

# Open Data

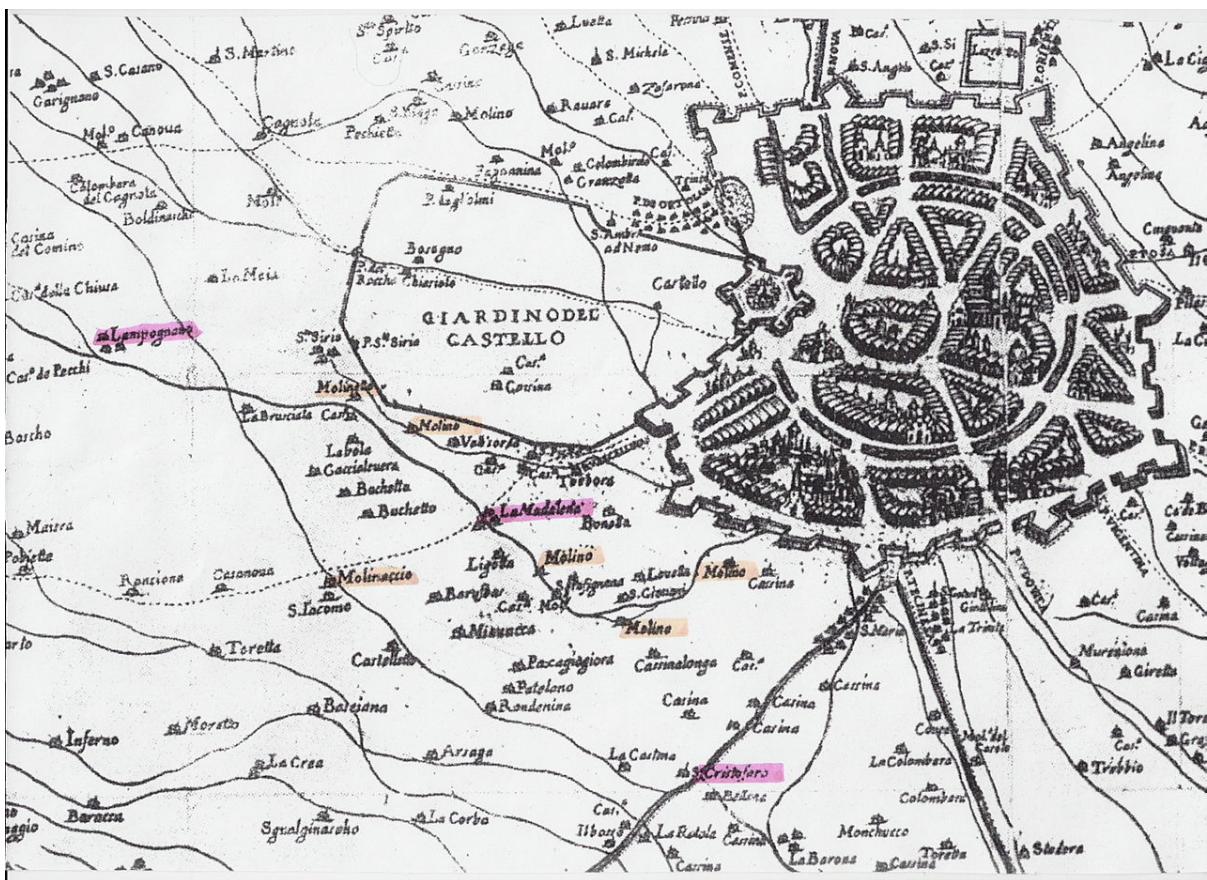
- Perché parliamo di dati?
- Perché parliamo di "open"?

## **Una volta i dati erano**

- Elenchi:
  - Liste
  - Vocabolari
  - Enciclopedie
- Numeri
- Mappe



Da una lapide in S. Pietro a Roma



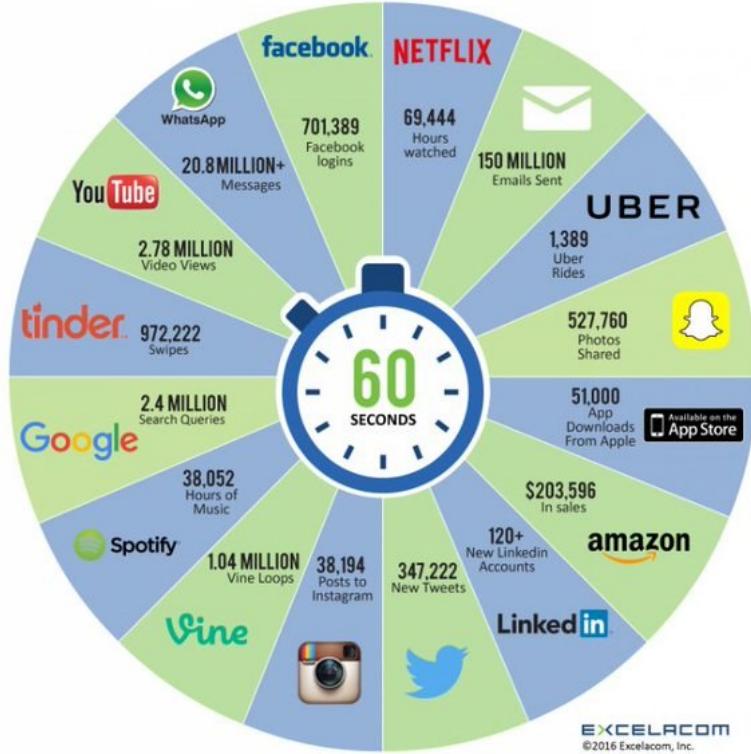
## Dopo l'avvento del digitale tutto è un dato

- Elenchi
- Numeri
- Testi (la Bibbia è un libro o un testo?)
- Musica
- Immagini
- Telefonate
- ...

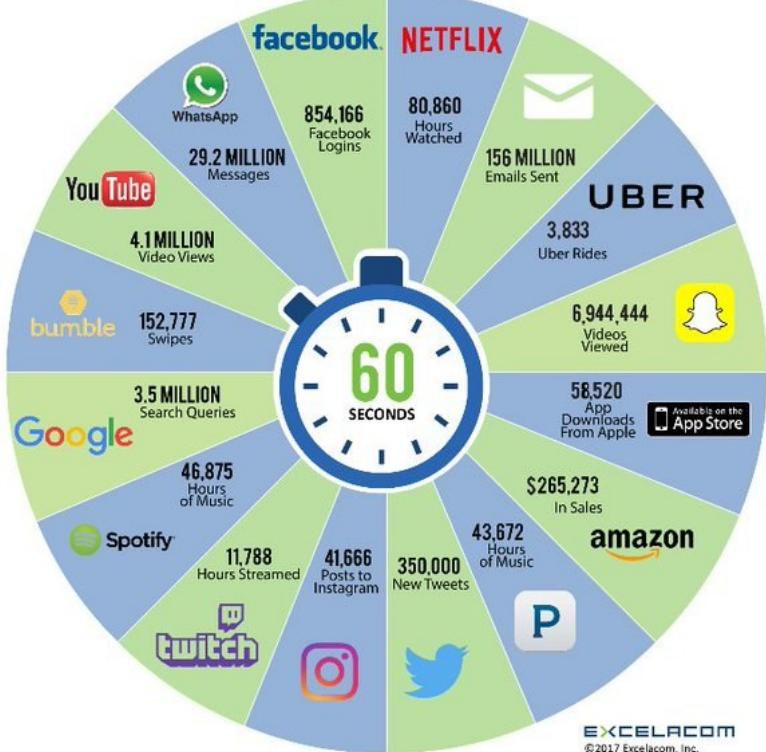
## **Se è un dato, lo possiamo**

- Memorizzare
- Analizzare
- Confrontare
- Utilizzare

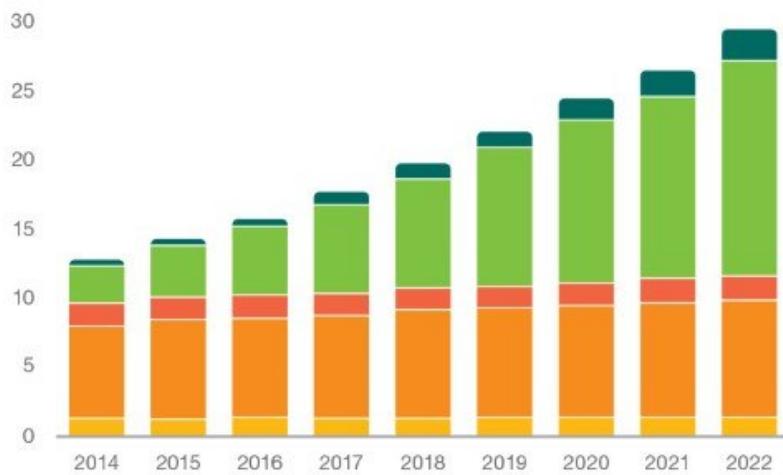
# 2016 What happens in an INTERNET MINUTE?



