

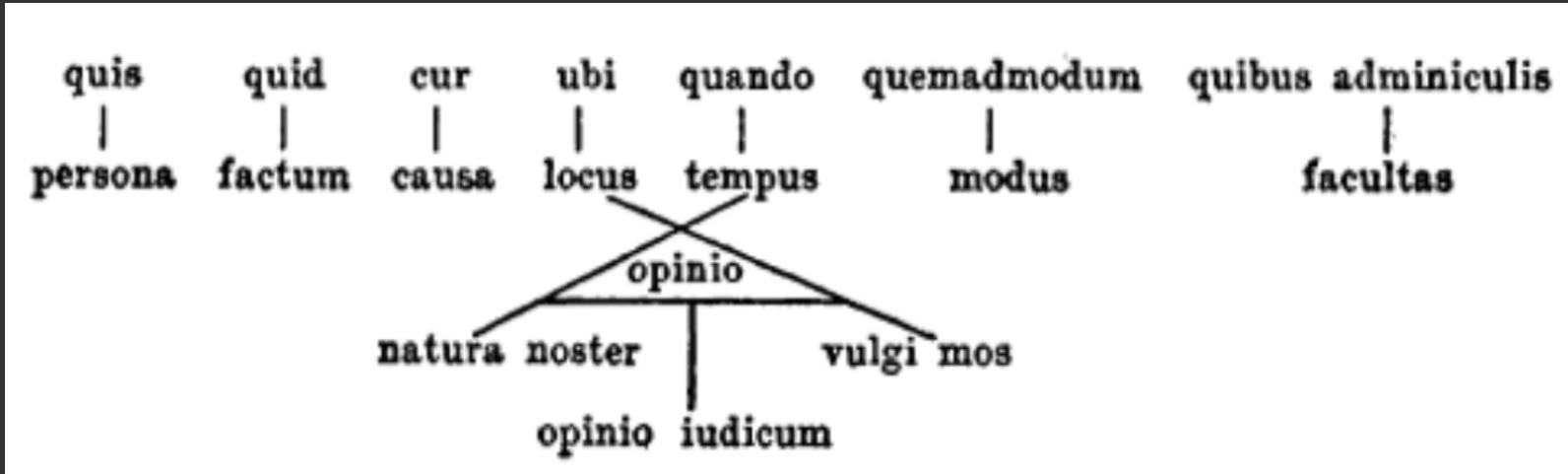


Spreadsheets are your friends

(and your data will love them)

Cristian Consonni
Fondazione Bruno Kessler
20 dicembre 2013
School of data, Trento

Struttura della notizia

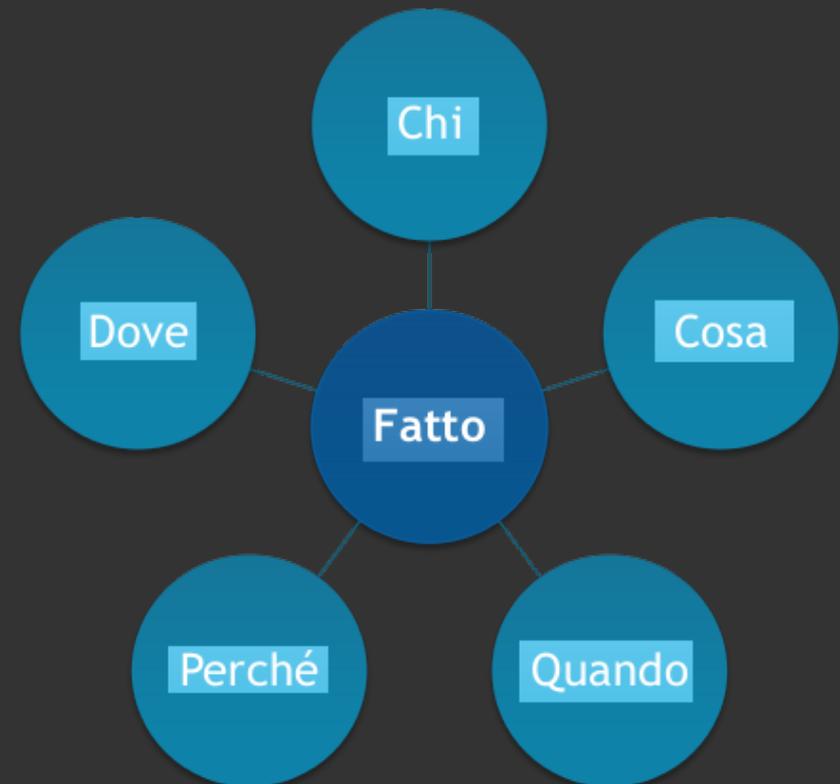


Le 5 “W”:

- ◆ Who is it about?
- ◆ What happened?
- ◆ When did it take place?
- ◆ Where did it take place?
- ◆ Why did it happen?

Struttura della notizia

- Ogni aspetto di una notizia può essere tradotto in un dato ↔ un dato può essere incorporato in un aspetto della notizia;
- Ogni colonna è una dimensione dei dati;
- I dati devono aiutare a rispondere alle domande precedenti;



«The problem I have is that the worldview that my students have correspond to reality in the world the year their teachers were born»

<http://www.gapminder.org/videos/ted-us-state-department/>

Perché i dati?



«Software is what the 21st century is made of.

What steel was to the economy of the 20th century.

What steel was to the power of the 20th century

*What steel was to the politics of the 20th century,
software is now.*

*It's the crucial building block, the component out of
which everything else is made.*

*And when I speak of everything, else I mean, of course,
freedom.»*

Tratto da:

“Why Political Liberty Depends on
Software Freedom More Than Ever”

Eben Moglen @ 2011 FOSDEM
conference in Brussels on Feb 5,
2011

Esercitazione

D A T A
P I P E L I N E

Data pipeline: summary

- Data pipeline I: acquisition
- Data pipeline II: cleaning
- Data pipeline III: analysis
- Data pipeline IV: visualizing

- Human-readable

Dati leggibili facilmente da un umano, per esempio, una pagina di Wikipedia.

Internet

Da Wikipedia, l'enciclopedia libera.

Wikidata: [Internet \(Q75\)](#), rete mondiale di reti di computer ad accesso pubblico
Alias: Nessuno

 *Disambiguazione* – Se stai cercando la tipologia di rete di computer, vedi **Internet (informatica)**.

Internet^[1] (contrazione della locuzione inglese *interconnected networks*, ovvero "reti interconnesse") è una rete mondiale di reti di computer ad accesso pubblico, attualmente rappresentante il principale mezzo di comunicazione di massa, che offre all'utente una vasta serie di contenuti potenzialmente informativi e servizi.



- Machine-readable

«Formats that are machine readable are ones which are able to have their data extracted by computer programs easily. [...] Common machine-readable file formats are CSV files.»

da <http://schoolofdata.org/handbook/appendix/glossary/#term-machine-readable>

• *Data acquisition: obiettivo finale*

La data acquisition consiste nell'ottenere dei dati in formato *machine-readable*

Metodi:

- Scaricare dataset da portali open-data (facile)
- Scraping di pagine web (medio)
- Scraping di PDF (difficile)

Acquisition: good questions

- Chi ha prodotto i dati? Un ente pubblico? Un'azienda? (affidabilità)
- Come sono stati prodotti i dati? Il processo di raccolta dati è documentato?
- È possibile ottenere gli stessi dati (o almeno dati simili) in altri modi? È possibile confrontare dati di dettaglio con dati aggregati?

Accedi | Iscriviti

OPENdata TRENTINO beta

Dataset Organizzazioni Categorie Apps Informazioni FAQ

Dati Aperti del Trentino. Tutti i dati che cercavi del Sistema Trentino.

DOP, IGT, SGP
i prodotti tipici che danno sapore alla Provincia di Trento
Clicca qui per scoprire di più

Tags popolari

- società
- settori economici
- mercato del lavoro
- popolazione
- agricoltura
- servizi
- silvicoltura
- pesca
- sistema economico s...
- istruzione e formaz...

Cerca dati

Cerca i dati, e ottieni gli aggiornamenti per i dataset a cui sei interessato.

