UNIVERSITÀ DEGLI STUDI DI UDINE
hic sunt futura

La Scienza dei Dati

Eddy Maddalena

Dipartimento di Matematica, Informatica e Fisica (DMIF), Università di Udine

Liceo Scientifico Bertoni, Udine, 18 aprile 2023

UNIVERSITÀ DEGLI STUDI DI UDINE
hic sunt futura

Eddy Maddalena



Contatti:

- eddy.maddalena@uniud.it
- eddymaddalena.net

Docente:

- Fondamenti di Scienza dei dati e laboratorio (triennale)
- Progettazione e Analsi orientata agli Oggetti (magistrale)

Ricercatore:

- Crowdsourcing, Information Retrieval, Human Computation, Gamification, Citizen Science, Smart Cities, ..

Scaletta

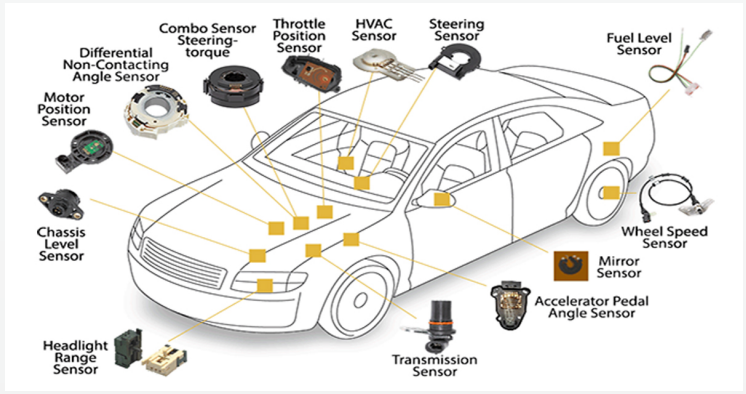
- Presentazioni - 2 minuti
- Panoramica sulla Data Science - 30 minuti
- Panoramica su R ed RStudio - 18 minuti

I dati

> "Un dato è una descrizione elementare codificata di un'informazione, un'entità, di un fenomeno, di una transazione, di un avvenimento o di altro."
> Wikipedia

> "I dati sono informazione fattuale (come misuarazioni o statistiche), usuate come base per ragionamenti, discussioni e calcoli"
> Joe Martin

