UNIVERSITÀ DEGLI STUDI DI MILANO
Facoltà di Studi Umanistici
Corso di Laurea in Lettere Moderne

LE STORIE DEI DATI
FRA GIORNALISMO E DIVULGAZIONE SCIENTIFICA

Tesi di Laurea di:
Valerio BERRA
Matr. n. 827155

Relatore: Chiar. mo Prof. Francesco TISSONI
Correlatore: Chiar. mo Prof. Stefano MONTANELLI

Anno Accademico 2014-2015

Indice

Appendice 2

Introduzione

I dati sono ovunque.

Che si tratti della distribuzione del reddito, delle persone nate in un anno o dei tweet pubblicati in un giorno è sempre più semplice imbattersi in numeri che raccontano quello che accade attorno a noi.

I dati infatti non sono altro che la descrizione di un fenomeno e registrarli è diventato sempre più semplice da quando molti aspetti della nostra vita sono stati affidati all'informatica. Nel tempo la loro mole è cresciuta a tal punto da rendere necessaria una nuova metrica. A partire dall'inizio degli anni 2000 dalla genomica e dall'astrofisica è stata importata nel linguaggio comune la formula *Big Data* che ben descrive l'enorme massa di numeri a nostra disposizione.

Perché questi dati acquistino valore è necessario però che qualcuno si occupi di raccoglierli, analizzarli e soprattutto comunicarli.

Il modo più naturale per trasmettere questo tipo di informazioni sembra essere la data visualization, la traduzione di una serie di numeri in immagini più facilmente comprensibili. Un processo da cui nascono le infografiche che troviamo nei giornali, nei report aziendali o nelle pubblicità.

Con l'aumentare dei dati a disposizione, queste forme di visualizzazione sono diventate sempre più articolate tanto da arrivare a costruire una sorta di linguaggio in grado di creare anche discorsi complessi. Si sta parlando qui di un ambito di studio ibrido in cui si mischiano competenze diverse, dalla statistica alla grafica, e dove anche gli studi umanistici potrebbero trovare il loro spazio.

È certo infatti che la visualizzazione dei dati sia l'unico modo per comunicare le informazioni in essi contenuti?

Con l'obiettivo di rispondere a questa domanda sono stati sviluppati nel seguente elaborato due percorsi. Il primo riguarda l'aspetto strettamente teorico della questione mentre il secondo si serve di un caso pratico per studiare i meccanismi che regolano la data visualization.

Per cominciare ad orientarsi in questo terreno è stato necessario capire attraverso una catalogazione dove vengono utilizzate le infografiche. Ad ogni ambito individuato sono stati associati diversi esempi in modo tale da comprendere tutti i passaggi che trasformano una complessa serie di dati grezzi in una visualizzazione.

Dopo aver definito per quali scopi si utilizza questo linguaggio si è passati quindi ad un'analisi degli elementi che lo compongono, dai simboli grafici fino ai modelli più articolati

Le competenze acquisite attraverso questo lavoro di ricerca sono state messe alla prova con un progetto di divulgazione scientifica in collaborazione con il Crisp, il Centro di Ricerca per Servizi di Pubblica Utilità alla Persona dell'Università degli Studi di Milano-Bicocca.

Questo percorso ha permesso non solo di capire meglio come utilizzare certe forme di visualizzazione ma anche di arrivare a definire un flusso di lavoro per passare dalla raccolta dei dati alla loro divulgazione.

Per avere una visione più ampia e aggiornata di queste tematiche in appendice all'elaborato sono state poi raccolte interviste a giornalisti, analisti e semiologi che negli ultimi anni hanno concentrato la loro attenzione sui dati e sul modo in cui questi vengono comunicati.

Capitolo 1
Storia, classificazione e prospettive della data visualization

1. Contrastare il pregiudizi con i dati. Il progetto *Ignorance* di Hans Rosling.

"The first way to think about the future is to know about the present"
Hans Rosling[1]

Il progetto TED[2] è stato fondato nel 1984 con lo scopo di raccogliere e divulgare "ideas worth spreading", idee che vale la pena diffondere. I suoi fondatori Richard Saul Wurman e Harry Marks hanno cominciato così ad organizzare una conferenza all'anno a Vancouver, chiamando sul palco persone con un'idea o una storia interessante da raccontare. Già dall'inizio degli anni '90 il numero degli ospiti e quello degli spettatori iniziò a crescere tanto che le conferenze furono organizzate anche in altre città del mondo. Ora sono migliaia gli interventi realizzati sotto il logo di TED, molti dei quali sono stati raccolti sul loro sito.

Nel giungo del 2014 questa rassegna ha fatto tappa a Berlino e qui si è presentato anche Hans Rosling, un medico che insegna Salute Globale presso l'università svedese Karolinska Institutet.
Il suo intervento è cominciato ponendo al pubblico queste tre domande.

1. Quanto è cambiato il numero di morti all'anno per disastri naturali durante l'ultimo secolo?
2. Per quanto tempo, in media, è andata a scuola una donna di 30 anni?
3. Negli ultimi 20 anni come è cambiata la percentuale di persone che vivono in estrema povertà?

Ad ogni quesito erano associate tre risposte che gli spettatori potevano scegliere con un dispositivo collegato al computer di Rosling. Il risultato della votazione veniva così visualizzato in diretta.

Dopo aver lasciato il tempo per rispondere il medico svedese ha fornito le soluzioni corrette, svelando quanto il pubblico fosse poco consapevole di quello che stava accadendo nel mondo. In nessun caso infatti la maggior parte dei presenti ha indicato l'opzione giusta, anzi la percentuale di persone che hanno risposto correttamente non ha mai superato il 20%.

A questo punto Hans Rosling ha svelato i risultati che avrebbe ottenuto da un altro tipo di platea.
Se infatti si fosse recato allo zoo per porre la stessa domanda ad un branco di scimpanzé avrebbe registrato delle risposte più corrette.
I primati avrebbero risolto i loro dubbi schiacciando i tasti del loro dispositivo a caso, attribuendo quindi lo stesso valore a tutte le opzioni. In questo processo anarchico almeno il 33% avrebbe così scelto la risposta giusta.

Partendo da questa provocazione Hans Rosling ha potuto così illustrare nel suo intervento altri esperimenti in cui diversi campioni di persone sono stati sottoposti a domande riguardanti i livelli di salute, ricchezza e povertà nel mondo. Raramente la maggior parte degli intervistati è riuscita fornire la risposta esatta o anche solo ad avvicinarsi agli otiimi risultati ottenuti dal suo ipotetico branco di scimpanzé.

