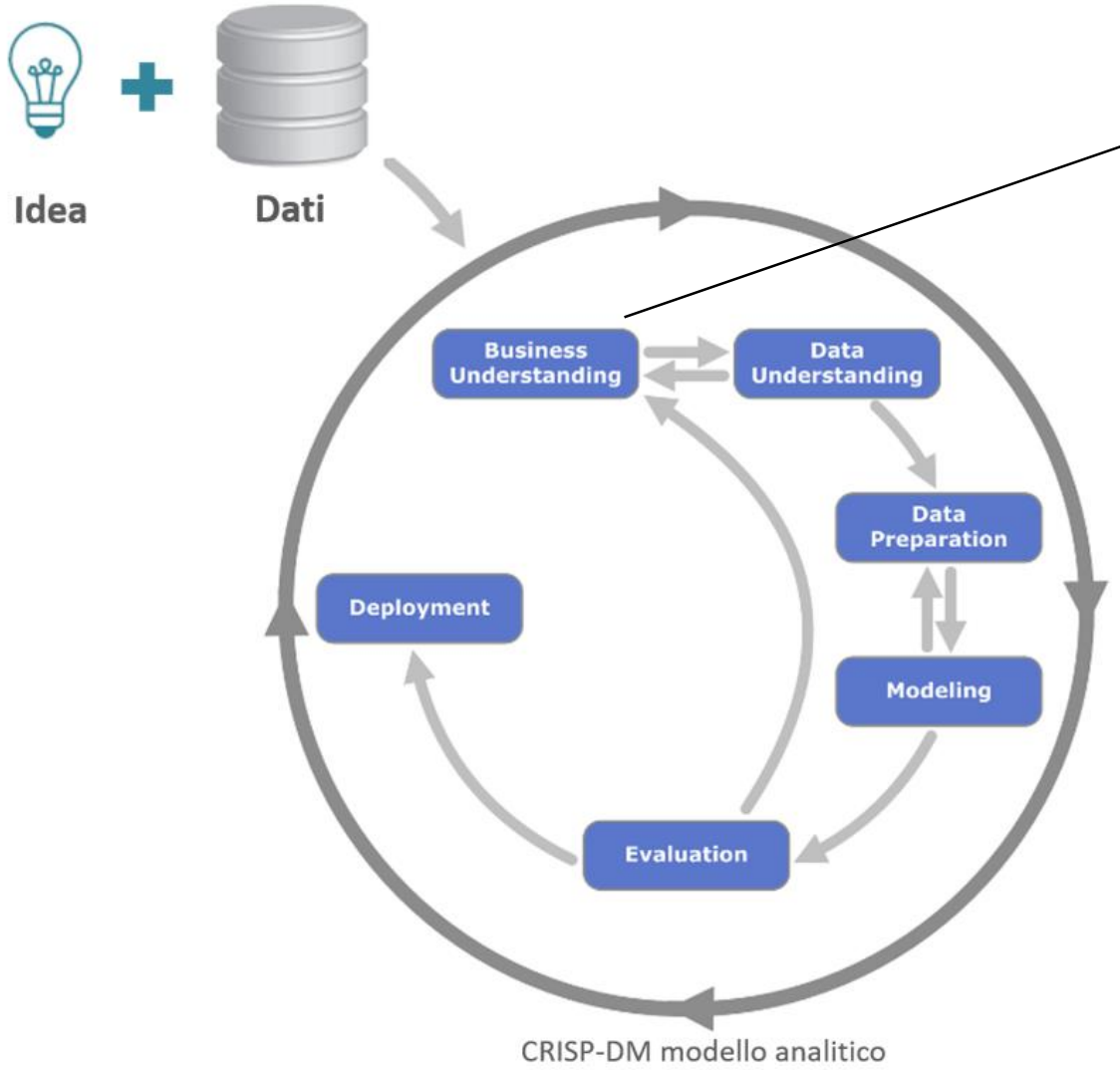


Preparazione
Feature engineering
Esplorazione

Analisi
Modello
Training

Distribuzione
Consumo
Monitoraggio

Data Science, un processo iterativo



CRISP-DM modello analitico

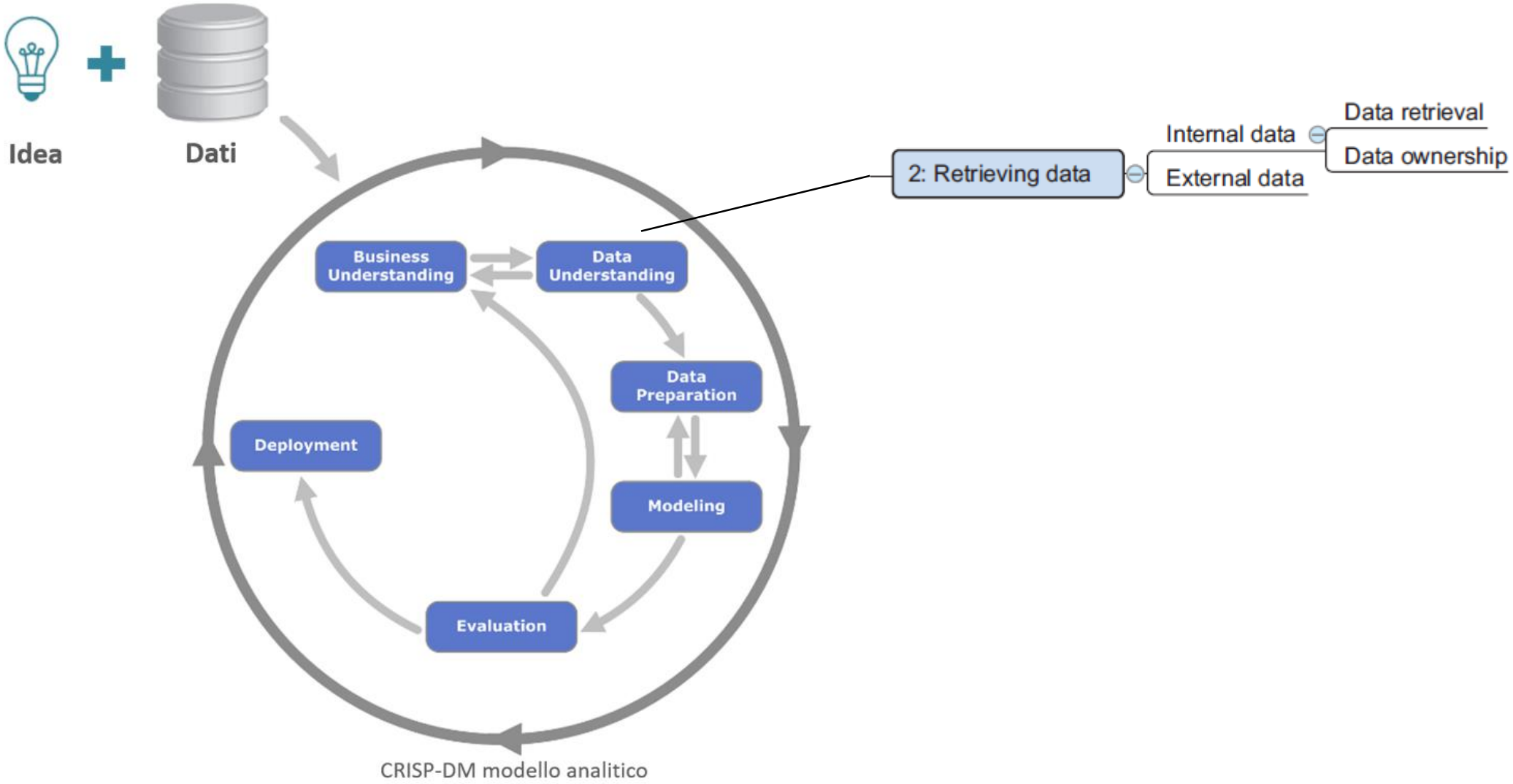
1: Setting the research goal

Define research goal

Create project charter

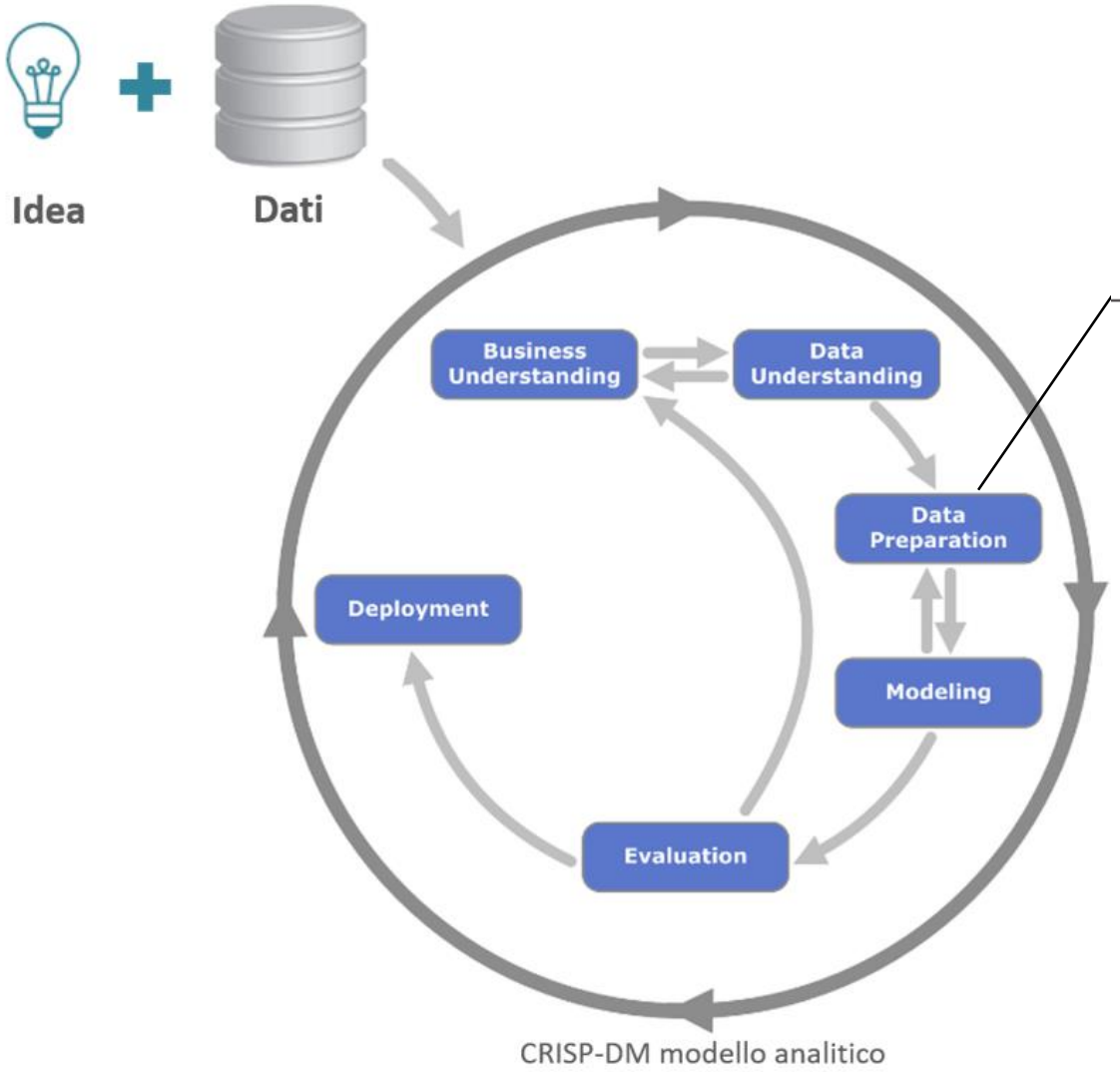
- A clear research goal
- The project mission and context
- How you're going to perform your analysis
- What resources you expect to use
- Proof that it's an achievable project, or proof of concepts
- Deliverables and a measure of success
- A timeline