I dati digitali



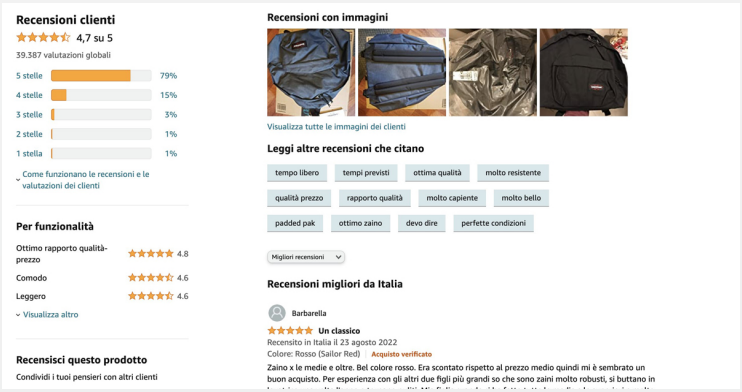
Bourns

Roomba

Ambient Weather

Instagram

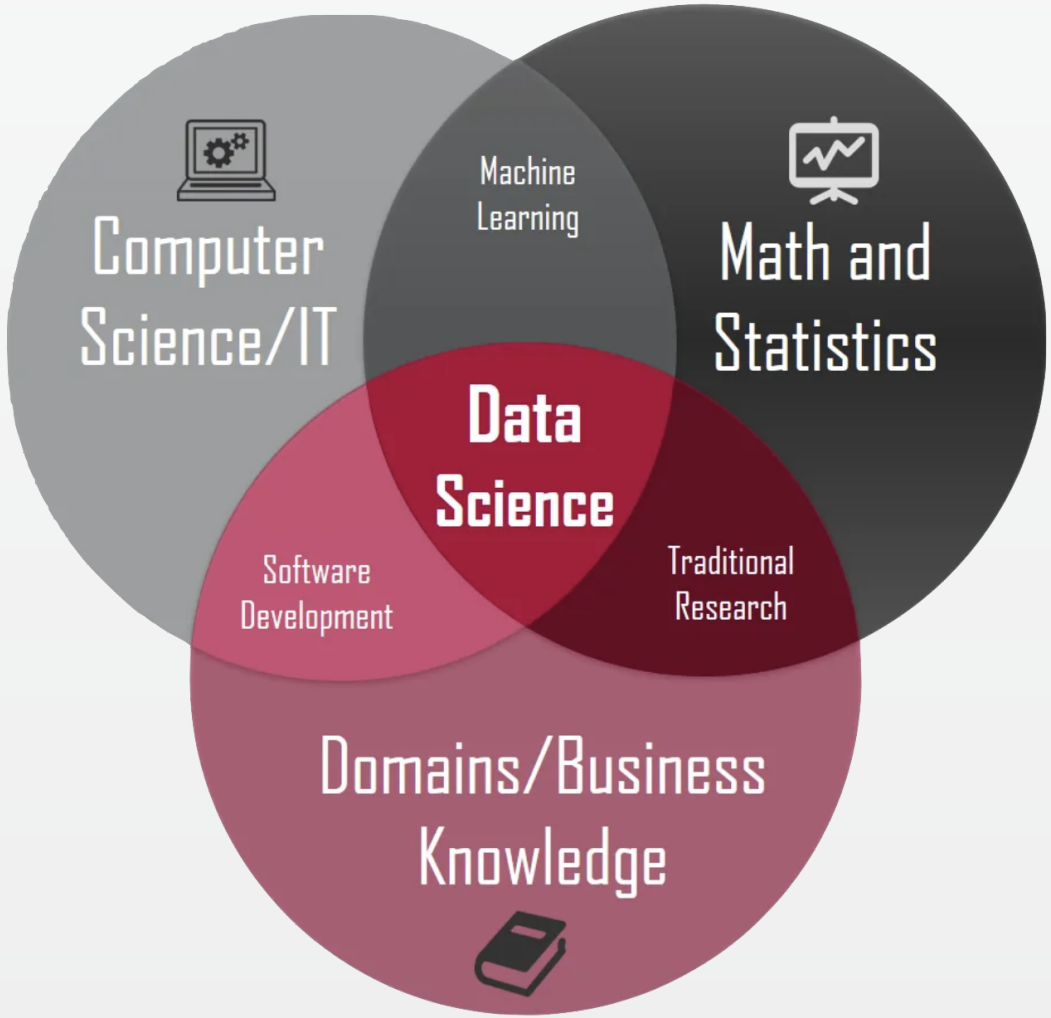
Google Home



Amazon Shop

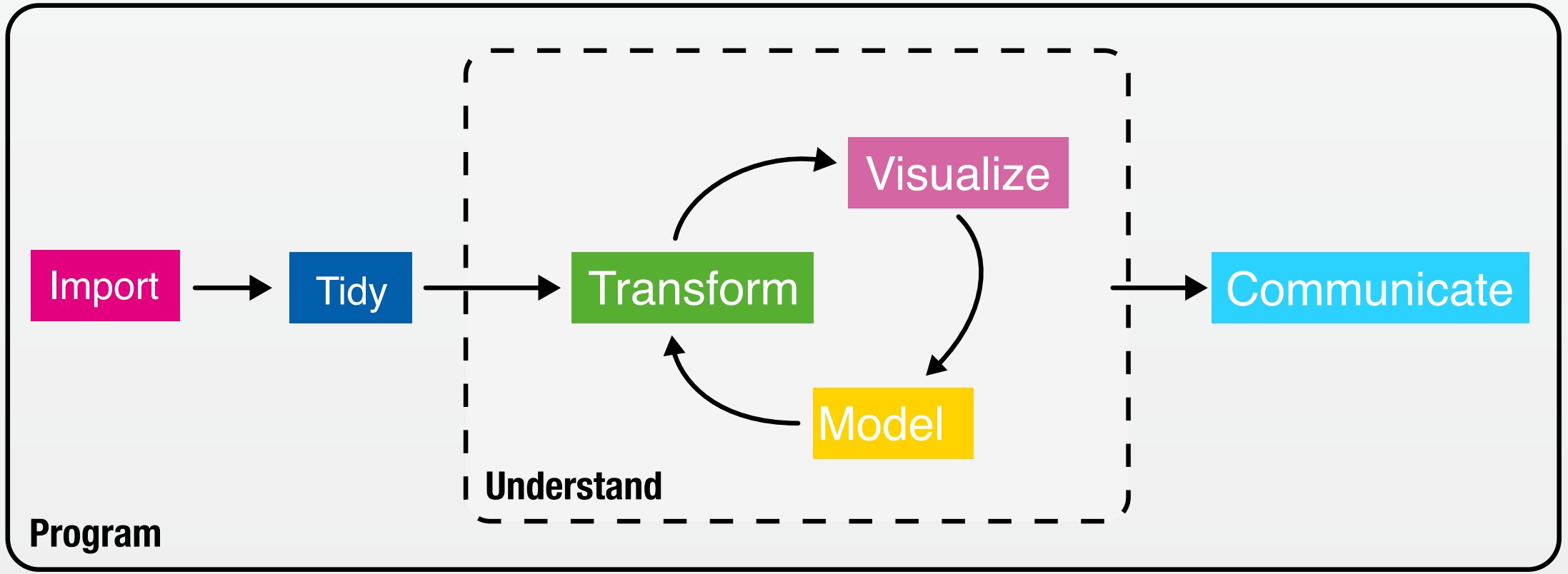
Lo Scienziato dei Dati

- Lavora con esperti di dominio
- Background multidisciplinare: statistica, programmazione, data visualization, comunicazione, ..
- Guadagna mediano USA 2022: $54k all'anno [1]
- Stipendio di un Data Scientist USA 2022: $123k all'anno [2]


Computer Science/IT
Machine Learning
Math and Statistics
Data Science
Software Development
Traditional Research
Domains/Business Knowledge

Il workflow della Data Science

Il ciclo di vita della Data Science va dall'import dei dati fino alla comunicazione dei risultati



Import

- Inzialmente si importano i dati in R
- Tipicamente si caricano i dati presenti in un file, un database o delle web API e li si trasforma in un data frame di R

Esempio di file CSV (Comma-separated values)

```
City, Country, Population, Lat, Lng, Capital
Udine, Italy, 99341, 46.0693000, 13.2371500, FALSE
Trento, Italy, 117417, 46.0678700, 11.1210800, FALSE
Ljubljana,Slovenia, 279631, 46.0510800, 14.5051300, TRUE
Lugano, Switzerland, 63932, 46.0100800, 8.9600400, FALSE
```

Tidy

- Una volta importati i dati è buona norma **pulirli** e **riordinarli**
- Pulire i dati significa immagazzinarli in una forma consistente alla semantiche del dataset e al modo in cui questo è immagazzinato
- Pulire ed organizzare i dati è importante. Lavorare su una struttura consistente permette di focalizzarsi sulle proprie domande o ipotesi
    - evitando di sprecare tempo ed energie a riorganizzare continuamente i propri dati

Tidy: La forma normale

- Un dataframe pulito avrà:
  - una variabile per ogni colonna
  - un'osservazione per ogni riga

VARIABILE

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
41	5.0	3.5	1.3	0.3	setosa
42	4.5	2.3	1.3	0.3	setosa
43	4.4	3.2	1.3	0.2	setosa
44	5.0	3.5	1.6	0.6	setosa
45	5.1	3.8	1.9	0.4	setosa
46	4.8	3.0	1.4	0.3	setosa
47	5.1	3.8	1.6	0.2	setosa
48	4.6	3.2	1.4	0.2	setosa
49	5.3	3.7	1.5	0.2	setosa
OSSERVAZIONE 50	5.0	VALORE 3.3	1.4	0.2	setosa
51	7.0	3.2	4.7	1.4	versicolor
52	6.4	3.2	4.5	1.5	versicolor