A spiegare le motivazioni di queste dinamiche è arrivato sul palco Ola Rosling, statistico e figlio di Hans. Per prima cosa ha spiegato quali meccanismi psicologici rendono i preconcetti un nemico così duro da combattere per chi si occupa di informazione. La parte più interessante del suo

1 H. Rosling, 2014, *How not to be ignorant about the world*. La lezione completa si può trovare sul sito del progetto TED all'indirizzo: https://www.ted.com/talks/hans_and_ola_rosling_how_not_to_be_ignorant_about_the_world?language=en#t-439355
2 https://www.ted.com

intervento è arrivata però quando si è concentrato su Gapminder[3], il sito fondato insieme al padre per diffondere dati corretti sullo stato della popolazione mondiale. Un'idea che ha preso forma nel *The Ignorance Project*, come si può leggere nel manifesto pubblicato su www.gapminder.org.

"The mission of Gapminder Foundation is to fight devastating ignorance with a fact-based worldview that everyone can understand. We started the Ignorance Project to investigate what the public know and don't know about basic global patterns and macro-trends. We use surveys to ask representative groups of people simple questions about key-aspects of global development"[4]

"Everyone can understand" è la formula chiave di questa presentazione. Perché chiunque possa capire questi dati i due fondatori hanno scelto di comunicarli attraverso una serie di grafici dinamici che sfruttano delle animazioni per analizzare lo sviluppo di un fenomeno in un periodo di tempo definito. Ci sono così lavori che mostrano come nel corso della storia cambiato il rapporto fra Prodotto Interno Lordo e aspettativa di vita in ogni Paese del mondo, oppure quali sono i fattori che hanno permesso all'Asia di crescere a livello economico.

In questo modo la Gapminder Foundation mostra perfettamente come si possono fondere due fattori: utilizzare i dati per fornire la visione complessiva di un fenomeno e creare visualizzazioni che siano semplici da fruire e piacevoli da vedere.

The Ignorance Project può essere quindi il primo passo per introdurre questo capitolo perché chiarisce subito come si possano trasmettere attraverso dei grafici informazioni complesse. Nelle prossime pagine, partendo da una breve introduzione storica, verranno presentati gli ambiti in cui vengono utilizzati i processi di data visualization.

3 http://www.gapminder.org
4 http://www.gapminder.org/ignorance

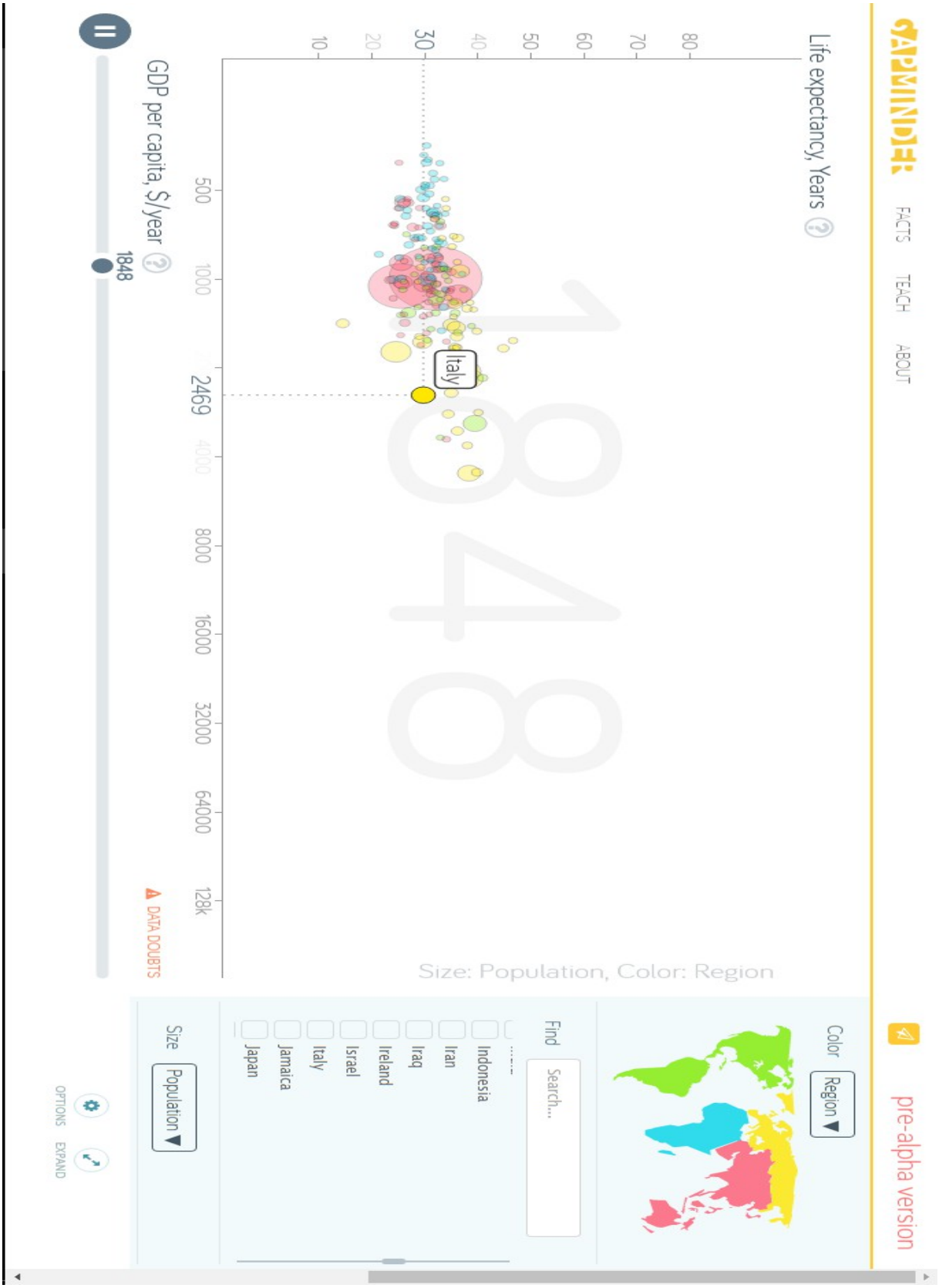


Illustrazione 1: Esempio di un progetto di Gapminder. Qui vengono incrociati sullo stesso grafico i dati relativi all'aspettativa di vita e Pil pro capite in ogni nazione. Nella barra che definisce la cronologia è possibile sia selezionare l'anno che dare il via all'animazione. In questa schermata è stata presa come esempio l'Italia. La visualizzazione completa si può tovare su http://www.gapminder.org/tools/bubbles#_.