# 2017 What happens in an INTERNET MINUTE?



Connected devices (billions)



	2016	2022	CAGR
Wide-area IoT	0.4	2.1	30%
Short-range IoT	5.2	16	20%
PC/laptop/tablet	1.6	1.7	0%
Mobile phones	7.3	8.6	3%
Fixed phones	1.4	1.3	0%
	16 billion	29 billion	10%

Fonte: Ericsson



# People now watch 1 billion hours of YouTube per day

Darrell Etherington @etherington / Feb 28, 2017

 Comment



## Dati testuali

- Contengono testi, normalmente in "linguaggio naturale".
- Se non sono direttamente narrazioni o descrizioni, sono complicati da gestire automaticamente
- Per questo normalmente vengono *aumentati* con l'introduzione di *markup* o *tag* (il più famoso è l'*hashtag*)

## Dati numerici

- Contengono numeri, interi o con la virgola
- Di solito descrivono una grandezza
- Unità di misura?
- Base?
- Importante sapere quali separatori per i decimali e le migliaia
- Sono "facili" da gestire automaticamente
- Per questo spesso si cerca di ridurre tutto a un numero

## Dati tabulari

- Composti da tabelle di righe e colonne
- Le colonne identificano gli attributi
- Le righe identificano entità
- Per identificare relazioni:
  - sono necessarie più tabelle
  - O più ripetizioni
- Sono "facili" da gestire automaticamente
- I DataBase storicamente sono formati da tabelle (Entity-Relationship)
- Le colonne sono un primo livello informativo
- Tutte le entità devono avere lo stesso numero di attributi

## Dati a oggetti gerarchici

- Composti da oggetti o liste di oggetti che contengono ciascuno
  - attributi
  - altri oggetti o liste di oggetti
- Permettono di gestire relazioni (es. appartenenza) in modo più evidente
- Sono gestibili automaticamente
- La loro gestione è meno efficiente dal punto di vista computazionale
- Ogni entità può avere un numero di attributi diverso

# Dati geografici

- Composti da entità che abbiano associate delle *coordinate*
- Utilizzati nei GIS Geographical Information Systems
- Problema dei sistemi di riferimento e delle proiezioni
  - <http://spatialreference.org/> (<http://spatialreference.org/>)
  - <https://courses.washington.edu/gis250/lessons/projection/> (<https://courses.washington.edu/gis250/lessons/projection/>)
  - <http://epsg.io/?q=Italy> (<http://epsg.io/?q=Italy>)
- Normalmente sono coordinate
  - Geografiche, in gradi es [EPSG4326](https://epsg.io/4326) (<https://epsg.io/4326>) / WGS84 (GPS, Google Earth, ...)
  - Geometriche, in metri es [EPSG3857](https://epsg.io/3857) (<https://epsg.io/3857>) o [EPSG3003](https://epsg.io/3003) (<https://epsg.io/3003>) (Monte Mario)

Un tool: QGIS <https://www.qgis.org> (<https://www.qgis.org>).

## *Open data*

- Cos'è il copyright?
  - Quali sono le norme nel nostro paese?
  - Cosa coprono?
- Come si applica il copyright quando il bene è **duplicabile senza danno**?

## Partiamo dal *software*

- Inizialmente il software era tutto libero
- Uno dei primi a farne un problema è stato Henry William "Bill" Gates
- Uno dei primi a contestare questa cosa è stato RMS o Richard Matthew Stallman

February 3, 1976

An Open Letter to Hobbyists

To me, the most critical thing in the hobby market right now is the lack of good software courses, books and software itself. Without good software and an owner who understands programming, a hobby computer is wasted. Will quality software be written for the hobby market?

Almost a year ago, Paul Allen and myself, expecting the hobby market to expand, hired Monte Davidoff and developed Altair BASIC. Though the initial work took only two months, the three of us have spent most of the last year documenting, improving and adding features to BASIC. Now we have 4K, 8K, EXTENDED, ROM and DISK BASIC. The value of the computer time we have used exceeds \$40,000.

The feedback we have gotten from the hundreds of people who say they are using BASIC has all been positive. Two surprising things are apparent, however. 1) Most of these "users" never bought BASIC (less than 10% of all Altair owners have bought BASIC), and 2) The amount of royalties we have received from sales to hobbyists makes the time spent of Altair BASIC worth less than \$2 an hour.

Why is this? As the majority of hobbyists must be aware, most of you steal your software. Hardware must be paid for, but software is something to share. Who cares if the people who worked on it get paid?

Is this fair? One thing you don't do by stealing software is get back at MITS for some problem you may have had. MITS doesn't make money selling software. The royalty paid to us, the manual, the tape and the overhead make it a break-even operation. One thing you do do is prevent good software from being written. Who can afford to do professional work for nothing? What hobbyist can put 3-man years into programming, finding all bugs, documenting his product and distribute for free? The fact is, no one besides us has invested a lot of money in hobby software. We have written 6800 BASIC, and are writing 8080 APL and 6800 APL, but there is very little incentive to make this software available to hobbyists. Most directly, the thing you do is theft.

What about the guys who re-sell Altair BASIC, aren't they making money on hobby software? Yes, but those who have been reported to us may lose in the end. They are the ones who give hobbyists a bad name, and should be kicked out of any club meeting they show up at.

I would appreciate letters from any one who wants to pay up, or has a suggestion or comment. Just write me at 1180 Alvarado SE, #114, Albuquerque, New Mexico, 87108. Nothing would please me more than being able to hire ten programmers and deluge the hobby market with good software.

*Bill Gates*  
Bill Gates  
General Partner, Micro-Soft

*The feedback we have gotten from the hundreds of people who say they are using BASIC has all been positive. Two surprising things are apparent, however, 1) Most of these "users" never bought BASIC (less than 10% of all Altair owners have bought BASIC), and 2) The amount of royalties we have received from sales to hobbyists makes the time spent on Altair BASIC worth less than \$2 an hour.*

*Why is this? As the majority of hobbyists must be aware, most of you steal your software. Hardware must be paid for, but software is something to share. Who cares if the people who worked on it get paid?*

*Is this fair? One thing you don't do by stealing software is get back at MITS for some problem you may have had.*

*Bill Gates  
General Partner, Micro-Soft*

## Quesione di *licenze*

- Un bene immateriale non lo possiedi, ti viene *dato in licenza*
- La licenza indica quello che ti è permesso e quello che ti è vietato
- In particolare, tutte le licenze ti vietano di
  - Duplicare (tranne che per conservazione)
  - Prestare
  - ...

*Copyright e Copyleft*

# Le licenze *Open Source*

- La  
GPL

Ma anche...

- La MIT
- La Apache
- La freemium
- La shareware
- La freeware
- La public  
domain
- La WTFPL

## **Quali licenze sono considerate *Open Source*?**

- Ci aiuta la **OSI Open Source Initiative** <https://opensource.org/licenses> (<https://opensource.org/licenses>).
- Ci permette anche di scegliere una licenza <https://choosealicense.com/> (<https://choosealicense.com/>).