Transform

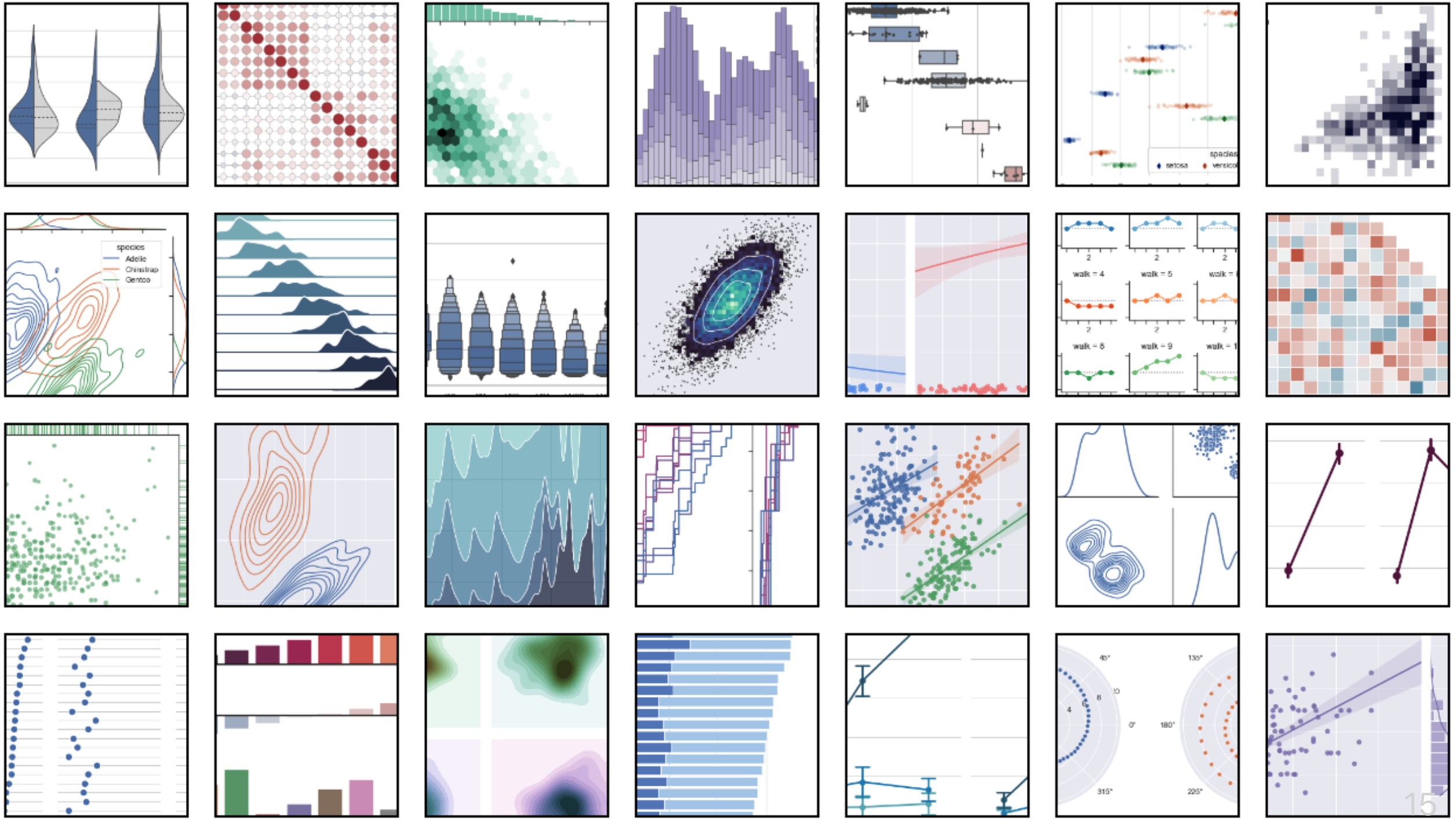
- Dati puliti e organizzati possono essere **trasformati** (effettuando delle **query**)
- Le trasformazioni includono:
  - Concentrarsi sulle osservazioni di interesse (ad es., tutte le persone di una città; i dati raccolti in un mese specifico in un form)
  - Creare nuove variabili in funzione di variabili esistenti (ad es., calcolare la velocità da distanze e tempi)
  - Utilizzare svariate funzioni statistiche (ad es.,conteggi, somme, medie, mediane)
- Assieme tidying e transforming vengono anche detti **wrangling**, dato che spesso si deve "bisticciare" con i dati per trasformali in forma natulare

Visualize e model

- Completate le fasi precedenti, ci si può concentrare su due forme di generazione delle conoscenza:
  - **Visualizzazione**
  - **Modellizzazione**
- Questi strumenti hanno ruoli complementari, con relativi punti di forza e di debolezza
- In ogni analisi reale si itera tra le due più e più volte

Visualize

- La **visualizzazione** è un attività umana fondamentale
- Una buona visualizzazione mostra cose insaspettate, o solleva nuove domande relative ai dati
- Una buona visualizzazione può anche suggerire che ci si sta ponendo le domande sbagliate, o che si devono usare dati diversi
- Le visualizzazioni **non scalano** particolarmente bene a causa della limitata capacità umana nell'interpretarle


species
Adelie
Chinstrap
Gentoo
walk = 4
walk = 5
walk = 8
walk = 9

Model

- I **modelli** sono degli strumenti complementari alla visualizzazione
- Si pongono l'obbiettivo di fornire un riassunto dei dati a dimensionalità ridotta
- Idealmente, un modello cattura un *segnale* vero (ad es., un pattern generato da un fenomeno d'interesse) e ignora il *rumore* (ad es., una variazione casuale al quale non siamo interessati)
- I modelli sono fondamentalmente modelli matematici o computazionali, pertanto scalano molto bene
- Tuttavia, "*la mappa non è il territorio*": ogni modello fa delle assunzioni; questo fa la differenza tra realtà e modellizzazione della realtà

Communicate

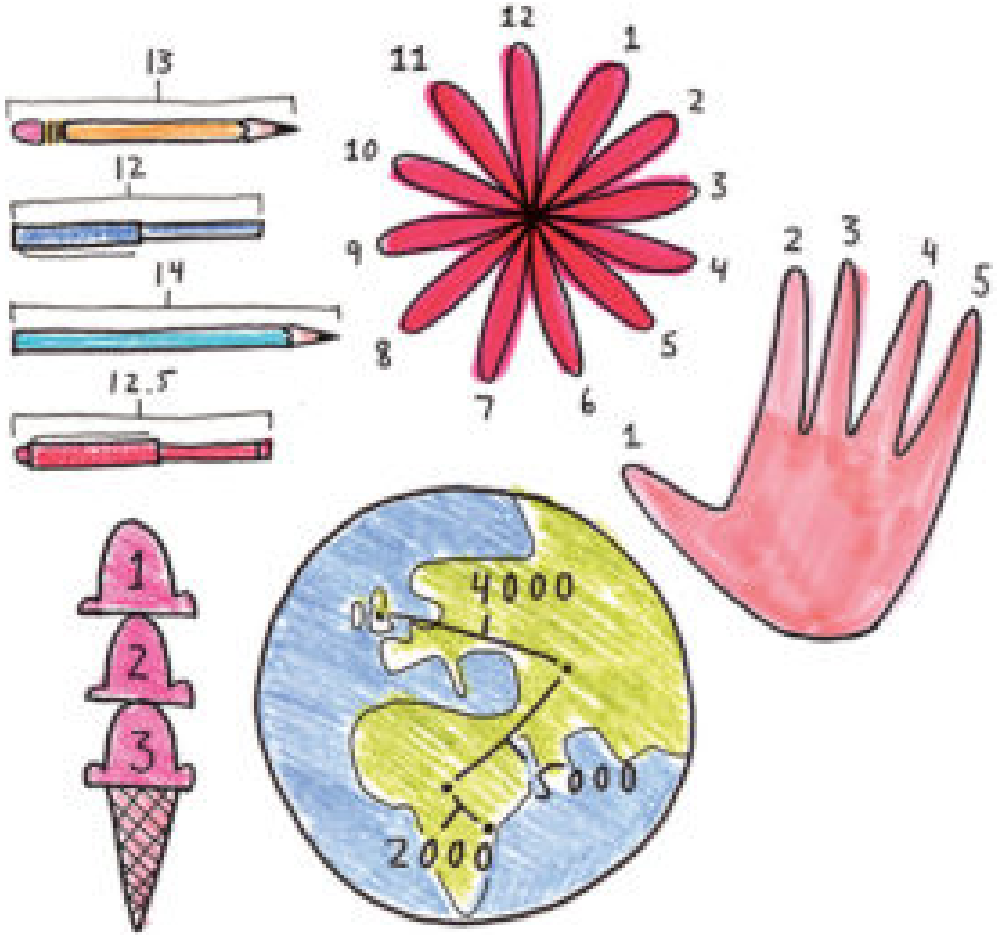
- L'ultima fase della Data Science è la **comunicazione**
- Questa è una parte cruciale di ogni progetto di analisi
- Non importa quanto:
  - buono sia il modello
  - chiare siano le visualizzazioni
  - agevolmente questi consentono di comprendere i dati
  - se poi non li si comunica efficacemente ad altri e al se stessa/o del futuro

INTRODUZIONE

CHE COSA SONO I DATI?