2. Storia delle infografiche

Trasformare i numeri in forme e colori. Una traduzione semantica prima che grafica sempre più facile da portare a termine. Più cresce il numero di software dedicati a questo tipo di operazione più basse sono le capacità informatiche richieste per utilizzarli. Tableau Public[5], Gephi[6], Piktochart[7] e Cartodb[8] sono solo alcuni degli esempi che mostrano quanto ora sia semplice prendere una tabella di dati e riportarla su una mappa, trasformala in barre, cerchi, torte e pittogrami. Certo questi sono gli anni in cui si stanno sviluppando strumenti del genere perché questi sono gli anni in cui sono necessari. I dati piovono ovunque. Vengono comunicati dai governi e messi a disposizione del pubblico, vengono creati dagli utenti dei social network o dalle applicazioni che colorano i menù dei nostri smartphone. Addirittura ora si possono studiare dati in campi dove prima era possibile solo fare delle stime. Si può misurare l'esatto tempo passato a leggere un libro in formato elettronico o quello impiegato per compiere un viaggio in treno. Si è verificato un aumento esponenziale della quantità di dati a disposizione tanto che è stato necessario coniare il termine Big Data per descrivere meglio questi enormi dataset pieni di cifre su qualsiasi aspetto della nostra vita.

E forse proprio perché questo argomento è intrecciato a doppio filo con la contemporaneità diventa difficile pensare che nel XII secolo ci fosse già chi munito di righello e matita pensava a come visualizzare i dati in suo possesso.

Di questi precedenti parla anche Simon Rogers che nel 2009 ha fondato il datablog del quotidiano inglese *The Guardian*[9], uno spazio dedicato esclusivamente al giornalismo basato sull'analisi di grosse quantità di dati. Nel 2013 ha scritto un libro intitolato *Fact are Sacred* per raccogliere lavori e riflessioni. Sulle pagine iniziali si può trovare una lista di dieci cose che il lettore potrà imparare sul data journalism leggendo il suo volume. Ecco la prima.

“It may be trendy but it's not new”[10]

Anche se oggi le infografiche stanno conoscendo un momento di diffusione su vasta scala la loro storia ha radici molto profonde.

Tracce di un nuovo linguaggio. Il generale Yu, Jhon Snow, Charles Minard e Eric Burgess.

Qualche cenno alla storia della visualizzazione dei dati si può trovare in *The Visual Display of Quantitative Information*[11], un volume scritto da Edward Rolf Tufte, matematico statunitense esperto di infografiche.

I primi esempi di combinazione fra cartografia e tecniche statistiche si riscontrano nel XVII secolo anche se esiste un reperto che si colloca ancora più lontano del tempo. Risale infatti al XII secolo una mappa redatta da uno sconosciuto cartografo cinese che traccia i movimenti del generale Yu il Grande con una precisione impensabile per i colleghi europei della stessa epoca. A rendere ancora più straordinaria quest'opera è poi il fatto che l'originale non sia scritto su carta ma inciso su una tavoletta di pietra. Oltre le notevoli doti da cartografo e la sua precisione come incisore, questo

5 https://public.tableau.com
6 https://gephi.org
7 http://piktochart.com
8 https://cartodb.com
9 http://www.theguardian.com/data
10 S. Rogers, *Facts are Sacred*, 2013
11 E. R. Tufte, *The Visual Display of Quantitative Information*, 2001

sconosciuto autore potrebbe quindi essere ricordato anche per aver creato una delle prime mappe tematiche mai registrate. Qui infatti il territorio geografico diventa lo sfondo per rappresentare una serie di eventi.

Illustrazione 2: Fonte: http://mapdesign.icaci.org/tag/topographic/

Alla fine del XX secolo troviamo però un esempio ancora più chiaro di come la data visualization possa servire a comprendere un fenomeno.

Illustrazione 3: Fonte: https://en.wikipedia.org/wiki/John_Snow

Questa mappa è stata disegnata da Jhon Snow, un medico che in quegl'anni cercava di capire come si stesse propagando l'epidemia di colera. I pallini disegnanti nei quartieri rappresentano le morti avvenute a causa di questa malattia in un quartiere di Londra nel mese di settembre del 1854. Le

pompe d'acqua invece sono indicate da una croce. Come si può capire dalla disposizione dei due elementi, la fonte d'acqua inquinata che stava diffondendo l'epidemia in questa zona della metropoli si trovava a Board Street, qui infatti è dove si concentra il maggior numero di pallini. Questa visualizzazione quindi riesce a comunicare subito quello che difficilmente si sarebbe scoperto in un altro modo, come commenta Edward Tufte.

"Of course the link between the pump and the disease might have been revealed by computation and analysis without graphics, with some good luck and hard work. But, here at least, graphical analysis testifies about the data far more efficiently then calculation".

Quello di Jon Snow è un esempio diventato tanto famoso che gli sviluppatori di Cartodb.com hanno addirittura deciso di dedicargli un tutorial. Una delle mappe che si possono creare nella sezione *Learn* del sito è infatti la versione contemporanea e a colori di quella creata dal medico inglese. Pur mantenendo gli stessi valori e ovviamente la stessa porzione di territorio, nella visualizzazione che si può creare con questo sito di mapping vengono aggiunti i colori e dei cerchi per indicare le pompe d'acqua.

Illustrazione 4: Fonte: http://docs.cartodb.com/tutorials/conditional_styling/

Il terzo caso preso in esame è un grande classico per quanto riguarda i grafici. Ed è uno dei primi esperimenti in cui all'interno di una stessa tavola vengono rappresentante più valori.
Charles Joseph Minard è un ingegnere francese che nel 1869 decide di rappresentare la disfatta dell'esercito napoleonico nella campagna di Russia del 1812. Il generale Bonaparte era partito dal confine polacco con un contingente di 422 000 uomini nel giugno del 1812, quando a settembre raggiunse Mosca le sue truppe erano già scese a 100 000 unità e di queste solo 10 000 riuscirono a superare il viaggio di ritorno e l'inverno russo per tornare a casa.