Cerca

Ultime modifiche

Retribuzione personale provinciale
Retribuzione del personale della Provincia Autonoma di Trento per qualifica
csv-semicolon delimited

Incentivi LP 14/80
L'archivio contiene i dati relativi agli interventi di efficientamento energetico e di produzione di energia rinnovabili incentivati dalla

Tweet

Federico Sannicolò @sannicolof
Sto installando @OpenRefine per prepararmi workshop @SchoolOfData di domani a Trento con ansia di partecipare! @DatiTrentinoit
↳ Ritwittato da dati.trentino.it
Espandi

dati.trentino.it @DatiTrentinoit
Quota 600 per dati.trentino.it Retribuzioni personale provinciale dati.trentino.it/dataset/retrib... #opendatatre #opendataitaly

Twitta a @DatiTrentinoit

Accedi | Iscriviti

OPENdata TRENTINO beta

Dataset Organizzazioni Categorie Apps Informazioni FAQ

Dati Aperti del Trentino. Tutti i dati che cercavi del Sistema Trentino.

Organizzazioni / PAT S. Gestione Strade / Riassunto rilievo traffico ...

Riassunto rilievo traffico automatico (stazioni fisse) anno 2011

Sostenitori **0**

Organizzazione

PAT S. Gestione Strade
Servizio Gestione Strade
L'obiettivo generale del Servizio Gestione Strade è quella di garantire la mobilità dell'utenza stradale sul territorio trentino in condizioni di... leggi di più

Categorie
Mobilità

Riassunto rilievo traffico automatico (stazioni fisse) anno 2011

Servizio Gestione Strade - Rilievo traffico anno 2011 -

Nel corso dell'anno 2011 si è svolto un censimento della circolazione stradale usando dei sensori automatici di rilevamento traffico posizionati in punti strategici della rete viaria provinciale. Il risultato di questo censimento è disponibile presso il Servizio Gestione Strade e contiene, per ogni passaggio memorizzato, le seguenti informazioni:

- Ora di transito del veicolo
- Direzione di marcia
- Tipologia del veicolo (9 categorie)
- Velocità di transito sul sensore

Questo dataset rappresenta solo una sintesi giornaliera del dato originale disponibile presso il Servizio Gestione Strade previa sottoscrizione modulo di richiesta dati./delibera PAT 2689 dd 27-10-2000)

La titolarità piena ed esclusiva del dataset "Riassunto rilievo traffico automatico (stazioni fisse) anno 2011" è della Provincia Autonoma di Trento (Licenziante), ai sensi della L. 633/41 e s.m.i.

La Provincia Autonoma di Trento autorizza la libera e gratuita consultazione, estrazione, riproduzione e modifica dei dati in essa contenuti da parte di chiunque (Licenziatario) vi abbia interesse per qualunque fine, purché nel rispetto dei termini della licenza Creative Commons - Attribuzione 2.5 Italia (testo integrale: <http://creativecommons.org/licenses/by/2.5/it/legalcode>).

L'attribuzione dovrà fornire una menzione adeguata di:

- Autore originale e/o titolare dei diritti: **Dossi Marco**
- Terze parti designate, se esistenti
- Nome della Banca Dati: **Riassunto rilievo traffico automatico (stazioni fisse) anno 2011**
- per utenti Intranet PAT e comuni PAT : <http://172.17.36.96/TRAFFICO/pratiche.html> ove ciò sia ragionevolmente possibile;

CSV (formato testo)

http://dati.trentino.it/it/storage/f/2013-11-11T155543/riassunto_dati_traffico_anno_2011.csv

```
riassunto_dati_traffico_anno_2011.csv (~/Scrivania/schoolofdata/data) - gedit
File Modifica Visualizza Cerca Strumenti Documenti Aiuto
Apri Salva Annulla
riassunto_dati_traffico_anno_2011.csv x
1 Codice Punto, strada, km, localita, Data, Totale, Motocicli, Autovetture e monovolumi, Autovetture e monovolumi con
  rimorchio, Furgoni, Autocarro medio (fino a 8_7 m), Autocarro grande (da 8_7 m), Autocarro con rimorchio, Trattore con semirimorchio, Autobus
2 101, SP 1,2+300, Calceranica, 01/01/2011 0:00:00, 4272.00, 44.00, 4133.00, 1.00, 70.00, 9.00, 2.00, 1.00, 1.00, 11.00
3 101, SP 1,2+300, Calceranica, 02/01/2011 0:00:00, 5342.00, 38.00, 5167.00, 7.00, 93.00, 19.00, 3.00, 1.00, 1.00, 13.00
4 101, SP 1,2+300, Calceranica, 03/01/2011 0:00:00, 5855.00, 41.00, 5488.00, 15.00, 191.00, 62.00, 26.00, 2.00, 9.00, 21.00
5 101, SP 1,2+300, Calceranica, 04/01/2011 0:00:00, 6457.00, 53.00, 6006.00, 21.00, 184.00, 95.00, 53.00, 8.00, 17.00, 20.00
6 101, SP 1,2+300, Calceranica, 05/01/2011 0:00:00, 6540.00, 51.00, 6118.00, 22.00, 209.00, 78.00, 23.00, 8.00, 8.00, 23.00
7 101, SP 1,2+300, Calceranica, 06/01/2011 0:00:00, 4222.00, 60.00, 4025.00, 8.00, 84.00, 22.00, 12.00, 0.00, 1.00, 10.00
8 101, SP 1,2+300, Calceranica, 07/01/2011 0:00:00, 5859.00, 49.00, 5503.00, 17.00, 158.00, 63.00, 38.00, 2.00, 9.00, 20.00
9 101, SP 1,2+300, Calceranica, 08/01/2011 0:00:00, 5834.00, 48.00, 5540.00, 17.00, 141.00, 39.00, 18.00, 0.00, 10.00, 21.00
10 101, SP 1,2+300, Calceranica, 09/01/2011 0:00:00, 4422.00, 50.00, 4262.00, 3.00, 84.00, 8.00, 2.00, 3.00, 3.00, 7.00
11 101, SP 1,2+300, Calceranica, 10/01/2011 0:00:00, 6210.00, 43.00, 5710.00, 6.00, 256.00, 100.00, 42.00, 7.00, 19.00, 27.00
12 101, SP 1,2+300, Calceranica, 11/01/2011 0:00:00, 6477.00, 56.00, 5921.00, 20.00, 252.00, 113.00, 52.00, 12.00, 24.00, 27.00
13 101, SP 1,2+300, Calceranica, 12/01/2011 0:00:00, 6495.00, 62.00, 5899.00, 21.00, 310.00, 113.00, 35.00, 9.00, 22.00, 24.00
14 101, SP 1,2+300, Calceranica, 13/01/2011 0:00:00, 6456.00, 56.00, 5866.00, 20.00, 287.00, 128.00, 36.00, 8.00, 25.00, 30.00
15 101, SP 1,2+300, Calceranica, 14/01/2011 0:00:00, 7148.00, 77.00, 6490.00, 31.00, 306.00, 128.00, 54.00, 7.00, 25.00, 30.00
16 101, SP 1,2+300, Calceranica, 15/01/2011 0:00:00, 6475.00, 96.00, 6107.00, 8.00, 165.00, 54.00, 11.00, 2.00, 11.00, 21.00
17 101, SP 1,2+300, Calceranica, 16/01/2011 0:00:00, 5414.00, 75.00, 5217.00, 6.00, 90.00, 14.00, 3.00, 1.00, 0.00, 8.00
18 101, SP 1,2+300, Calceranica, 17/01/2011 0:00:00, 6225.00, 57.00, 5622.00, 15.00, 292.00, 129.00, 48.00, 8.00, 23.00, 31.00
19 101, SP 1,2+300, Calceranica, 18/01/2011 0:00:00, 6597.00, 51.00, 5874.00, 33.00, 370.00, 138.00, 66.00, 10.00, 24.00, 31.00
20 101, SP 1,2+300, Calceranica, 19/01/2011 0:00:00, 6416.00, 73.00, 5791.00, 24.00, 307.00, 118.00, 30.00, 9.00, 21.00, 23.00
21 101, SP 1,2+300, Calceranica, 20/01/2011 0:00:00, 6724.00, 71.00, 6053.00, 23.00, 312.00, 129.00, 53.00, 10.00, 24.00, 31.00
```



LibreOffice 4.1
The Document Foundation

**Usiamo
LibreOffice:**
www.libreoffice.org

FREE OFFICE SUITE

LibreOffice 4: The free office suite the community has been dreaming of for twelve years.