Ogni pianta, ogni persona e ogni interazione in cui siamo coinvolti può essere mappata, quantificata e misurata. Tutte misurazioni sono ciò che noi chiamiamo dati.

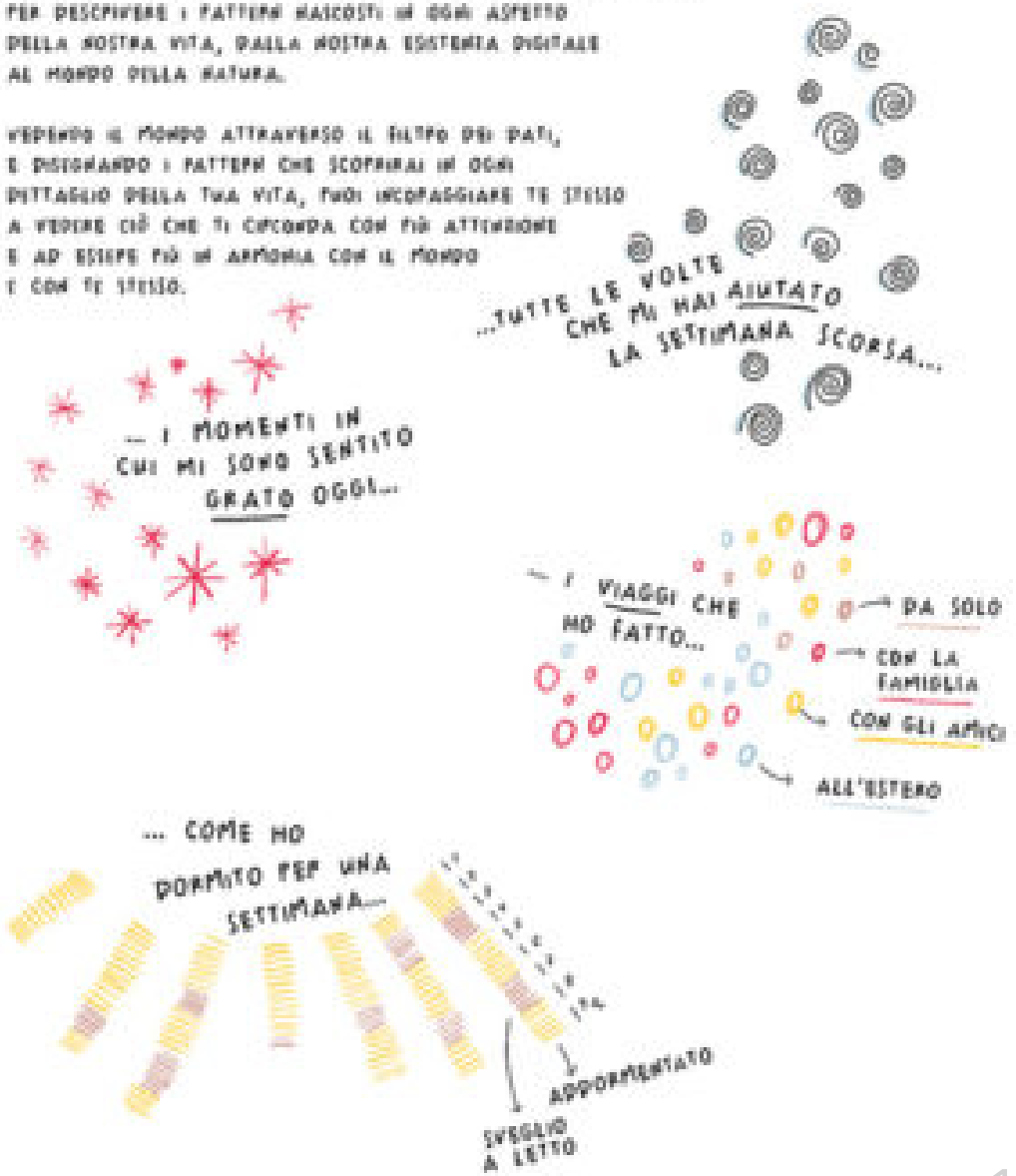
Una volta che imparerai a individuare questi numeri invisibili, inizierai a vederli ovunque, in ogni cosa.