Illustrazione 5: Fonte https://en.wikipedia.org/wiki/Charles_Joseph_Minard

Per disegnare una delle più grandi disfatte della storia militare europea Minard scelse di rappresentare non solo le costanti perdite dei battaglioni ma anche il percorso compiuto in territorio russo e addirittura i picchi di freddo toccati durante l'inverno. Il risultato è quello di una mappa geografica della regione interessata dal conflitto su cui compaiono due linee diverse che tendono ad assottigliarsi. Una, quella marrone, definisce l'avanzata dell'esercito napoleonico fino a Mosca, l'altra, quella nera, il ritorno. Lungo queste due linee compaiono anche i nomi delle città attraversate e il numero delle unità rimaste. Un piccolo grafico, posto sotto questa mappa, mostra anche la costante diminuzione della temperatura durante la ritirata svelando come ad ogni abbassamento registrato corrisponda un maggiore numero di perdite.

Ecco tutte le variabili identificate da Tufte in questa visualizzazione:

- la dimensione dell'esercito
- la sua posizione su un piano bidimensionale
- la direzione dell'esercito
- la temperatura in diverse giornate durante la ritirata da Mosca

Per la complessità dei dati presi in considerazione e allo stesso tempo per la sua potenza narrativa il grafico di Minard non può che incassare questo giudizio dal matematico statunitense:

“It may well be the best statiscal graphic ever drawn”.

L'ultimo frammento tratto dalla storia della data visualization in questo momento si trova circa 12 miliardi di chilometri dalla Terra[12].

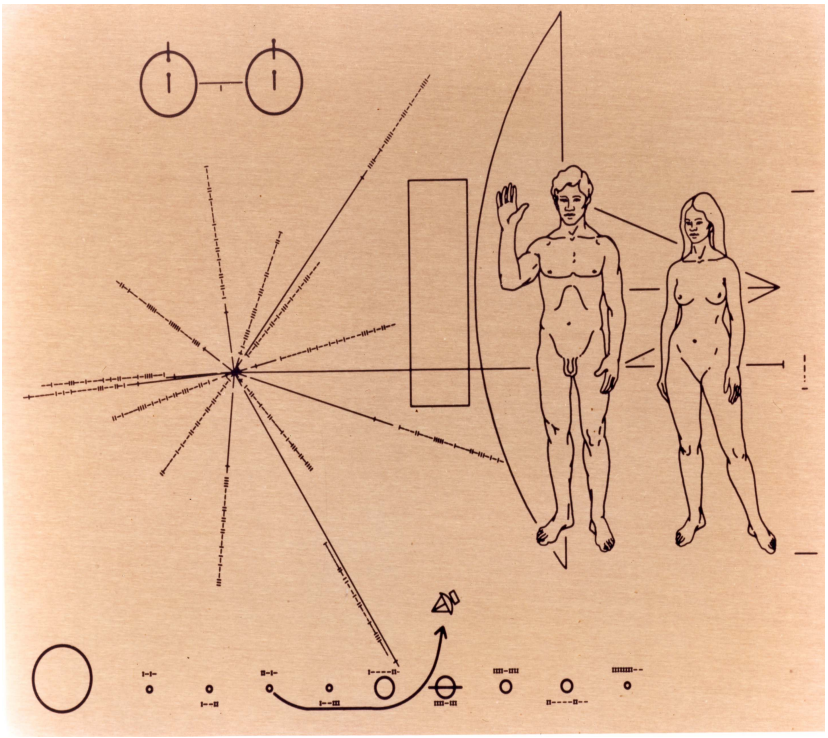


Illustrazione 6: Fonte: https://it.wikipedia.org/wiki/Placca_dei_Pioneer

12 F. Tissoni, *Social network, Comunicazione e marketing*, 2014

Nel 1972 e nel 1973 le sonde Pioner 10 e Pioner 11 sono state lanciate dalla NASA[13] per diventare i primi oggetti costruiti dall'uomo a superare i confini del sistema solare. Considerata la loro destinazione, il giornalista Eric Burgess pensò che queste missioni fossero un'ottima opportunità per diffondere nello spazio un messaggio della civiltà terrestre.

Le informazioni contenute sulla placca sono rivolte quindi ad una specie aliena, dotata di minime conoscenze scientifiche.

Sono tre le notizie che si possono ricavare dalle linee incise sulla placca. La prima è sulla morfologia degli essere umani. Un uomo e una donna sono rappresentati nudi con dietro la sagoma della sonda, così da fornire un riferimento per le nostre dimensioni. La seconda notizia contenuta è la provenienza dell'oggetto spaziale, nella fascia in basso infatti è contenuto uno schema del sistema solare ed è indicata la rotta percorsa per uscirne. L'ultima informazione riguarda invece la posizione della Terra. In questo caso è stato utilizzato un linguaggio basato su linee e pulsar che dovrebbe definire le coordinate del sistema solare e il momento in cui la sonda è stata lanciata.

Sono state avanzate diverse critiche a questo progetto. Alcune sono legittime e riguardano l'uso di segni propri della cultura umana, come le linee e le frecce, altre invece sono più discutibili e si interrogano sull'opportunità di mandare immagini di corpi nudi nello spazio. In ogni caso questa placca offre un esempio delle potenzialità di cui vengono investite le infografiche, scelte in questo caso per comunicare con esseri viventi con cui ancora non siamo entrati in contatto.

Se mai le nostre placche di alluminio sbarcheranno su pianeti abitati da altri essere vivente sarebbe davvero interessante sapere quali informazioni siano state in grado di trasmettere.

L'arrivo dei Big Data. Le Cinque V di Bernard Marr

La storia del data design non segue però una linea regolare. C'è un momento in cui esplode qualcosa, un momento in cui tutte le tecniche legate alla visualizzazione delle informazioni contenute nei numeri diventano uno strumento indispensabile per comprendere il mondo.
All'inizio degli anni 2000 prima in ambito scientifico e poi nell'immaginario collettivo comincia a farsi strada il termine Big Data.
I suoi primi utilizzi si possono trovare in due campi: la genomica e l'astrofisica. Si tratta di due branche del sapere paradossalmente agli antipodi: lo studio microscopico dei genomi che definiscono le creature viventi e quello macroscopico delle dinamiche dell'universo. È proprio in questi campi di ricerca che vengono messi a punto strumenti in grado di registrare una quanità prima inimmaginabile di dati.

Il numero di informazioni registrare diventa così talmente ampio che c'è bisogno di una nuova metrica. Una formula che permetta di capire subito quanto sia colossale la quantità di informazioni a disposizione. Nasce quindi il termine Big Data.
Negli ultimi anni l'uso di queste due parole è diventato sempre più comune tanto che l'economista Bernard Marr ha deciso di definire le caratteristiche necessarie a capire quando davvero ci si trova davanti a dei Big Data[14].
Il metodo che descrive è quello delle Cinque V, cinque parole che cominciano per V e inquadrano queste enormi collezioni di dati.