DOWNLOAD
LIBREOFFICE
NOW!

import nel foglio di calcolo (II)

Aprire il CSV con LibreOffice Calc:
parte la procedura guidata

Importazione testo - [riassunto_dati_traffico_anno_2011.csv]

Importa

Tipo di carattere: Unicode (UTF-8)

Lingua: Predefinita - Italiano (Italia)

Dalla riga: 1

Opzioni di sillabazione

Larghezza fissa

Separato

Tabulazione Virgola Altri

Punto e virgola Spazio

Raggruppa i separatori di campo Separ. di testo: "

Altre opzioni

Campo tra virgolette come testo

Individua numeri speciali

Campi

Tipo colonna: Standard

	Standard	Standard	Standard	Standard	Standard	Standard
1	Codice Punto	strada	km	localita	Data	Totale
2	101	SP 1	2+300	Calceranica	01/01/2011 0:00:00	4272.00
3	101	SP 1	2+300	Calceranica	02/01/2011 0:00:00	5342.00
4	101	SP 1	2+300	Calceranica	03/01/2011 0:00:00	5855.00
5	101	SP 1	2+300	Calceranica	04/01/2011 0:00:00	6457.00
6	101	SP 1	2+300	Calceranica	05/01/2011 0:00:00	6540.00
7	101	SP 1	2+300	Calceranica	06/01/2011 0:00:00	4222.00
8	101	SP 1	2+300	Calceranica	07/01/2011 0:00:00	5850.00

OK

Annulla

?

import nel foglio di calcolo (III)

riassunto_dati_traffico_anno_2011.csv - LibreOffice Calc

File Modifica Visualizza Inserisci Formato Strumenti Dati Finestra ?

Arial 10

A1

	A	B	C	D	E	F	G	H	I	J
	Codice Punto	strada	km	localita	Data	Totale	Motocicli	Autovetture e monovolumi	Autovetture e monovolumi con rimorchio	Furgoni
1	101 SP 1	2+300	Calceranica	01/01/2011 0:00:00	4272.00	44.00	4133.00	1.00	70.00	9.00
2	101 SP 1	2+300	Calceranica	02/01/2011 0:00:00	5342.00	38.00	5167.00	7.00	93.00	19.00
3	101 SP 1	2+300	Calceranica	03/01/2011 0:00:00	5855.00	41.00	5488.00	15.00	191.00	66.00
4	101 SP 1	2+300	Calceranica	04/01/2011 0:00:00	6457.00	53.00	6006.00	21.00	184.00	92.00
5	101 SP 1	2+300	Calceranica	05/01/2011 0:00:00	6540.00	51.00	6118.00	22.00	209.00	78.00
6	101 SP 1	2+300	Calceranica	06/01/2011 0:00:00	4222.00	60.00	4025.00	8.00	84.00	22.00
7	101 SP 1	2+300	Calceranica	07/01/2011 0:00:00	5859.00	49.00	5503.00	17.00	158.00	63.00
8	101 SP 1	2+300	Calceranica	08/01/2011 0:00:00	5834.00	48.00	5540.00	17.00	141.00	38.00
9	101 SP 1	2+300	Calceranica	09/01/2011 0:00:00	4422.00	50.00	4262.00	3.00	84.00	8.00
10	101 SP 1	2+300	Calceranica	10/01/2011 0:00:00	6210.00	43.00	5710.00	6.00	256.00	10.00
11	101 SP 1	2+300	Calceranica	11/01/2011 0:00:00	6477.00	56.00	5921.00	20.00	252.00	11.00
12	101 SP 1	2+300	Calceranica	12/01/2011 0:00:00	6495.00	62.00	5899.00	21.00	310.00	12.00
13	101 SP 1	2+300	Calceranica	13/01/2011 0:00:00	6456.00	56.00	5966.00	20.00	287.00	12.00
14	101 SP 1	2+300	Calceranica	14/01/2011 0:00:00	7148.00	77.00	6490.00	31.00	306.00	12.00
15	101 SP 1	2+300	Calceranica	15/01/2011 0:00:00	6475.00	96.00	6107.00	8.00	165.00	54.00
16	101 SP 1	2+300	Calceranica	16/01/2011 0:00:00	5414.00	75.00	5217.00	6.00	90.00	17.00
17	101 SP 1	2+300	Calceranica	17/01/2011 0:00:00	6225.00	57.00	5622.00	15.00	292.00	12.00
18	101 SP 1	2+300	Calceranica	18/01/2011 0:00:00	6597.00	51.00	5874.00	33.00	370.00	12.00
19	101 SP 1	2+300	Calceranica	19/01/2011 0:00:00	6416.00	73.00	5791.00	34.00	297.00	11.00
20	101 SP 1	2+300	Calceranica	20/01/2011 0:00:00	6724.00	71.00	6051.00	29.00	321.00	12.00
21	101 SP 1	2+300	Calceranica	21/01/2011 0:00:00	7094.00	61.00	6436.00	27.00	312.00	12.00
22	101 SP 1	2+300	Calceranica	22/01/2011 0:00:00	6476.00	79.00	6196.00	6.00	151.00	44.00
23	101 SP 1	2+300	Calceranica	23/01/2011 0:00:00	5194.00	66.00	5034.00	9.00	66.00	7.00
24	101 SP 1	2+300	Calceranica	24/01/2011 0:00:00	6330.00	63.00	5699.00	11.00	296.00	10.00
25	101 SP 1	2+300	Calceranica	25/01/2011 0:00:00	6669.00	79.00	5964.00	20.00	313.00	10.00
26	101 SP 1	2+300	Calceranica	26/01/2011 0:00:00	6470.00	74.00	5793.00	17.00	320.00	10.00
27	101 SP 1	2+300	Calceranica	27/01/2011 0:00:00	6698.00	56.00	5987.00	24.00	355.00	10.00
28	101 SP 1	2+300	Calceranica	28/01/2011 0:00:00	7132.00	72.00	6450.00	23.00	310.00	10.00
29	101 SP 1	2+300	Calceranica	29/01/2011 0:00:00	6574.00	92.00	6193.00	23.00	156.00	5.00
30	101 SP 1	2+300	Calceranica	30/01/2011 0:00:00	4976.00	76.00	4792.00	0.00	73.00	12.00
31	101 SP 1	2+300	Calceranica	31/01/2011 0:00:00	6320.00	73.00	5687.00	25.00	288.00	12.00
32	101 SP 1	2+300	Calceranica	01/02/2011 0:00:00	6600.00	58.00	5939.00	29.00	288.00	10.00
33	101 SP 1	2+300	Calceranica	02/02/2011 0:00:00	6712.00	67.00	6055.00	25.00	312.00	10.00
34	101 SP 1	2+300	Calceranica	03/02/2011 0:00:00	6780.00	76.00	6092.00	17.00	290.00	10.00
35	101 SP 1	2+300	Calceranica	04/02/2011 0:00:00	7555.00	66.00	6811.00	30.00	339.00	10.00
36	101 SP 1	2+300	Calceranica	05/02/2011 0:00:00	6619.00	115.00	6249.00	22.00	153.00	40.00
37	101 SP 1	2+300	Calceranica	06/02/2011 0:00:00	5777.00	123.00	5503.00	8.00	108.00	2.00
38	101 SP 1	2+300	Calceranica	07/02/2011 0:00:00	6349.00	82.00	5693.00	21.00	290.00	10.00
39	101 SP 1	2+300	Calceranica	08/02/2011 0:00:00	6541.00	77.00	5833.00	16.00	308.00	10.00
40	101 SP 1	2+300	Calceranica	09/02/2011 0:00:00	6617.00	70.00	5984.00	24.00	279.00	10.00
41	101 SP 1	2+300	Calceranica	10/02/2011 0:00:00	6646.00	90.00	5964.00	47.00	272.00	14.00
42	101 SP 1	2+300	Calceranica	11/02/2011 0:00:00	7294.00	96.00	6589.00	31.00	294.00	10.00
43	101 SP 1	2+300	Calceranica	12/02/2011 0:00:00	6884.00	125.00	6438.00	33.00	174.00	60.00
44	101 SP 1	2+300	Calceranica	13/02/2011 0:00:00	5003.00	55.00	4854.00	2.00	67.00	11.00
45	101 SP 1	2+300	Calceranica	14/02/2011 0:00:00	6773.00	95.00	6089.00	16.00	314.00	12.00

Foglio 1 / 1

- Nuovo
- Apri... Ctrl+O
- Documenti recenti
- Procedure guidate
- Chiudi
- Salva Ctrl+S
- Salva con nome... Ctrl+Maiusc+S
- Salva tutto
- Ricarica
- Versioni...
- Esporta...
- Esporta nel formato PDF...
- Invia
- Proprietà...
- Firme digitali...
- Modelli
- Anteprima nel browser web
- Anteprima di stampa
- Stampa... Ctrl+P
- Impostazioni stampante...
- Esci Ctrl+Q

Salviamo una copia.
Best practice: conservare sempre i
 dati originali!