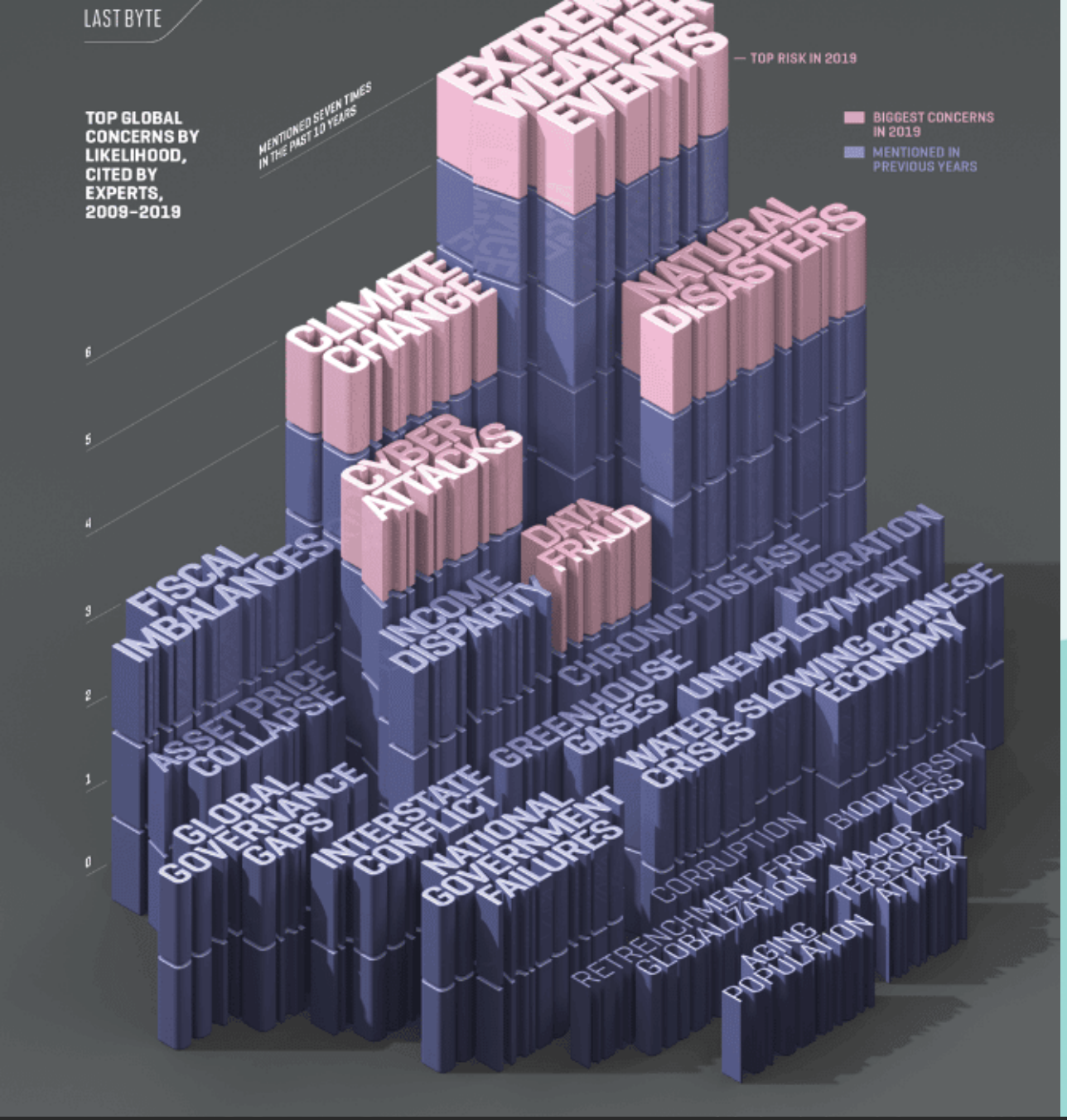


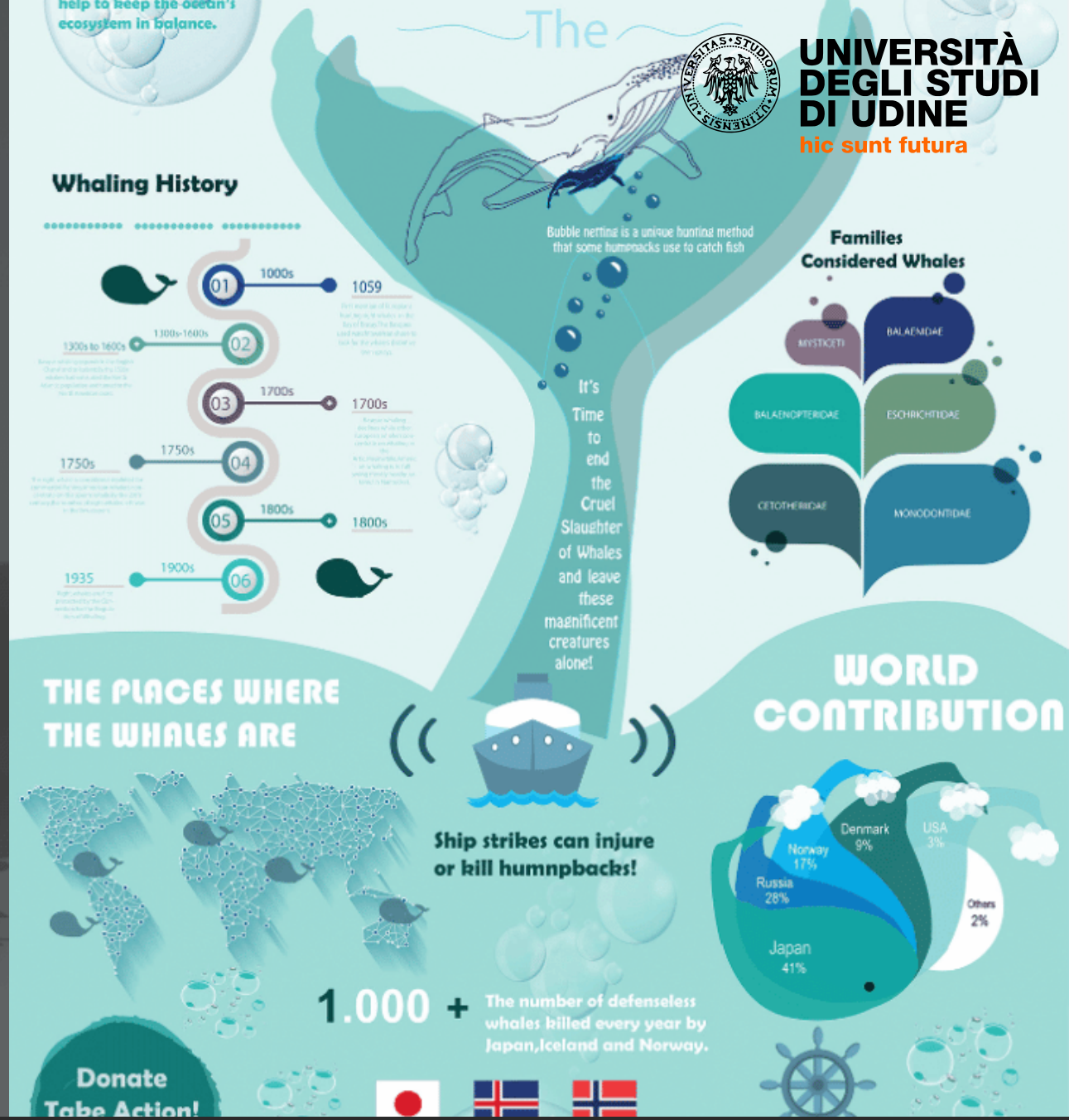
PERCHÉ SONO IMPORTANTI?

I DATI RACCOLTI NELLA VITA QUOTIDIANA POSSONO OFFRIRCI UNO SPACCATO DEL MONDO, PROPRIO COME UNA FOTO PUÒ CATTURARE MOMENTI NEL TEMPO. INOLTRE, POSSONO ESSERE USATI PER DESCRIVERE I PATTERN NASCOSTI IN OGNI ASPETTO DELLA NOSTRA VITA, DALLA NOSTRA ESISTENZA DIGITALE AL MONDO DELLA NATURA.

VEDENDO IL MONDO ATTRAVERSO IL FILTRO DEI DATI, E DISEGNANDO I PATTERN CHE SCOPRIRAI IN OGNI DETTAGLIO DELLA TUA VITA, PUOI INCORAGGIARE TE STESSO A VEDERE CIÒ CHE TI CIRCONDA CON PIÙ ATTENZIONE E AD ESSERE PIÙ IN ARMONIA CON IL MONDO E CON TE STESSO.