Data Science, un processo iterativo

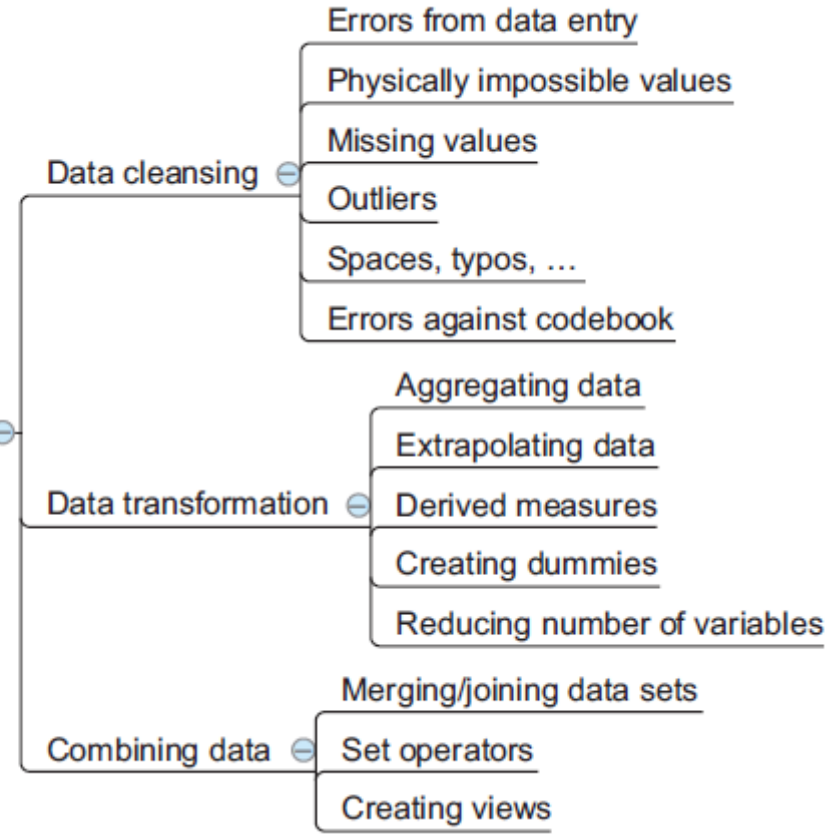


CRISP-DM modello analitico

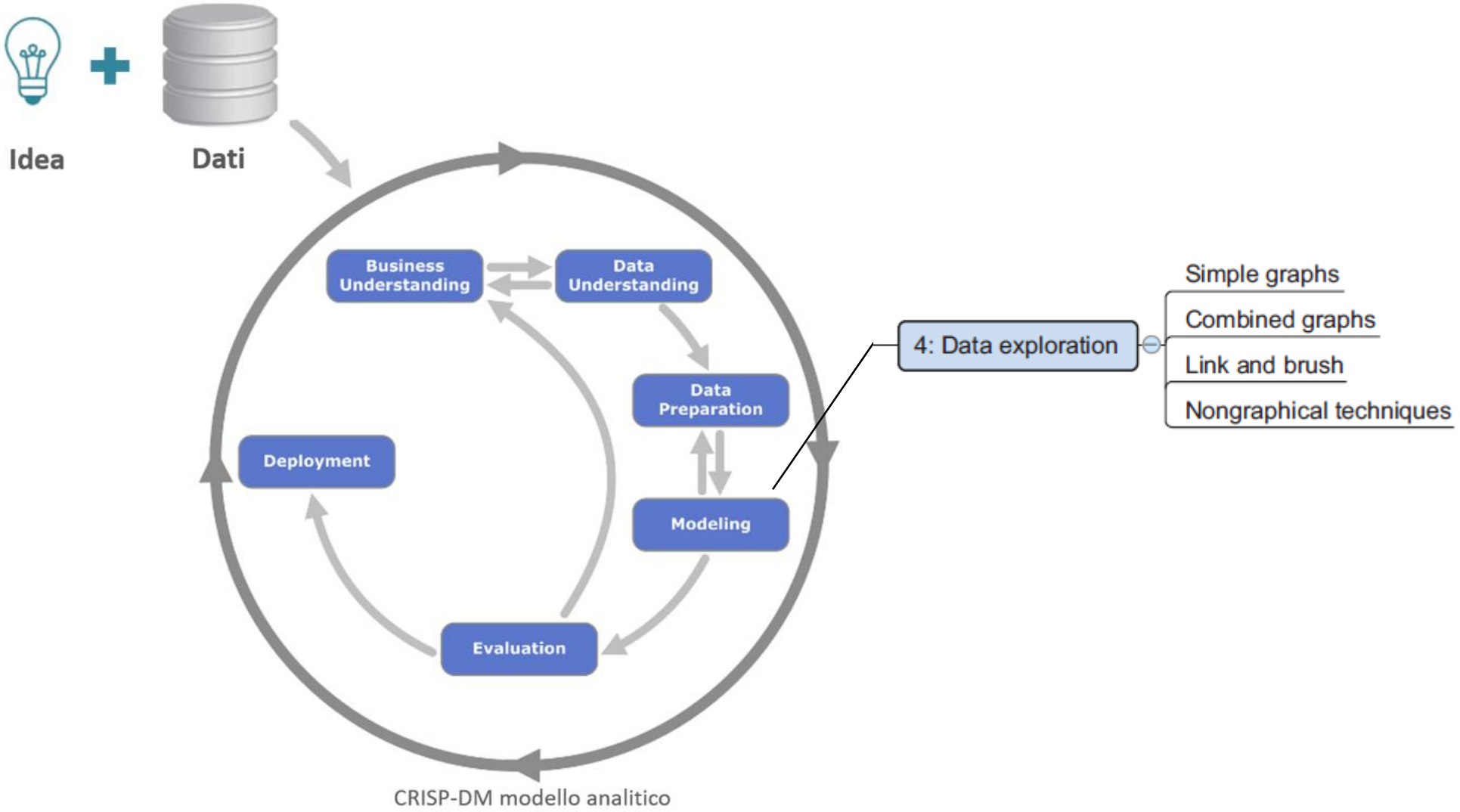
Data Science, un processo iterativo



CRISP-DM modello analitico

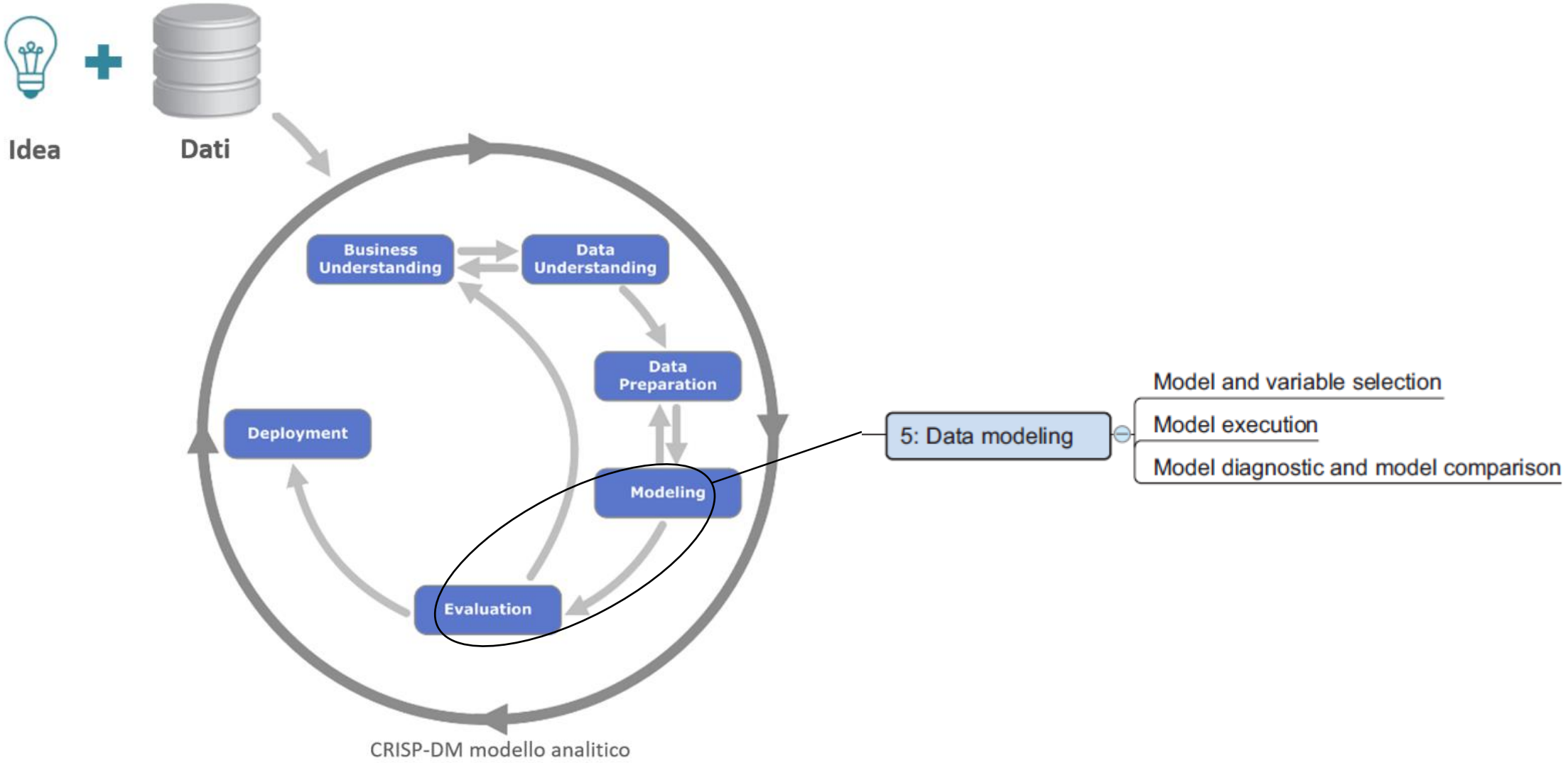


Data Science, un processo iterativo



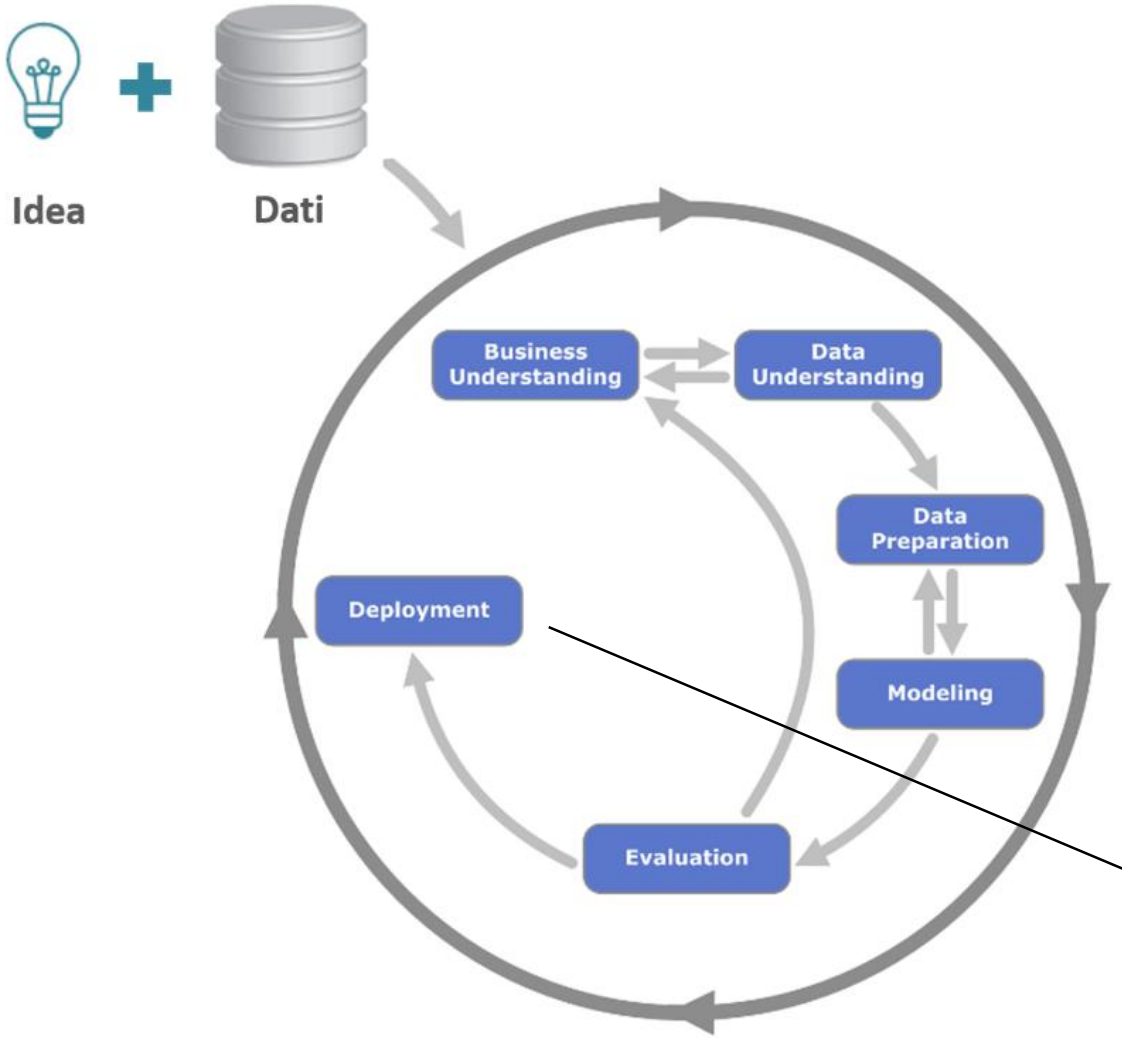
CRISP-DM modello analitico

Data Science, un processo iterativo



CRISP-DM modello analitico

Data Science, un processo iterativo



CRISP-DM modello analitico

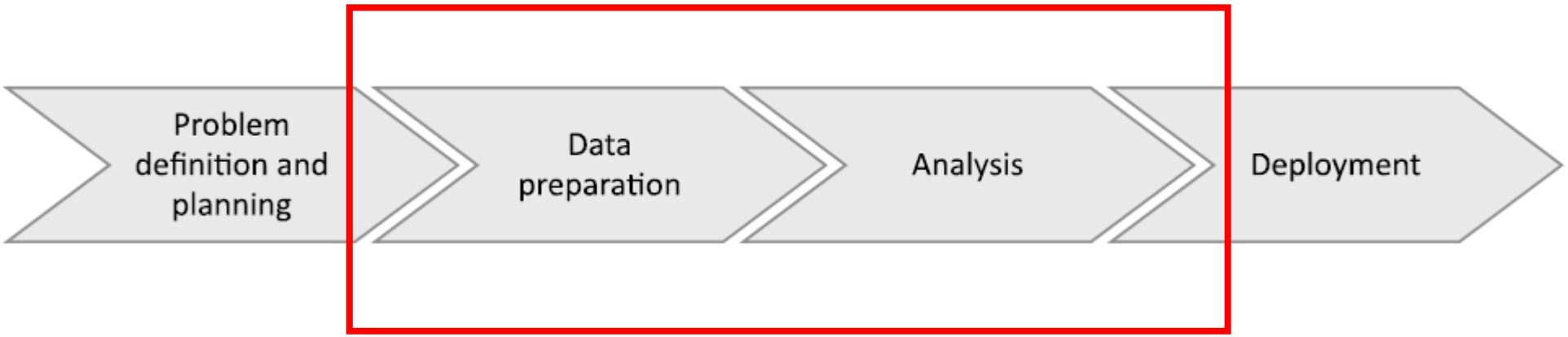
6: Presentation and automation

Presenting data

Automating data analysis

3. Analisi dei dati come parte del processo della DS

Analisi dei dati



- Preparazione dei dati: 70% - 80% del totale del lavoro
- Analisi
  - Descrittiva
  - Esplorativa
  - Predittiva

Preparazione dei dati

Integrazione

- Combinare datasets
- Creare views

Pulizia

- Errori
- Valori impossibili
- Valori mancanti
- Outliers
- Inconsistenze
- Duplicati
- Rimuovere colonne

Trasformazione

- Scaling
- Feature engineering
- Creare dummies
- Riduzione di variabili
- Raggruppare categorie
- Aggregazione

Get a free car valuation

Find out how much a car is worth

See what you could get if you sold your car yourself or part exchanged.
Or get a guide price if you're looking to buy.

Enter registration
e.g. AB12CDE

Enter mileage
e.g. 10000

Get a valuation

Don't know the registration? Select the car's make and model.

Simple is anything *but* easy...

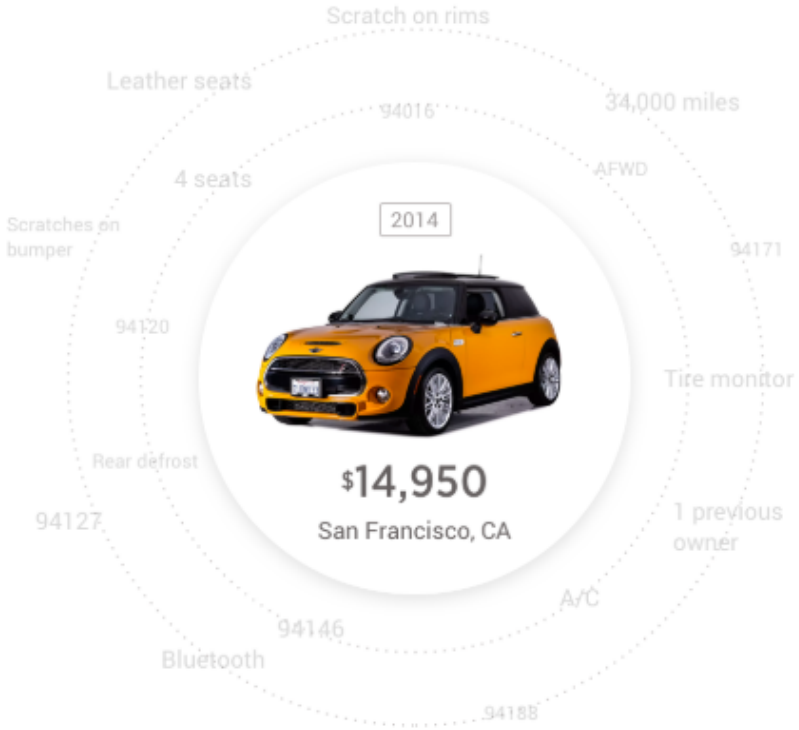
Marketplace & consumer products

Logistics platforms

Pricing engine

Because no two cars are the same, pricing is a huge algorithmic challenge that's at the core of our business. We use machine learning to accurately price vehicles while accounting for fluctuations in demand, availability, and local-market conditions.

Targeted inventory & pricing



Caso di uso: car pricing

Tabella macchine (216 x 26)

Attribute:	Attribute Range:
symboling:	-3, -2, -1, 0, 1, 2, 3.
normalized-losses:	continuous from 65 to 256.
make_id:	continuous from 1 to 22.
fuel-type:	diesel, gas.
aspiration:	std, turbo.
num-of-doors:	four, two.
body-style:	hardtop, wagon, sedan, hatchback, convertible.
drive-wheels:	4wd, fwd, rwd.
engine-location:	front, rear.
wheel-base:	continuous from 86.6 120.9.
length:	continuous from 141.1 to 208.1.
width:	continuous from 60.3 to 72.3.
height:	continuous from 47.8 to 59.8.
curb-weight:	continuous from 1488 to 4066.
engine-type:	dohc, dohcv, l, ohc, ohcf, ohcv, rotor.
num-of-cylinders:	eight, five, four, six, three, twelve, two.
engine-size:	continuous from 61 to 326.
fuel-system:	1bbl, 2bbl, 4bbl, idi, mfi, mpfi, spdi, spfi.
bore:	continuous from 2.54 to 3.94.
stroke:	continuous from 2.07 to 4.17.
compression-ratio:	continuous from 7 to 23.
horsepower:	continuous from 48 to 288.
peak-rpm:	continuous from 4150 to 6600.
city-mpg:	continuous from 13 to 49.
highway-mpg:	continuous from 16 to 54.
price:	continuous from 5118 to 45400.

Tabella fabbriche (22 x 2)

Attribute:	Attribute Range:
make_id:	continuous from 1 to 22.
make:	alfa-romero, audi, bmw, chevrolet, dodge, honda, isuzu, jaguar, mazda, mercedes-benz, mercury, mitsubishi, nissan, peugot, plymouth, porsche, renault, saab, subaru, toyota, volkswagen, volvo

Number of Attributes: 26 total
16 continuous (numerico, comma flottante)
1 integer (numerico, intero)
9 nominal (= categorico)

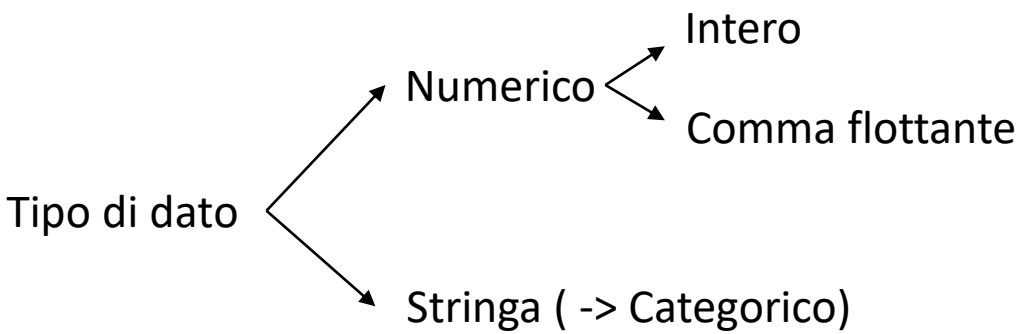
Caso di uso: car pricing

Tabella macchine (216 x 26)

symboling	216 non-null	int64
normalized-losses	175 non-null	float64
make-id	216 non-null	int64
fuel-type	216 non-null	string
aspiration	216 non-null	string
num-of-doors	214 non-null	string
body-style	216 non-null	string
drive-wheels	216 non-null	string
engine-location	216 non-null	string
wheel-base	216 non-null	float64
length	216 non-null	float64
width	216 non-null	float64
height	216 non-null	float64
curb-weight	216 non-null	int64
engine-type	216 non-null	string
num-of-cylinders	216 non-null	string
engine-size	216 non-null	int64
fuel-system	216 non-null	string
bore	212 non-null	float64
stroke	212 non-null	float64
compression-ratio	216 non-null	float64
horsepower	214 non-null	float64
peak-rpm	214 non-null	float64
city-mpg	216 non-null	int64
highway-mpg	216 non-null	int64
price	212 non-null	float64

Tabella fabbriche (22 x 2)

make-id	22 non-null	int64
make	22 non-null	string



Caso di uso: car pricing

Integrazione dei dati

Tabella macchine

make-id	....	price
1	...	13495
1	...	16500
2	...	13950
2	...	17450
2	...	15250
...	...	...

216 x 26

Tabella fabbriche

make-id	make
1	alfa-romero
2	audi
3	bmw
4	chevrolet
5	dodge
...	...

22 x 2

make-id	make	....	price
1	alfa-romero	...	13495
1	alfa-romero	...	16500
2	audi	...	13950
2	audi	...	17450
2	audi	...	15250
...	...	...	...

216 x 27

Caso di uso: car pricing

Esempi di errori da sanare

Inconsistenze

Data entry error

make-id	normalized-losses	num-of-doors	body-style	curb-weight	num-of-cylinders	horsepower	city-mpg	highway-mpg	price
1		two	convertible	2548	four	111	21	27	13495
1		two	hatchback	-2823	six	154	19	26	16500
2	164	four	sedan	2337	four	102	9999	9999	13950
2	164	four	sedan	2824	?	115	18	22	17450
2		two	sedan	2507	?	110	19	25	15250
2	158	four	sedan	2844	?	110	19	25	17710
2		four	wagon	2954	five	110	19	25	18920
2	158	four	sedan	3086	five	NaN	17	20	23875
2		two	hachback	3053	five	160	16	22	
3	192	two	sedan	2395	four	101	23	29	16430
3	192	four	sedan	2395	four	NaN	23	29	16925
3	188	two	sedan	2710	six	121	21	28	20970
3	188	two	sedan	2710	six	121	21	28	20970
3	188	four	sedan	2765	six	121	21	28	21105
3		four	sedan	3230	six	182	16	22	30760
3		two	sedan	3380	six	NaN	16	22	41315
3		four	sedan	3505	six	182	15	20	36880
4	121	two	hatchback	1488	three	48	47	53	5151

Duplicati

Blanks = Dati mancanti

Dati mancanti

make-id	make
1	alfa-romero
2	audi
3	bmw
4	chevrolet
5	dodge
6	honda
7	isuzu
8	jaguar
9	mazda
10	mercedes-benz
11	mercury
12	mitsubishi
13	nissan
14	peugot
15	plymouth
16	porsche
17	renault
18	saab
19	subaru
20	toyota
21	volkswagen
22	volvo

Caso di uso: car pricing

Cosa fare con i dati mancanti?

- Rimuovere le righe
- Sostituire con 0’s
- Interpolare linearmente
- Forward/Backward fill (riempire con il valore precedente/seguente)
- Imputation (metodo statistico per trovare il valore più probabile)

In [17]:
```
macchine.head(10)
```

Out[17]:

	symboling	normalized-losses	make-id	fuel-type	aspiration	num-of-doors	body-style	drive-wheels	engine-location	wheel-base	length	width	height	curb-weight	engine-type	num-of-cylinders	engine-size	sy
0	3	NaN	1	gas	std	two	convertible	rwd	front	88.6	168.8	64.1	48.8	2548	dohc	four	130	
1	3	NaN	1	gas	std	two	convertible	rwd	front	88.6	168.8	64.1	48.8	2548	dohc	four	130	
2	1	NaN	1	gas	std	two	hatchback	rwd	front	94.5	171.2	65.5	52.4	2823	ohcv	six	152	
3	2	164.0	2	gas	std	four	sedan	fwd	front	99.8	176.6	66.2	54.3	2337	ohc	four	109	
4	2	164.0	2	gas	std	four	sedan	4wd	front	99.4	176.6	66.4	54.3	2824	ohc	five	136	
5	2	NaN	2	gas	std	two	sedan	fwd	front	99.8	177.3	66.3	53.1	2507	ohc	five	136	
6	1	158.0	2	gas	std	four	sedan	fwd	front	105.8	192.7	71.4	55.7	2844	ohc	five	136	
7	1	NaN	2	gas	std	four	wagon	fwd	front	105.8	192.7	71.4	55.7	2954	ohc	five	136	
8	1	158.0	2	gas	turbo	four	sedan	fwd	front	105.8	192.7	71.4	55.9	3086	ohc	five	131	
9	0	NaN	2	gas	turbo	two	hatchback	4wd	front	99.5	178.2	67.9	52.0	3053	ohc	five	131	

In [7]:
```
macchine.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 216 entries, 0 to 215
Data columns (total 26 columns):
symboling            216 non-null int64
normalized-losses    175 non-null float64
make-id              216 non-null int64
fuel-type            216 non-null object
aspiration           216 non-null object
```

4. Tipi di analisi

Alcuni tipi di analisi dei dati

- Descrittiva: Sintesi dei dati, univariate e bivariate charts
- Esplorativa: Relazione fra le variabili, correlazione, outliers
- Inferenziale: Inferenza dei parametri di una popolazione a partire da un campione
- Predittiva: Applicazione di algoritmi per predire risultati (regressione, classificazione)

Alcuni tipi di analisi dei dati

Statistica descrittiva:

- Moda

  3, 4, 5, 6, 7, 7, 7, 8, 8, 9 -> **7**

- Mediana

  3, 4, 7, 2, 3, 7, 4, 2, 4, 7, 4 -> 2, 2, 3, 3, 4, **4**, 4, 4, 7, 7, 7

- Media

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

3, 4, 5, 7, 7, 8, 9, 9, 9 -> 61/9 = **6.78**

Alcuni tipi di analisi dei dati

Statistica descrittiva:

- Campo di variazione

  3, 4, 5, 6, 7, 7, 7, 8, 8, 8, 9 -> **6**

- Quartile

  3, 4, 7, 2, 3, 7, 4, 2, 4, 7, 4 -> 2, 2, 3, 3, 4, **4**, 4, 4, 7, 7, 7 = Q2 (mediana)

  2, 2, **3**, 3, 4, 4, 4, 4, 7, 7, 7 = Q1

  2, 2, 3, 3, 4, 4, 4, 4, **7**, 7, 7 = Q3

Alcuni tipi di analisi dei dati

Statistica descrittiva:

- Varianza

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

3, 4, 4, 5, 5, 5, 6, 6, 6, 7, 7, 8, 9

$\bar{x} = 5.8$ -> s^2 = 34.32/(13-1) = **2.86**

- Deviazione standard

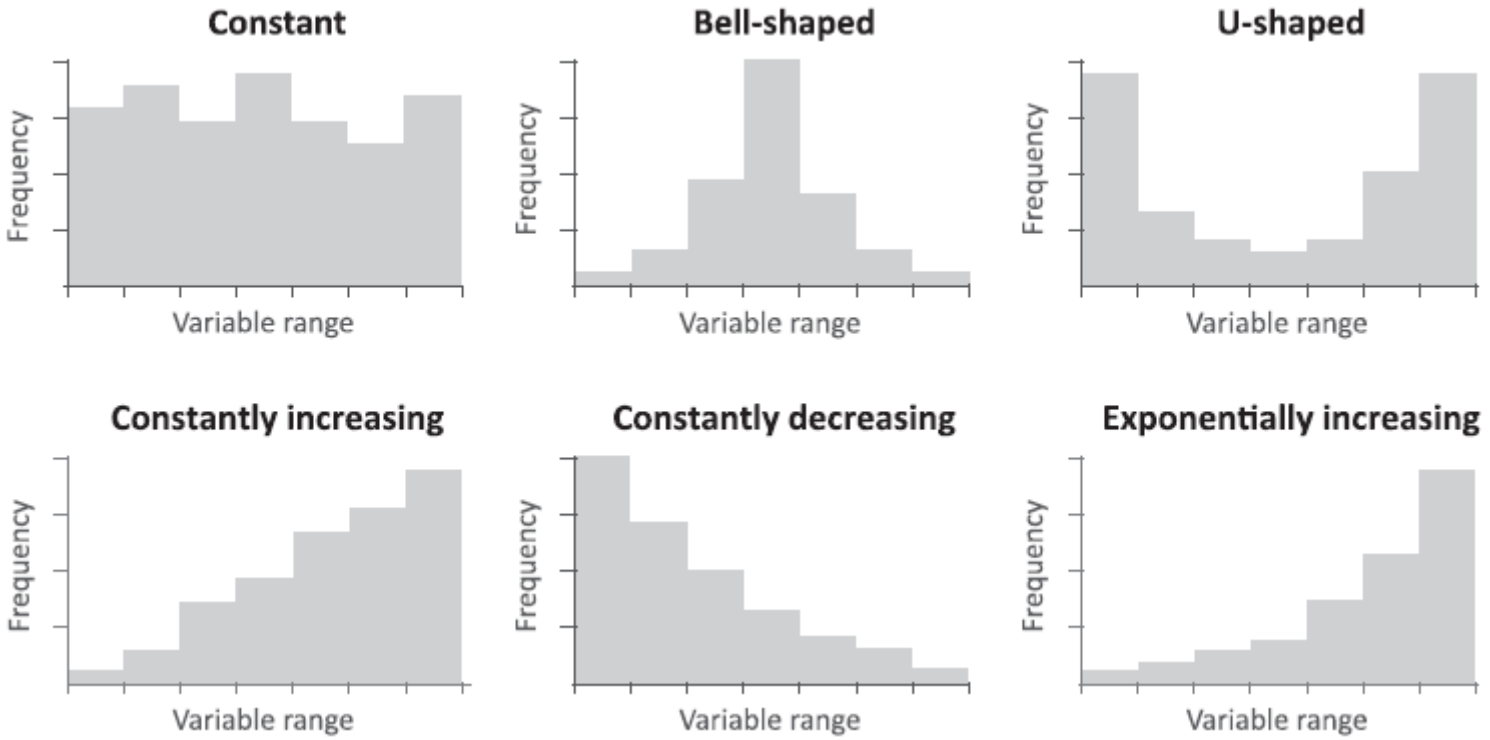
$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

s = **1.69**

Alcuni tipi di analisi dei dati

Statistica descrittiva:

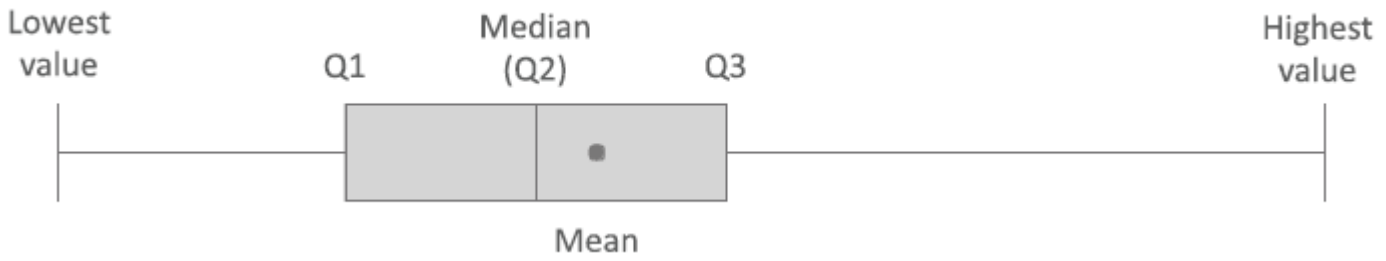
- Istogramma

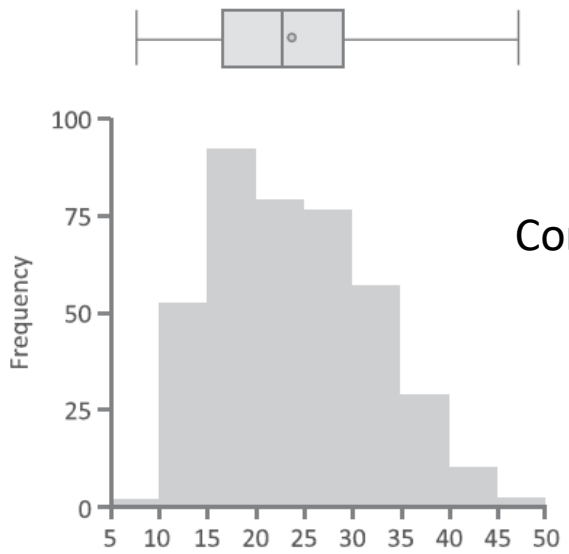


Alcuni tipi di analisi dei dati

Statistica descrittiva:

- Box plot



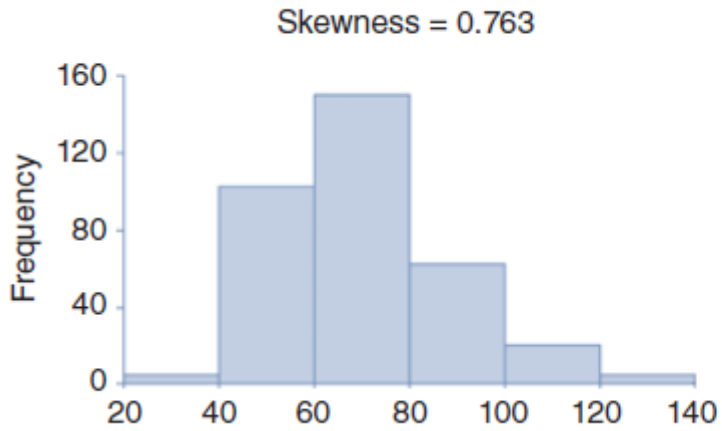


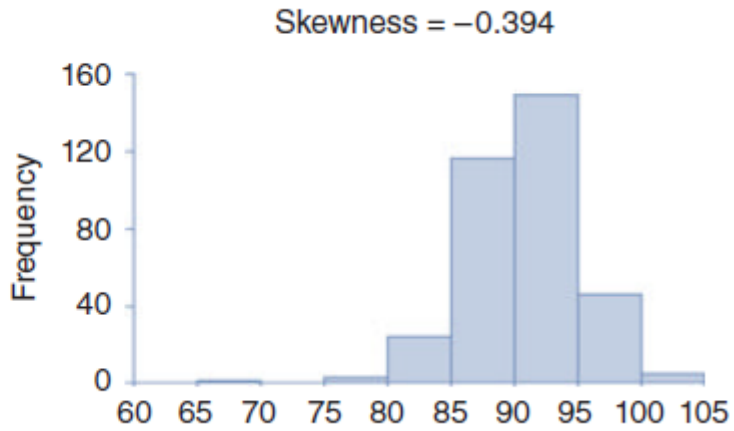
Confronto istogramma – box plot

Alcuni tipi di analisi dei dati

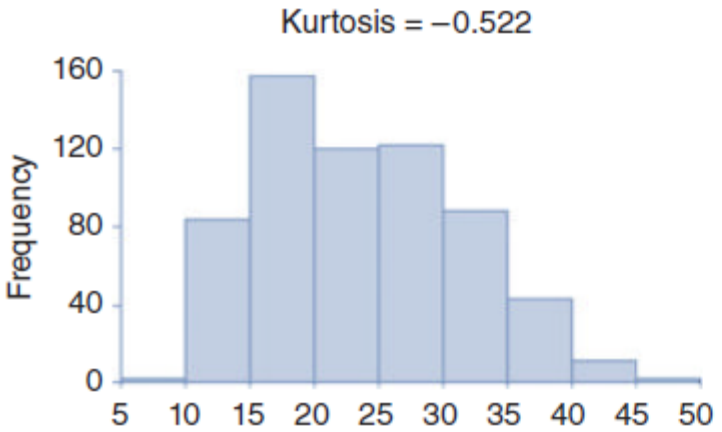
Statistica descrittiva:

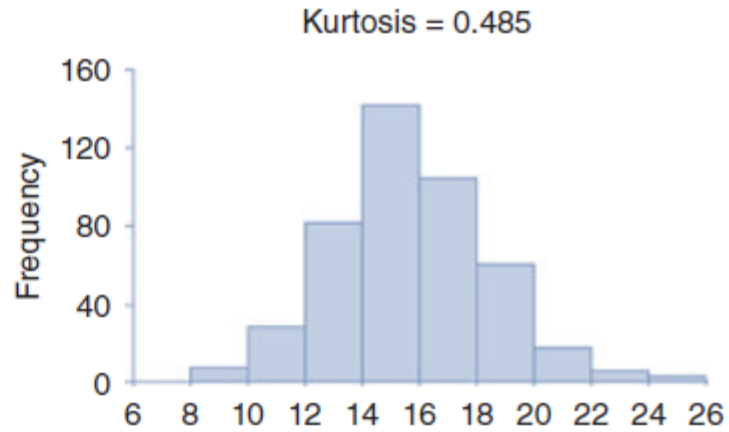
- Indice di asimmetria





- Curtosi

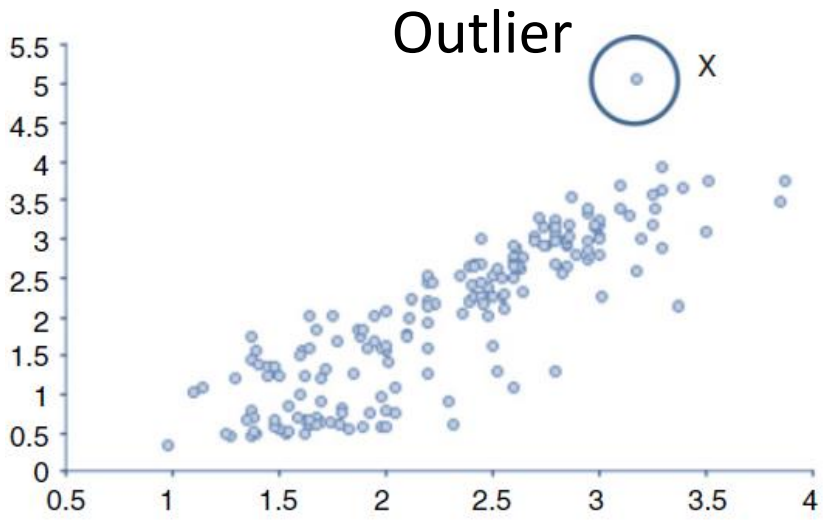




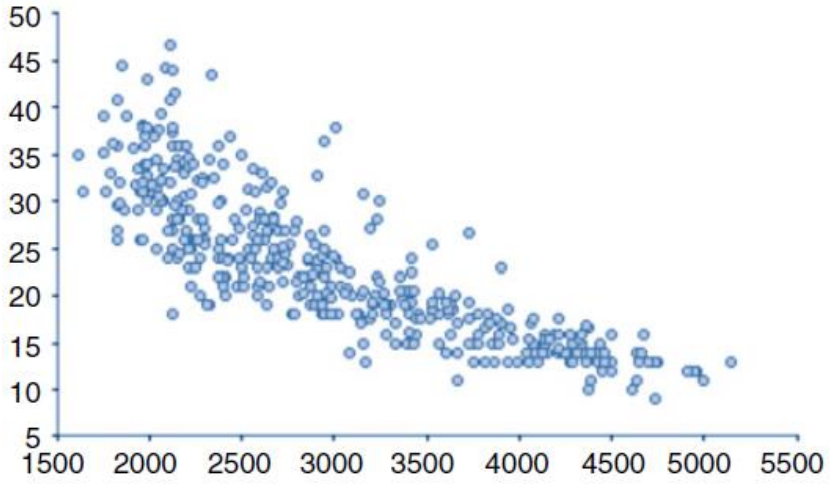
Alcuni tipi di analisi dei dati

Analisi esplorativa:

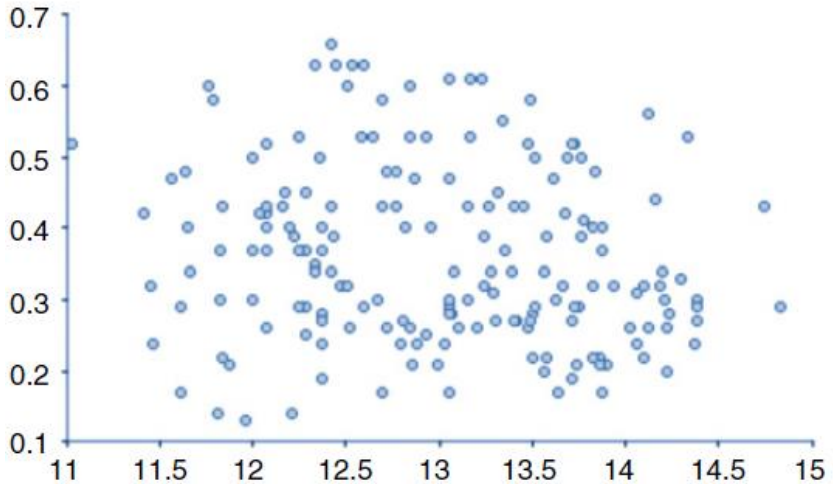
- Grafici a dispersione



Relazione lineare positiva



Relazione non-lineare negativa



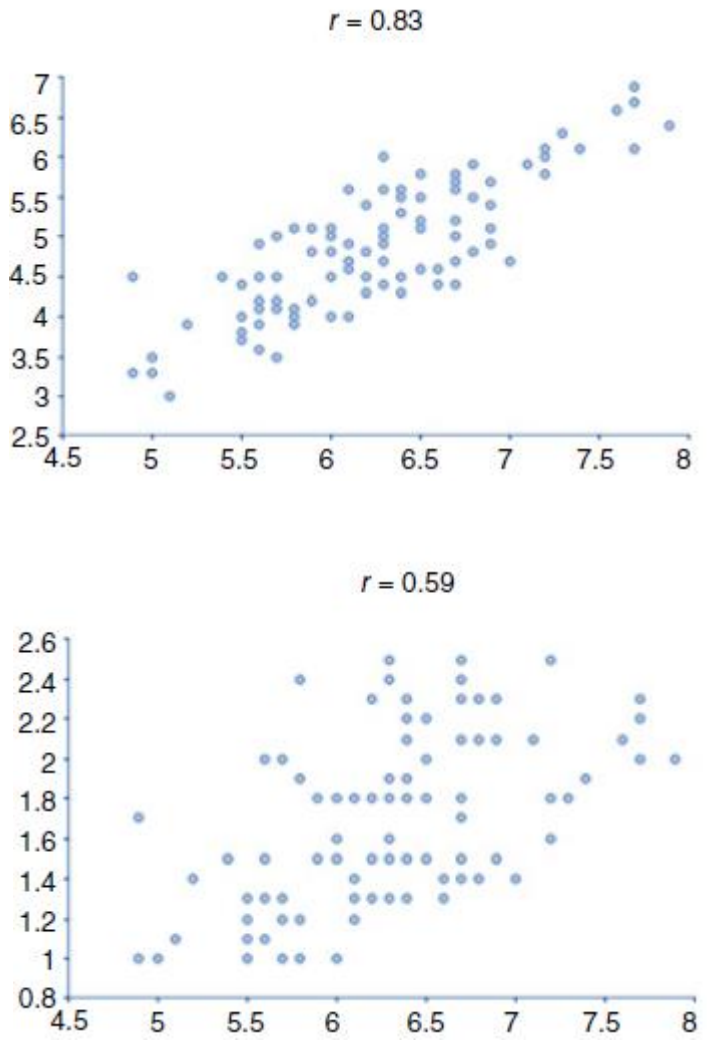
Nessuna relazione

Alcuni tipi di analisi dei dati

Analisi esplorativa:

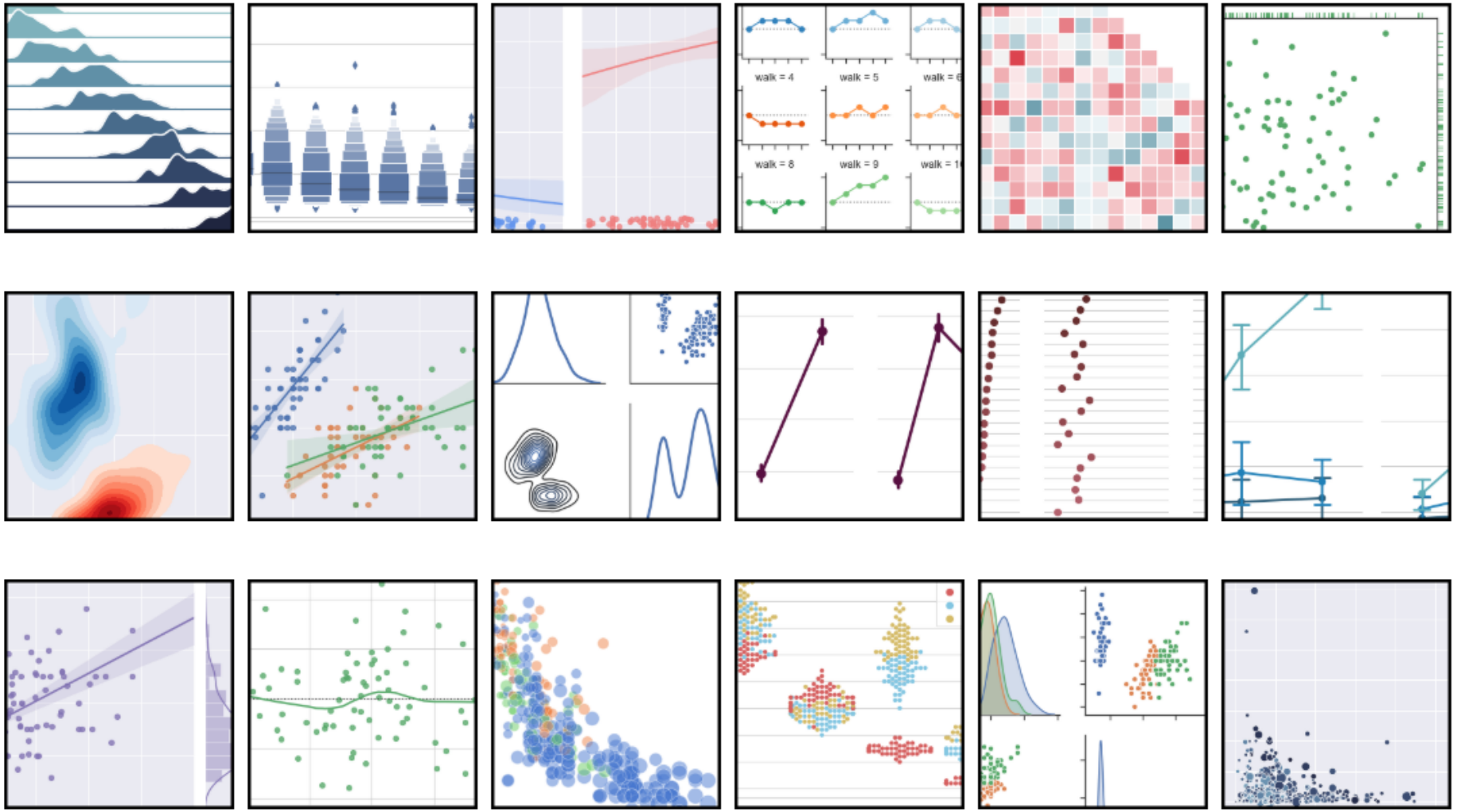
- Correlazione Pearson

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$



Alcuni tipi di analisi dei dati

Analisi esplorativa:



https://seaborn.pydata.org/examples/index.html

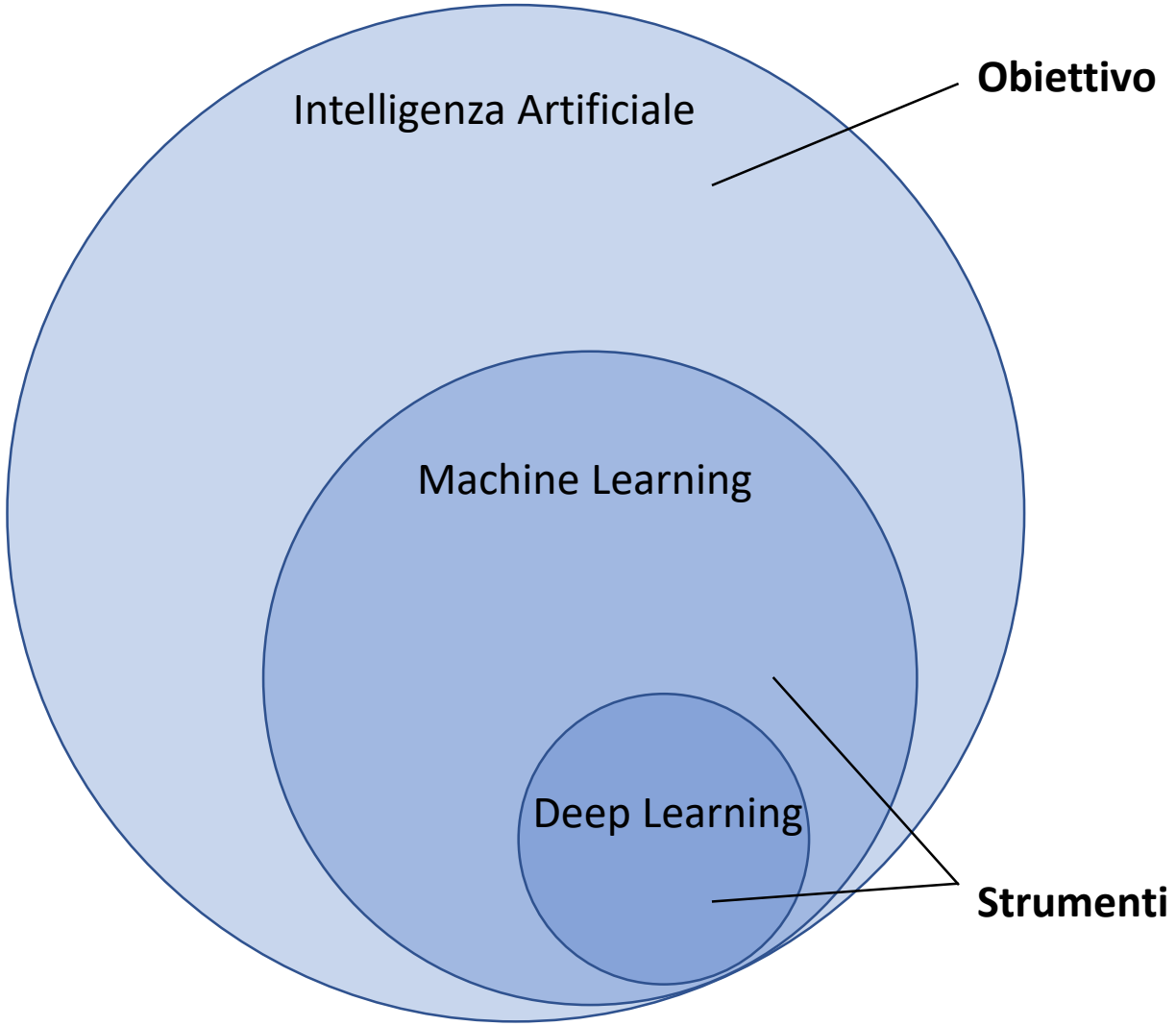
Alcuni tipi di analisi dei dati

Analisi predittiva: **<u>Machine Learning</u>**

Machine Learning è la tecnica che permette ai computer di apprendere a partire dai dati.

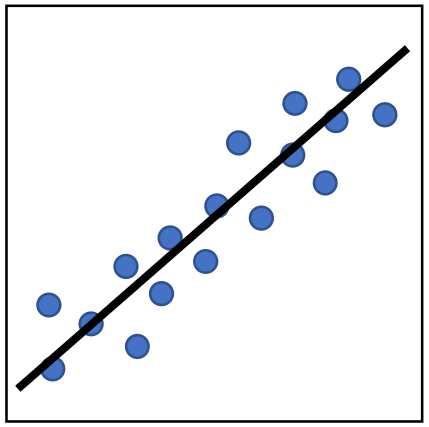
Apprendere a fare cosa?

- Determinare un valore numerico
- Classificare un obietto tra 2 o più classi
- Aggruppare dati simili
- Raccomandare un’azione
- Predire un dato futuro
- Rilevare un evento non normale

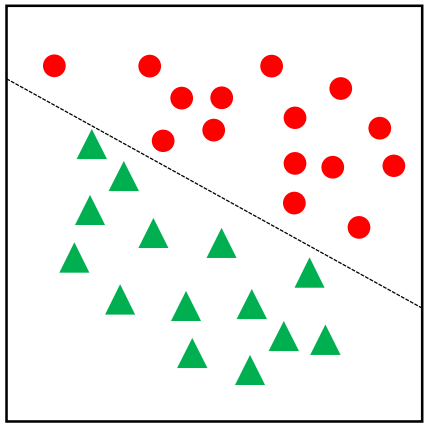


Machine Learning

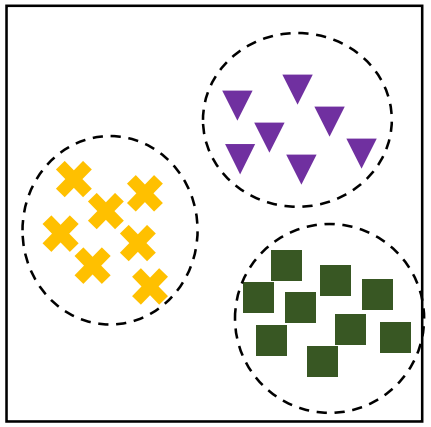
Analisi predittiva



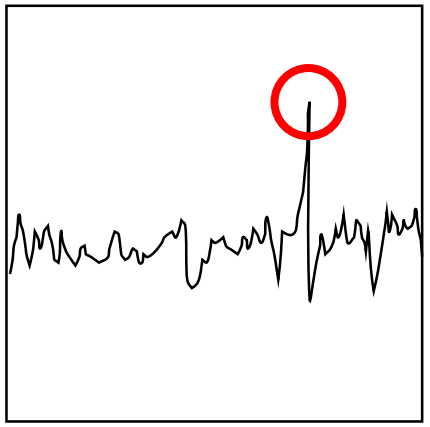
Regressione



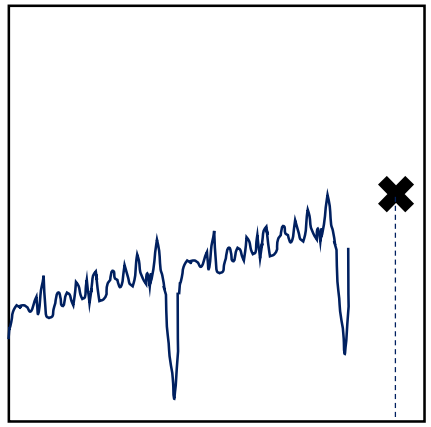
Classificazione



Clustering



Anomaly detection

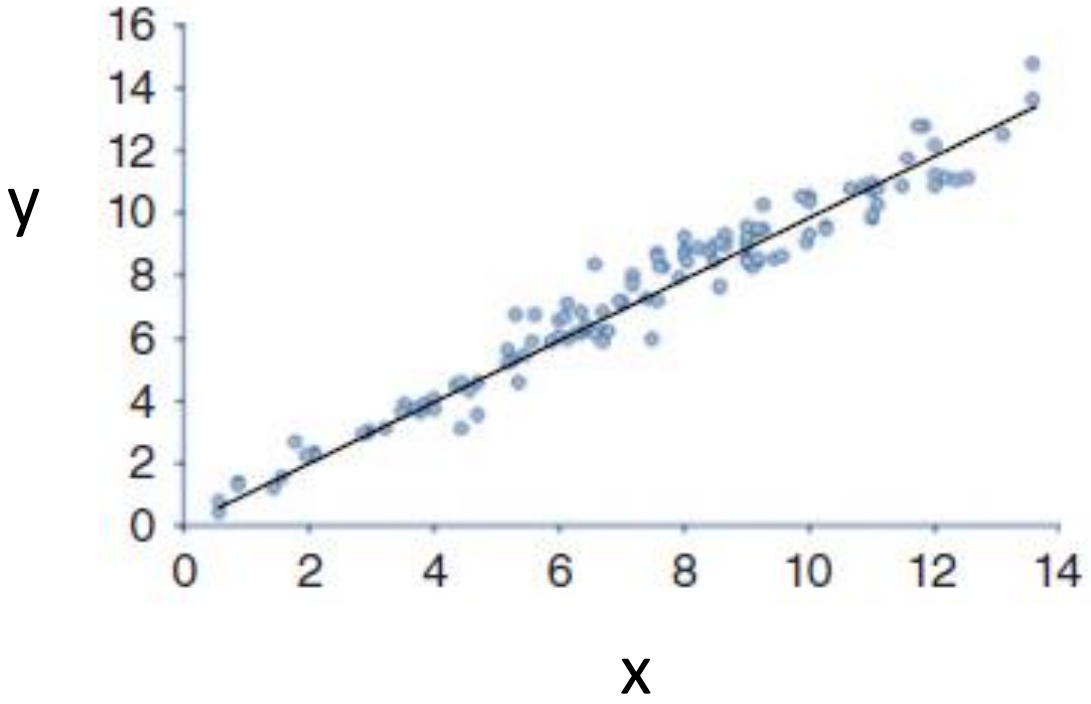


Time-series forecasting

Machine Learning

Esempio: Regressione lineare

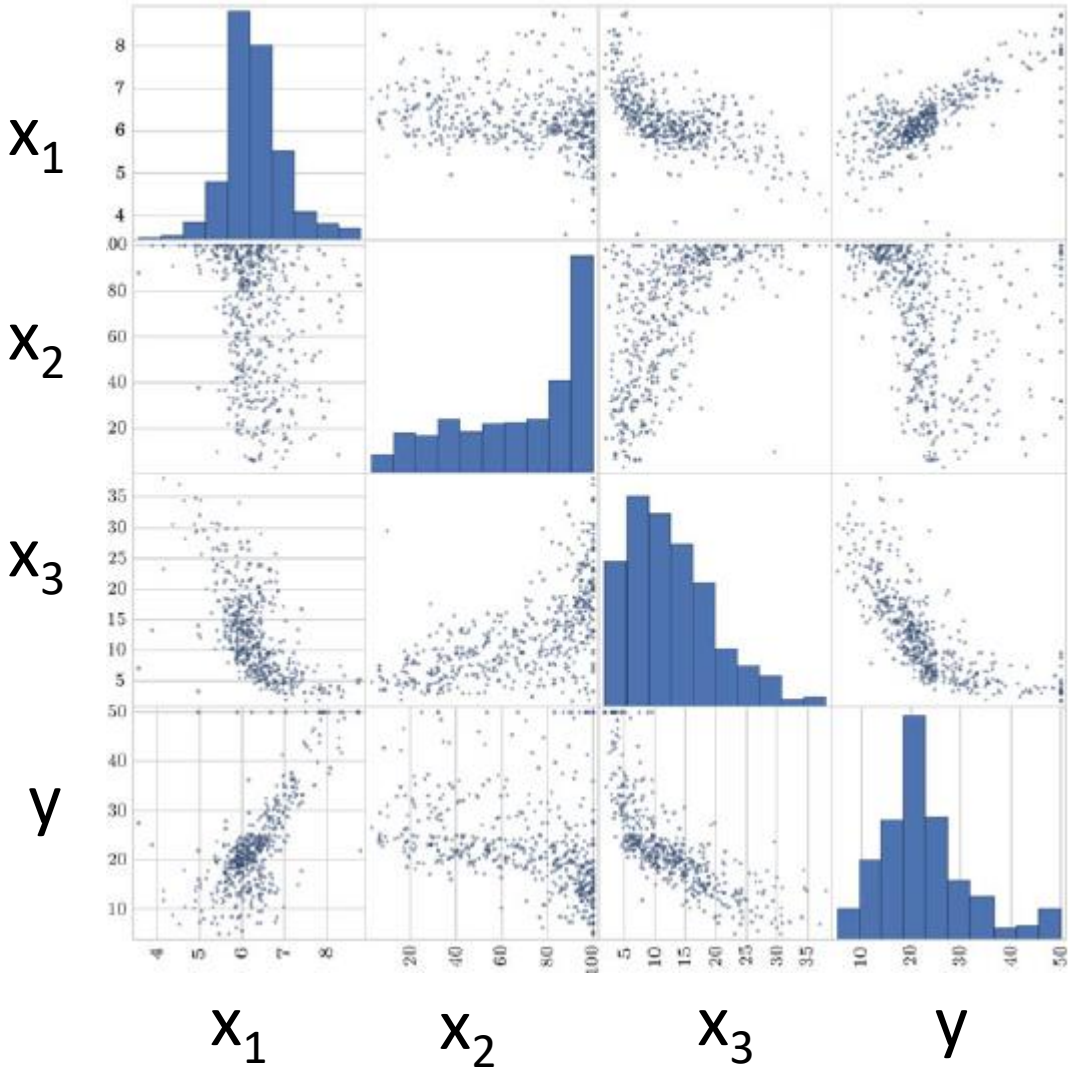
$$y = f(x)$$



Machine Learning

Esempio: Regressione lineare

$$y = f(x_1, x_2, x_3)$$



Machine Learning

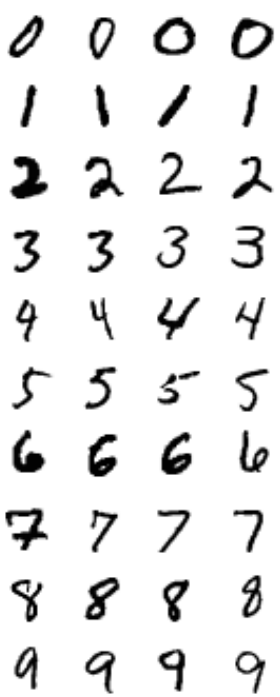
Esempio: Classificazione multi-class (riconoscimento di immagini)

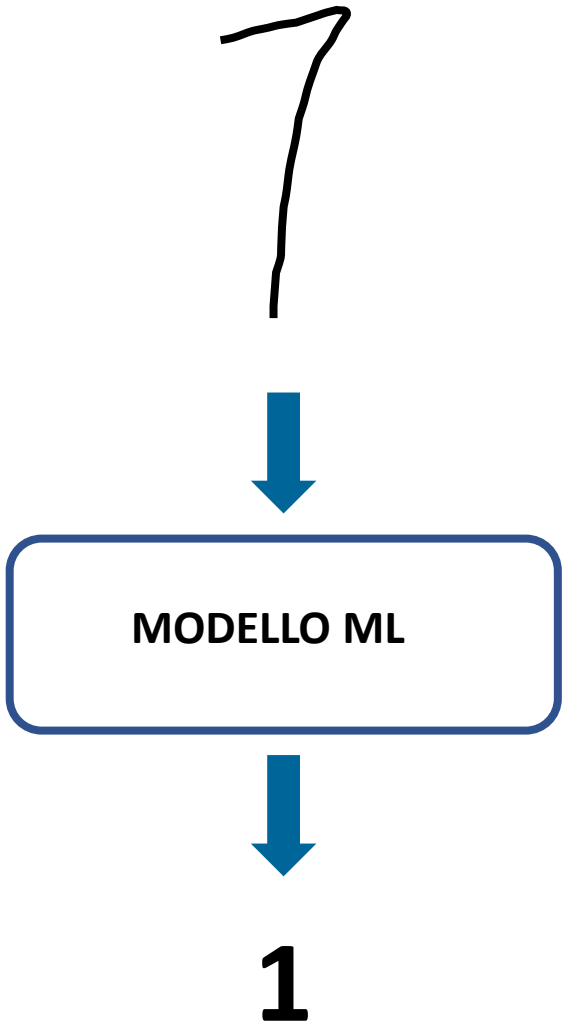
Machine Learning

Esempio: Classificazione multi-class (riconoscimento di immagini)

Training dataset

Test dataset



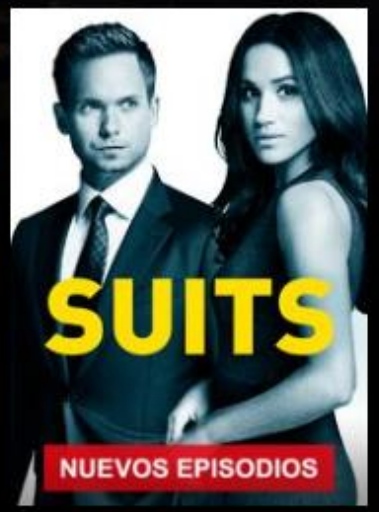


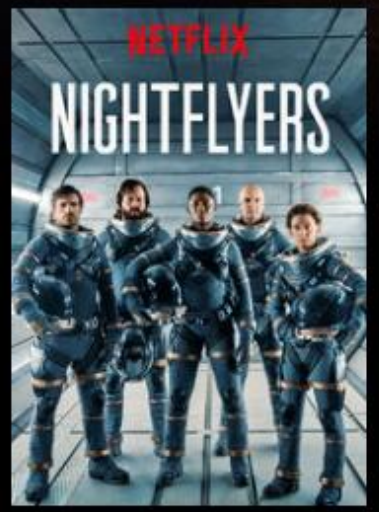
5. Applicazioni, casi reali

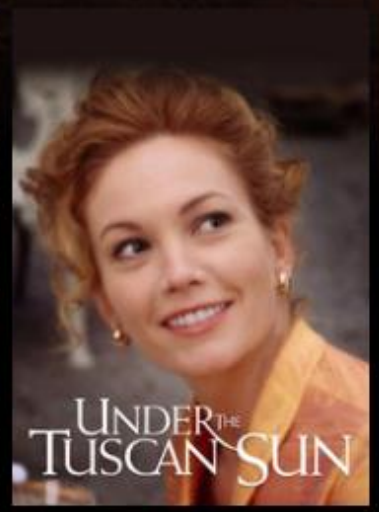
tflix

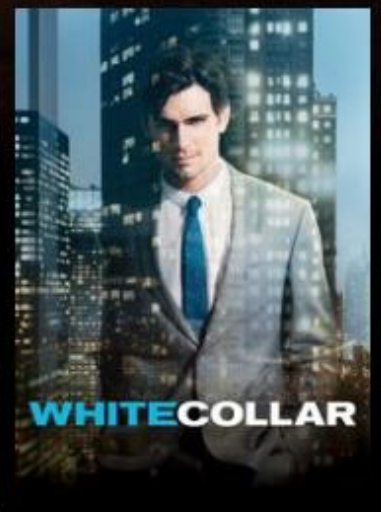
5 NETFLIX

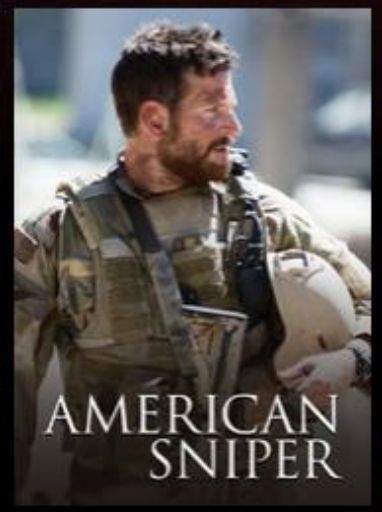
Nuestra selección para Ariel











Populares en Netflix

IBM

Marketplace

Search IBM Marketplace

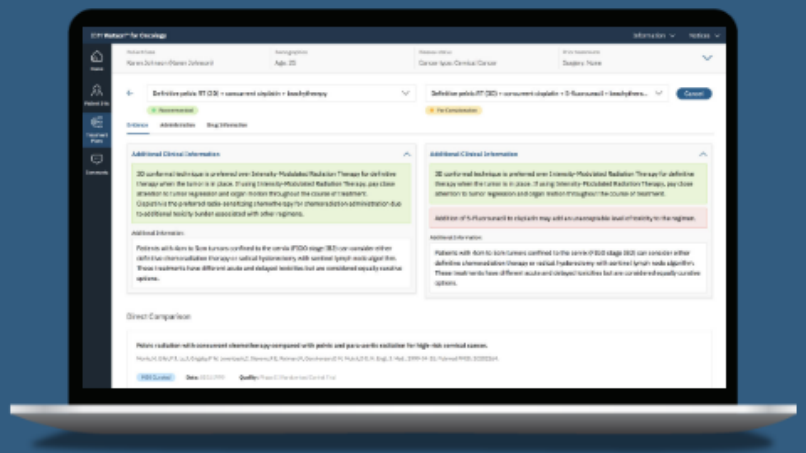
IBM Watson for Oncology

Overview | Details | Pricing | Resources | Support | Watch the video

IBM Watson for Oncology

Watson for Oncology helps physicians quickly identify key information in a patient's medical record, surface relevant evidence and explore treatment options.