## **Cos'è una *licenza virale*?**

## Come si applica tutto questo ai dati?

- I dati sono *software*, ma di un tipo particolare
- Oppure potremmo dire che il software è un dato di tipo particolare

*The Open Definition states: "A piece of content or data is open if anyone is free to use, reuse, and redistribute it – subject only, at most, to the requirement to attribute and share-alike."*

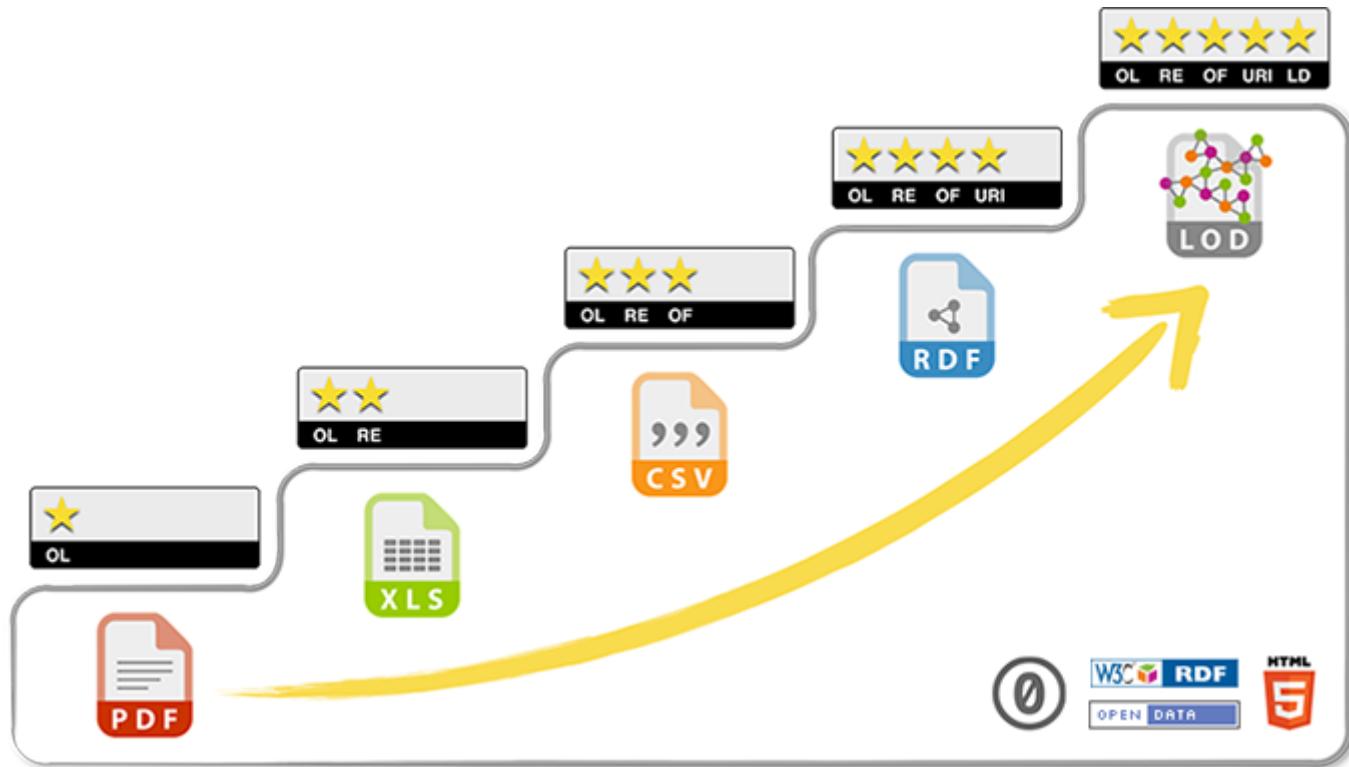
## **Esistono *licenze per i dati***

- Fra le più famose la **Creative Commons (CC)**
- Ma anche qui ci aiuta la **OKFN Open Knowledge Foundation**  
<https://licenses.opendefinition.org/> (<https://licenses.opendefinition.org/>).
- Esiste un **selezionatore di licenze** <https://creativecommons.org/choose/?lang=it> (<https://creativecommons.org/choose/?lang=it>).
- Esiste un **motore di ricerca di contenuti open**  
<https://search.creativecommons.org/> (<https://search.creativecommons.org/>).
- In Italia esiste la **IODL Italian Open-Data License**  
<https://www.dati.gov.it/content/italian-open-data-license-v20>  
(<https://www.dati.gov.it/content/italian-open-data-license-v20>).

## **La qualità dei dati**

Tim Berners-Lee, l'inventore del World Wide Web, ha definito una classifica dei tipi di Open Data <https://www.w3.org/DesignIssues/LinkedData.html> (<https://www.w3.org/DesignIssues/LinkedData.html>) basandosi su 5 criteri:

1. **OL** Open License
2. **RE** machine REadable
3. **OF** Open Format
4. **URI** Uniform Resource Identifier
5. **LD** Linked Data



# Formati e metadati

## Un formato è un *contratto*

- Comunicazione: trasmittente, ricevente, messaggio
- Perché la comunicazione avvenga correttamente, il messaggio dev'essere compreso
- Il *formato* definisce che aspetto avrà il messaggio, in modo da poterlo decodificare
  - Nella lingua parlata la Lingua è un formato
  - Nella lingua scritta l'alfabeto è un formato
  - I formati digitali puntano alla *non ambiguità*

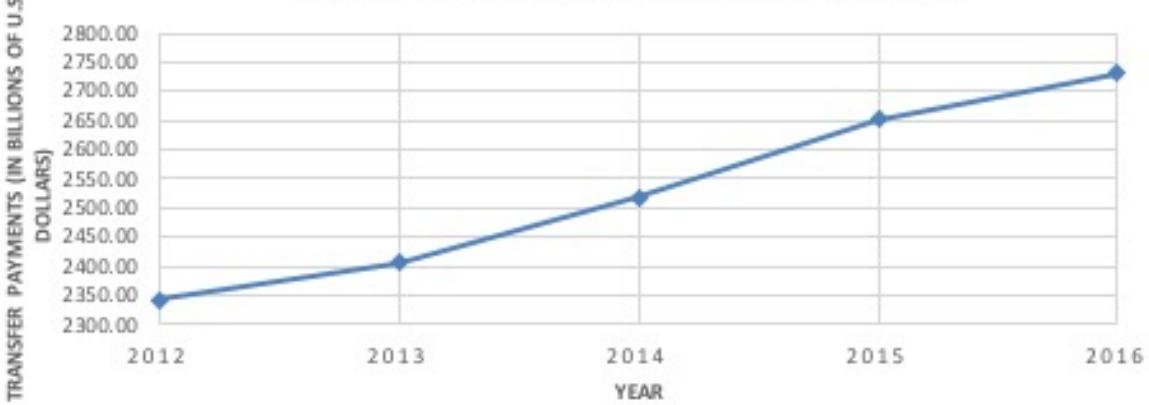
## I *metadati* sono dati ulteriori che "aumentano" i dati

- Non sono parte del messaggio in sé
- Ma ne aiutano:
  - La comprensione
  - La catalogazione
  - I collegamenti
- Le note di un libro sono metadati
- NB anche i metadati sono dati. E hanno un formato...