13 National Aeronautics and Space Administration. È l'agenzia spaziale degli U.S.A.
14 B. Marr, *Big Data: Using Smart Big Data, Analytics and Metrics To Make Better Decision and Improve Performance*, 2015

Volume

Il primo aspetto si riferisce alla grande quantità di dati che vengono generati ogni secondo. L'esempio più immediato sono i dati del web. Fra le mail, i tweet, le foto e i video che vengono prodotti ogni secondo è necessario ricorrere ad unità di misura ben superiore ai Terabytes, si parla qui di Zettabytes o addirittura Brontobytes[15]. Solo su facebook ogni giorno vengono inviati 10 miliardi di messaggi e vengono pubblicate 350 milioni di foto. Il sito InternetLiveState[16] offre in tempo reale dei contatori che monitorano questi dati e vengono azzerati solo alla fine di ogni giornata.

Velocity

Si riferisce alla velocità con cui i dati vengono generati e spostati. Basti pensare all'immediatezza con cui i sistemi di trading colgono i segnali forniti dai mercati per capire se è meglio vendere o comprare un pacchetto di azioni. I Big Data permettono di analizzare i dati mentre sono generati.

Variety

Sono quasi illimitati i terreni in cui si possono raccogliere i dati. Non esistono più solo quelli strutturati, come possono essere i dati provenienti dal mercato finanziario, perfettamente ordinabili all'interno di tabelle. Ora l'80% dei dati esistenti non sono strutturati ossia non seguono un rigoroso modello di classificazione ma sono costituti da elementi difficilmente collocabili all'interno di una tabella, come foto e video. La tecnologia per analizzare i Big Data si occupa ora anche di questo. Capire come sia possibili definire con gli stessi criteri dati che si presentano in forme completamente diverse.

Veracity

Con una quantità così estesa di dati, la qualità e l'accuratezza sono meno controllabili ma è sempre possibile analizzarli. In questi casi infatti l'ampio volume di dati a disposizione riesce ad offrire comunque informazioni significative sulle tendenze più importanti.

Value

I Big Data possono diventare fondamentali se si riesce a dare loro valore. Se si riesce non solo a conoscerli ma anche a sfruttarli per il processo decisionale di un'azienda, di un'amministrazione pubblica o di un'intera democrazia.

L'arrivo dei Big Data ha segnato così un passaggio fondamentale per la data visualization. Questa infatti non è diventata solo uno strumento per trasmettere delle informazioni in modo più immediato di prima ma è diventata addirittura l'unico modo in cui certe informazioni possono essere comprese.

15 Per avere un metro di paragone, ecco la scala di grandezza delle unità di misura per quantificare i byte: Kilobyte= 10^3, Megabyte = 10^6, Gigabyte = 10^9, Terabyte = 10^{12}, Petabyte = 10^{15}, Exabyte = 10^{18}, Zettabyte = 10^{21}, Yottabyte = 10^{24}, Brontobyte = 10^{27}

16 http://www.internetlivestats.com

3. Quando serve la data visualization. Proposta di un criterio di classificazione

Scoprire fonti d'acqua inquinate, illustrare la storia di un esercito o comunicare con civiltà aliene. In una breve escursione sulla storia della data visualization è stato possibile capire quanto ampie siano le possibilità di questa forma di comunicazione.

Ma qual è la differenza fra i grafici visti nel capitolo precedente e quali utilizzati adesso?

Il confine fra la storia e il presente è segnato dall'avvento dei computer. É solo con l'introduzione di software in grado di generare automaticamente delle visualizzazioni partendo da una tabella di numeri che è stato possibile affrontare dataset sempre più ampi e complessi. Nelle prossime pagine dunque verranno presentati alcuni esempi di come le dataviz possano essere utilizzate in vari ambiti del sapere.

Risulterebbe però caotico lanciarsi in una rassegna di questo tipo senza aver definito almeno una rotta con cui orientarsi. Una rotta che potrebbe essere costruita sullo scopo per cui queste visualizzazioni sono state create. Passando in rassegna qualsiasi raccolta di data visualization, come può essere il volume *Atlas of Science*[17] di Katy Börner, si possono definire tre motivi per cui costruire una dataviz:

- **Visualizzare i dati per capirli**. È questa la funzione più utilizzata in ambito scientifico. La visualizzazione dei dati ha qui lo scopo di permettere una comprensione chiara e immediata delle informazioni che si hanno a disposizione. Il dataset viene rappresentato nella sua complessità permettendo a chi lo osserva di confermare o smentire le proprie tesi e formulare nuove ipotesi. In questo caso la data visualization non è necessariamente la tappa conclusiva di un lavoro ma uno strumento di comprensione.

- **Visualizzare i dati per ammirarli**. Approcciarsi ai dati con finalità puramente artistiche potrebbe apparire un contrasto stridente dato che nel sentire comune arte e matematica sono agli antipodi della conoscenza. Ci sono invece degli esperimenti che guardano alla visualizzazione dei dati come un processo espressivo, caratterizzato prima di tutto da un respiro estetico.

- **Visualizzare i dati per raccontarli**. Nei lavori raccolti in questa categoria lo scopo della visualizzazione non è solo rappresentare dei dati ma raccontare una storia. Scegliere delle informazioni, definire l'ordine in cui presentarle e arrivare anche a delle conclusioni. Gli ambiti in cui questo tipo di processi vengono utilizzati di più sono il data journalism e la divulgazione scientifica.

Le categorie non sono esclusive. Ogni visualizzazione contiene tutte e tre queste tensioni ma è proprio chiarendo di volta in volta qual è la più significativa che si può utilizzare questa classificazione.

17 K. Börner, *Atlas of Science, Visualizing What We Know*, 2010

4. Visualizzare per capire. Bihanic e le forme dei dati

La prima categoria è sicuramente quella che raccoglie al suo interno il maggior numero di esempi, anzi potremmo dire che creare delle forme per capire dei dati è scintilla alla base di ogni visualizzazione.

Forse infatti i primi esempi di data visualization con cui si entra in contatto servono proprio a questo. Basti pensare ai regoli di plastica con cui nei primi mesi di scuola i bambini imparano ad eseguire i calcoli più semplici. Cosa sono se non numeri che prendono forma?

Questo tema è stato affrontato ampiamente da David Bihanic, designer dell'University of Valenciennes and Hainaut-Cambries che nel 2015 ha pubblicato un saggio dal titolo *Giving Shape to Data*[18].
Nel suo elaborato suggerisce un paragone che definisce bene il rapporto che intercorre fra una serie di numeri e le forme che li rappresentano.