Watson for Oncology | Contact Us



Join us at Think 2019 | February 12-15 | San Francisco

Get $200 USD off the conference rate with any purchase in the IBM Marketplace

What Watson for Oncology can do for your organization

Let's talk

All things Alexa

Instantly connect to Alexa to play music, control your smart home, and get information, news, weather, and more using just your voice.

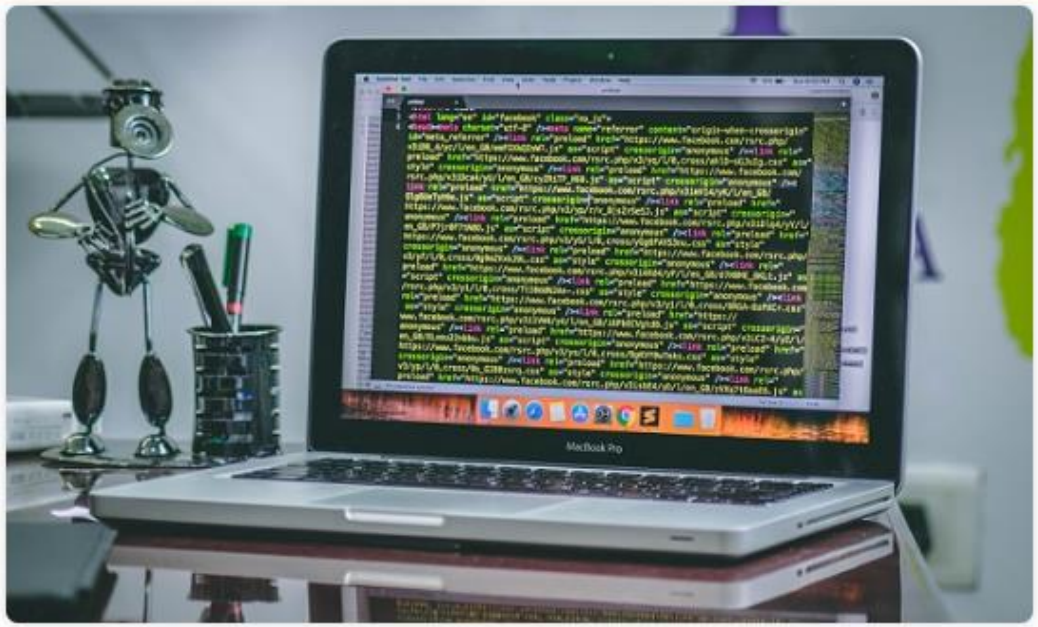
Shop devices | Use Alexa

Amazon Devices


FORMATION LAP
ROLEX
PREDICTED PIT STOP / TYRE STRATEGY
LAP 18-28
1 STOP U S
1 STOP SS S
LAP 28-38
TYRES
U SS S
Insights

Algoritmi e immobiliare, non più due mondi sconosciuti!

Scritto da Maria Fruscione



Casavo, l'algoritmo decide che il prezzo è giusto e la casa si vende in 30 giorni

L'INIZIATIVA DI UNA STARTUP CHE HA CONVINTO IL MERCATO: LA SUA RACCOLTA FONDI HA SUPERATO I 9 MILIONI. LA SOCIETÀ ARRIVA ANCHE AD ACQUISTARE IN PRIMA PERSONA L'IMMOBILE PER RESTARE NEI TEMPI GARANTITI. ECCO COME FUNZIONA

Giorgio Tinacci ad di Casavo

Barbara Ardù

Roma

Velocizzare il mercato delle vendite immobiliari nelle grandi città. E' quanto si è messo in testa Giorgio Tinacci, 27 anni, ex consulente di Boston Consulting Group, che a ottobre ha lanciato Casavo, startup immobiliare che nel giro di sei mesi ha raccolto 9,49 milioni di euro in Italia. La sua idea ha dunque convinto gli investitori italiani abituati a puntare sulle startup. Meglio se innovative e Casavo lo è. Utilizza un sofisticato processo di valutazione basato su algoritmi che triangolano diverse fonti dati e che alla fine permette di determinare con precisione il valore di una proprietà. Il prezzo "giusto" si potrebbe definire, dato il quartiere, la grandezza e altre variabili. Fissato il valore della casa, e questa è la novità, Casavo si impegna ad acquistare in un tempo che non va oltre i 30 giorni. Certo a un prezzo che è standard, magari un po' più basso da quello che forse il proprietario si aspetterebbe, ma sempre il più alto possibile. «Noi ci prendiamo l'incertezza - afferma Giorgio Tinacci, ad di Casavo - perché va da sé che l'immobile acquisito potrebbe essere venduto a un valore più elevato, ma questa certezza non è data e soprattutto noi eliminiamo al venditore lo stress dell'attesa». Per ora le case vendute sono state dieci, ma dovrebbero arrivare a 60 entro l'anno. E superare l'obiettivo non sembra così difficile, anche perché le agenzie immobiliari piuttosto che nemici, in realtà sono i loro migliori alleati. «Ne abbiamo già 200 convenzionate - spiega Tinacci - perché siamo una risorsa anche per loro». E se per i privati non è prevista alcuna commissione, le agenzie pagano un prezzo per usufruire della piattaforma e vendere tramite Casavo». Perché lo fanno? «Perché la velocità è tutto. Sono tanti i motivi che inducono qualcuno a voler vendere in poco tempo. E per un'agenzia poter offrire al suo cliente l'opzione Casavo è una aggiunta».

In un mercato, quello immobiliare dove la volatilità dei prezzi è alta, c'è poca trasparenza e i tempi di vendita sono più elevati di quelli della media Ue (tre mesi contro i nostri sei), Casavo potrebbe offrire certezza e trasparenza. Due qualità che fanno comodo anche alle agenzie, per proporre al cliente una vendita veloce. A lavorare nella startup milanese sono al momento in 12, tutti giovani e con formazione tecnologica. In autunno Casavo approderà a Roma. La data non è ancora fissata, ma lo sbarco è dato per certo. E secondo Giorgio Tinacci i numeri arriveranno presto. «Basta vedere il successo avuto da Open Door, che lavora come noi, ma da ben prima di noi negli Usa o alle altre realtà che stanno crescendo in Europa».

Che cos'è un algoritmo?

Nell'era della digitalizzazione, conoscere la risposta a questa domanda è necessario. La gran parte degli oggetti e degli strumenti di cui disponiamo e utilizziamo ogni giorno, si basano su algoritmi.

"L'algoritmo è un procedimento che risolve un determinato problema attraverso un numero finito di passi elementari, chiari e non ambigui, in un tempo ragionevole."

Ma perchè parlare di algoritmi, e **perchè sono così importanti anche nel mercato immobiliare?**

Una delle tante novità di Casavo è proprio l'utilizzo di un algoritmo. Vediamo ora in che modo.

In un'epoca come questa, qualsiasi innovazione in grado di velocizzare i processi e rendere più semplice, intuitivo e veloce il processo per il cliente, viene accolta e vista di buon occhio.

Progetti di Data Science

Manufacturing

- Manutenzione predittiva
- Analisi di sensori in tempo reale
- Previsione consumo di energia

Retail

- Previsione di vendita
- Ottimizzazione dei prezzi
- Offerte personalizzate
- Ottimizzazione delle campagne di marketing
- Previsione domanda di noleggio

Banking

- Previsione rischio di credito
- Rilevamento e prevenzione delle frodi online
- Previsione prezzo di azioni

Healthcare

- Predizione flusso di pazienti
- Previsione durata degenza
- Previsione di malattie
- Diagnosi precoce

Anche: Educazione, entertainment, business, ecc…

https://gallery.azure.ai/

Progetti di Data Science

Manutenzione preventiva nell'industria

Analisi di dati da sensori, impostazioni operative e cronologia della manutenzione di macchine in funzione, per svolgere il monitoraggio condition-based e fare previsioni di guasti e manutenzione predittiva.

Comportamento dei consumatori basato su dati di geolocalizzazione

Studio del comportamento offline dei consumatori tramite l'analisi di dati di geolocalizzazione mobile, per la segmentazione, identificazione di abitudini e preferenze di potenziali clienti, utile per l'avvio di campagne di marketing personalizzate.

Previsione del consumo energetico

Analisi di dati riguardanti il consumo di energia elettrica in Italia durante un numero di anni, e applicazione di algoritmi di Machine Learning per la previsione del consumo nell'anno successivo.

Progettazione e ottimizzazione di dispositive elettronici

Algoritmi di Machine Learning applicati alla progettazione di dispositivi (specificamente celle solari) con caratteristiche ottimizzate, utilizzando dati estratti dalla modellizzazione e simulazione fisica.

* Progetti eseguiti da Ariel Cedola

6. Strumenti

Alcuni strumenti

Alcuni strumenti

TensorFlow

Open source software library for high performance numerical computation.

Azure Cognitive Toolkit

Free, open-source, commercial-grade toolkit to train deep learning algorithms optimized for speech.

Pytorch

Scientific computing framework that puts GPUs first.

scikit-learn

Simple and efficient tools for data mining and data analysis

ML.NET

.NET based solution for building Machine Learning models

Onnx

An open format to represent deep learning models.

Caffe2

Lightweight, modular, and scalable deep learning framework.

MxNet

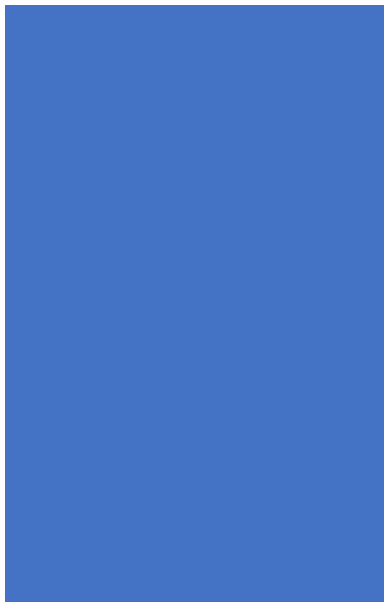
A flexible and efficient library for deep learning.

Chainer

A powerful, flexible, and intuitive framework for neural networks.

https://azure.microsoft.com/en-us/overview/ai-platform/

Alcuni strumenti

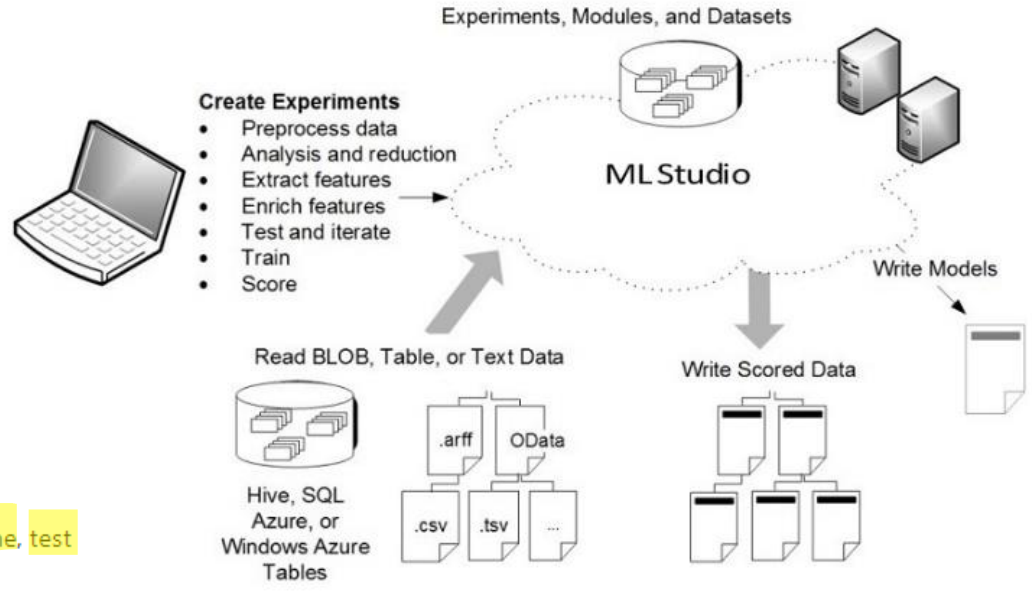


Azure AI

Artificial intelligence productivity for virtually every developer and scenario

Machine Learning Studio

- Sviluppo con trascinamento senza server
- Sperimentazione intuitiva senza codice
- Distribuisci servizi Web in pochi minuti



Cos'è Machine Learning Studio?

Azure Machine Learning Studio offre un'area di lavoro visiva e interattiva per eseguire facilmente le operazioni di compilazione, test e iterazione di un modello di analisi predittiva. È possibile trascinare la selezione di ***set di dati*** e ***moduli*** di analisi in un'area di disegno interattiva, collegandoli tra loro per ottenere un *esperimento* da eseguire in Machine Learning Studio. Per eseguire l'iterazione della progettazione del modello, modificare l'esperimento, salvare eventualmente una copia e ripeterne l'esecuzione. Quando si è pronti, è possibile convertire l'*esperimento di training* in un *esperimento predittivo* e quindi pubblicarlo come un *servizio Web* in modo che altri utenti possano accedere al modello.

Quando utilizzarlo?

Azure Machine Learning Studio è adatta a sperimentare modelli di apprendimento automatico in modo semplice e rapido quando gli algoritmi di apprendimento automatico predefiniti sono sufficienti per le soluzioni dell'utente.

Panoramica delle funzionalità

https://studio.azureml.net

Guest Access Workspace: Free trial access without logging in.
Free Workspace: Free persisted access, no Azure subscription needed.
Standard Workspace: Full access with SLA under an Azure subscription.

Anomaly Detection
- One-class Support Vector Machine
- Principal Component Analysis-based Anomaly Detection
- Time Series Anomaly Detection*

Classification

Two-class Classification
- Averaged Perceptron
- Bayes Point Machine
- Boosted Decision Tree
- Decision Forest
- Decision Jungle
- Logistic Regression
- Neural Network
- Support Vector Machine

Multi-class Classification
- Decision Forest
- Decision Jungle
- Logistic Regression
- Neural Network
- One-vs-all

Clustering
- K-means Clustering

Recommendation
- Matchbox Recommender

Regression
- Bayesian Linear Regression
- Boosted Decision Tree
- Decision Forest
- Fast Forest Quantile Regression
- Linear Regression
- Neural Network Regression
- Ordinal Regression
- Poisson Regression

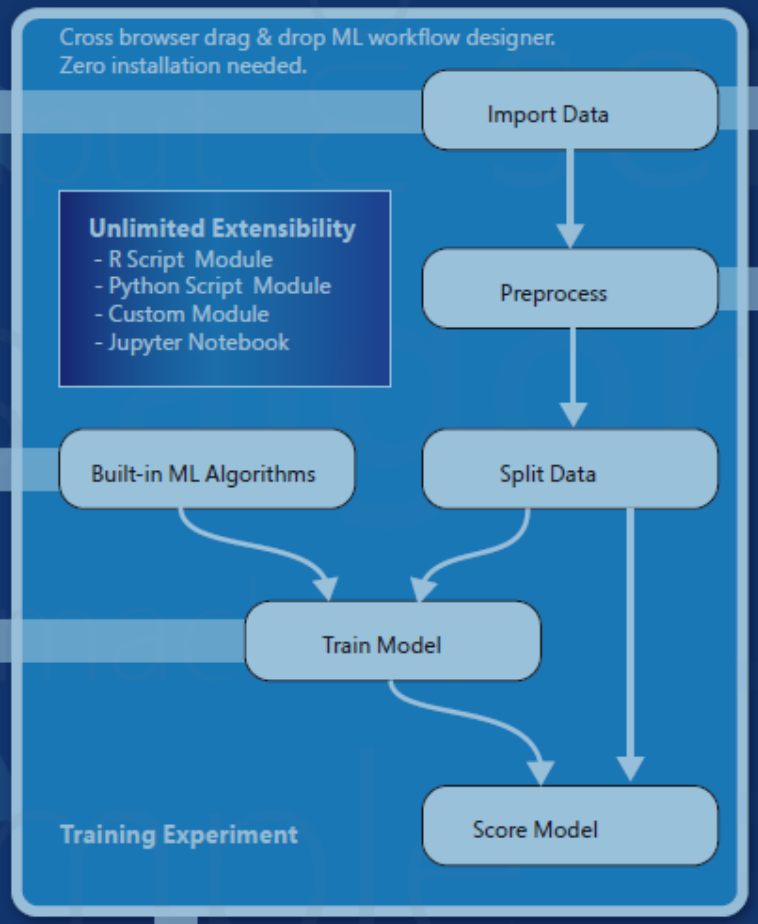
Statistical Functions
- Descriptive Statistics
- Hypothesis Testing T-Test
- Linear Correlation
- Probability Function Evaluation

Text Analytics
- Feature Hashing
- Named Entity Recognition
- Vowpal Wabbit

Computer Vision
- OpenCV Library

Data/Model Visualization
- Scatterplots
- Bar Charts
- Box plots
- Histogram
- R and Python Plotting Libraries
- REPL with Jupyter Notebook
- ROC, Precision/Recall, Lift
- Confusion Matrix
- Decision Tree*

Training
- Cross Validation
- Retraining
- Parameter Sweep



Data Source
- Azure Blob Storage
- Azure SQL DB
- Azure SQL DW*
- Azure Table
- Desktop Direct Upload
- Hadoop Hive Query
- Manual Data Entry
- OData Feed
- On-prem SQL Server*
- Web URL (HTTP)

Data Format
- ARFF
- CSV
- SVMLight
- TSV
- Excel
- ZIP

Data Preparation
- Clean Missing Data
- Clip Outliers
- Edit Metadata
- Feature Selection
- Filter
- Learning with Counts
- Normalize Data
- Partition and Sample
- Principal Component Analysis
- Quantize Data
- SQLite Transformation
- Synthetic Minority Oversampling Technique

Enterprise Grade Cloud Service
- SLA: 99.95% Guaranteed Up-time
- Azure AD Authentication
- Compute at Large Scale
- Multi-geo Availability
- Regulatory Compliance*