## I formati *non readable*

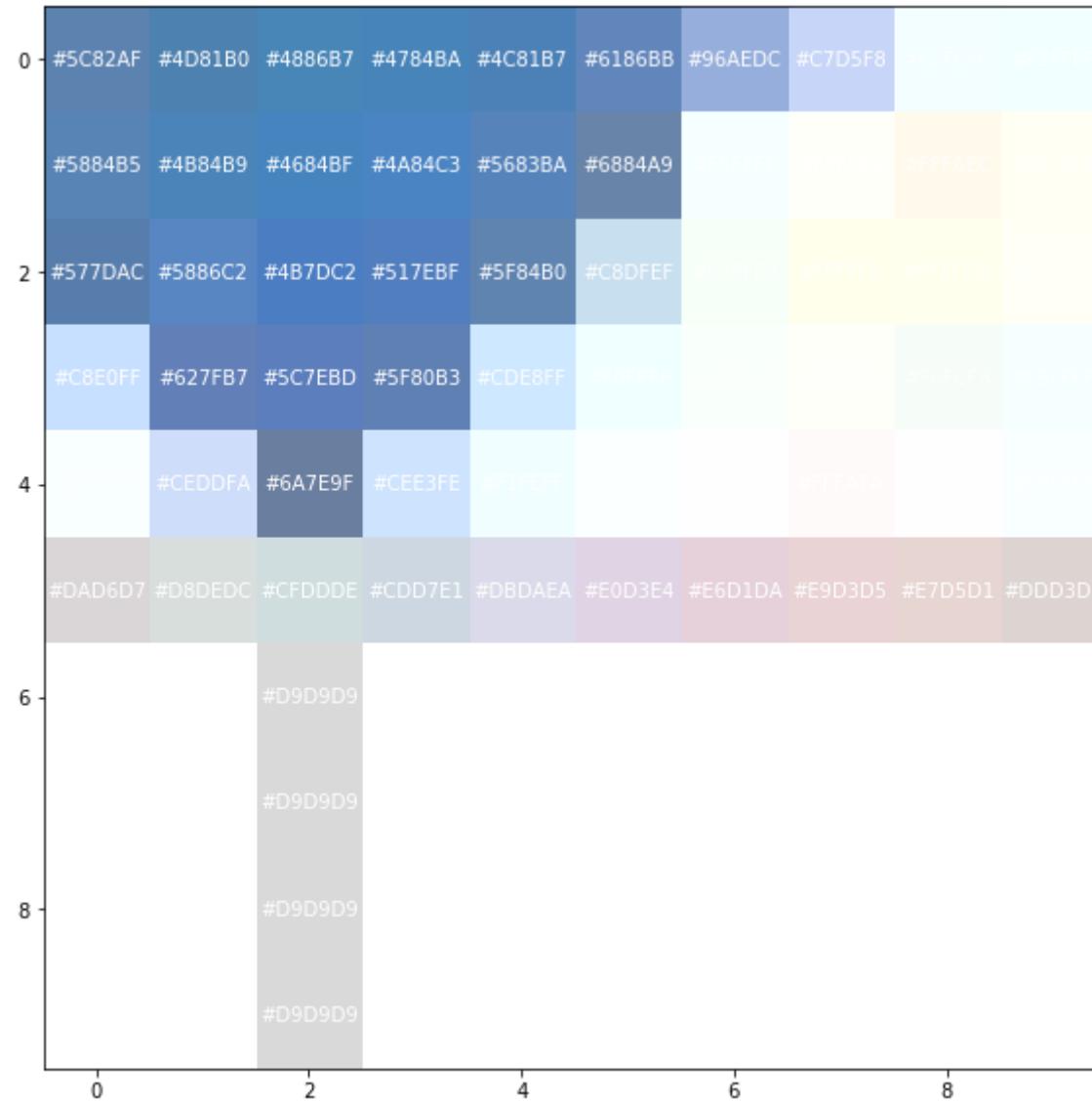
- Raster
- PDF (soprattutto raster)
- In generale i **formati proprietari**

## FEDERAL GOVERNMENT TRANSFER PAYMENTS IN THE UNITED STATES SINCE 2012



In [59]: fig

Out[59]:



# **PDF *non* è un formato readable**

- <https://webproto.si.unimib.it/ADP/>  
(https://webproto.si.unimib.it/ADP/)



UNIVERSITÀ DEGLI STUDI DI MILANO – BICOCCA  
PIAZZA DELL'ATENEO NUOVO, 1 - 20128 MILANO

REP.Nr. 202/2018  
Registrato il 24/01/2018

POSTIMV

Visto: L. Capo Settore  
AFFISSO ALL'ALBO  
Dal 24/01/2018 al 30/06/2018

Visto: I. Capo Settore

IL RETTORE

Universita' degli Studi  
di Milano - Bicocca  
Progetto MIUR-Rete Bio-OPeR (55387/2012)  
VOLTI 2013 cod. 20130122023  
Città: IL CO  
PCTC INFORMATONI E SERVIZI PER I STUDI  
P. IVA: 01546360156 - P. BANCA: ANTTAB  
C. RESEGUENTE: RPT

- |                    |  |
|--------------------|--|
| <b>Vista</b>       | la Legge 7.8.1990, n. 241 "Nuove norme sul procedimento amministrativo" e successive modifiche;  |
| <b>Vista</b>       | la Legge 19.11.1990, n. 341 "Riforme degli ordinamenti didattici universitari";  |
| <b>Vista</b>       | la Legge 2.6.1999, n. 264, come modificata dalla Legge 8 gennaio 2002, "Norme in materia di accessi ai corsi universitari";  |
| <b>Visto</b>       | il decreto legislativo del 30 giugno 2003, n. 196 'Codice in materia di dati personali';   |
| <b>Visto</b>       | il Decreto del Ministro dell'Università e della Ricerca del 16 marzo 2007 "Determinazione delle classi di laurea magistrale";  |
| <b>Visto</b>       | il D.M. 22 ottobre 2004, n. 270 'Modifiche al regolamento recante norme concernenti l'autonomia didattica degli atenei';   |
| <b>Vista</b>       | la Legge 30 dicembre 2010, n. 240 "Norme in materia di organizzazione delle università, di personale accademico e redazionale, nonché delega al Governo per incentivare la qualità e l'efficienza del sistema universitario";  |
| <b>Visti</b>       | il Regolamento didattico di Ateneo e il Regolamento degli Studenti di Ateneo;  |
| <b>Vieti</b>       | il parere espresso dal Consiglio della Scuola di Economia e Statistica il 10 gennaio 2017 o le deliberazioni dell'11 febbraio 2017 del Consiglio del Dipartimento di Economia, Metodi quantitativi, Strategie d'Impresa, del Consiglio del Dipartimento di Scienze Economico-Aziendali e Diritto per l'Economia e del Consiglio del Dipartimento di Statistica e Metodi Quantitativi, avvenuti per oggetto l'approvazione delle procedure di selezione per i corsi di laurea triennali di Economia e i corsi di laurea magistrale in Marketing e mercati globali e in Scienze economico-azientali per l'anno accademico 2017/2018; |
| <b>Viste</b>       | le deliberazioni del 23 gennaio 2017, n. 1°, del Senato Accademico e del 24 gennaio 2017, n. 52, del Consiglio di Amministrazione di questa Università re attive alla determinazione della programmazione locale, ai sensi dell'art. 2, etterna a) della legge 264/99 dei corsi di laurea e laurea magistrale già a Scuola di Economia e Statistica per l'anno accademico 2017/2018;   |
| <b>Visto</b>       | l'art. 5 (norme relative all'accesso) del Regolamento didattico del corso di laurea magistrale in Scienze economico-aziendali I per l'a.a. 2017/2018, che limita l'accesso al corso a 200 posti, dei quali 1 per studenti cinesi del progetto 'Marco Polo' e 3 per studenti extra Ue non residenti in Italia;  |
| <b>Visto</b>       | Il Bando per l'accesso al corso di laurea magistrale in Scienze economico-aziendali per l'a.a. 2017/2018, allegato al D.R. n. 43341/17 del 13 luglio 2017;   |
| <b>Verificato</b>  | che, dopo la pubblicazione delle graduatorie degli ammessi e la chiusura dei termini previsti per le iscrizioni, gli studenti immatricolati risultano essere 180;  |
| <b>Considerato</b> | che risultano scoperti 40 posti per l'accesso al primo anno;   |

## I formati *readable*

Più o meno *strutturati*...