“Architecture inspire the best analogy: these piles of data are like stones waiting to be put toward the erection of a building whose fate depends solely on the project and the work that ensues”

I dati sono il materiale con cui costruire un edificio. Il modo in cui vengono assemblati dipende però dal progetto, dalla mano dell'architetto che ha tracciato su carta come dovranno accostarsi gli uni agl'altri.
L'architetto di Bihanic è il designer e l'edificio da costruire è la visualizzazione che appare sullo schermo del suo computer. E qui sono almeno tre concetti su cui è interessante fermarsi.

I dati sono le pietre. Per quanto possa essere fervida l'immaginazione di chi si trova a creare una dataviz, il punto di partenza è sempre una tabella di numeri. Progettare una visualizzazione non vuol dire quindi lasciarsi inspirare da una serie di cifre per creare un prodotto esteticamente valido ma piuttosto lavorare una materia grezza per trasformarla in un prodotto finito.

Il risultato è un progetto. Qualsiasi grafico che venga prodotto partendo da un una serie di dati non è semplicemente la versione più colorata di una tabella. Siamo sempre davanti ad una costruzione, operata per rendere più comprensibile qualcosa che altrimenti sarebbe difficile da capire ma di certo non trasparente.

Il progetto è tracciato da qualcuno. Quando ci si trova davanti ad una visualizzazione non è importante solo capire il processo che ha portato alla sua creazione ma anche lo scopo di chi lo ha realizzato. Sapere a quali domande si è cercato di rispondere rende più chiare le risposte.

La riflessione di Bihanic si sviluppa proprio sulla necessità di creare delle forme per capire i numeri. La data visualization diventa quindi un linguaggio indispensabile non solo per comunicare dei dati ma prima di tutto per capirne il senso, per poterli immaginare e quindi per poter cominciare a riflettere. Creare una dataviz vuol dire in questo senso creare una oggetto che sia più semplice osservare.

Ed è proprio in virtù di questa semplicità che Bihanic suggerisce di abbandonare le visualizzazioni complesse e cercare forme che siano facilmente decifrabili.

18 D. Bihanic, *Giving Shape To Data*, in D. Bihanic *New Challenges for Data Design*, 2015

“Of course , we perceive and interact better wen things are orderly, balanced, and unified than when they are dismantled or disparate. What matters most is not how the shapes are arranged, but rather the quality of the shape themself”.
Così il ricercatore dell'università francese definisce quella che possiamo considerare la forma primordiale della data visualization, una forma in cui ogni elemento grafico, dal colore alle dimensioni, è orientato a rendere comprensibile il dato di partenza.

Uno sforzo di semplicità quindi, finalizzato a creare uno strumento in grado di diffondere e accrescere la conoscenza. Nelle ultime parole del suo saggio Bihanic lascia quasi un mandato ai designer che dovranno occuparsi di questa materia.

“The approach of datta design basically aims at creatively exploring and investigating new ways and methods of representing, visualizing, and processing computer data. To do so, data designer are not refraining from forging other visions and paradigms that radically cut ties with tradition, rules, and designs, which, up to now. Have been adamantly defended and accepted as true. Their ultimate goal is to optimize all research, including those of a seemingly trivial nature, for the purposes of rebuilding the fundation of all knowledge”.

Se queste parole definiscono un orizzonte ideale, non è difficile capire che questi criteri di semplicità e chiarezza vengano adottati soprattutto nell'ambito scientifico dove la visualizzazione di una serie di dati spesso non è il risultato finale di una ricerca ma semplicemente uno strumento per riuscire a comprendere un fenomeno.

Ordinare il sapere. La *Map of Science* di Klanvas e Boyak

Richard Klanvas e Kevin W. Boyak sono due accademici che condividono lo stesso interesse: capire attraverso gli articoli pubblicati nei giornali scientifici come è strutturata la conoscenza, in quali ambiti si sta concentrando la ricerca e quali discipline sono collegate fra loro. E per raggiungere questo obiettivo hanno deciso di utilizzare una dataviz. È da qui che nasce il progetto *Map of Science*[19].

Per raggiungere questo obiettivo tra il gennaio 2001 e il dicembre 2005 hanno collezionato 7,2 milioni di articoli scientifici pubblicati su oltre 16 000 riviste.

Davanti ad una tale mole di dati, la prima operazione per cominciare ad orientarsi è stata quella di definire un criterio di classificazione e la loro scelta è ricaduta sulla letteratura in comune.
Le riviste che citavano una stessa letteratura scientifica simile sono state così raccolte sotto una stessa disciplina e ogni disciplina è stata rappresentata nella dataviz da un cerchio. Più il cerchio è grande e più il numero di riviste che si occupano di quella disciplina è ampio.
Per definire i collegamenti tra i cerchi sono poi andati a vedere quale percentuale di letteratura scientifica ogni disciplina condividesse con le altre. Anche in questo caso è stata inserita una variabile dimensionale. I cerchi sono stati collegati da linee, più la linea è spessa e più la percentuale di letteratura condivisa è ampia.
Dopo aver definito questi due elementi, attraverso un algoritmo tutti i cerchi sono stati distribuiti su una superficie sferica e raggruppati in base ai loro collegamenti. Due discipline completamente scollegate venivano così rappresentate in punti opposti della sfera.

La mappa riprodotta a pagina 21 è stata adatta ad una superficie a due dimensioni. Le distanze fra i

19 http://www.mapofscience.com/

diversi punti non sono quindi propriamente esatte e gli estremi della mappa devono considerarsi come uniti.
Il risultato finale è la creazione di una sorta di planisfero del mondo accademico, dove esistono continenti del sapere, raggruppamenti di discipline con molti collegamenti tra loro.

Oltre a mostrare la complessità di queste dinamiche, la mappa offre anche la possibilità di capire in quali discipline collaborano fra di loro studiosi provenienti da ambiti diversi.

Prendendo spunto dai dati raccolti i due autori hanno anche proposto delle previsioni su quali ambiti saranno collegati tra loro nei prossimi anni.

"If the structure of scienze show below is moving toward stability, we would expect connecredness between nieghboring fields to increase, and connecredness between distant fields to decrease. We found the opposite, suggesting that the underluing structure is unstable and likely to change dramatically over the next decade"[20].

Klanvas e Boyak hanno così costruito anche altre sei piccole mappe in cui, attraverso una serie di frecce bianche e nere definiscono gli andamenti futuri dei principali ambiti di studio. Un'utile traccia per capire dove investire per sviluppare conoscenze o creare collegamenti ancora inediti.