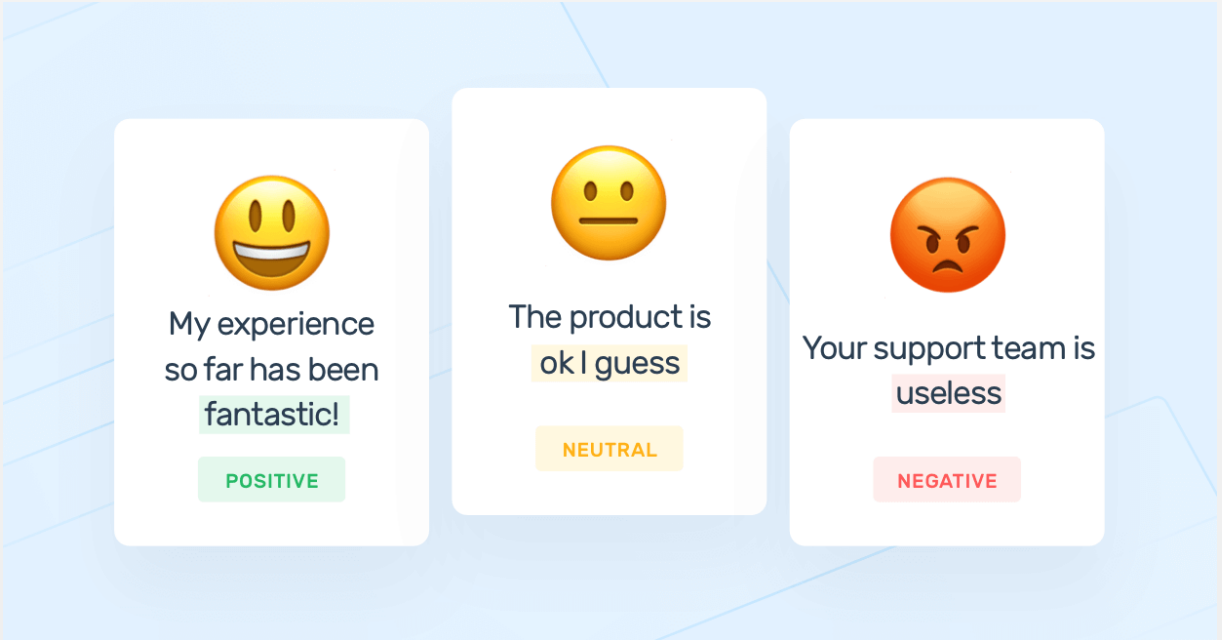

LAST BYTE
TOP GLOBAL CONCERNS BY LIKELIHOOD, CITED BY EXPERTS, 2009–2019
MENTIONED SEVEN TIMES IN THE PAST 10 YEARS
— TOP RISK IN 2019
BIGGEST CONCERNS IN 2019
MENTIONED IN PREVIOUS YEARS
EXTREME WEATHER EVENTS
CLIMATE CHANGE
NATURAL DISASTERS
CYBER ATTACKS
DATA FRAUD
FISCAL IMBALANCES
INCOME DISPARITY
CHRONIC DISEASE
MIGRATION
ASSET PRICE COLLAPSE
GREENHOUSE GASES
UNEMPLOYMENT
WATER CRISES
SLOWING CHINESE ECONOMY
GLOBAL GOVERNANCE GAPS
INTERSTATE CONFLICT
CORRUPTION
BIODIVERSITY LOSS
NATIONAL GOVERNMENT FAILURES
RETRENCHMENT FROM GLOBALIZATION
MAJOR TERRORIST ATTACK
AGING POPULATION
0
1
2
3
4
5
6


help to keep the ocean's ecosystem in balance.
The
UNIVERSITÀ DEGLI STUDI DI UDINE
hic sunt futura
Whaling History
01 1000s 1059
02 1100s-1600s 1300s to 1600s
03 1700s 1700s
04 1750s 1750s
05 1800s 1800s
06 1900s 1935
Bubble netting is a unique hunting method that some humpbacks use to catch fish
It's Time to end the Cruel Slaughter of Whales and leave these magnificent creatures alone!
Families Considered Whales
MYSTICETI
BALAENIDAE
BALAENOPTERIDAE
ESCHRICHTIDAE
CETOTHERIIDAE
MONODONTIDAE
THE PLACES WHERE THE WHALES ARE
Ship strikes can injure or kill humnpbacks!
WORLD CONTRIBUTION
Norway 17%
Denmark 9%
USA 3%
Russia 28%
Others 2%
Japan 41%
1.000 +
The number of defenseless whales killed every year by Japan,Iceland and Norway.
Donate
Take Action!