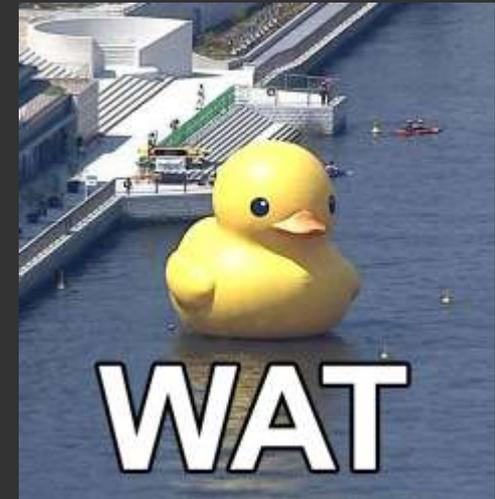
Data type (I)

Facciamo delle somme

28304	177	S	15/12/2012 0:00:00	1223.00	34.00	1105.00	8.00
28305	177	S	16/12/2012 0:00:00	2132.00	35.00	2026.00	7.00
28306	177	S	17/12/2012 0:00:00	1243.00	32.00	1110.00	2.00
28307	177	S	18/12/2012 0:00:00	1387.00	28.00	1213.00	7.00
28308	177	S	19/12/2012 0:00:00	1496.00	37.00	1319.00	7.00
28309	177	S	20/12/2012 0:00:00	1397.00	28.00	1244.00	8.00
28310	177	S	21/12/2012 0:00:00	1468.00	57.00	1278.00	10.00
28311	177	S	22/12/2012 0:00:00	1620.00	50.00	1507.00	7.00
28312	177	SS	23/12/2012 0:00:00	2187.00	61.00	2083.00	4.00
28313	177	SS	24/12/2012 0:00:00	1356.00	34.00	1252.00	7.00
28314	177	SS	25/12/2012 0:00:00	1181.00	31.00	1134.00	0.00
28315	177	SS	26/12/2012 0:00:00	1367.00	45.00	1299.00	4.00
28316	177	SS	27/12/2012 0:00:00	1465.00	32.00	1349.00	5.00
28317	177	SS	28/12/2012 0:00:00	1718.00	41.00	1559.00	15.00
28318	177	SS	29/12/2012 0:00:00	1804.00	57.00	1677.00	15.00
28319	177	SS	30/12/2012 0:00:00	2106.00	54.00	2005.00	4.00
28320	177	SS	31/12/2012 0:00:00	1816.00	57.00	1687.00	5.00
28321	TOTALE				0		0

Facciamo delle somme

28304	177	S	15/12/2012 0:00:00	1223.00	34.00	1105.00	8.00
28305	177	S	16/12/2012 0:00:00	2132.00	35.00	2026.00	7.00
28306	177	S	17/12/2012 0:00:00	1243.00	32.00	1110.00	2.00
28307	177	S	18/12/2012 0:00:00	1387.00	28.00	1213.00	7.00
28308	177	S	19/12/2012 0:00:00	1496.00	37.00	1319.00	7.00
28309	177	S	20/12/2012 0:00:00	1397.00	28.00	1244.00	8.00
28310	177	S	21/12/2012 0:00:00	1468.00	57.00	1278.00	10.00
28311	177	S	22/12/2012 0:00:00	1620.00	50.00	1507.00	7.00
28312	177	SS	23/12/2012 0:00:00	2187.00	61.00	2083.00	4.00
28313	177	SS	24/12/2012 0:00:00	1356.00	34.00	1252.00	7.00
28314	177	SS	25/12/2012 0:00:00	1181.00	31.00	1134.00	0.00
28315	177	SS	26/12/2012 0:00:00	1367.00	45.00	1299.00	4.00
28316	177	SS	27/12/2012 0:00:00	1465.00	32.00	1349.00	5.00
28317	177	SS	28/12/2012 0:00:00	1718.00	41.00	1559.00	15.00
28318	177	SS	29/12/2012 0:00:00	1804.00	57.00	1677.00	15.00
28319	177	SS	30/12/2012 0:00:00	2106.00	54.00	2000.00	4.00
28320	177	SS	31/12/2012 0:00:00	1816.00	57.00	1687.00	5.00
28321	TOTALE				0		0



È un problema di rappresentazione dei numeri da cui discende un problema con il formato dei dati.

Data type (II)

Formatta celle

Protezione celle

Numeri Carattere Effetti carattere Allineamento Tipografia asiatica Bordo Sfondo

Categoria	Formato	Lingua
Numero	General	Inglese (USA)
Percentuale	-1234	Inglese (USA)
Valuta	-1234.12	Inglese (Zimbabwe)
Data	-1,234	Interlingua
Orario	-1,234.12	Irlandese
Scientifico	-1,234.12	Islandese
Frazione	(1,234)	Italiano (Italia)
Valore booleano	(1,234.12)	Italiano (Svizzera)

Opzioni

Posizioni decimali: Valori negativi in rosso

Zeri iniziali: Separatore delle migliaia

Codice del formato:

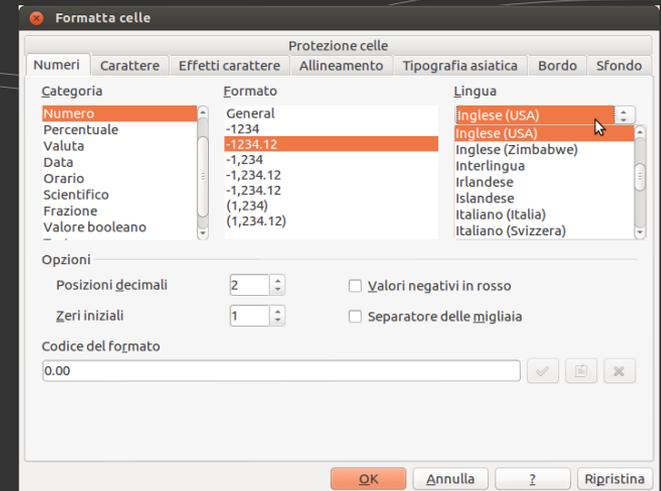
Modifica/Trova e sostituisci

Espressione regolare:
trova: $^{\wedge}.*\$$ → sostituisci: $\&$

Applicare ai valori.

È possibile poi tornare alla
lingua italiana (“.” → “,”)

(Oppure si può importare direttamente
con l'impostazione in inglese)



<http://www.regular-expressions.info/tutorial.html>



«Some people, when confronted with a problem, think "I know, I'll use regular expressions." Now they have two problems.»

Jamie Zawinski, alt.religion.emacs
(http://en.wikiquote.org/wiki/Jamie_Zawinski)

Filtraggio dei dati

The screenshot shows the Microsoft Excel interface with the 'Dati' menu open. The 'Filtro' option is highlighted, and its sub-menu is visible, showing options like 'Filtro automatico', 'Filtro standard...', and 'Filtro speciale...'. The spreadsheet data is as follows:

	A	B	C	D	
1	Codice Punto	strada	km	localita	
2	101	SP 1	2+300	Calceranica	01
3	101	SP 1	2+300	Calceranica	02
4	101	SP 1	2+300	Calceranica	03
5	101	SP 1	2+300	Calceranica	04
6	101	SP 1	2+300	Calceranica	04
7	101	SP 1	2+300	Calceranica	05
8	101	SP 1	2+300	Calceranica	06
9	101	SP 1	2+300	Calceranica	07/01/2012 0:00:00
10	101	SP 1	2+300	Calceranica	08/01/2012 0:00:00
11	101	SP 1	2+300	Calceranica	09/01/2012 0:00:00
12	101	SP 1	2+300	Calceranica	10/01/2012 0:00:00
13	101	SP 1	2+300	Calceranica	11/01/2012 0:00:00
14	101	SP 1	2+300	Calceranica	12/01/2012 0:00:00
15	101	SP 1	2+300	Calceranica	13/01/2012 0:00:00
16	101	SP 1	2+300	Calceranica	14/01/2012 0:00:00
17	101	SP 1	2+300	Calceranica	15/01/2012 0:00:00
18	101	SP 1	2+300	Calceranica	16/01/2012 0:00:00

Filtro e ordinamento

Filtro standard

Criteri filtro

Operatore	Nome di campo	Condizione	Valore
	Totale	>	0
	-nessuno-	=	
	-nessuno-	=	
	-nessuno-	=	

Più opzioni ? OK

Filtri condizionali dei dati

Dati/Ordina ...