Le mappe qui presentate sono quelle codificate nel 2007, negli ultimi anni il progetto è andato avanti arricchendo il dataset e visualizzando informazioni sempre più complesse.

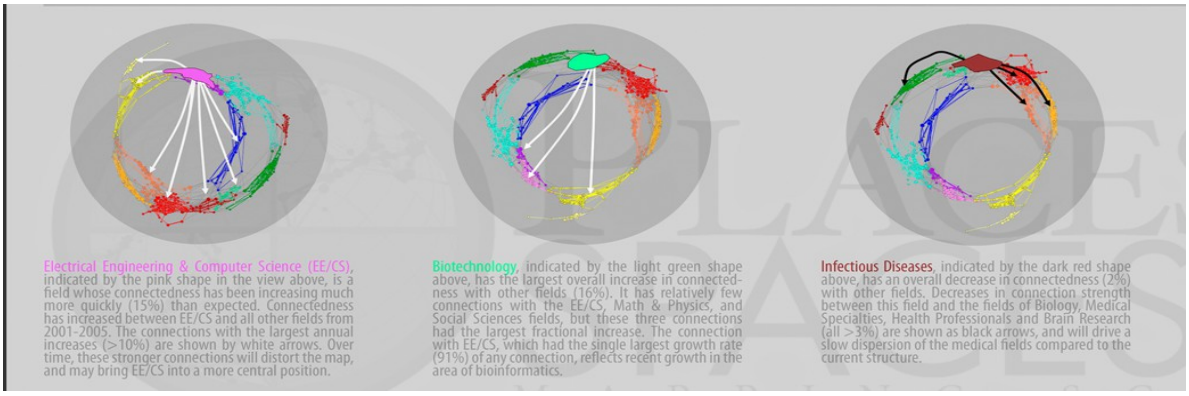


Illustrazione 7: Fonte: http://scimaps.org/maps/map/maps_of_science_fore_50/detail

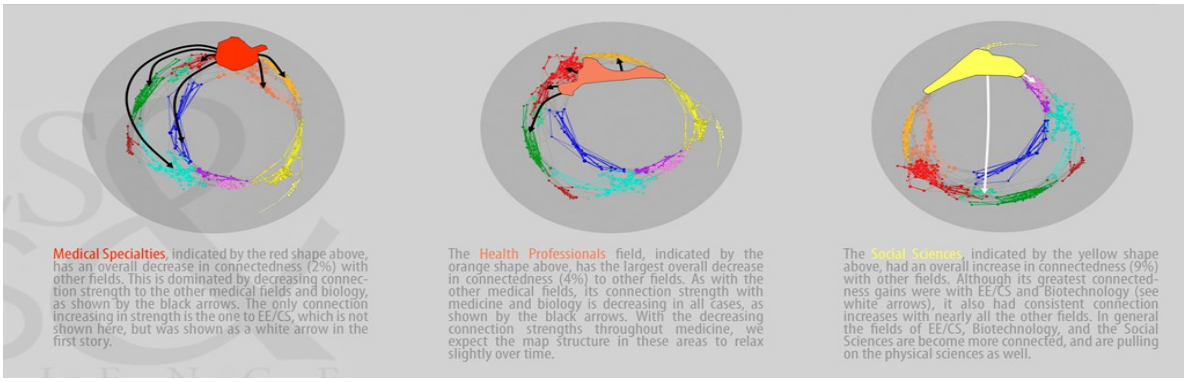


Illustrazione 8: Fonte: http://scimaps.org/maps/map/maps_of_science_fore_50/detail

20 http://scimaps.org/maps/map/maps_of_science_fore_50/detail

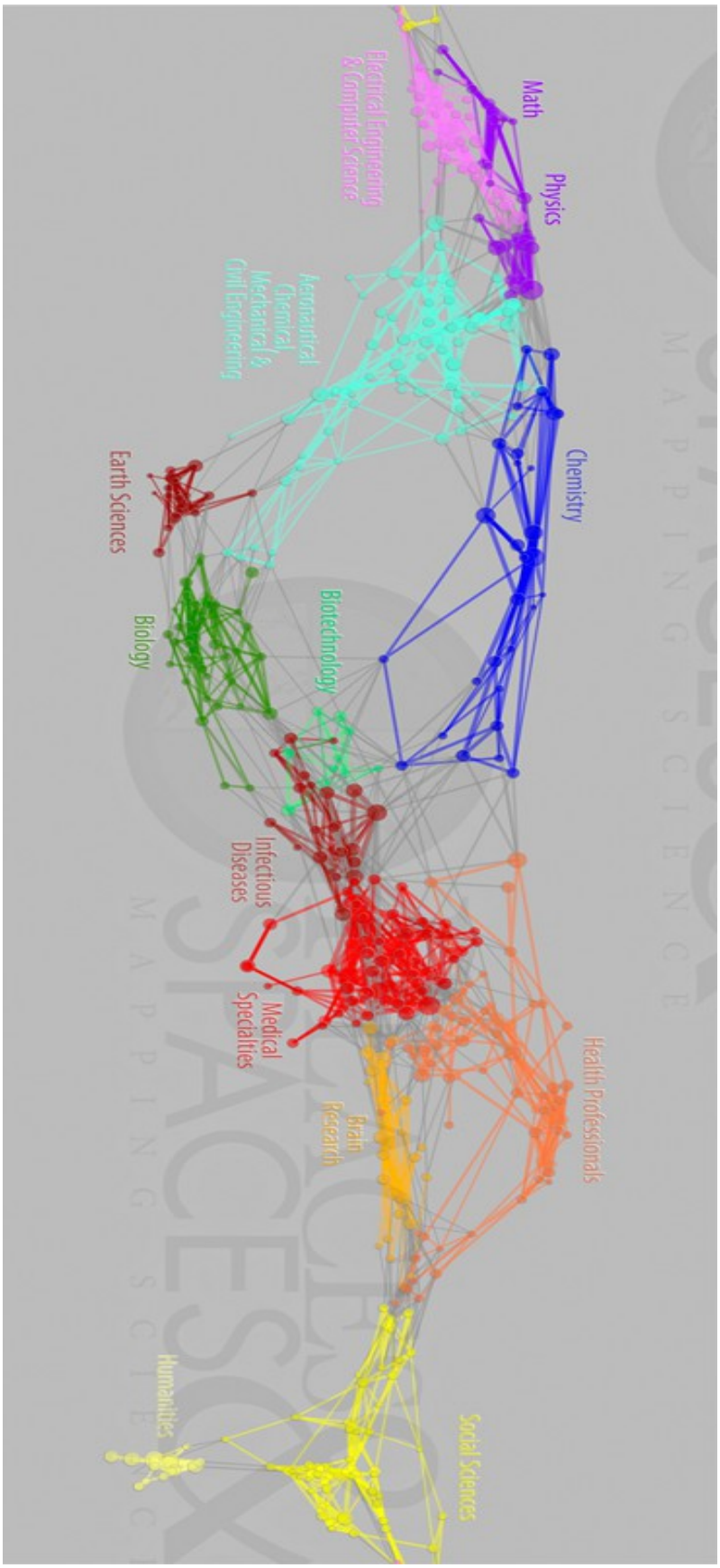


Illustrazione 9: Fonte: http://scimaps.org/maps/map/maps_of_science_fore_50/detail