- TXT
- Formati **binari**
  - hanno più problemi dovuti per esempio all'*endianness*
  - sono molto più compatti
  - richiedono una macchina per essere interpretati da un umano
- Formati **human readable**
  - possono essere interpretati (con diversa difficoltà) da un umano
  - sono meno compatti
  - nella pratica, con la compressione il problema dello spazio è molto gestibile

## Esempi di formati *human readable*

- CSV
- XML
- JSON e GeoJSON
- Frictionless datapackage di OKFN
- LOD:
  - JSON-LD
  - RDG
- PDF e OCR e Tabula

# TXT

- Il formato di testo è machine readable, ma in generale *non strutturato*.
- Non è necessariamente una cosa negativa: dipende da cosa rappresenta
- Restano alcuni problemi, legati soprattutto:
  - A come si va "a capo"
  - A come si codificano i caratteri, vedi:
    - <https://www.asciiitable.com/> (<https://www.asciiitable.com/>)
    - <http://www.unicode.org/> (<http://www.unicode.org/>)
    - [https://www.w3schools.com/charsets/ref\\_html\\_utf8.asp](https://www.w3schools.com/charsets/ref_html_utf8.asp)  
[\(\[https://www.w3schools.com/charsets/ref\\\_html\\\_utf8.asp\]\(https://www.w3schools.com/charsets/ref\_html\_utf8.asp\)\)](https://www.w3schools.com/charsets/ref_html_utf8.asp)

# CSV (Comma) Separated Values

- Formato **testuale**
- Formato **tabulare**
- Il separatore può essere la virgola, ma spesso è usato il punto e virgola, la barra verticale, ...
- Alcuni problemi del TXT restano
- Il separatore può confondersi con il separatore dei decimali
- Non è indicato il **tipo delle colonne**
- Non si possono gestire gerarchie
- Parametri importanti:
  - Separatore
  - Carattere che circonda le celle di testo (potrebbero contenere il separatore)
  - (eventuali) linee di Header

## **Andiamo a vedere un CSV**

<https://www.dati.lombardia.it/Solidariet-/Albo-Cooperative-Sociali-al-31-12-2015/tuar-wxya/data> (<https://www.dati.lombardia.it/Solidariet-/Albo-Cooperative-Sociali-al-31-12-2015/tuar-wxya/data>)

```
In [83]: coopSocialiF=open('download/Albo_Cooperative_Sociali_al_31.12.2015.csv','r')

for i in range(4):
    print (coopSocialiF.readline())
```

DENOMINAZIONE, COMUNE, PROVINCIA, TELEFONO, EMAIL, AREA, SERVIZI, ATTIVITA, TIPOLOGIE  
PERSONE SVANTAGGIATE, WGS84\_Y, WGS84\_X, location

ZEROGRAFICA SOCIETA' COOPERATIVA SOCIALE,Milano,MI,3387621099,,,Tipografia e  
stampa,Detenuti,45.5253903,9.0936669,"(45.5253903, 9.0936669)"

STELLA MARIS - SOCIETA' COOPERATIVA SOCIALE ONLUS,Mantova,MN,0376151018,info@co  
opstellamaris.it,"Anziani, Disabili/Handicappati, Famiglia, Psichiatria","Alt  
ro, Assistenza Domiciliare integrata Anziani, Assistenza Domiciliare integrata  
Disabili, Assistenza domiciliare pazienti psichiatrici",,45.1518998,10.782216  
5,"(45.1518998, 10.7822165)"

L'AQUILONE SOCIETA' COOPERATIVA SOCIALE,Sesto Calende,VA,03311830570,laquilon  
e.scs@laquilonescs.it,"Disabili/Handicappati, Disagio Giovanile, Educativa, Fami  
glia, Minori, Prevenzione del disagio, Stranieri","Animazione culturale e ter  
ritoriale, Assistenza Pre e Post Scolastica, Assistenza domiciliare minori, As  
sistenza scolastica ""ad personam"", Attività Extra scolastiche a minori, Cen  
tro di Aggregazione Giovanile, Educatori di Strada, Formazione - Consulenza - P  
rogettazione, Informagiovani - Spazio giovani, Progettazione servizi speriment  
ali",,45.7249613,8.6322301,"(45.7249613, 8.6322301)"

In [85]: coopSociali.head()

Out[85]:

	DENOMINAZIONE	COMUNE	PROVINCIA	TELEFONO	
0	ZEROGRAFICA SOCIETA' COOPERATIVA SOCIALE	Milano	MI	3387621099	NaN
1	STELLA MARIS - SOCIETA' COOPERATIVA SOCIALE ONLUS	Mantova	MN	0376151018	info@coopstell...
2	L'AQUILONE SOCIETA' COOPERATIVA SOCIALE	Sesto Calende	VA	03311830570	laquilonescs@l...
3	IL BUON PASTORE SOCIETA' COOPERATIVA SOCIALE -...	Milano	MI	3391607328	vincenzotufanc...

	DENOMINAZIONE	COMUNE	PROVINCIA	TELEFONO	
4	PROGETTO E LAVORO - SOCIETA' COOPERATIVA SOCIALE	Brescia	BS	0302524763	NaN

- Proviamo ad aprire il file su Google Drive. Usiamo ad esempio  
`IMPORTDATA("https://www.dati.lombardia.it/api/views/tuar-wxya/rows.csv?accessType=DOWNLOAD")`

# XML (eXtensible Markup Language)

- Linguaggio di *markup*: testo + tag
  - Es <persona><nome>Matteo</nome><cognome>Fortini</cognome></persona>
- Gerarchico
- Può rappresentare qualunque tipo di dato
- Spesso associato a una DTD Document Type Definition
- Linguaggio XQuery per l'analisi
- Molti formati di file sono basati su XML, es ODS/ODT,DOCX/XLSX, ...
- Formato molto importante: RSS

```
In [125]: coopSocialiFXML=open('download/Albo_Cooperative_Sociali_al_31.12.2015.xml','r')

coopSocialiFXMLTree=etree.parse(coopSocialiFXML)

print(etree.tostring(coopSocialiFXMLTree,pretty_print=True)[:1000].decode("asci
i"))

<response>
  <row>
    <row _id="3010" _uuid="556A6125-B90E-45D3-8FAE-7A6D88D53E0A" _position="30
10" _address="https://www.dati.lombardia.it/resource/tuar-wxya/3010">
      <denominazione>ZEROGRAFICA SOCIETA' COOPERATIVA SOCIALE</denominazione>
      <comune>Milano</comune>
      <provincia>MI</provincia>
      <telefono>3387621099</telefono>
      <attivita_>Tipografia e stampa</attivita_>
      <tipologie_persone_svantaggiate>Detenuti</tipologie_persone_svantaggiate
>
      <wgs84_y>45.5253903</wgs84_y>
      <wgs84_x>9.0936669</wgs84_x>
      <location human_address="{"address": "Via XX Settembre, 10, 20131 Milano, MI", "city": "Milano", "state": "Lombardia", "zip": "20131"}" latitude="45.5253903" longitude="9.0936669" needs_recoding="false"/>
    </row>
    <row _id="3011" _uuid="2899B26F-8A5B-47AD-9412-55A68C108590" _position="30
11" _address="https://www.dati.lombardia.it/resource/tuar-wxya/3011">
      <denominazione>STELLA MARIS - SOCIETA' COOPERATIVA SOCIALE ONLUS
```

- Anche questo può essere aperto in Google Drive con  
`IMPORTXML("https://www.dati.lombardia.it/api/views/tuar-wxya/rows.xml?accessType=DOWNLOAD"; "/response/row/row")`
- NB Notare che abbiamo dovuto aggiungere una *query xpath*
  - Il pregio è che cambiando la query possiamo selezionare solo una parte, esempio  
`/response/row/row/denominazione`