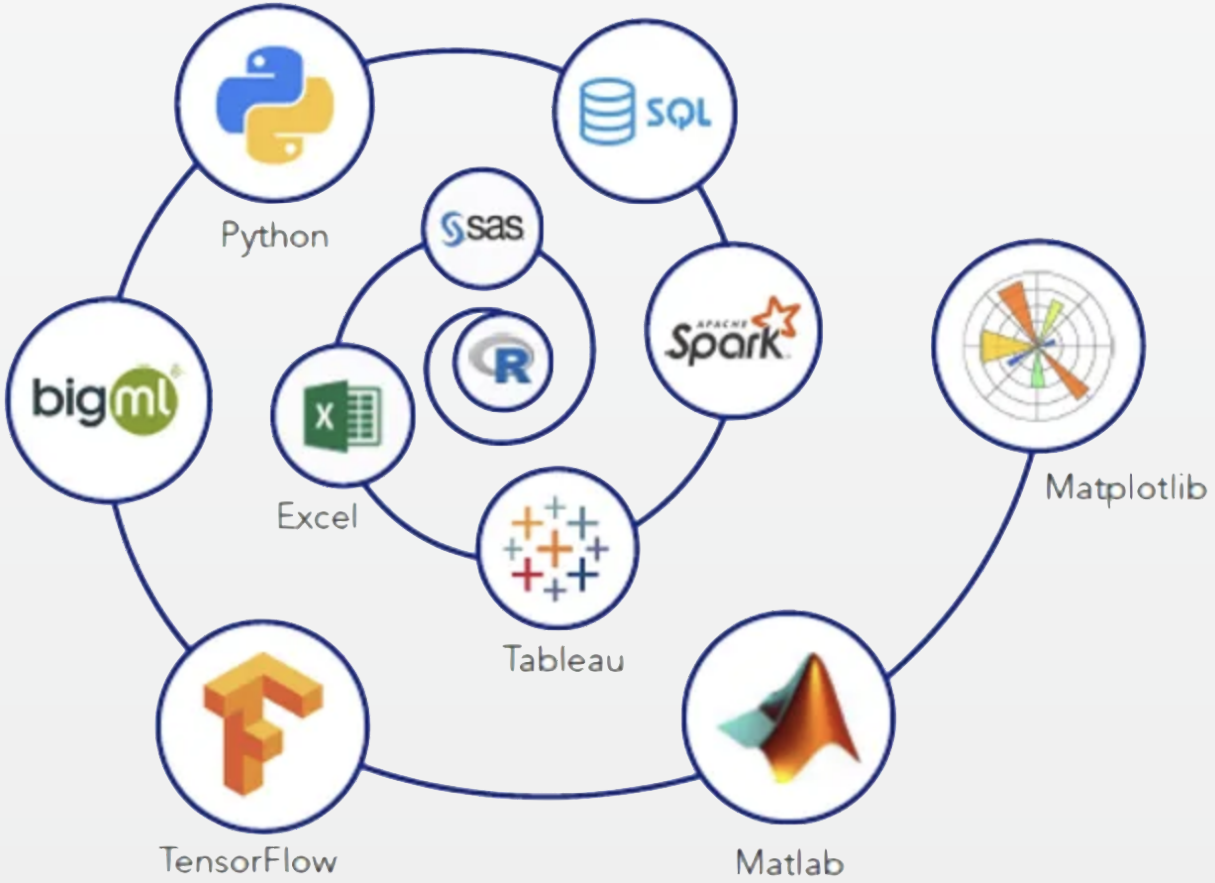

BIG
DATA
INFORMATION
TECHNOLOGIES
MANAGEMENT
ANALYSIS
RESEARCH
PETABYTES
STORAGE
LARGE
SETS
MILLION
SEARCH
INTERNET
LEARNING
VOLUME
PROCESS
GOVERNMENT
SOFTWARE
SAN
NAS
SOCIAL
DEVELOPMENT
TECHNOLOGY
SCIENCE
FUTURE
NEW
BILLION
PARALLEL
INITIATIVE
ECONOMIC
TERABYTES
BASED
WORLD
EVERY
SIZE
EXABYTES
DATABASES
PROCESSING
ANALYTICS
SIMULATIONS
FRAMEWORK
BUSINESS
BETWEEN
STATISTICS
VISUALIZATION
APPLICATIONS
INFRASTRUCTURE
AMOUNT
REAL
STRUCTURE
WORK
RATE
HIGH
USERS
DEPARTMENT
QUERIES
SET
DETERMINE
WELL
DEFINITION
USING
MASSIVE
FACTOR
ONLINE
DAY
MANAGE
WORLDWIDE
SYSTEMS
MAKING
INTEGRATION
RESULTS
YEARS
DRIVE
REPLICATION
HANDLE
SOURCES
CHALLENGES
LIMITS
COMPLEX
HAND
CAPACITY
ANNOUNCED
PAST
ONE
OUT
INCLUDE
NETWORKS
HYPOTHESIS
END
MOST
NEARLY
MAPREDUCE
CRITIQUES
APPROACH
MANY
PRIVACY
INTERNATIONAL
IMPLEMENTATION
BECOMES
SECOND
NATIONAL
INTELLIGENCE
FLOW
APPROACHES
PROCESSES
PRIVATE
TERMS
YEAR
CAPABILITIES
SUCCESSFUL
NEEDED
USES
LESS
HUGE
COMPARED
SHARED
VARIOUS
TIMES
ARCHITECTURE
PARADIGM
FUNDING
INSIGHT
TOOLS
CENTER
RECORDS
COST
DISTRIBUTED
RELEVANT
USED
PER
PROJECT
ANNUAL
NOW
VARIETY
MARKET
SENSOR
TIME
WOULD
RELATED
TRAFFIC
PEOPLE
CURRENTLY
USE
STATE
CRITIQUE
HIGHER
ALGORITHMS
COMPANIES

NLP (Natural Language Processing)

- sottobranca di linguistica, informatica e intelligenza artificiale
- elaborare e analizzare grande quantità di dati di linguaggio naturale
- Applicazioni: **Sentiment analysis** e **Text Summarization**



Tool per la Scienza dei dati



R

R è un ambiente software libero per effettuare calcoli statistici e produrre visualizzazioni grafiche. È utile in tutte le fasi del workflow. Punti di forza:

1. **Capacità**: offre un vasto insieme di funzionalità
2. **Comunità**: vanta una comunità di utenti molto numerosa
3. **Prestazioni**: è piuttosto veloce (se eseguito in memoria centrale)

Pagina web: r-project.org

RStudio

RStudio è un ambiente di sviluppo integrato (IDE) per R che include:

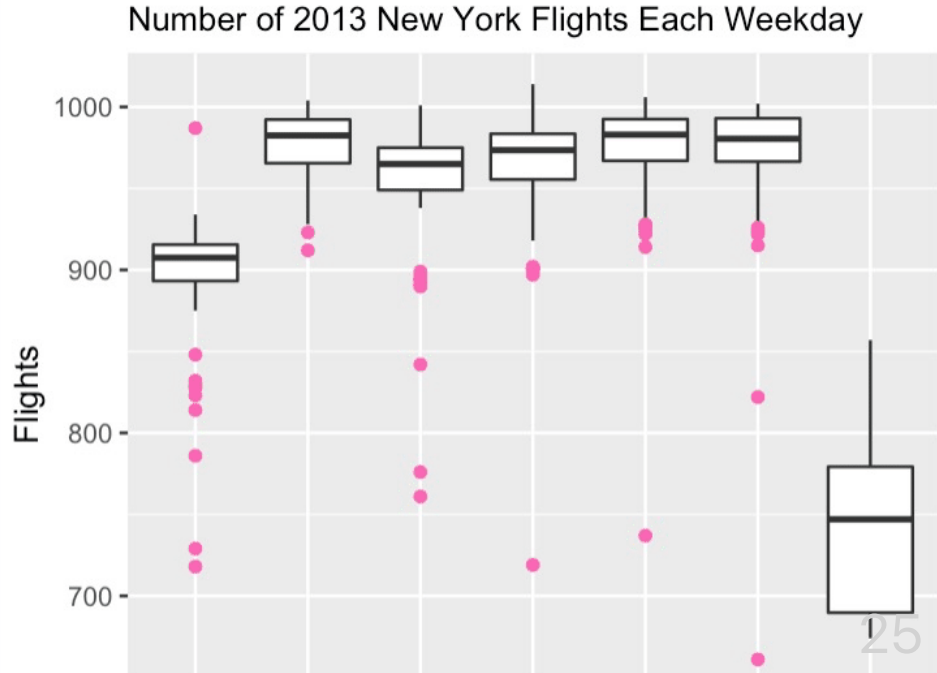
- un terminale
- un editor di testo con colorazione della sintassi che supporta la diretta esecuzione del codice
- strumenti per la visualizzazione dei grafici
- uno storico
- un debugger
- supporto alla gestione dello spazio di lavoro

```r
library(nycflights13)     ## package containing flights dataset
library(lubridate)
library(dplyr)
library(ggplot2)

head(flights, n = 3)
daily <- flights %>%
  mutate(date = make_date(year, month, day)) %>%
  count(date) %>%
  mutate(wday = wday(date, label = TRUE))
head(daily, n = 3)
ggplot(daily, aes(wday, n)) +
  geom_boxplot(outlier.colour = "hotpink") +
  labs(x = "Weekday", y = "Flights",
       subtitle = "Number of 2013 New York Flights Each Weekday")
```

```
# A tibble: 3 x 19
   year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time arr_delay carrier
  <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>     <dbl> <chr>
1  2013     1     1      517            515         2      830            819        11 UA
2  2013     1     1      533            529         4      850            830        20 UA
3  2013     1     1      542            540         2      923            850        33 AA
# … with 9 more variables: flight <int>, tailnum <chr>, origin <chr>, dest <chr>, air_time <dbl>,
#   distance <dbl>, hour <dbl>, minute <dbl>, time_hour <dttm>
> daily <- flights %>%
+   mutate(date = make_date(year, month, day)) %>%
+   count(date) %>%
+   mutate(wday = wday(date, label = TRUE))
> head(daily, n = 3)
# A tibble: 3 x 3
  date           n wday
  <date>     <int> <ord>
1 2013-01-01   842 Tue
2 2013-01-02   943 Wed
3 2013-01-03   914 Thu
```

```
Data
daily                365 obs. of 3 variables
    $ date: Date[1:365], format: "2013-01-01" "2013-01-02" …
    $ n   : int [1:365] 842 943 914 915 720 832 933 899 902…
    $ wday: Ord.factor w/ 7 levels "Sun"<"Mon"<"Tue"<..: 3 …
```