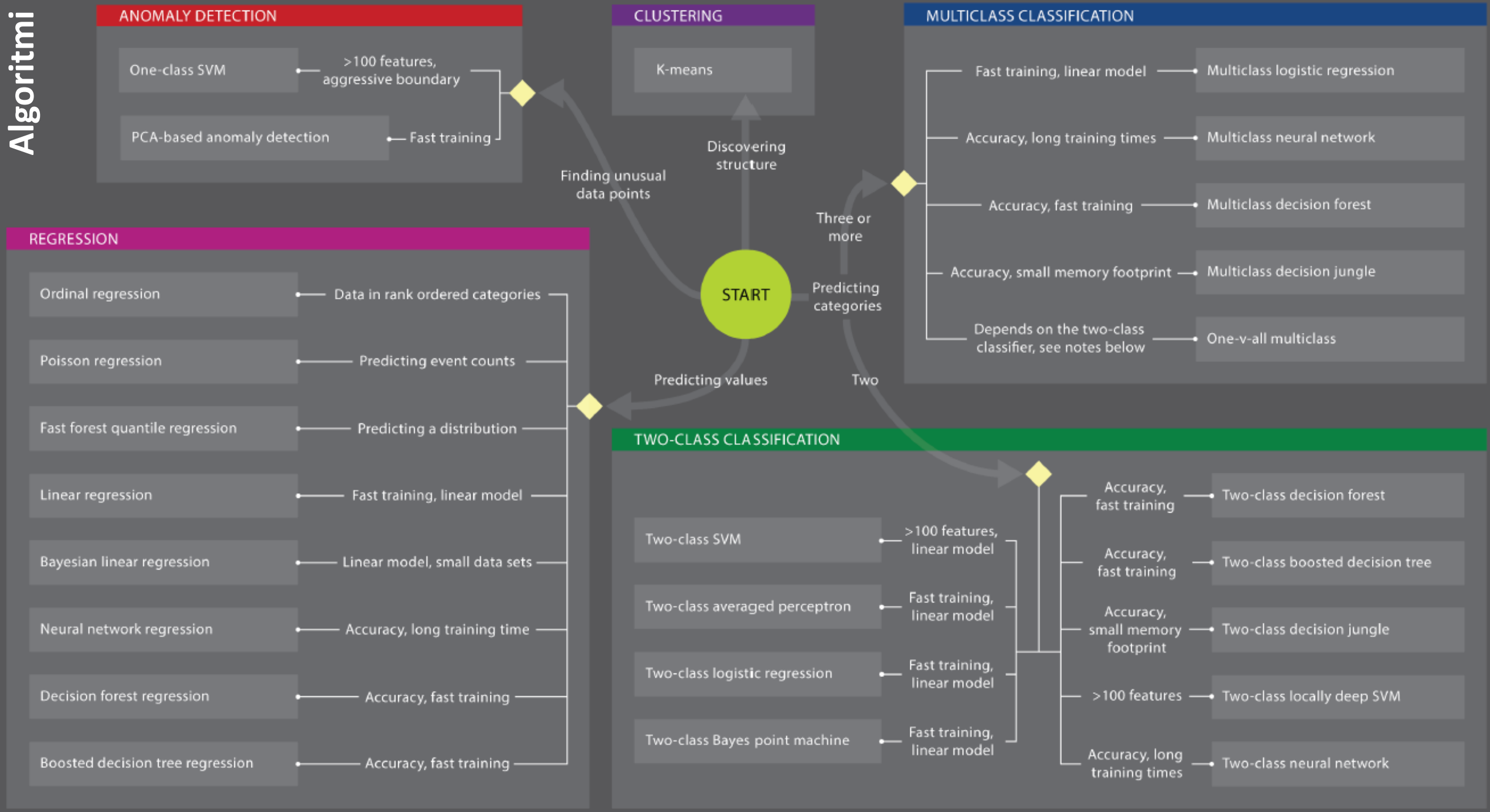
One-click Operationalization

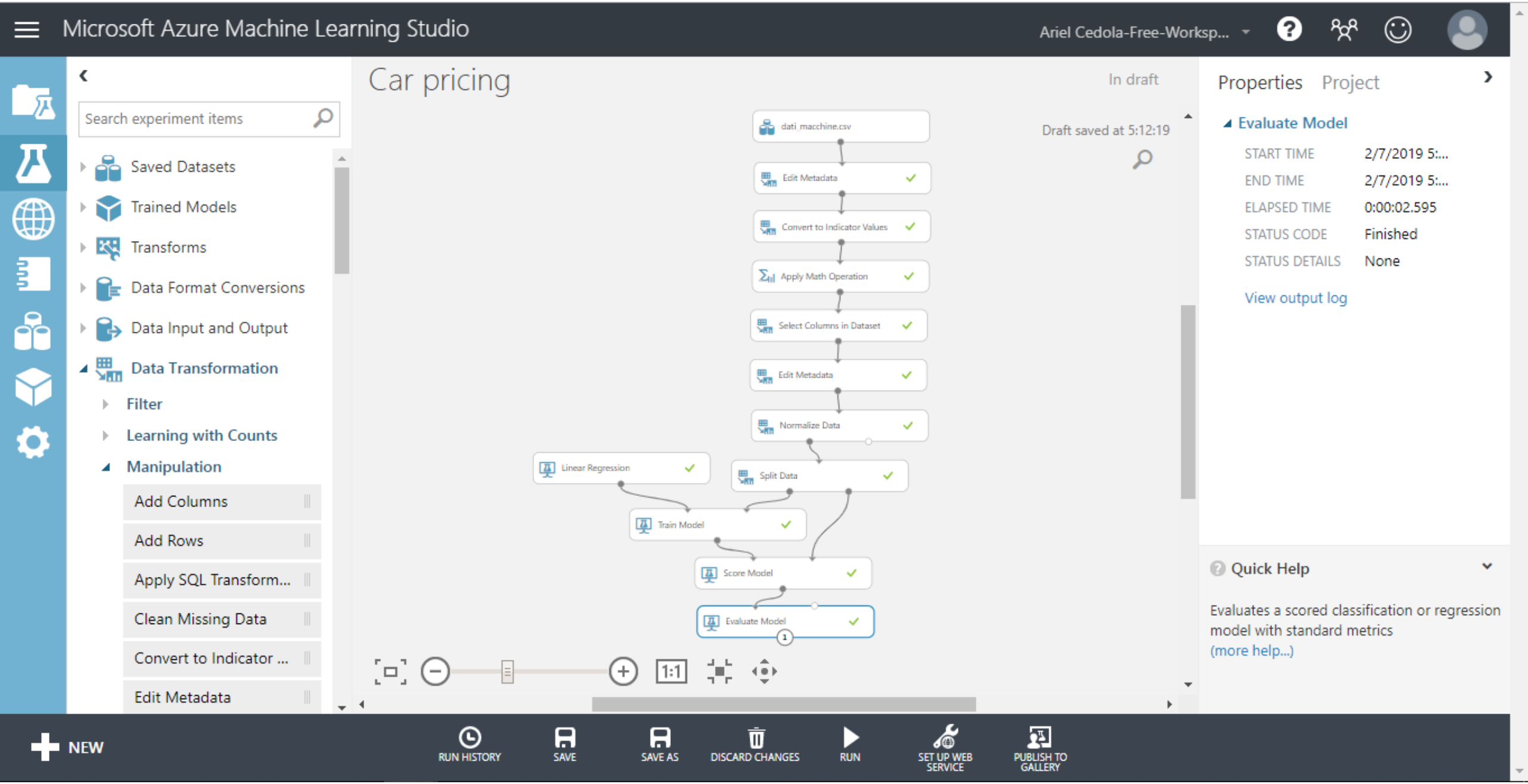
Predictive Experiment

Make Prediction with Elastic APIs
- Request-Response Service (RRS)
- Batch Execution Service (BES)
- Retraining API

Community
- Gallery (http://gallery.azureml.net)
- Samples & Templates
- Workspace Sharing and Collaboration
- Live Chat & MSDN Forum Support

Algoritmi


ANOMALY DETECTION
One-class SVM
>100 features, aggressive boundary
PCA-based anomaly detection
Fast training
CLUSTERING
K-means
MULTICLASS CLASSIFICATION
Fast training, linear model
Multiclass logistic regression
Accuracy, long training times
Multiclass neural network
Accuracy, fast training
Multiclass decision forest
Accuracy, small memory footprint
Multiclass decision jungle
Depends on the two-class classifier, see notes below
One-v-all multiclass
Finding unusual data points
Discovering structure
Three or more
START
Predicting categories
Predicting values
Two
REGRESSION
Ordinal regression
Data in rank ordered categories
Poisson regression
Predicting event counts
Fast forest quantile regression
Predicting a distribution
Linear regression
Fast training, linear model
Bayesian linear regression
Linear model, small data sets
Neural network regression
Accuracy, long training time
Decision forest regression
Accuracy, fast training
Boosted decision tree regression
Accuracy, fast training
TWO-CLASS CLASSIFICATION
Two-class SVM
>100 features, linear model
Two-class averaged perceptron
Fast training, linear model
Two-class logistic regression
Fast training, linear model
Two-class Bayes point machine
Fast training, linear model
Accuracy, fast training
Two-class decision forest
Accuracy, fast training
Two-class boosted decision tree
Accuracy, small memory footprint
Two-class decision jungle
>100 features
Two-class locally deep SVM
Accuracy, long training times
Two-class neural network


Microsoft Azure Machine Learning Studio
Ariel Cedola-Free-Worksp...
Car pricing
In draft
Draft saved at 5:12:19
Search experiment items
Saved Datasets
Trained Models
Transforms
Data Format Conversions
Data Input and Output
Data Transformation
Filter
Learning with Counts
Manipulation
Add Columns
Add Rows
Apply SQL Transform...
Clean Missing Data
Convert to Indicator ...
Edit Metadata
dati_macchine.csv
Edit Metadata
Convert to Indicator Values
Apply Math Operation
Select Columns in Dataset
Edit Metadata
Normalize Data
Linear Regression
Split Data
Train Model
Score Model
Evaluate Model
1:1
Properties
Project
Evaluate Model
START TIME 2/7/2019 5:...
END TIME 2/7/2019 5:...
ELAPSED TIME 0:00:02.595
STATUS CODE Finished
STATUS DETAILS None
View output log
Quick Help
Evaluates a scored classification or regression model with standard metrics
(more help...)
NEW
RUN HISTORY
SAVE
SAVE AS
DISCARD CHANGES
RUN
SET UP WEB SERVICE
PUBLISH TO GALLERY

7. Lab 1: Identificare i dati disponibili da cui estrarre valore

Riflettere su un progetto di Data Analysis/Data Science

- Scegliere il settore
- Identificare l’obbietivo, cosa vorrei analizzare o predire?
- Determinare i dati necesari, sono a disposizione?
- Raggionare su come eseguire il processo (strumenti preferiti, metodi, ecc)

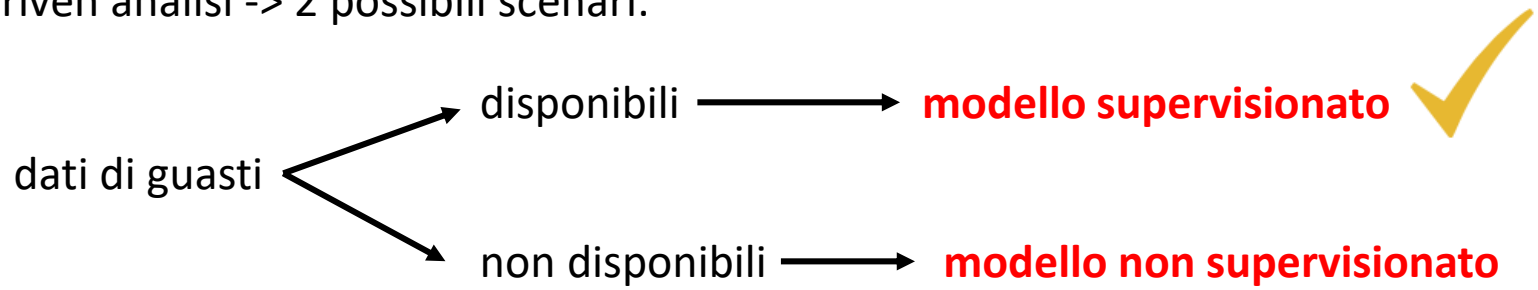
8. Lab 2: Analisi basico con Excel

9. Lab 3: Analisi predittiva

Caso di studio: Dati di telemetria di macchine

Alcune considerazioni:

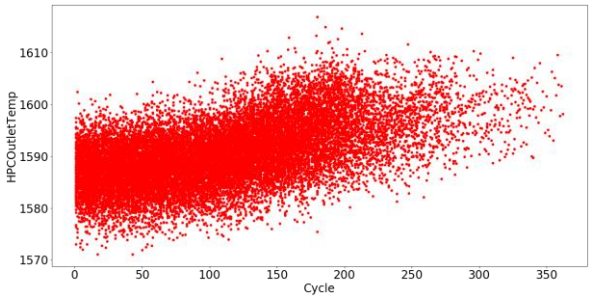
- Telemetria: Variabili in gran numero, dati complesse, impossibile trovare un pattern senza **Machine Learning**
- Data-driven analisi -> 2 possibili scenari:

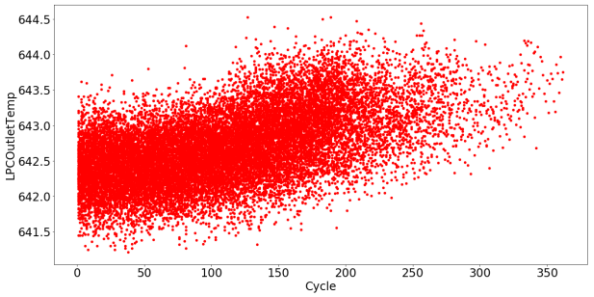


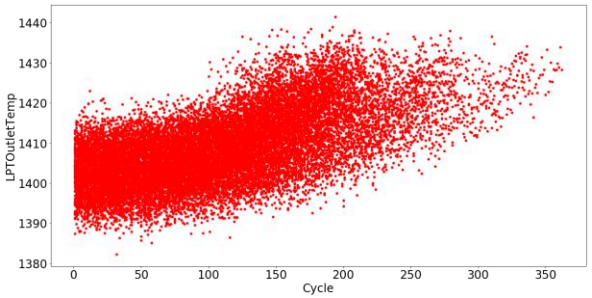
- È molto difficile trovare dati per dimostrare o valutare la qualità di un modello di Machine Learning per la manutenzione predittiva -> **simulazione**
- **Caso di studio**:
  - dati simulati: 7 sensori (temperatura, velocità, pressione, ecc)
  - 120 macchine
  - dati versus n. ciclo fino al fallimento

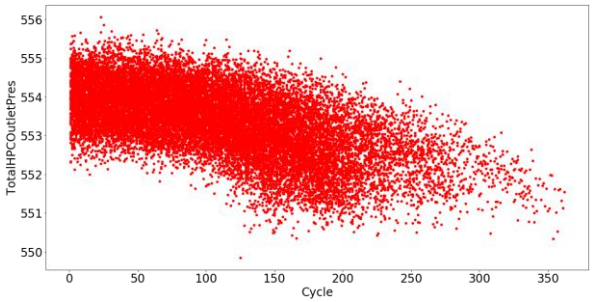
Applicazione alla Manutenzione predittiva

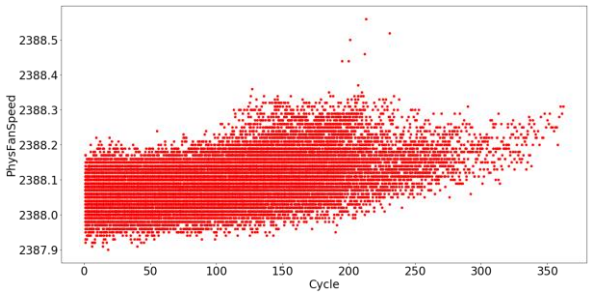
Esempio: Simulazione dati di telemetria

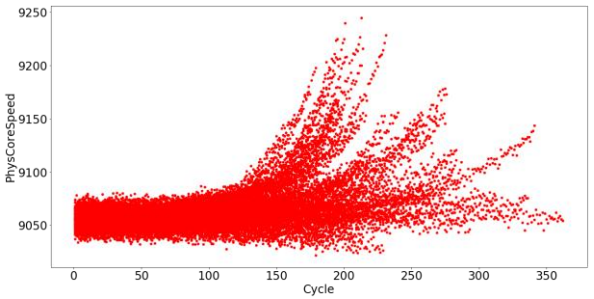


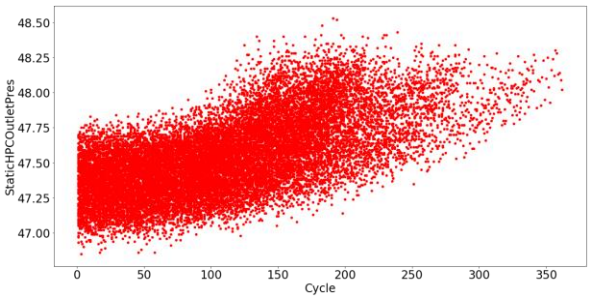


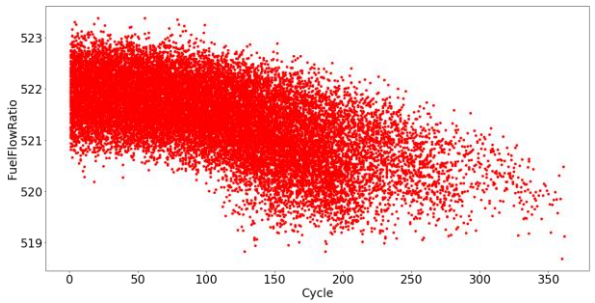


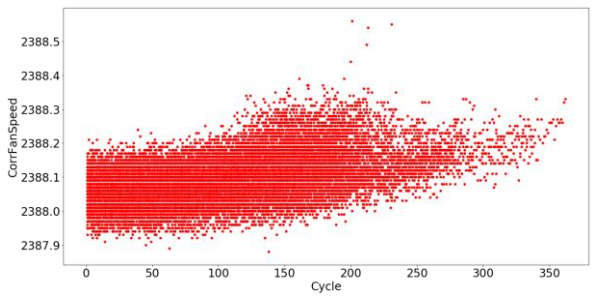


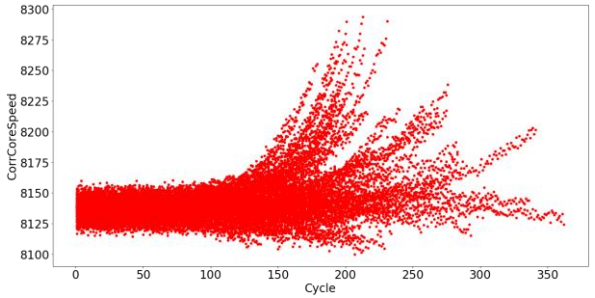


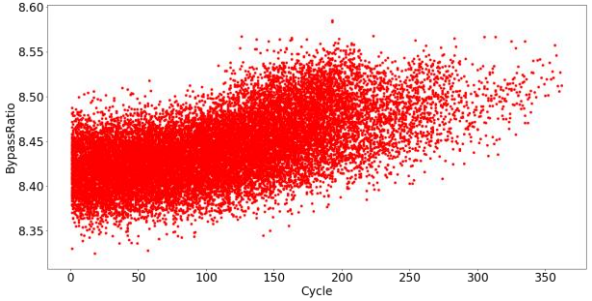


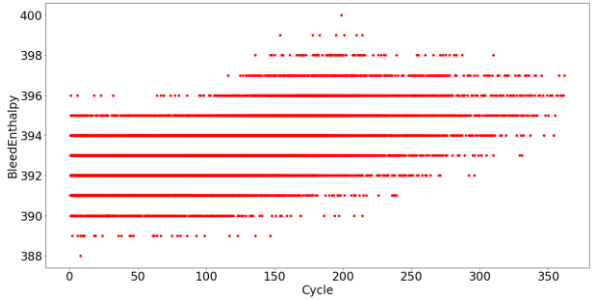












Dati

N. macchina

Ciclo di lavoro

7 features

Etichetta: 2 classi

ID_macchina	cycle	s1	s2	s3	s4	s5	s6	s7	TTF
1	1	641.29858442562	554.02089125960936	47.459055320542369	383	1393.5009321684715	39.680307010852218	8129.2682163045165	long
1	2	641.82611449314106	554.10393426885048	47.143694456083033	389	1384.8209493964694	39.524294855373263	8129.4994533488489	long
1	3	641.72354582532819	554.31148511662445	47.291815212358678	387	1392.2295643452298	39.559548426839548	8127.2567367128577	long
1	4	641.63278399134037	554.02947011918127	46.668360381600309	387	1392.5669835134211	39.417021487362859	8128.8773468733016	long
1	5	641.91618591623762	553.88653085789485	47.06424651884376	388	1395.7082859296627	39.264181377776687	8129.3389653306394	long
1	6	641.64423208664925	553.48584874507037	47.252416964735687	387	1383.6049928120624	39.6462772656237	8130.2847131247463	long
1	7	641.20059681516113	553.80150122022474	47.03141207215355	387	1389.8320913965231	39.553963135376328	8127.67418490626	long
1	8	641.82730223802037	554.00813335551175	46.811837156603538	390	1377.1111340218331	39.299037628410211	8129.6417577346874	long
1	9	641.906375790902	554.01981519095659	47.189924876456168	389	1386.1794441055927	39.463608741516161	8125.4365331480967	long
1	10	641.38287483940871	553.90265137195	46.9624408937961	386	1400.5214018943705	39.390020868412336	8127.9557459939606	long
1	11	641.95132177175992	553.8650005058181	46.8957832025357	387	1379.9142696697309	39.489306175978811	8128.2465646320461	long
1	12	642.18091490680945	553.96575702340874	47.50541225059262	390	1390.9774532839178	39.463109044634813	8128.7179580879256	long
1	13	641.08279532002973	553.88930064944543	47.369606392966375	386	1388.0650011713744	39.52228939840127	8127.4143774726354	long
1	14	641.55087332764981	553.55513942862922	47.272908808645909	388	1389.8189583199219	39.385814444437877	8124.4833065112816	long
1	15	641.48739396820542	554.14558743078874	47.198306336346981	388	1390.3828971272897	39.408578730003406	8124.6075169308251	long
1	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....
1	167	643.51261369811516	552.42406246932808	47.716441012360889	396	1440.4740108913506	38.73563151045942	8190.1887076339508	short
1	168	643.87056191135127	552.434595872027	47.709093048562778	401	1433.9827334541965	38.650046661051888	8191.1861016488292	short
1	169	643.23704131654631	552.4544604920643	48.069929650851776	396	1427.3035587017137	38.623188454969693	8193.37174085445	short
1	170	643.85082411604185	551.71364876893279	48.195195712399112	403	1440.4740524285826	38.52855977059604	8196.2511804068727	short
1	171	643.48235654977782	552.45612983417266	48.011135876492261	399	1430.7425344561716	38.616729590675767	8199.8268392256177	short
2	1	641.23534003787938	554.50094253932218	47.103028231014335	390	1368.7572789859591	39.439287988991424	8148.6382784299785	long
2	2	641.57671718743734	554.37029578300314	47.199347360758104	392	1387.7520539323375	39.046030161170371	8147.8273779481588	long
2	3	642.01074103074166	554.2140791790797	47.107910505749828	390	1383.5038134473962	39.217800261972563	8150.8075948129745	long
2	4	642.40114030592986	554.37150375143278	47.098556344956883	389	1386.946288317786	39.042801638736748	8148.06446753053	long
2	5	642.22366835330013	554.56477878517342	47.308179353062577	392	1397.665546467437	39.300046267923186	8149.0871807221592	long
.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....
120	173	642.92546630563982	553.11074104328793	47.734538985765461	395	1409.1075450190958	38.758709474610264	8083.883484409268	short
120	174	643.03819481224753	552.74221378785239	47.73513907217081	395	1409.0381690424226	38.715144601534519	8085.0597536322175	short
120	175	642.52138217487243	553.02454827228757	47.59190829004261	397	1427.0763603492305	38.677510826810185	8083.645116568815	short
120	176	643.4881495080881	552.312948320938	47.6483167476248	397	1421.618436944678	38.72689851443414	8081.5276522046643	short
120	177	643.76483558245116	551.95505981787551	47.817768731119855	395	1426.721688315446	38.695753948692968	8084.2020338112616	short