## RSS

- RSS è un dialetto di XML
- permette di distribuire aggiornamenti in modo semplice
- è pensato per facilitare la gestione automatizzata

<https://validator.w3.org/feed/docs/rss2.html>  
[\(https://validator.w3.org/feed/docs/rss2.html\)](https://validator.w3.org/feed/docs/rss2.html)

Esempi:

- <http://www.protezionecivile.gov.it/jcms/it/rss.wp>  
[\(http://www.protezionecivile.gov.it/jcms/it/rss.wp\)](http://www.protezionecivile.gov.it/jcms/it/rss.wp)
- <http://www.radio.rai.it/rss/podcast/rssradio.jsp?channel=RF2&n=&id=16269>  
[\(http://www.radio.rai.it/rss/podcast/rssradio.jsp?channel=RF2&n=&id=16269\)](http://www.radio.rai.it/rss/podcast/rssradio.jsp?channel=RF2&n=&id=16269)

```
In [130]: podcastTree=etree.parse("http://www.radio.rai.it/rss/podcast/rssradio.jsp?channel=RF2&n=&id=16269")
print(etree.tostring(podcastTree,pretty_print=True)[:1000].decode("utf-8"))
```

```
<rss xmlns:itunes="http://www.itunes.com/dtds/podcast-1.0.dtd" version="2.0">
  <channel>
    <title>pascal</title>
    <link>http://www.radio2.rai.it/dl/portaleRadio/Programmi/Page-0e06ff4d-fb20-4ef1-a675-0be791150831.html?set=ContentSet-c49babfb-8ac7-44b3-b461-9166ef14e18c&type>Main</link>
      <itunes:subtitle>Pascal &#232; il programma di Matteo Caccia in onda su Radio2 dal luned&#236; al venerd&#236; alle 22.30 che racconta storie di vita.</itunes:subtitle>
      <description>Pascal &#232; il programma di Matteo Caccia in onda su Radio2 dal luned&#236; al venerd&#236; alle 22.30 che racconta storie di vita. Episodi grandi o piccoli, stravolgenti o minimi, momenti che hanno modificato per sempre la nostra vita o che, anche se di poco, l&#146;hanno indirizzata. Storie che sono il termometro della temperatura di ognuno di noi e che in parte raccontano chi siamo.
      Quando eravamo bambini e scoprivamo un enorme segreto, quando quella notte in viaggio ci perdemmo, quando la incontrammo e
```

## Con i feed RSS possiamo collegare servizi usando ad esempio

- Un lettore di feed (es. Thunderbird)
- IFTTT <https://ifttt.com/feed> (<https://ifttt.com/feed>)
- Zapier <https://zapier.com/apps/rss/integrations> (<https://zapier.com/apps/rss/integrations>)

# **JSON (JavaScript Object Notation)**

<https://www.json.org/json-it.html> (<https://www.json.org/json-it.html>)

- Nasce per essere più sintetico di XML
- Rimane gerarchico

```
In [131]: coopSocialiFJSON=open('download/Albo_Cooperative_Sociali_al_31.12.2015.json',  
'r')  
  
coopSocialiFJSONtree=json.load(coopSocialiFJSON)  
  
print(json.dumps({"data":coopSocialiFJSONtree["data"][:2]},indent=4, sort_keys  
=True))
```

```
{  
    "data": [  
        [  
            3010,  
            "556A6125-B90E-45D3-8FAE-7A6D88D53E0A",  
            3010,  
            1456838409,  
            "52",  
            1456838409,  
            "52",  
            null,  
            "ZEROGRAFICA SOCIETA' COOPERATIVA SOCIALE",  
            "Milano",  
            "MI",  
            "3387621099",  
            null,  
            null,  
            null,  
            "Tipografia e stampa",  
            "Detenuti",  
            "45.5253903",  
            "9.0936669",  
            [  
                {"\\"address\\":\\"\\",\\"city\\":\\"\\",\\"state\\":\\"\\",\\"zip  
\":\\\"\\"},  
                "45.5253903",  
                "9.0936669",  
                null,
```

```
        false
    ],
    [
        3011,
        "2899B26F-8A5B-47AD-9412-55A68C108590",
        3011,
        1456838409,
        "52",
        1456838409,
        "52",
        null,
        "STELLA MARIS - SOCIETA' COOPERATIVA SOCIALE ONLUS",
        "Mantova",
        "MN",
        "0376151018",
        "info@coopstellamaris.it",
        "Anziani, Disabili/Handicappati, Famiglia, Psichiatria",
        "Altro, Assistenza Domiciliare integrata Anziani, Assistenza Domiciliare integrata Disabili, Assistenza domiciliare pazienti psichiatrici",
        null,
        null,
        "45.1518998",
        "10.7822165",
        [
            "{\"address\":\"\\\", \"city\":\"\\\", \"state\":\"\\\", \"zip
\\\":\\\"\\\"}\",
            "45.1518998",
            "10.7822165",
            null,
            false
        ]
    ]
}
```

# GeoJSON

<http://geojson.org/> (<http://geojson.org/>).

- È un'estensione di JSON creata per contenere *dati geografici*, ovvero

- Punti
- Linee
- Poligoni

che abbiano delle *coordinate geografiche*

*The coordinate reference system for all GeoJSON coordinates is a geographic coordinate reference system, using the World Geodetic System 1984 (WGS 84) [WGS84] datum, with longitude and latitude units of decimal degrees.*

## (Tabular) Data Package

- Dal progetto Frictionless data ([frictionlessdata.io](https://frictionlessdata.io)) di OKFN Open Knowledge Foundation
- <https://frictionlessdata.io/docs/tabular-data-package/> (<https://frictionlessdata.io/docs/tabular-data-package/>).
- CSV+JSON con uno schema di metadati per descrivere il contenuto del CSV

## Formati per LD (Linked Data)

- Basati su *triple* soggetto,predicato,oggetto
- Vengono create delle *ontologie* che definiscono relazioni e attributi in domini particolari
- Vengono utilizzate degli URI univoci che permettono di collegare basi dati fra di loro
- <https://lod-cloud.net/> (<https://lod-cloud.net/>)
- <http://lodview.it/> (<http://lodview.it/>). Esempio (<http://lodview.it/lodview/?IRI=http%3A%2F%2Fdbpedia.org%2Fresource%2FLondon&sparql=http%3A%2F%2F>)  
NB cliccare sui cerchietti in alto a destra per la visualizzazione grafica