Ordina

Criteri Opzioni

Chiave di ordinamento 1

Totale Crescente Decrescente

Chiave di ordinamento 2

- non definito - Crescente Decrescente

Chiave di ordinamento 3

- non definito - Crescente Decrescente

OK Annulla ? Ripristina

Acquisizione: pivot tables

riassunto_dati_traffico_anno_2012.ods - LibreOffice Calc

File Modifica Visualizza Inserisci Formato Strumenti **Dati** Finestra ?

Arial 10

A1 \sum = Codice Punto

	A	B	C	D		H
1	Codice Punto	strada	km	localita	De	
2	101	SP 1	2+300	Calceranica	01	Autoveicoli e e
3	101	SP 1	2+300	Calceranica	02	monovolumi
4	101	SP 1	2+300	Calceranica	03	Autoveicoli rimorchi
5	101	SP 1	2+300	Calceranica	04	
6	101	SP 1	2+300	Calceranica	04	
7	101	SP 1	2+300	Calceranica	05	
8	101	SP 1	2+300	Calceranica	06/01/2012 0:00:00	0,00
9	101	SP 1	2+300	Calceranica	07/01/2012 0:00:00	0,00
10	101	SP 1	2+300	Calceranica	08/01/2012 0:00:00	0,00

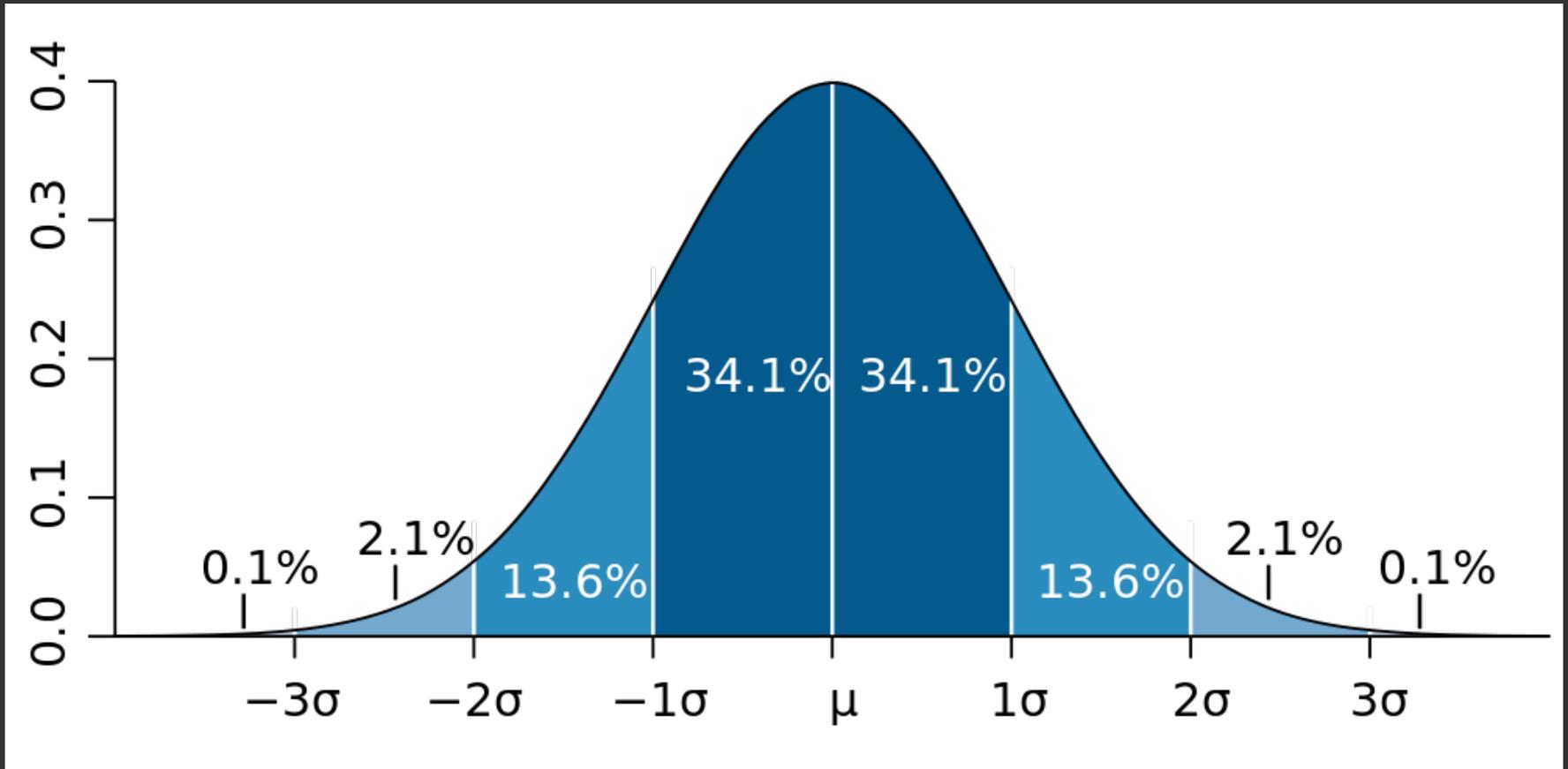
Definisci area...
Seleziona area...
Ordina...
Filtro
Formulario...
Subtotali...
Validità...
Operazioni multiple...
Testo a colonne...
Consolida...
Raggruppa e struttura
Tabella pivot
Aggiorna area

Crea...
Aggiorna
Elimina

Funzioni di base

- Matematiche
 - SOMMA
 - MEDIA
 - CONTA.SE
- Logiche
 - SE
- Testo
 - CONCATENA
 - STRINGA.ESTRAI
- Statistiche
 - DEV.ST.POP

(intermezzo statistico)



https://commons.wikimedia.org/wiki/File:Standard_deviation_diagram.svg

CC-BY-SA 2.5 by Mwtoews

Tabelle pivot

riassunto_dati_traffico_anno_2012.ods - LibreOffice Calc

File Modifica Visualizza Inserisci Formato Strumenti **Dati** Finestra ?

Arial 10

A1 \sum = Codice Punto

	A	B	C	D		H
1	Codice Punto	strada	km	localita	De	
2	101	SP 1	2+300	Calceranica	01	Autove
3	101	SP 1	2+300	Calceranica	02	e e
4	101	SP 1	2+300	Calceranica	03	monovolu
5	101	SP 1	2+300	Calceranica	04	mi
6	101	SP 1	2+300	Calceranica	04	rimorch
7	101	SP 1	2+300	Calceranica	05	
8	101	SP 1	2+300	Calceranica	06/01/2012 0:00:00	0,00
9	101	SP 1	2+300	Calceranica	07/01/2012 0:00:00	0,00
10	101	SP 1	2+300	Calceranica	08/01/2012 0:00:00	0,00

Definisci area...
Seleziona area...
Ordina...
Filtro
Formulario...
Subtotali...
Validità...
Operazioni multiple...
Testo a colonne...
Consolida...
Raggruppa e struttura
Tabella pivot
Aggiorna area

Crea...
Aggiorna
Elimina

Open Data - Chromium

Open Data x

www.aci.it/laci/studi-e-ricerche/dati-e-statistiche/open-data.html

Wikipedia Disk... Wikipedia:Wiki... Pundit-Timeline... download.geof... MySQL Python... Osmose - Open... Deployment — ... bottle-werkzeu...

Automobile Club d'Italia

Cerca nel sito

Home L'ACI Il Club Servizi Area Soci

La Federazione ACI per lo Sport **Studi e ricerche** Sicurezza stradale Driving in Italy Altri contatti URP PEC

Sei in [Home](#) / [L'ACI](#) / [Studi e ricerche](#) / [Dati e statistiche](#) / [Open Data](#)

Open Data

"I dati statistici della presente sezione sono liberamente fruibili da chiunque nel rispetto dei termini previsti dalla licenza di utilizzo Creative Commons CC-BY-ND 3.0 (cfr. <http://creativecommons.org/licenses/by-nd/3.0/it/>)"

"Le tavole sono esposte in formato aperto *.ods (per il cui utilizzo, qualora occorra, è necessario scaricare la suite Openoffice all'indirizzo <http://www.openoffice.org/it/>)"

- [Autoritratto](#) (File ZIP, 6.16 MB)
- [Autotrend](#) (File PDF, 60 Kb)
- [Annuario Statistico](#) (File PDF, 61 Kb)
- [Localizzazione degli incidenti stradali](#) (File PDF, 61 Kb)
- [Localizzazione degli incidenti stradali su strade provinciali](#) (File PDF, 61 Kb)
- [Fringe Benefit](#) (File PDF, 61 Kb)

[Scarica Adobe Reader](#)