Markdown

- Markdown è un linguaggio di markup con una sintassi del testo semplice
- progettata per essere convertita in HTML e in molti altri formati

https://markdownlivepreview.com/ →

Markdown Live Preview

```
# Markdown syntax guide

## Headers

# This is a Heading h1
## This is a Heading h2
###### This is a Heading h6

## Emphasis

*This text will be italic*
_This will also be italic_

**This text will be bold**
__This will also be bold__

_You **can** combine them_

## Lists

### Unordered

* Item 1
* Item 2
* Item 2a
* Item 2b

### Ordered

1. Item 1
1. Item 2
1. Item 3
  1. Item 3a
  1. Item 3b

## Images

![This is a alt text.](/image/sample.png "This is a sample image.")

## Links
```


Markdown syntax guide

Headers

This is a Heading h1

This is a Heading h2

This is a Heading h6

Emphasis

This text will be italic
This will also be italic

This text will be bold
This will also be bold

*You **can** combine them*

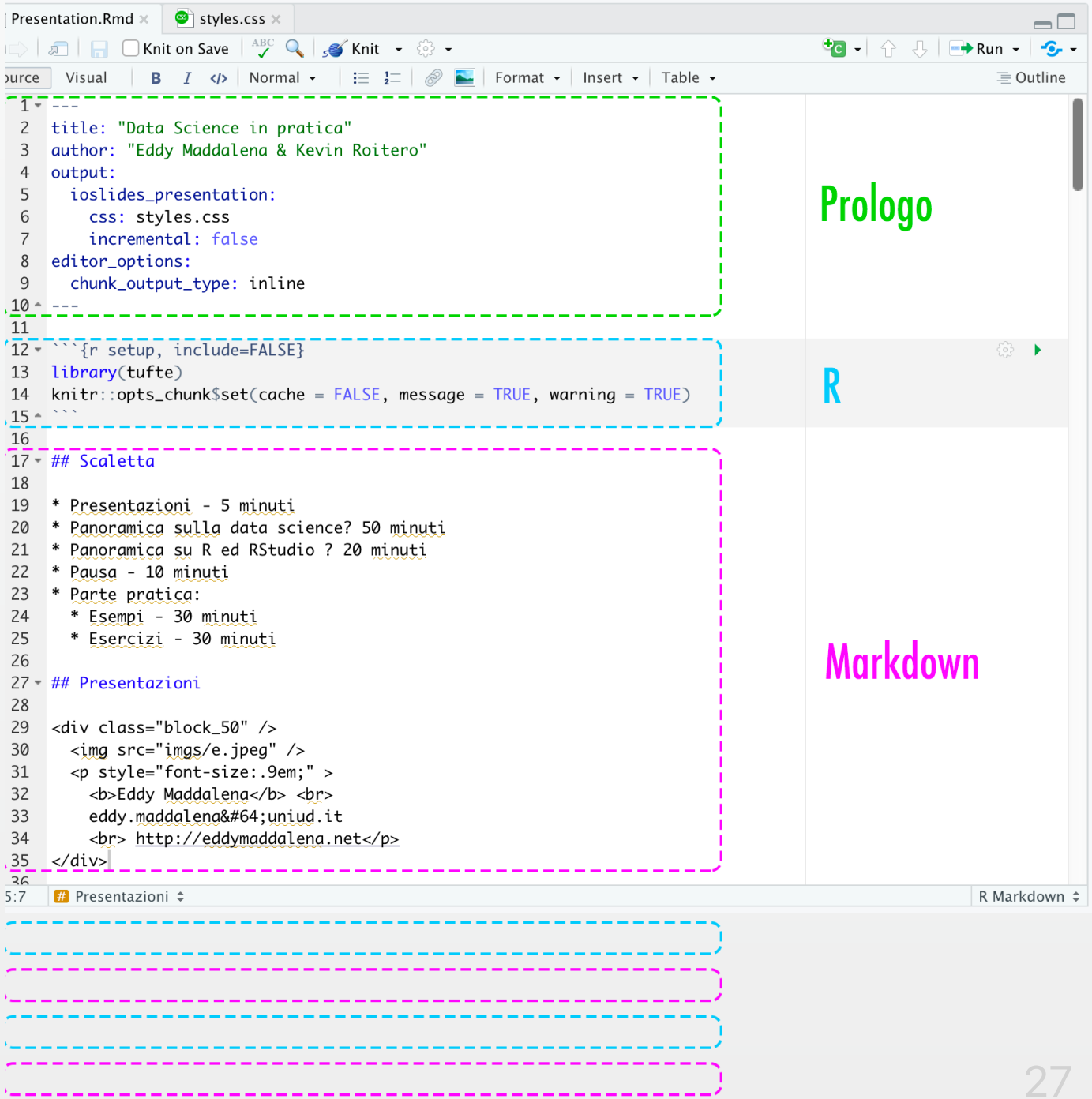
Lists

Unordered

- Item 1
- Item 2
- Item 2a
- Item 2b

R and RMarkdown

- Combiniamo assieme codici R ed markdown
- Programmiamo e presentiamo contemporaneamente



kaggle

Search datasets

Filters

Create

Home

Competitions

Datasets

Code

Discussions

Courses

More

Your Work

RECENTLY VIEWED

Video Game Sales

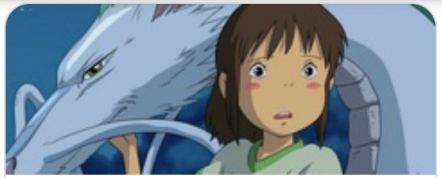
Videogame Sales

Minecraft Composting ...

Fortnite Statistics Corr...

COD Players data visu...

All datasets | Computer Science | Education | Classification | Computer Vision | NLP | Data Visualization | Pre-Trained Model



Animated Movies IMDb

Tarun R Jain · Updated 6 days ago

Usability **10.0** · 10 kB

1 File (CSV)

14

Airbnb-NYC-Cleaned

sandeep majumdar · Updated 6 days ago

Usability **10.0** · 7 MB

1 File (CSV)

17

Top 6 Economies in the world by GDP

Charan Chandrasekaran · Updated 6 da...

Usability **9.1** · 22 kB

1 File (CSV)

9

Chicago Bicycle Rent Usage

Gunnar N. · Updated 19 days ago

Usability **9.1** · 455 MB

28 Files (CSV)

18

Popular Datasets

See All

Data Science Job Salaries

Ruchi Bhatia · Updated 3 months ago

Usability **10.0** · 8 kB

1 File (CSV)

AMEX data - integer dtypes - parquet format

raddar · Updated 3 months ago

Usability 4.4 · 4 GB

2 Files (other)

House Rent Prediction Dataset

Sourav Banerjee · Updated 11 days ago

Usability **10.0** · 84 kB

2 Files (CSV, other)

McDonald's India : Menu Nutrition Dataset

Deep Contractor · Updated a month ago

Usability **9.4** · 5 kB

1 File (CSV)

Riferimenti

[1] https://www.thebalancecareers.com/

[2] https://builtin.com/salaries/data-analytics/data-scientist