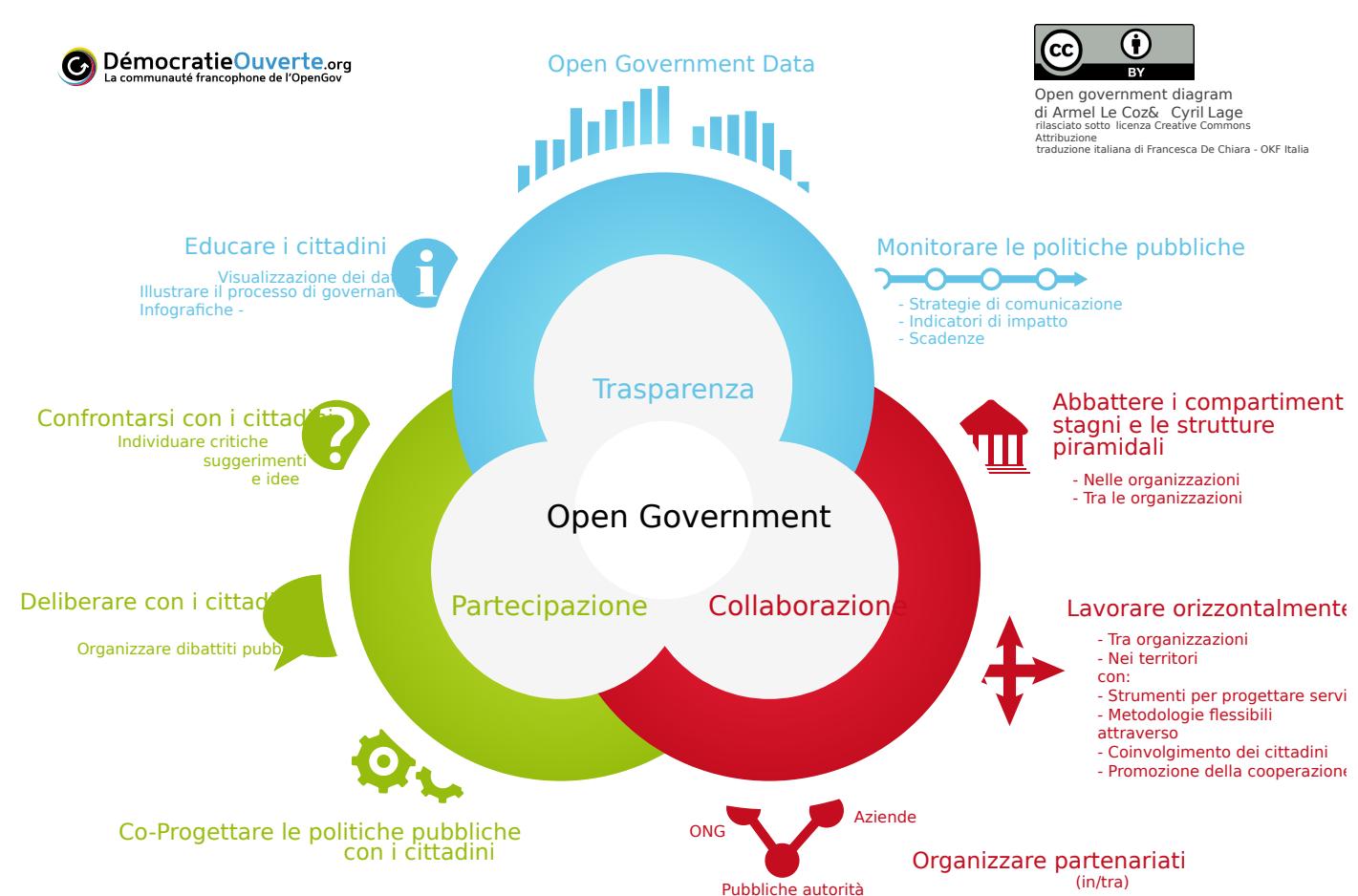
# Dove troviamo open data?

- Pubblici
- Volontari
- Privati

# Dati Pubblici

- I dati pubblici hanno due motivazioni fondamentali:
  - La **trasparenza**
  - La **restituzione** (nel senso che sono dati finanziati da denaro pubblico)

# Trasparenza e *open government*



# La trasparenza

Need to know  
diventa  
**Right to know**

Ho *diritto di sapere*, non ti devo dimostrare la mia necessità.

*The Three Laws of Open Government Data:*

- *If it can't be spidered or indexed, it doesn't exist*
- *If it isn't available in open and machine readable format, it can't engage*
- *If a legal framework doesn't allow it to be repurposed, it doesn't empower*

*David Eaves*

## Trasparenza in Italia

- Svezia 1766, "Freedom of the Press Act" nella Costituzione
- Finlandia 1951 "Laki yleisten asiakirjain julkisuudesta 9.2.1951/83"
- **Italia:**
  - Decreto "trasparenza" 33/2013 (<http://www.normattiva.it/uri-res/N2Ls?urn:nir:stato:decreto.legislativo:2013-03-14;33!vig=>)
  - Decreto legislativo "FOIA" 97/2016 (<http://www.foia.gov.it/foia/>)

# Trasparenza in Italia

Open by default

*Art. 52 del CAD - Codice Amministrazione Digitale*

- 1. I dati e i documenti che i soggetti di cui all'articolo 2, comma 2, pubblicano, con qualsiasi modalità, senza l'espressa adozione di una licenza di cui all'articolo 2, comma 1, lettera h), del decreto legislativo 24 gennaio 2006, n. 36, si intendono rilasciati come dati di tipo aperto ai sensi all'articolo 1, comma 1, lettere l-bis) e l-ter), del presente Codice, ad eccezione dei casi in cui la pubblicazione riguardi dati personali.*

## Quali esempi di Open data pubblici italiani?

### Dati pubblici per trasparenza

- Governo: [Italiasicura](http://italiasicura.governo.it)  
(<http://italiasicura.governo.it>)
- Parlamento:
  - [Camera](http://dati.camera.it/it/) (<http://dati.camera.it/it/>)
  - [Senato](http://dati.senato.it/sito/home) (<http://dati.senato.it/sito/home>)
- [Ministero dell'Interno](http://dait.interno.gov.it/) (<http://dait.interno.gov.it/>).

## Quali esempi di Open data pubblici italiani?

### Dati pubblici per trasparenza

- MIT (<http://dati.mit.gov.it/catalog/dataset>)
- MIUR (<http://dati.istruzione.it/opendata/>)
- Ministero della Salute  
(<http://www.dati.salute.gov.it/dati/homeDataset.jsp>)
- Normattiva (<http://www.normattiva.it>)
- OpenCUP (<http://opencup.gov.it/>)

## Quali esempi di Open data pubblici italiani?

### Dati pubblici per trasparenza

- MISE (<http://www.sviluppoeconomico.gov.it/index.php/it/open-data>)
- MEF (<http://www.sviluppoeconomico.gov.it/index.php/it/open-data>)
- MiBACT (<http://www.beniculturali.it/mibac/export/MiBAC/sito-MiBAC/MenuPrincipale/Trasparenza/Open-Data/index.html>)
- SIOPE (<https://www.siope.it/Siope/> <http://soldipubblici.gov.it/it/home>)

## Quali esempi di Open data pubblici italiani?

### Dati pubblici perché finanziati con denaro pubblico

- ISTAT (<http://dati.istat.it/>) Preziosissimo Basi territoriali e variabili censuarie (<https://www.istat.it/it/archivio/104317>).
- ISPRA (<http://www.isprambiente.gov.it/it/banche-dati>).
- Copernicus (<http://copernicus.eu/data-access>).
- ARPA Lombardia (<https://www.datilombardia.it/Government/ARPA-LOMBARDIA-elenco-dataset-pubblicati/8ask-gxyr/data>).

## **Portali di dati pubblici nazionali**

- Dati Gov (<http://dati.gov.it/>)
- RNDT Repertorio Nazionale Dati Territoriali  
(<http://geodati.gov.it/geoportale/>)
- ANBSC (<http://www.benisequestraticonfiscati.it/>)

## Basi dati di interesse nazionale

- [Lista](https://www.agid.gov.it/it/dati/basi-dati-interesse-nazionale) (<https://www.agid.gov.it/it/dati/basi-dati-interesse-nazionale>) definita da AgID

## **Detto questo, come siamo messi?**

- Open Data Index  
(<https://index.okfn.org/>).

## Altri Open data

- [Lista da Nature](https://www.nature.com/sdata/policies/repositories) (<https://www.nature.com/sdata/policies/repositories>)
- [Awesome Public Datasets](https://github.com/awesomedata/awesome-public-datasets) (<https://github.com/awesomedata/awesome-public-datasets>)
- [Kaggle](https://www.kaggle.com/datasets) (<https://www.kaggle.com/datasets>)
- [Project Gutenberg](http://www.gutenberg.org/) (<http://www.gutenberg.org/>)

## **Dati "volontari"**

Molti progetti producono dati in modo collaborativo e volontario.

# Wikipedia

- Enciclopedia libera
- Diventa una base dati LOD:
  - DBpedia (<http://it.dbpedia.org/sparql>), proviamo anche il chatbot <http://chat.dbpedia.org/> (<http://chat.dbpedia.org/>).
  - Wikidata (<https://query.wikidata.org/>).