- [Regolamento di accesso telematico ai dati e servizi contenuti nel sito istituzionale dell'ACI](#) (file PDF, 72 KB)
- [Catalogo dei dati in formato Open](#) (file PDF, 84 Kb)

Studi e ricerche

- Rivista giuridica online
- Mobilità e sicurezza +
- Archivio +
- Dati e statistiche** -
- Annuario Statistico
- Autoritratto
- Auto Trend
- Veicoli e mobilità
- Incidentalità
- Open Data**
- Fondazione Filippo Caracciolo
- Link utili

Open Data - Chromium

(intermezzo sull'origine dei dati)

Open Data - Chromium

www.aci.it/laci/studi-e-ricerche/dati-e-statistiche/open-data.html

Automobile Club d'Italia

Home L'ACI Il Club Servizi

La Federazione ACI per lo Sport **Studi e ricerche** Sicurezza stradale Driving in Italy Altri contatti URP

Sei in Home / L'ACI / Studi e ricerche / Dati e statistiche / Open Data

Open Data

"I dati statistici della presente sezione sono liberamente fruibili da chiunque nel rispetto dei termini previsti dalla licenza di utilizzo Creative Commons CC-BY-ND 3.0 (cfr. <http://creativecommons.org/licenses/by-nd/3.0/it/>)"

"Le tavole sono esposte in formato aperto *.ods (per il cui utilizzo, qualora occorra, è necessario scaricare la suite Openoffice all'indirizzo <http://www.openoffice.org/it/>)"

- [Autoritratto](#) (File ZIP, 6.16 MB)
- [Autotrend](#) (File PDF, 60 Kb)
- [Annuario Statistico](#) (File PDF, 61 Kb)
- [Localizzazione degli incidenti stradali](#) (File PDF, 61 Kb)
- [Localizzazione degli incidenti stradali su strade provinciali](#) (File PDF, 61 Kb)
- [Fringe Benefit](#) (File PDF, 61 Kb)

[Scarica Adobe Reader](#)

• [Regolamento di accesso telematico ai dati e servizi contenuti nel sito istituzionale dell'ACI](#) (file PDF, 72 KB)

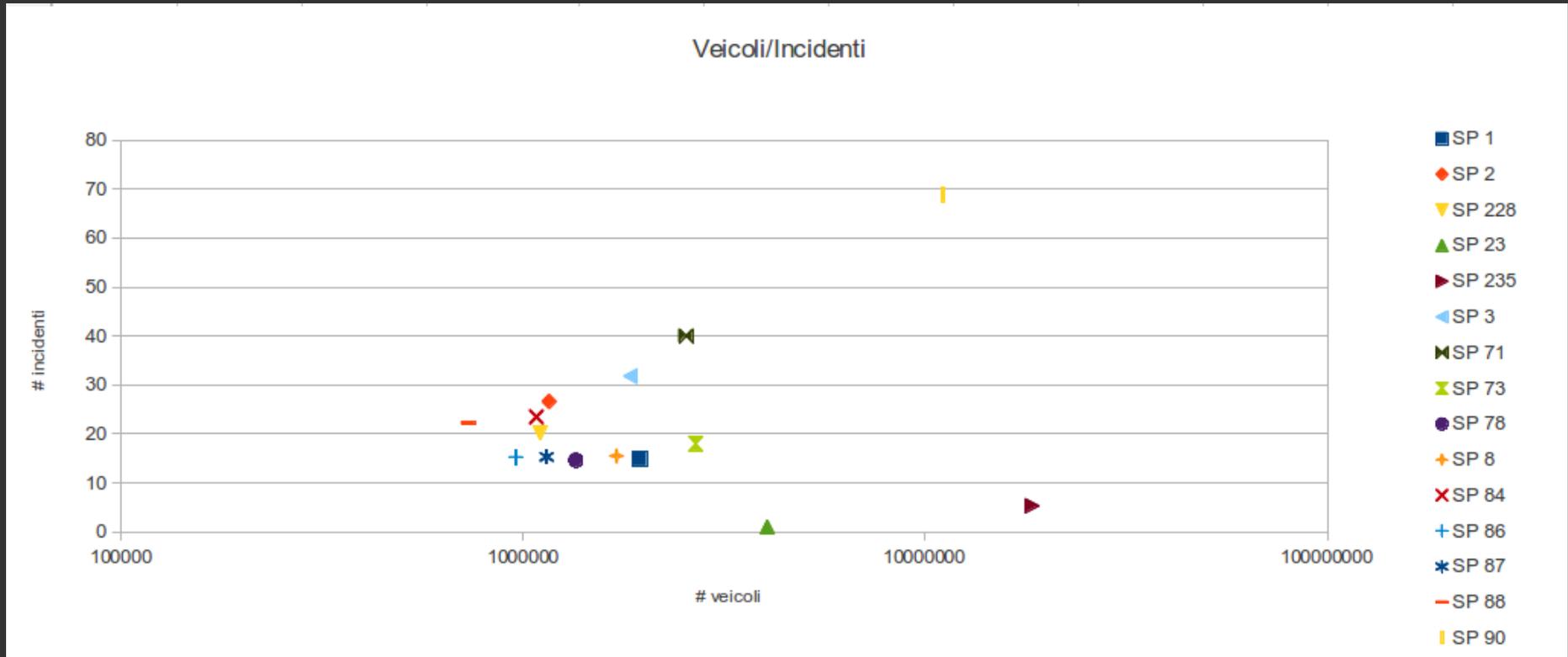
• [Catalogo dei dati in formato Open](#) (file PDF, 84 Kb)



#incidenti vs # veicoli

Nome strada	Veicoli	Incidenti
SP 1	1954918	14,866
SP 2	1162140	26,605
SP 228	1104131	20,217
SP 23	4051961	0,932
SP 235	18431427	5,3
SP 3	1854137	31,779
SP 71	2540191	40
SP 73	2690400	17,923
SP 78	1354838	14,592
SP 8	1709950	15,49
SP 84	1080852	23,457
SP 86	963420	15,207
SP 87	1147427	15,28
SP 88	731765	22,2
SP 90	11073415	68,868

un grafico



Data analysis: challenges

- Attenzione ai numeri piccoli
- Attenzione agli eventi rari
- Quali sono gli andamenti di lungo termine?
- Non lasciatevi trasportare dalle percentuali.
- Non lasciatevi trasportare dai numeri “ad effetto”

«The lesson from this is if it sound ridiculous, it probably is, and it needs to be checked thoroughly, which is not the easiest thing to do when you are on deadline.»

“Getting started with data journalism”, Claire Miller

Come salvare i propri dati

- Usare colori o strani font è inutile: *non fatelo!*
- È possibile esportare in CSV → nessun problema di compatibilità;
 - Si salva solo il foglio attivo
 - Non si salvano le formule o la formattazione!
- Utilizzando le funzionalità base (e salvando ne “vecchio” formato .xls, nel caso di Excel [97, 2000, XP, 2003], si riducono i problemi di compatibilità.
- Con formati aperti i problemi di compatibilità non si pongono! → I formati aperti sono *future proof*

Data cleaning: l'obiettivo

I dati devono essere spesso puliti per essere resi *omogenei*.

- Fase di preparazione dei dati
- Permette di creare visualizzazioni facilmente
- È un ottimo momento per iniziare a dare un'occhiata ai dati nel dettaglio

Raccolta di (alcuni) strumenti avanzati

- ✓ Raw <http://raw.densitydesign.org/>
- ✓ Datawrapper <http://datawrapper.de/>
- ✓ Google Fusion Tables <http://tables.googlelabs.com/>
- ✓ Geojson.io <http://geojson.io/>

Scraping (I): in generale

- A volte basta un semplice copia-incolla
- Se la pagina è strutturata è relativamente semplice.
- Si può considerare l'ipotesi di pagare un programmatore per ottenere i dati (“outsourcing”).