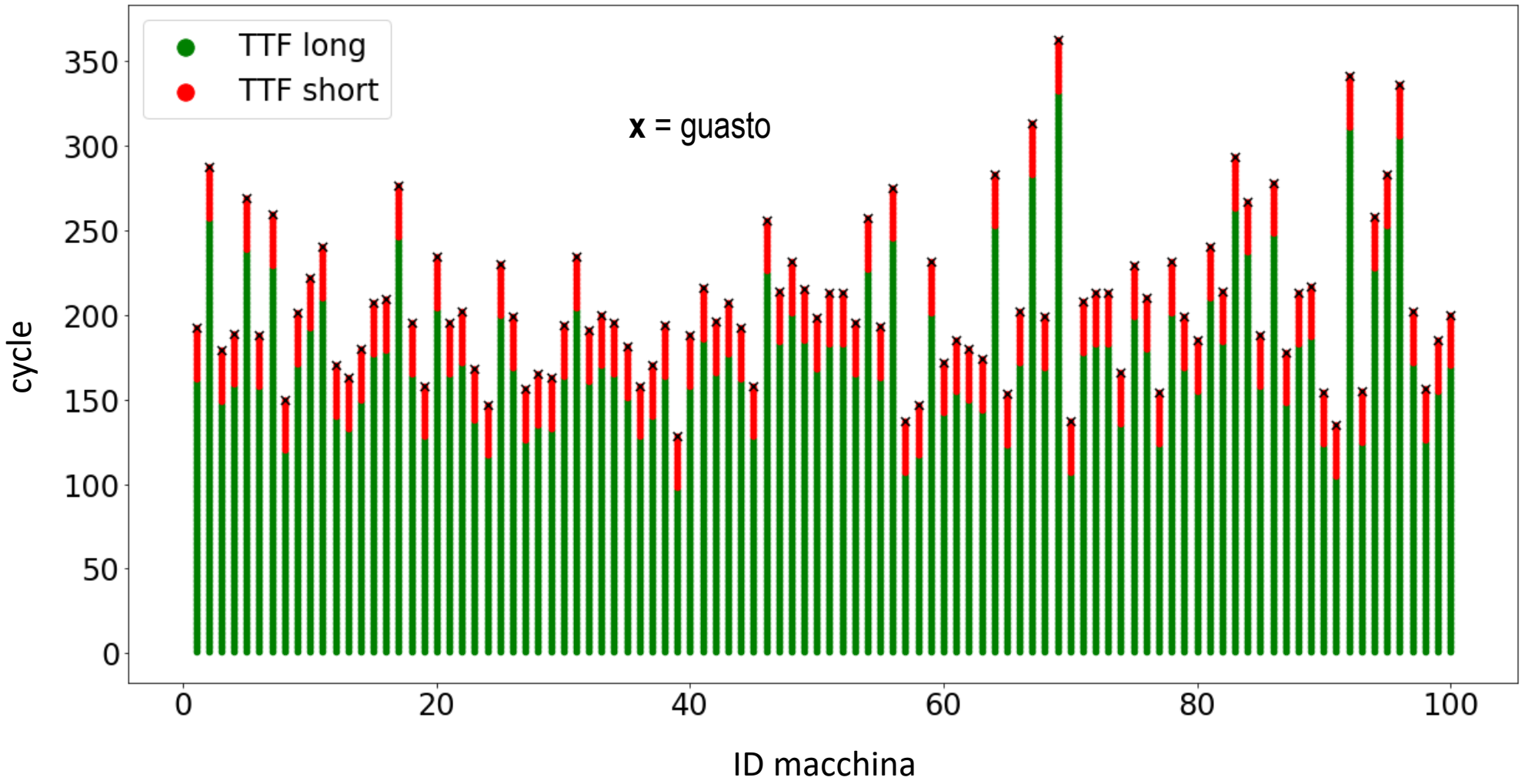
guasto

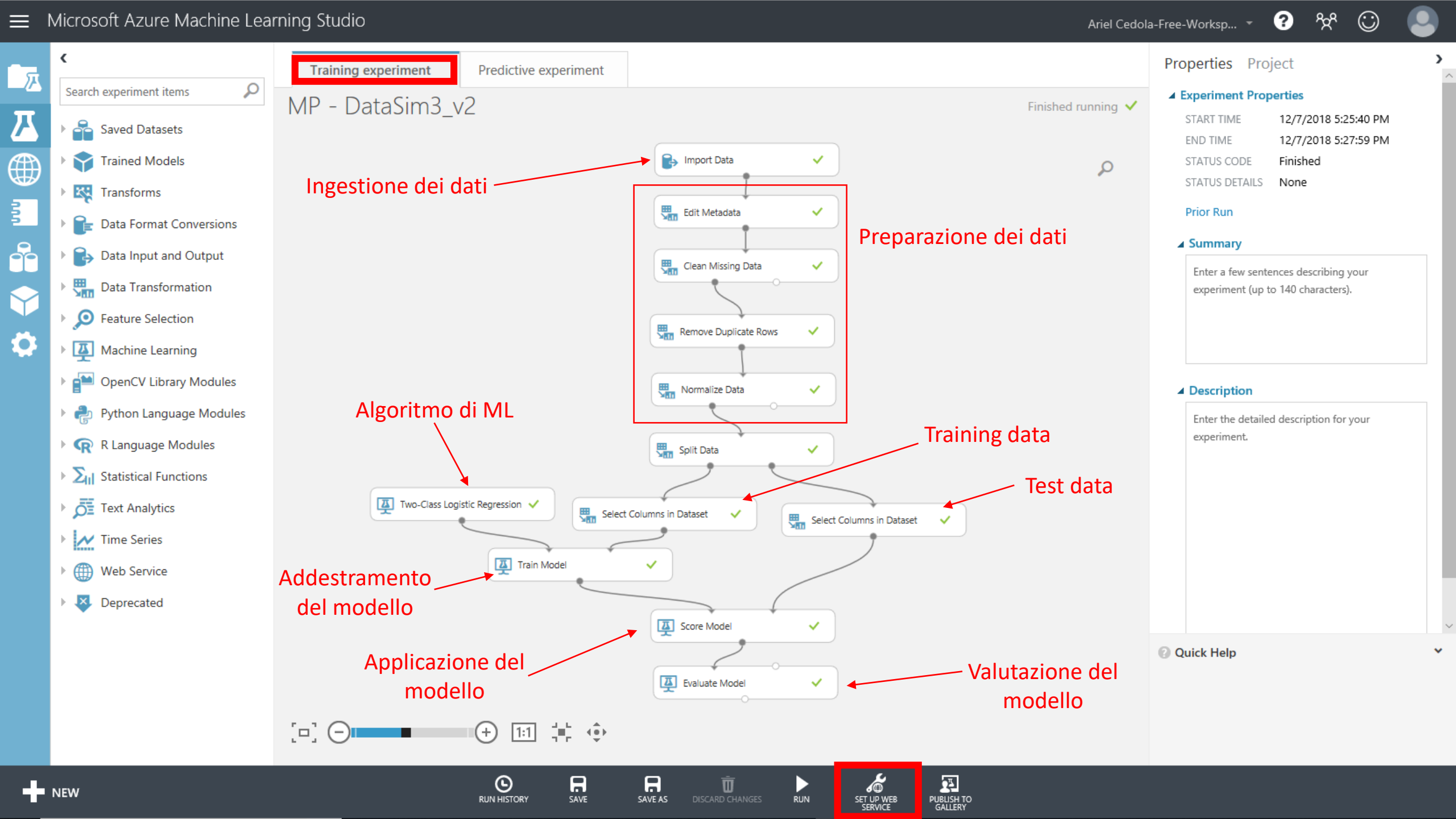
Dati

Time-to-failure (TTF)
- > 10 cicli -> **long**
- <= 10 cicli -> **short**

Training
ID 1-100

Test
ID 101-120




Microsoft Azure Machine Learning Studio
Ariel Cedola-Free-Worksp...
Search experiment items
Saved Datasets
Trained Models
Transforms
Data Format Conversions
Data Input and Output
Data Transformation
Feature Selection
Machine Learning
OpenCV Library Modules
Python Language Modules
R Language Modules
Statistical Functions
Text Analytics
Time Series
Web Service
Deprecated
Training experiment
Predictive experiment
MP - DataSim3_v2
Finished running
Import Data
Edit Metadata
Clean Missing Data
Remove Duplicate Rows
Normalize Data
Split Data
Two-Class Logistic Regression
Select Columns in Dataset
Select Columns in Dataset
Train Model
Score Model
Evaluate Model
Ingestione dei dati
Preparazione dei dati
Algoritmo di ML
Training data
Test data
Addestramento del modello
Applicazione del modello
Valutazione del modello
Properties
Project
Experiment Properties
START TIME 12/7/2018 5:25:40 PM
END TIME 12/7/2018 5:27:59 PM
STATUS CODE Finished
STATUS DETAILS None
Prior Run
Summary
Enter a few sentences describing your experiment (up to 140 characters).
Description
Enter the detailed description for your experiment.
Quick Help
NEW
RUN HISTORY
SAVE
SAVE AS
DISCARD CHANGES
RUN
SET UP WEB SERVICE
PUBLISH TO GALLERY

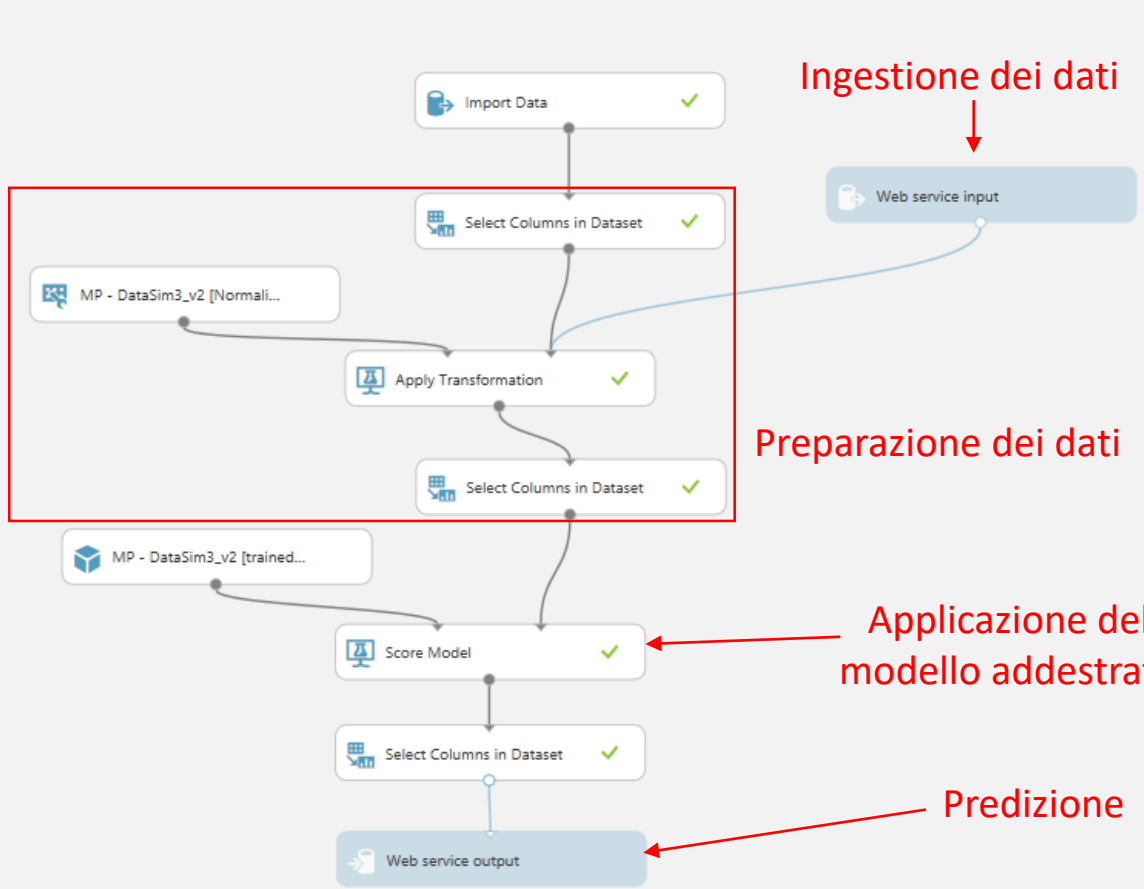
Search experiment items

- Saved Datasets
- Trained Models
- Transforms
- Data Format Conversions
- Data Input and Output
- Data Transformation
- Feature Selection
- Machine Learning
- OpenCV Library Modules
- Python Language Modules
- R Language Modules
- Statistical Functions
- Text Analytics
- Time Series
- Web Service
- Deprecated

Training experiment | Predictive experiment

MP - DataSim3_v2 [Predictive Exp.]

Finished running



Properties | Project

Experiment Properties

START TIME	12/7/2018 5:50:47 PM
END TIME	12/7/2018 5:52:08 PM
STATUS CODE	Finished
STATUS DETAILS	None

Go to web service

Prior Run

Summary

Enter a few sentences describing your experiment (up to 140 characters).

Description

Enter the detailed description for your experiment.

Quick Help

NEW | RUN HISTORY | SAVE | SAVE AS | DISCARD CHANGES | RUN | DEPLOY WEB SERVICE | PUBLISH TO GALLERY

mp - datasim3_v2 [predictive exp.]

DASHBOARD CONFIGURATION

General New Web Services Experience preview

Published experiment

View snapshot View latest

Description

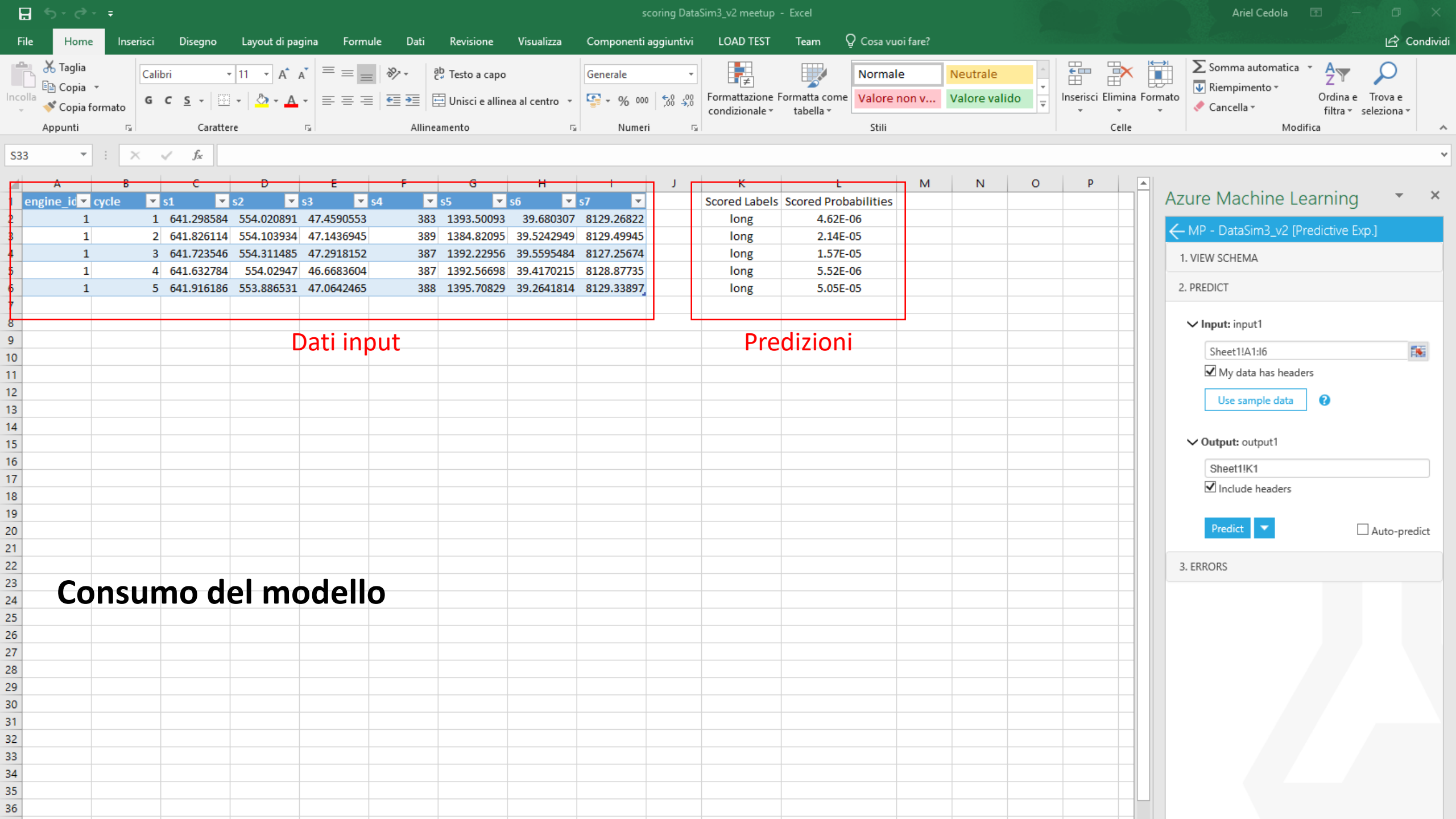
No description provided for this web service.