# OpenStreetMap

- Mappa collaborativa, ma anche DB geografico
- Licenza ODBL (share alike)
- Fornisce un motore di interrogazione: [Overpass](https://overpass-turbo.eu/) (<https://overpass-turbo.eu/>)
- Esistono tool per collaborare in emergenza: [HOT](https://www.hotosm.org/) (<https://www.hotosm.org/>)
- Esistono mappe tematiche per:
  - Bici/trekking: [OpenCycleMap](https://www.opencyclemap.org/) (<https://www.opencyclemap.org/>).
  - Disabilità motorie: [Wheelmap](https://wheelmap.org/) (<https://wheelmap.org/>?lat=45.4613424441503&lon=9.159498500000009&q=milan&zoom=12).
  - [Emergenze](https://wambachers-osm.website/emergency/) (<https://wambachers-osm.website/emergency/>#zoom=12&lat=45.469&lon=9.1914&layer=OpenStreetMap)
- Esistono tool per il routing: [OpenRouteService](https://maps.openrouteservice.org/) (<https://maps.openrouteservice.org/>)

## Altri dati da volontari

- [data.world](https://data.world/) (<https://data.world/>)
- [Internet Archive](http://www.archive.org/) (<http://www.archive.org/>)
- [Albo POP](http://albopop.it) (<http://albopop.it>)
- [Protezione Civile POP](http://www.protezionecivilepop.tk/) (<http://www.protezionecivilepop.tk/>)
- [TerremotoCentroItalia.info](http://www.terremotocentroitalia.info)  
(<http://www.terremotocentroitalia.info>)
- [ItaliaAFuoco.info](http://www.italiaafuoco.info) (<http://www.italiaafuoco.info>)
- [Politicamente Corretto](http://www.unapromessa.it) (<http://www.unapromessa.it>)

# Progetti che usano open data

- SOD - Spaghetti Open Data (<http://spaghettiopendata.org/>)
- Monithon (<http://www.monithon.it/>)
- Open Ricostruzione (<https://openricostruzione.regione.emilia-romagna.it/>)
- Open Bilanci (<https://openbilanci.it/>)
- Confiscati Bene (<http://www.confiscatibene.it/>)
- Contratti Pubblici (<https://contrattipubblici.org/>)

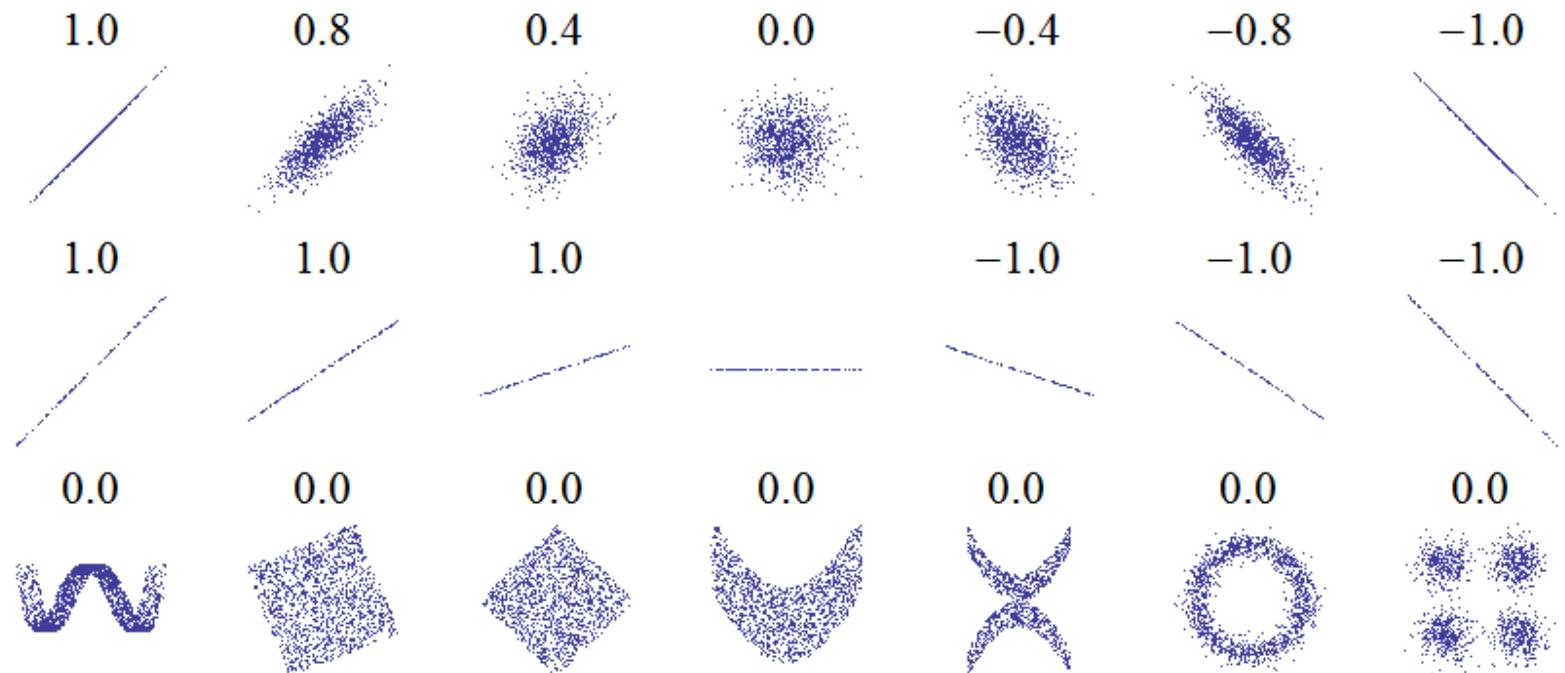
# **Analisi dei dati**

## Media, moda, mediana, varianza

- **Media:**  $\mu_X = \frac{\sum_{i=1}^N x_i}{N}$
- **Deviazione standard:**  $\sigma_X = \sqrt{\frac{\sum_i (x_i - \mu_X)^2}{n}}$
- **Moda:** valore più frequente
- **Mediana:** valore che ha il 50% di probabilità

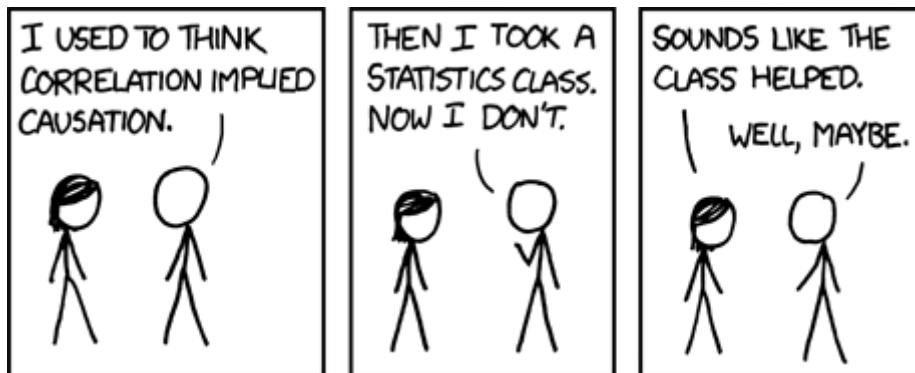
# Correlazione

$$-1 \leq \rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{\sum_{i=1}^n (x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^n (x_i - \mu_x)^2} \sqrt{\sum_{i=1}^n (y_i - \mu_y)^2}} \leq +1$$



## Correlation is not causation

- Spurious Correlations (<http://www.tylervigen.com/spurious-correlations>)



(XKCD)

## Spunti di riflessione

- Il nostro diritto digitale alla città (<http://cittadigitale.openpolis.it/>)
- Rosy Battaglia: Nova 24 su Open Government  
(<http://www.rosybattaglia.it/nova-il-sole-24-ore/>)

# Data Visualization

# Machine learning

# **Confusion matrix**

## Precisione di un test

# **Social Network Analysis**

# Come diventare produttori di dati?

- Drive
- Dropbox
- File statici
- Archive.org
- data.world

## **II** *Civic Hacking*