Scraping (I)

Sorgente HTML di una pagina:

```

1 <!DOCTYPE html>
2 <html lang="it" dir="ltr" class="client-nojs">
3 <head>
4 <meta charset="UTF-8" /><title>Internet - Wikipedia</title>
5 <meta name="generator" content="MediaWiki 1.22wmf16" />
6 <link rel="alternate" type="application/x-wiki" title="Modifica" href="/w/index.php?title=Internet&action=edit" />
7 <link rel="edit" title="Modifica" href="/w/index.php?title=Internet&action=edit" />
8 <link rel="shortcut icon" href="//bits.wikimedia.org/favicon/wikipedia.ico" />
9 <link rel="search" type="application/opensearchdescription+xml" href="/w/opensearch_desc.php" title="Wikipedia (it)" />
10 <link rel="EditURI" type="application/rsd+xml" href="//it.wikipedia.org/w/api.php?action=rsd" />
11 <link rel="copyright" href="//creativecommons.org/licenses/by-sa/3.0/" />
12 <link rel="alternate" type="application/atom+xml" title="Feed Atom di Wikipedia" href="/w/index.php?title=Speciale:UltimeModifiche&feed=atom" />
13 <link rel="canonical" href="http://it.wikipedia.org/wiki/Internet" />
14 <link rel="stylesheet" href="//bits.wikimedia.org/it.wikipedia.org/load.php?
debug=false&lang=it&modules=ext.gadget.Geonotice%7Cext.rtlcite%2Cwikihiro%7Cext.uls.nojs%7Cext.visualEditor.viewPageTarget.noscript%7Cmediawiki.
legacy.commonPrint%2Cshared%7Cmw.PopUpMediaTransform%7Cskins.vector&only=styles&skin=vector&*" />
15 <meta name="ResourceLoaderDynamicStyles" content="" />
16 <link rel="stylesheet" href="//bits.wikimedia.org/it.wikipedia.org/load.php?
debug=false&lang=it&modules=site&modules=site&only=styles&skin=vector&*" />
17 <style>a:lang(ar),a:lang(ckb),a:lang(kk-arab),a:lang(mzn),a:lang(ps),a:lang(ur){text-decoration:none}
18 /* cache key: itwiki:resourceLoader:filter:minify-css:7:d3ad324d383b55ef0ad58abad7f578f4 */</style>
19 <link rel="stylesheet" href="//bits.wikimedia.org/it.wikipedia.org/load.php?
debug=false&lang=it&modules=user&only=styles&skin=vector&user=CristianCantoro&version=20130430T154621Z&*" />
20
21 <script src="//bits.wikimedia.org/it.wikipedia.org/load.php?debug=false&lang=it&modules=startup&only=scripts&skin=vector&*"></script>
22 <script>if(window.mw){
23 mw.config.set({"wgCanonicalNamespace":"","wgCanonicalSpecialPageName":false,"wgNamespaceNumber":0,"wgPageName":"Internet","wgTitle":"Internet","wgCurRevisionId":61516000,"wgArticleId":2265,"wgIsArticle":true,"wgIsRedirect":false,"wgAction":"view","wgUserName":"CristianCantoro","wgUserGroups":["autopatrolled","**","user","autoconfirmed"],"wgCategories":["Informazioni senza fonte","Codice BNCF assente ma presente su Wikidata","Internet","Terminologia informatica"],"wgBreakFrames":false,"wgPageContentLanguage":"it","wgPageContentModel":"wikitext","wgSeparatorTransformTable":["",""],"wgDigitTransformTable":["",""],"wgDefaultDateFormat":"dmy","wgMonthNames":["","gennaio","febbraio","marzo","aprile","maggio","giugno","luglio","agosto","settembre","ottobre","novembre","dicembre"],"wgMonthNamesShort":["","gen","feb","mar","apr","mag","giu","lug","ago","set","ott","nov","dic"],"wgRelevantPageName":"Internet","wgUserId":195739,"wgUserEditCount":4983,"wgUserRegistration":1187910394000,"wgUserNewMsgRevisionId":null,"wgIsProbablyEditable":true,"wgRestrictionEdit":[],"wgRestrictionMove":[],"wgGlobalGroups":[],"wgVectorEnabledModules":{"collapsiblenav":true,"expandablesearch":false,"sectioneditlinks":false,"experiments":true},"wgWikiEditorEnabledModules":{"toolbar":true,"dialogs":true,"hidesig":true,"templateEditor":false,"templates":false,"preview":false,"previewDialog":false,"publish":false,"toc":false},"wgVisualEditor":{"isPageWatched":false,"magnifyClipIconURL":"//bits.wikimedia.org/static-1.22wmf16/skins/common/images/magnify-clip.png","pageLanguageCode":"it","pageLanguageDir":"ltr"},"wgGuidedTourHelpGuiUrl":"Aiuto:Visite guidate/guida","wgULSAcceptLanguageList":["it-it","it","en-us","en","fr"],"wgCategoryTreePageCategoryOptions":{"mode":"","hideprefix":20,"showcount":true,"namespaces":false},"Geo":{"city":"","country":""},"wgNoticeProject":"wikipedia","wgNoticeUserData":

```

ScraperWiki

The screenshot shows the ScraperWiki web interface. The browser address bar displays the URL: `https://classic.scraperwiki.com/scrapers/pdf_scraper_intro_4/edit/`. The page title is "Cristian Consonni / PDF Scraper Intro". A "Back to scraper overview" link is visible in the top right.

The main content area contains a Python script for scraping a PDF document. The script includes comments and code for importing libraries, setting the URL, fetching the PDF data, converting it to XML, and printing the results.

```

1 # Source:
2 # http://schoolofdata.org/2013/06/18/get-started-with-scraping-extracting-simple-tables-from-pdf-documents/
3 # by Tony Hirst
4
5 # 1. Add some necessary libraries
6 import scraperwiki
7 import urllib2, lxml.etree
8
9 # 2. The URL/web address where we can find the PDF we want to scrape
10 url = 'http://cdn.varner.eu/cdn-1ce36b6442a6146/Global/Varner/CSR/Downloads_CSR/Fabrikklister_VarnerGruppen_2013.pdf'
11
12 # 3. Grab the file and convert it to an XML document we can work with
13 pdfdata = urllib2.urlopen(url).read()
14 xmldata = scraperwiki.pdftoxml(pdfdata)
15 root = lxml.etree.fromstring(xmldata)
16
17 # 4. Have a peek at the XML (click the "more" link in the Console to preview it).
18 print lxml.etree.tostring(root, pretty_print=True)
19
20 # 5. How many pages in the PDF document?
21 pages = list(root)
22 print "There are", len(pages), "pages"
23
24 # 6. Iterate through the elements in each page, and preview them
25 for page in pages:
26     for el in page:
27         if el.tag == "text":
28             print el.text, el.attrib
29
30 exit(1)

```

Below the script, there are buttons for "Documentation" and "RUN". The "RUN" button is highlighted, and a "SAVE SCRAPER" button is visible in the top right of the script area. The text "Last saved 9 minutes ago" is displayed next to the save button.

The console output shows the following messages:

```

Starting run ...
Line 13 - pdfdata = urllib2.urlopen(url).read()
/usr/lib/python2.7/urllib2.py:126 -- urlopen((url='http://cdn.varner.eu/cdn-1ce36b6442a6146/Global/V ...more
HTTPError: HTTP Error 404: Not Found
Finished: 1.134 seconds elapsed
runfinished

```

Scraping (IV)

PDF:

«Scraping PDFs is a bit like cleaning drains with your teeth. It's slow, unpleasant, and you can't help but feel you're using the wrong tools for the job. [...] Why is scraping PDFs so hard? Well, the PDF standard was designed to do a particular job: describe how a document looks, anywhere and forever.»

Tratto da:

<http://blog.scrapewiki.com/2010/12/17/scraping-pdfs-now-26-less-unpleasant-with-scrapewiki/>

Tutorial per chi vuole cimentarsi con un po' di codice:

<http://schoolofdata.org/2013/06/18/get-started-with-scraping-extracting-simple-tables-from-pdf-documents/>

Quali sono i rischi quando si lavora con i dati

X le teorie si adattano ai dati, non viceversa.

*«Se le realtà non si adatta alla teoria, la realtà è sbagliata,»
(a volte erroneamente attribuita a Einstein)*

X correlazione non implica causalità.

*«Correlation doesn't imply causation, but it does waggle its eyebrows suggestively and gesture furtively while mouthing 'look over there'.»
<http://xkcd.com/552>*

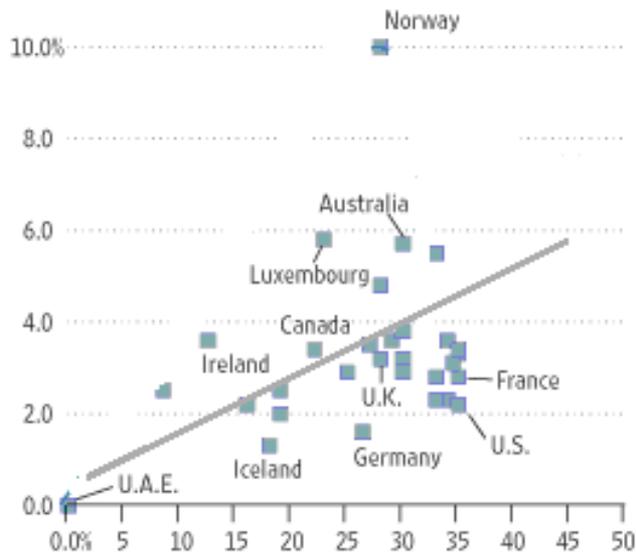
X i modelli teorici sono sempre validi entro certi limiti.