API key

HsOboJhKVzbXvN6ALsVK528sT/ImJvu1XRZ7i0EyYliQ7MdPmo5Gj+YMkVP6AIvGOmrItyCDApVkGyD7LBnBZw==

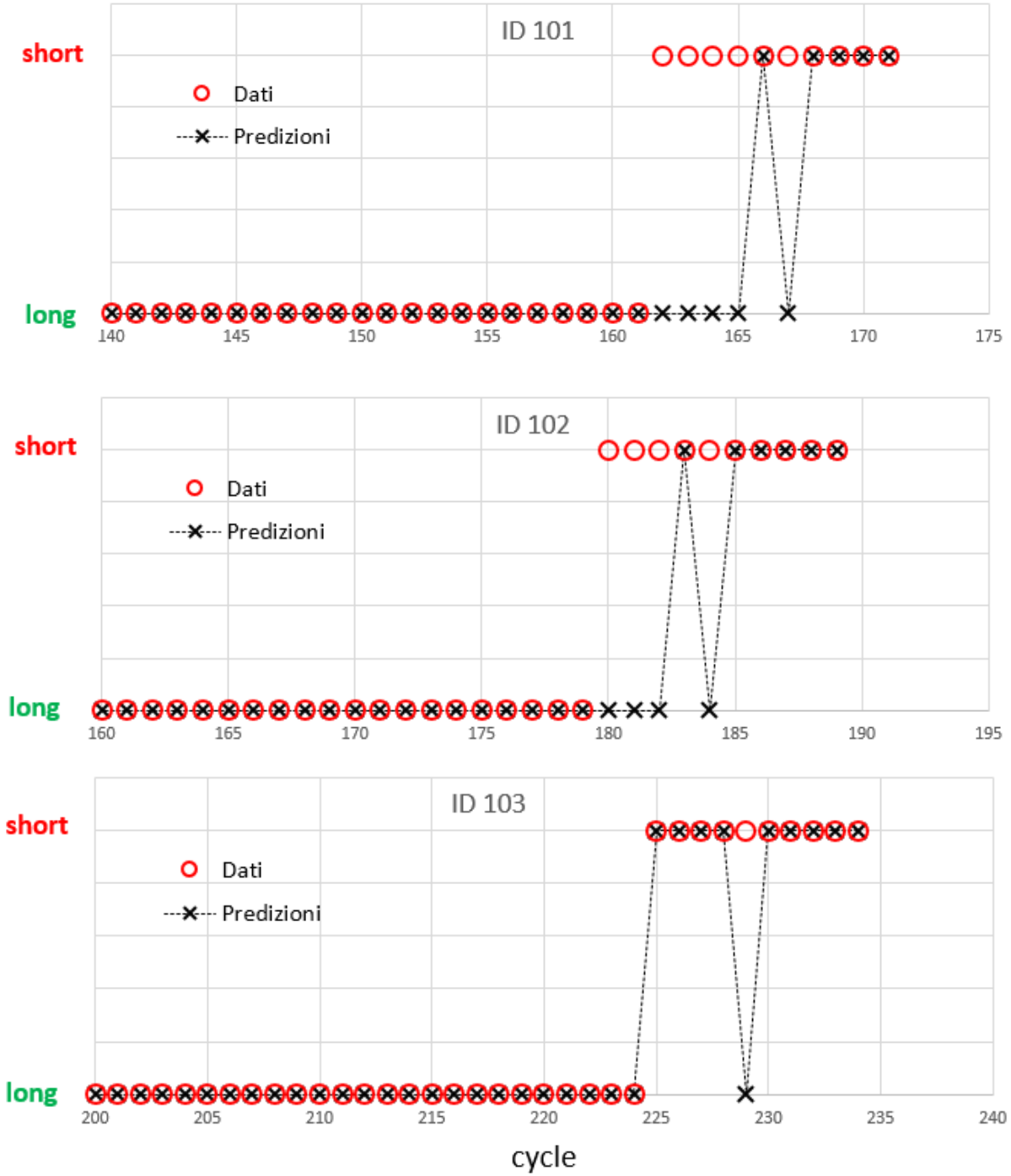
Default Endpoint

API HELP PAGE	TEST	APPS	LAST UPDATED
REQUEST/RESPONSE	Test Test preview	Excel 2013 or later | Excel 2010 or earlier workbook	12/7/2018 5:52:56 PM
BATCH EXECUTION	Test preview	Excel 2013 or later workbook	12/7/2018 5:52:56 PM

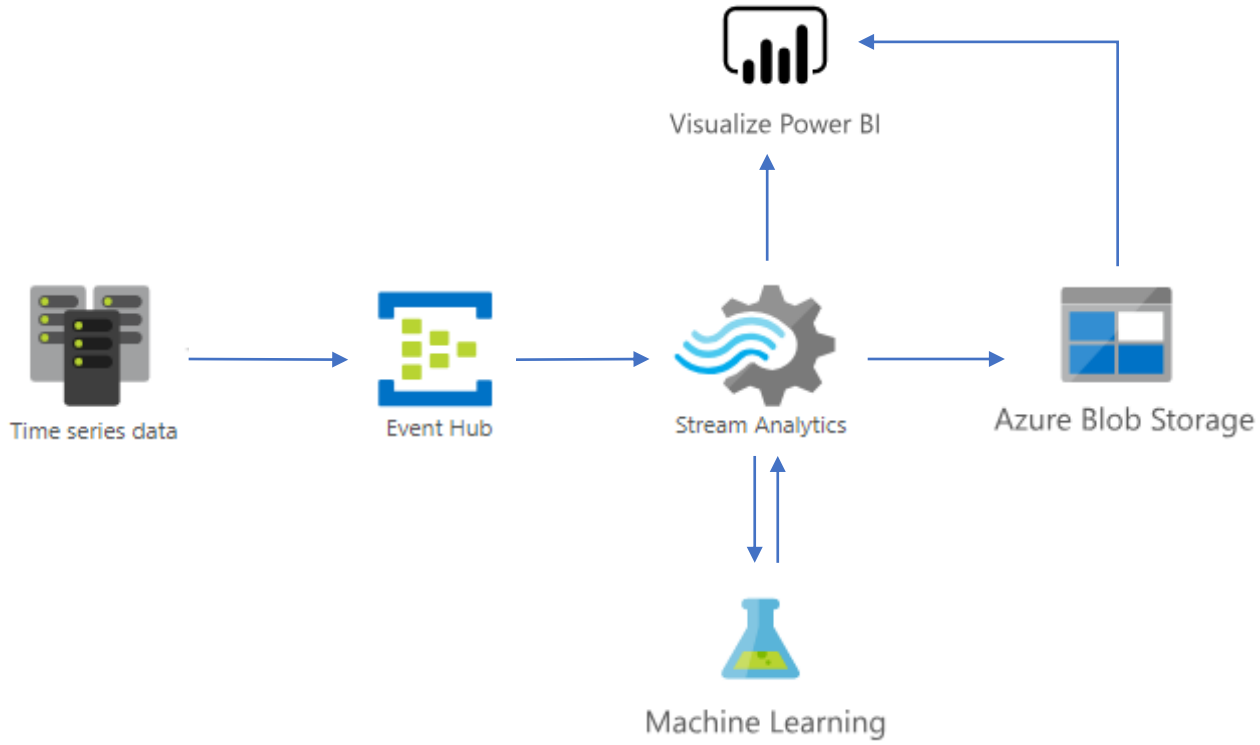

scoring DataSim3_v2 meetup - Excel
Ariel Cedola
File Home Inserisci Disegno Layout di pagina Formule Dati Revisione Visualizza Componenti aggiuntivi LOAD TEST Team Cosa vuoi fare? Condividi
Taglia Copia Copia formato Incolla
Calibri 11
Testo a capo
Unisci e allinea al centro
Generale
Formattazione condizionale
Formatta come tabella
Normale
Neutrale
Valore non v...
Valore valido
Inserisci Elimina Formato
Somma automatica
Riempimento
Cancella
Ordina e filtra
Trova e seleziona
Appunti Carattere Allineamento Numeri Stili Celle Modifica
S33
engine_id cycle s1 s2 s3 s4 s5 s6 s7
1 1 641.298584 554.020891 47.4590553 383 1393.50093 39.680307 8129.26822
1 2 641.826114 554.103934 47.1436945 389 1384.82095 39.5242949 8129.49945
1 3 641.723546 554.311485 47.2918152 387 1392.22956 39.5595484 8127.25674
1 4 641.632784 554.02947 46.6683604 387 1392.56698 39.4170215 8128.87735
1 5 641.916186 553.886531 47.0642465 388 1395.70829 39.2641814 8129.33897
Scored Labels Scored Probabilities
long 4.62E-06
long 2.14E-05
long 1.57E-05
long 5.52E-06
long 5.05E-05
Dati input
Predizioni
Consumo del modello
Azure Machine Learning
MP - DataSim3_v2 [Predictive Exp.]
1. VIEW SCHEMA
2. PREDICT
Input: input1
Sheet1!A1:I6
My data has headers
Use sample data
Output: output1
Sheet1!K1
Include headers
Predict
Auto-predict
3. ERRORS

Predizioni del modello

Esempio 3 macchine



Integrazione con altri servizi di Azure



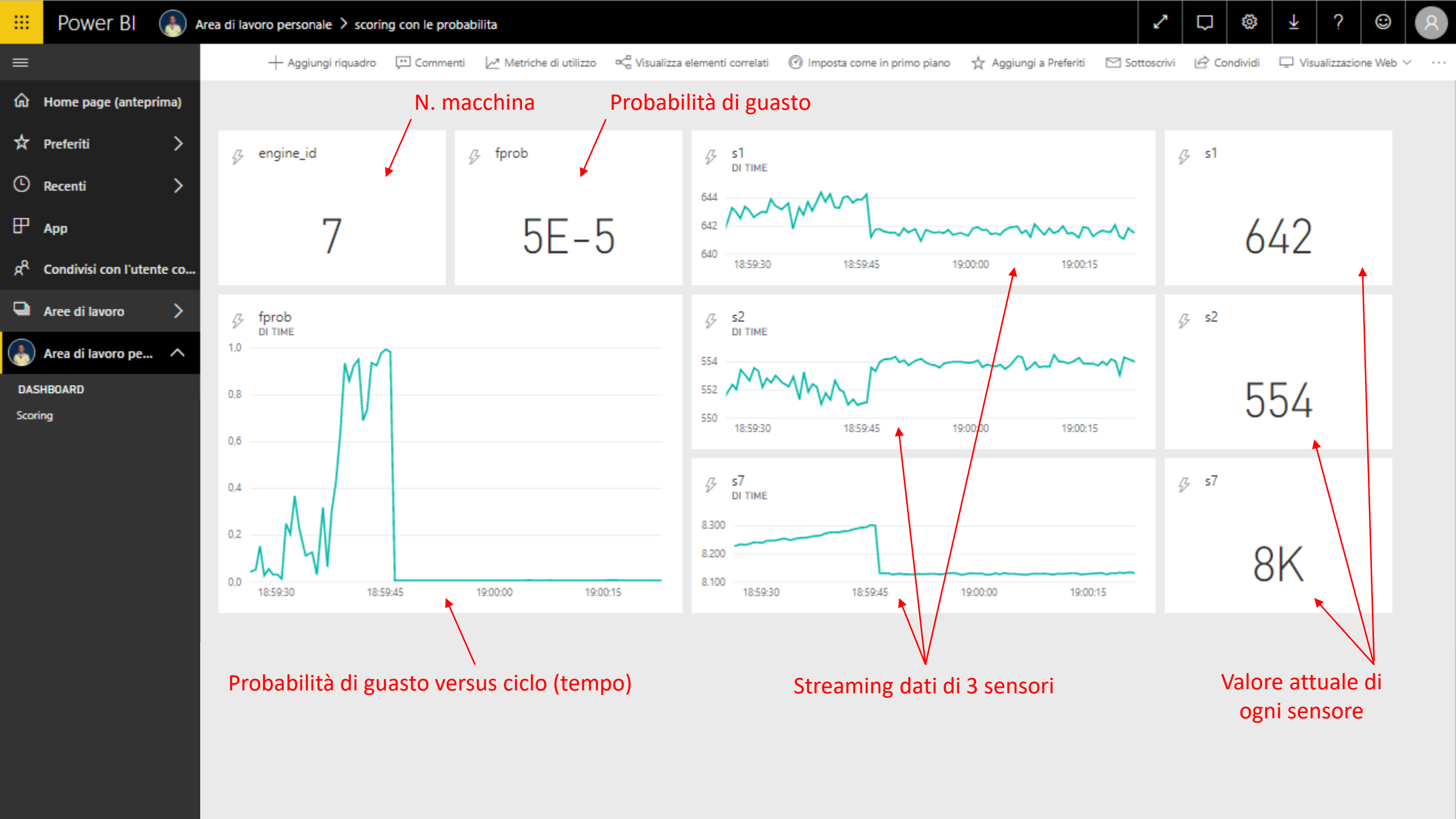
data_simulator_streaming.py | data_simulator_streaming_no_labels.py | data_simulator_streaming_no_labels 2.py ×

```
    noise_amplitude = 0.05*np.absolute(max_ciclo1 - min_ciclo1)
    matrix_features[:, i+2] = feature_dual(min_ciclo1, max_ciclo1, 0.2, N, N*0.2, noise_amplitude)


    # DATAFRAME
    df = pd.DataFrame(matrix_features)
    df[[0, 1, N_features + 2, N_features + 3]] = df[[0, 1, N_features + 2, N_features + 3]].astype('int32')

    dataset = dataset.append(df, ignore_index=True)

# REPLACE TTF LABELS
dataset.loc[:, N_features + 3].replace({0: 'long', 1: 'short'}, inplace=True)
header = ['engine_ID', 'cycle', 's1', 's2', 's3', 's4', 's5', 's6', 's7', 'RUL', 'TTF']
dataset.columns = header

# SAVE TO FILE
date_time_now = datetime.datetime.now().strftime("%d_%m_%Y_%H_%M_%S")
filename = 'C:/Users/Ariel/proyects/iDaq_Analytics/data_sim_' + date_time_now + '.csv'

# INFORMATION ABOUT THE CREATED EVENT HUBS NAMESPACE AND EVENT HUB
ADDRESS = 'amqps://eventhub1-idaq-test6.servicebus.windows.net/eventhub1'
USER = 'RootManageSharedAccessKey'
KEY = 'Myot11X7/Vu+CnRew/bXqaK+368vd18JEijs56biit8=' # Primary key

client = EventHubClient(ADDRESS, debug=False, username=USER, password=KEY)
sender = client.add_sender(partition="0")
client.run()

dataset_dict = dataset.iloc[:, 0:9].astype(object).to_dict(orient='records', into=OrderedDict)

for i in range(dataset.shape[0]):

    datax = json.dumps(dataset_dict[i])
    print(datax)
    time.sleep(0.1)

    sender.send(EventData(datax))

client.stop()
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL

2: Python

```
{"engine_ID": 10, "cycle": 121, "s1": 642.4245893547686, "s2": 552.2021039501753, "s3": 47.4381980150669, "s4": 390.0, "s5": 1391.3538192484948, "s6": 38.774215408115616, "s7": 8170.377157152344}
{"engine_ID": 10, "cycle": 122, "s1": 642.2530166117457, "s2": 553.708555172974, "s3": 47.31007707974808, "s4": 392.0, "s5": 1392.5159763246788, "s6": 39.20003840456184, "s7": 8169.343907693955}
{"engine_ID": 10, "cycle": 123, "s1": 642.3697140991245, "s2": 552.8828773823456, "s3": 47.242218282621614, "s4": 392.0, "s5": 1399.1099154471842, "s6": 38.77339177934689, "s7": 8167.888492203391}
{"engine_ID": 10, "cycle": 124, "s1": 642.5709701158892, "s2": 554.0247122579292, "s3": 47.16812931094532, "s4": 393.0, "s5": 1395.5082338200777, "s6": 39.225608346866, "s7": 8170.2361001031695}
{"engine_ID": 10, "cycle": 125, "s1": 642.089085009138, "s2": 553.1413952076717, "s3": 47.436448608476276, "s4": 392.0, "s5": 1404.954984452471, "s6": 38.78267635441342, "s7": 8175.44181226982}
{"engine_ID": 10, "cycle": 126, "s1": 642.7228298852547, "s2": 553.3174642811653, "s3": 47.536199087945185, "s4": 390.0, "s5": 1390.283076683137, "s6": 38.888069461696034, "s7": 8169.478652777631}
{"engine_ID": 10, "cycle": 127, "s1": 642.5266448173201, "s2": 552.5764489756208, "s3": 47.65289787694415, "s4": 393.0, "s5": 1392.9325089729684, "s6": 38.842909226646185, "s7": 8174.71992661619}
{"engine_ID": 10, "cycle": 128, "s1": 642.611101727085, "s2": 552.6060970780962, "s3": 47.45074912004954, "s4": 393.0, "s5": 1393.9899349144825, "s6": 39.0124747105222, "s7": 8172.268791710128}
{"engine_ID": 10, "cycle": 129, "s1": 642.719671301534, "s2": 552.5193812853728, "s3": 47.2315103327645, "s4": 392.0, "s5": 1399.7821325612451, "s6": 38.80831102380915, "s7": 8175.155198802237}
{"engine_ID": 10, "cycle": 130, "s1": 642.4447905655732, "s2": 552.6645364245795, "s3": 47.15250905629466, "s4": 391.0, "s5": 1401.7810919732149, "s6": 38.912591831651646, "s7": 8179.764819433443}
{"engine_ID": 10, "cycle": 131, "s1": 642.7414939908755, "s2": 553.484161873283, "s3": 47.36431312063939, "s4": 392.0, "s5": 1394.6868265461785, "s6": 38.9513754918869, "s7": 8176.035305649457}
{"engine_ID": 10, "cycle": 132, "s1": 642.8422092571269, "s2": 552.4976754915501, "s3": 47.372070641656336, "s4": 392.0, "s5": 1399.1009201605539, "s6": 38.900244880705145, "s7": 8179.076956903724}
{"engine_ID": 10, "cycle": 133, "s1": 642.8104217094929, "s2": 552.5518065392056, "s3": 47.45290482428114, "s4": 392.0, "s5": 1395.7767235395645, "s6": 39.0853831524983, "s7": 8174.513800002118}
{"engine_ID": 10, "cycle": 134, "s1": 642.8644689684527, "s2": 552.5957140306571, "s3": 47.293033405846785, "s4": 393.0, "s5": 1397.3751629897022, "s6": 38.78223966853994, "s7": 8182.7450713733015}
{"engine_ID": 10, "cycle": 135, "s1": 642.946594242107, "s2": 552.6708361008542, "s3": 47.373976993357054, "s4": 393.0, "s5": 1408.7342205665657, "s6": 38.61874966619529, "s7": 8175.9281246775345}
{"engine_ID": 10, "cycle": 136, "s1": 643.1342779200562, "s2": 552.4658417121817, "s3": 47.499381547683896, "s4": 393.0, "s5": 1402.54301025011, "s6": 38.846916234765004, "s7": 8184.348893680998}
{"engine_ID": 10, "cycle": 137, "s1": 642.9866678583081, "s2": 552.6428844931146, "s3": 47.48350228807191, "s4": 393.0, "s5": 1407.748262497885, "s6": 38.809881732937185, "s7": 8184.07504177756}
```


Power BI
Area di lavoro personale > scoring con le probabilita
Aggiungi riquadro
Commenti
Metriche di utilizzo
Visualizza elementi correlati
Imposta come in primo piano
Aggiungi a Preferiti
Sottoscrivi
Condividi
Visualizzazione Web
Home page (anteprima)
Preferiti
Recenti
App
Condivisi con l'utente co...
Aree di lavoro
Area di lavoro pe...
DASHBOARD
Scoring
N. macchina
Probabilità di guasto
engine_id
7
fprob
5E-5
s1
DI TIME
644
642
640
18:59:30
18:59:45
19:00:00
19:00:15
s1
642
fprob
DI TIME
1,0
0,8
0,6
0,4
0,2
0,0
18:59:30
18:59:45
19:00:00
19:00:15
s2
DI TIME
554
552
550
18:59:30
18:59:45
19:00:00
19:00:15
s2
554
s7
DI TIME
8.300
8.200
8.100
18:59:30
18:59:45
19:00:00
19:00:15
s7
8K
Probabilità di guasto versus ciclo (tempo)
Streaming dati di 3 sensori
Valore attuale di ogni sensore

Referenze

- MAKING SENSE OF DATA I - A Practical Guide to Exploratory Data Analysis and Data Mining; G. Myatt & W. Johnson, Wiley 2014
- Introduction to Data Science. A Python approach to concepts, techniques and applications; L. Igual & S. Seguí, Springer 2017
- Introducing Data Science, Big Data, machine learning and more, using Python tools; D. Cielen, A. Meysman, M. Ali, Manning 2016

FINE

Contatto:

Dott. Ariel **Cedola**
acedola@welol.it