*«Finché le leggi della matematica si riferiscono alla realtà, non sono certe, e finché sono certe, non si riferiscono alla realtà,»
Albert Einstein, *Sidelights on Relativity**

Rischi (1): adattare i dati alla teoria

Corporate Taxes and Revenue, 2004

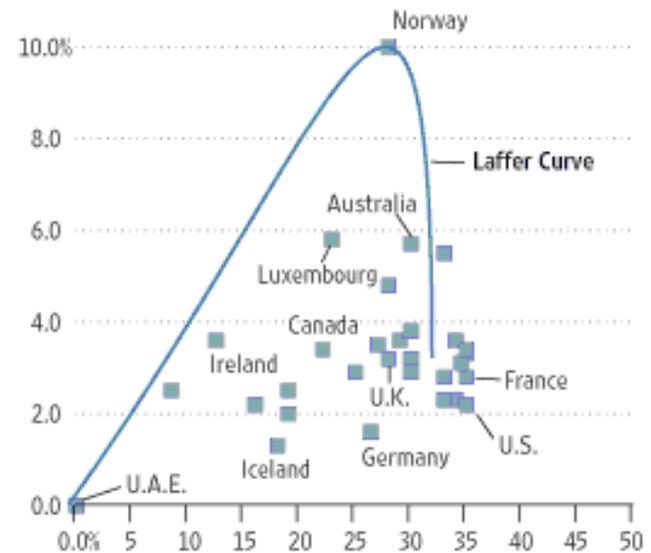
Left scale represents tax revenues as a percentage of GDP. Bottom scale represents central government corporate tax rates.



Sources: OECD Revenue Statistics, Kevin Hassett, American Enterprise Institute

Corporate Taxes and Revenue, 2004

Left scale represents tax revenues as a percentage of GDP. Bottom scale represents central government corporate tax rates.

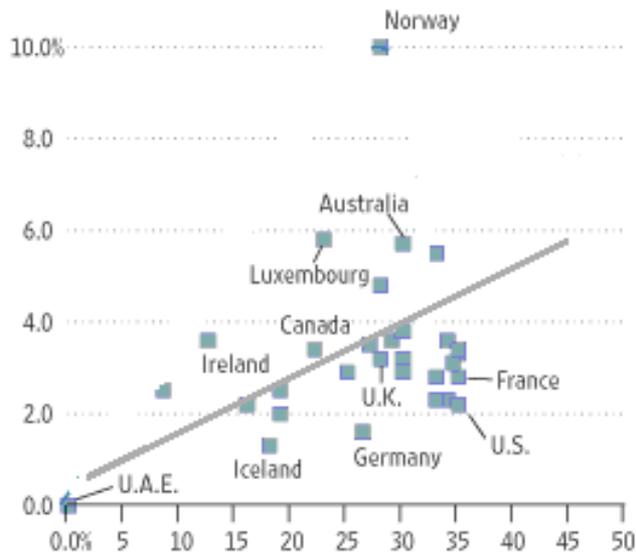


Sources: OECD Revenue Statistics, Kevin Hassett, American Enterprise Institute

Rischi (1bis): adattare i dati alla teoria

Corporate Taxes and Revenue, 2004

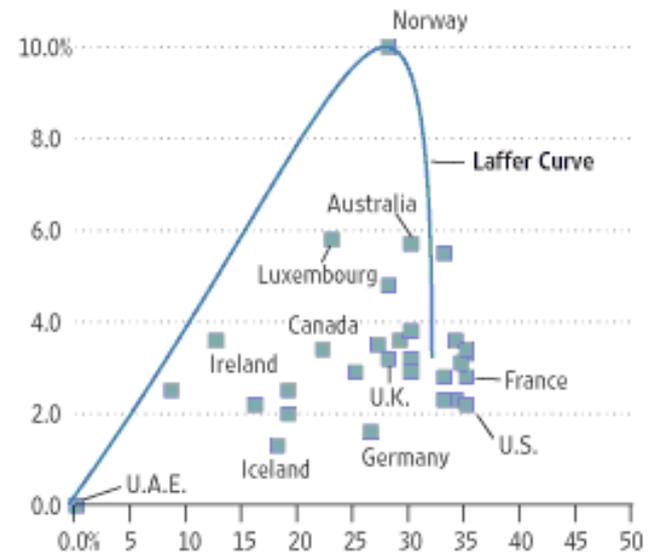
Left scale represents tax revenues as a percentage of GDP. Bottom scale represents central government corporate tax rates.



Sources: OECD Revenue Statistics, Kevin Hassett, American Enterprise Institute

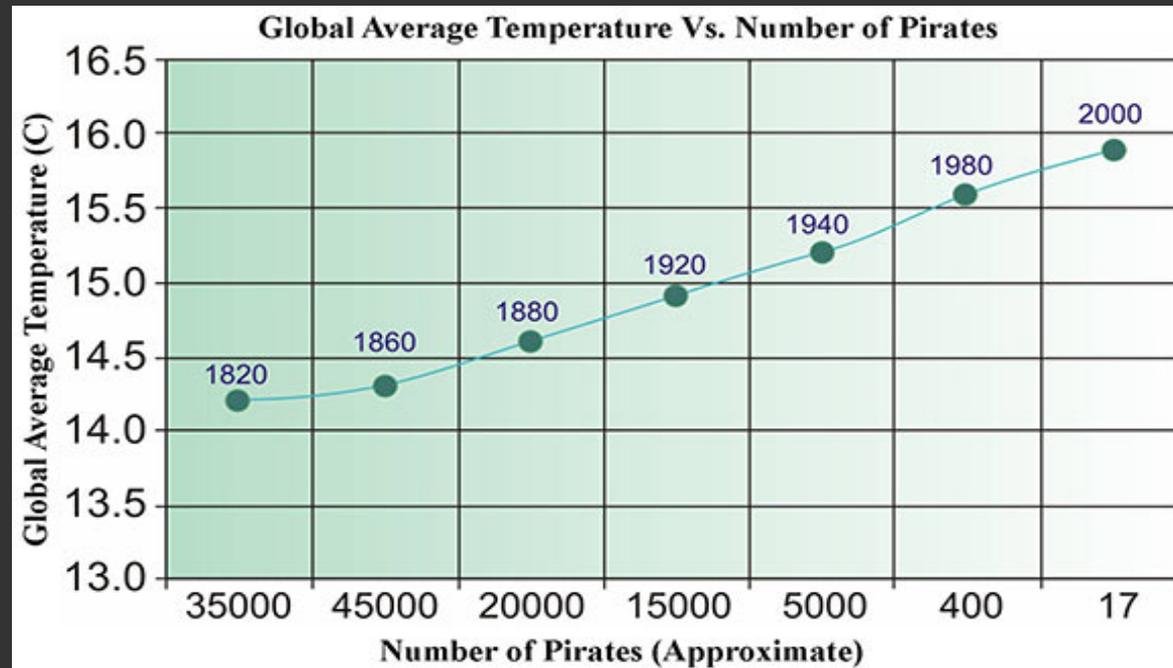
Corporate Taxes and Revenue, 2004

Left scale represents tax revenues as a percentage of GDP. Bottom scale represents central government corporate tax rates.



Sources: OECD Revenue Statistics, Kevin Hassett, American Enterprise Institute

Rischi (2): correlazione \rightarrow causalità? No!



<http://bressanini-lescienze.blogautore.espresso.repubblica.it/2013/02/15/mangia-cioccolato-e-vinci-il-premio-nobel/>

Cristian Consonni

Mail: consonni@fbk.eu

CristianCantoro →

{ skype, twitter, wiki*,
slideshare, ... }

Find this presentation on slideshare:
<http://www.slideshare.net/CristianCantoro>



Questa presentazione è abbondantemente ispirata a quella di Marco Montanari:

- <http://www.slideshare.net/sirmmo/rcs-27211305>

Questa presentazione è rilasciata con licenza
CC-BY-SA

- <http://creativecommons.org/licenses/by-sa/3.0/deed.it>



Except where otherwise noted, this work is licensed under

<http://creativecommons.org/licenses/by-sa/3.0/>