Barcellona Nascosta. Il progetto *atNight* di Santamaria-Varas e Martinez-Diez

Le tecniche di Data Visualization permettono di analizzare anche fenomeni che sfuggono alla nostra vista. Permettono di rendere visibile quello che visibile non è. Elaborando i dati dei social network è possibile ad esempio ricostruire la rete dei collegamenti fra gli utenti oppure i contenuti che quotidianamente pubblicano sui loro profili. Fenomeni che potrebbero essere visti solo singolarmente si possono mostrare insieme e fenomeni che lascerebbero la loro traccia solo su una striscia di codice informatico diventano visibili.

Mar Santamaria-Varas dell'Universitat Politecnica de Catalunya e Pablo Martinez-Diez del Design College of Barcelona hanno deciso di partire da questa premessa per svelare un aspetto intangibile della loro città.
È nato così nel 2013 il progetto *atNight*[21] che mostra cosa succede a Barcellona quando tramonta il sole e le sue vie vengono illuminate dai lampioni, dai fari dei taxi in corsa e dagli schermi dei telefoni. Un progetto che i due designer hanno accuratamente descritto in un articolo dal titolo *atNight: Nocturnal Landscape and Invisible Networks*[22].

Lo scopo dichiarato di *atNight* è quello di descrivere un territorio attraverso vari tipi di dati, fra cui spiccano quelli raccolti sotto la definizione di impronta digitale ossia la traccia che ogni utente lascia in rete, fatta di foto, tweet e geolocalizzazioni.

“Techinical advancements over the past decade have completely changed the way we sense, seize, use, plan and build present and future cities. Besides architecture of works. While physically experiecing the city, inhabitants also generate a digital footprint, a generous amount of data which describes people needs, beliefs and reaction”.

La base di partenza è stata fornita quindi dalle tradizionali mappe topografiche della città, alle quali sono stati aggiunti dati appartenenti a tre tipologie:

- **Open Data**. Informazioni provenienti da enti pubblici riguardanti la disposizione delle strade, la demografia delle aree abitate o l'utilizzo del suolo.

- **Mobilità e Energia**. Grazie ad una accordo con aziende pubbliche e private i due designer sono riusciti ad utilizzare dati che riguardavano il traffico stradale e il consumo energetico.

- **Digital Footprint**. L'ultima categoria si concentra sui dati geolocalizzati provenienti dai social network utilizzati da turisti e cittadini, è stato possibile raccogliere queste informazioni attraverso un sistema di API[23]. Le piattaforme monitorate sono state Flickr, Panoramico, Instagram e Twitter.

A questo punto è cominciata la fase più sperimentale del progetto in cui alle mappe topografiche della città sono stati aggiunti i dati ottenuti dalla ricerca. È stato un procedimento che si è sviluppato per tentativi. Prima sono state inserite tutte le registrazioni, rappresentate da punti che identificavano un solo fenomeno, poi i dati sono stati aggregati, tradotti in forme e definiti da

21 http://www.atnight.ws/

22 M. Santamaria-Varas e P. Martinez-Diez, *atNight: Nocturnal Landscape and Invisible Networks* in D. Bihanic *New Challenges for Data Design*, 2015

23 Le API, Application Programming Interface, sono proceduredisponibili al programmatore, di solito raggruppate a formare un set di strumenti specifici per l'espletamento di un determinato compito all'interno di un certo programma. Spesso con tale termine si intendono le librerie software disponibili in un certo linguaggio di programmazione.

varianti grafiche come il colore o le dimensioni.
Alla fine di questo processo sono state quindi identificate tre linee guida: Visual Structure, Mobilty Pattern e Activity.

Visual Structure

Questa è la categoria più ricca di dataviz, infatti al suo interno contiene altri tre ambiti di analisi.

Struttura del Territorio e Identità

Qui sono state realizzate due mappe gemelle: *Constellation Barcelona* e *Barcelona is Barcelona*. Entrambe prendono i dati dei social network e li riportano sulla cartina della città.
In *Costellation Barcelona* ogni tweet e ogni foto geolocalizzata diventa un punto bianco che sullo sfondo nero del territorio sembra disegnare una foto notturna scattata dal satellite.
La sua mappa gemella però affronta un'ulteriore scrematura. *Barcelona is Barcelona* non rappresenta con dei punti tutti i contenuti geolocalizzati ma solo quelli che contengono anche la parola “Barcelona”. Gli agglomerati di punti si collocano quindi in prossimità di luoghi d'interesse come la Sagrada Familia, il Parc Güell o la Rambla.
Nella prima visualizzazione vengono svelati i posti dove si utilizzano di più i social network mentre nella seconda ci sono quelli che cittadini e turisti ritengono più rappresentativi della città catalana.

Notte e Giorno

La serie di mappe presenti in questa sezione descrive due città diverse: quella che vive alla luce del giorno e quella che si illumina quando scende la sera.
Anche qui tutti i contenuti geolocalizzati dei social network sono rappresentati da punti ma questa volta vengono utilizzati due colori differenti. C'è infatti l'arancione per quello che viene pubblicato di giorno e il blu per quello che viene pubblicato di notte. Osservando *Barcelona Night and Day* ci sono punti della città che sono segnati sia di giorni che di notte e mentre altri, come le periferie, aspettano l'arrivo del buio per animarsi.

Città Visibile, Città da Vivere

L'ultimo ambito di questa prima parte del progetto riguarda la sentiment analysis, ossia l'analisi degli stati di umore espressi sul web dagli utenti. In *Visibile City, Living City* tutti i contenuti raccolti sono stati analizzati attraverso un algoritmo semantico che ha definito quali contenevano un'opinione negativa e quali una positiva.
In *Barcelona Sentiment/Night* e *Barcelona Sentiment/Day* invece, come per il resto del progetto, questa analisi prende in esame le differenze tra notte e giorno.

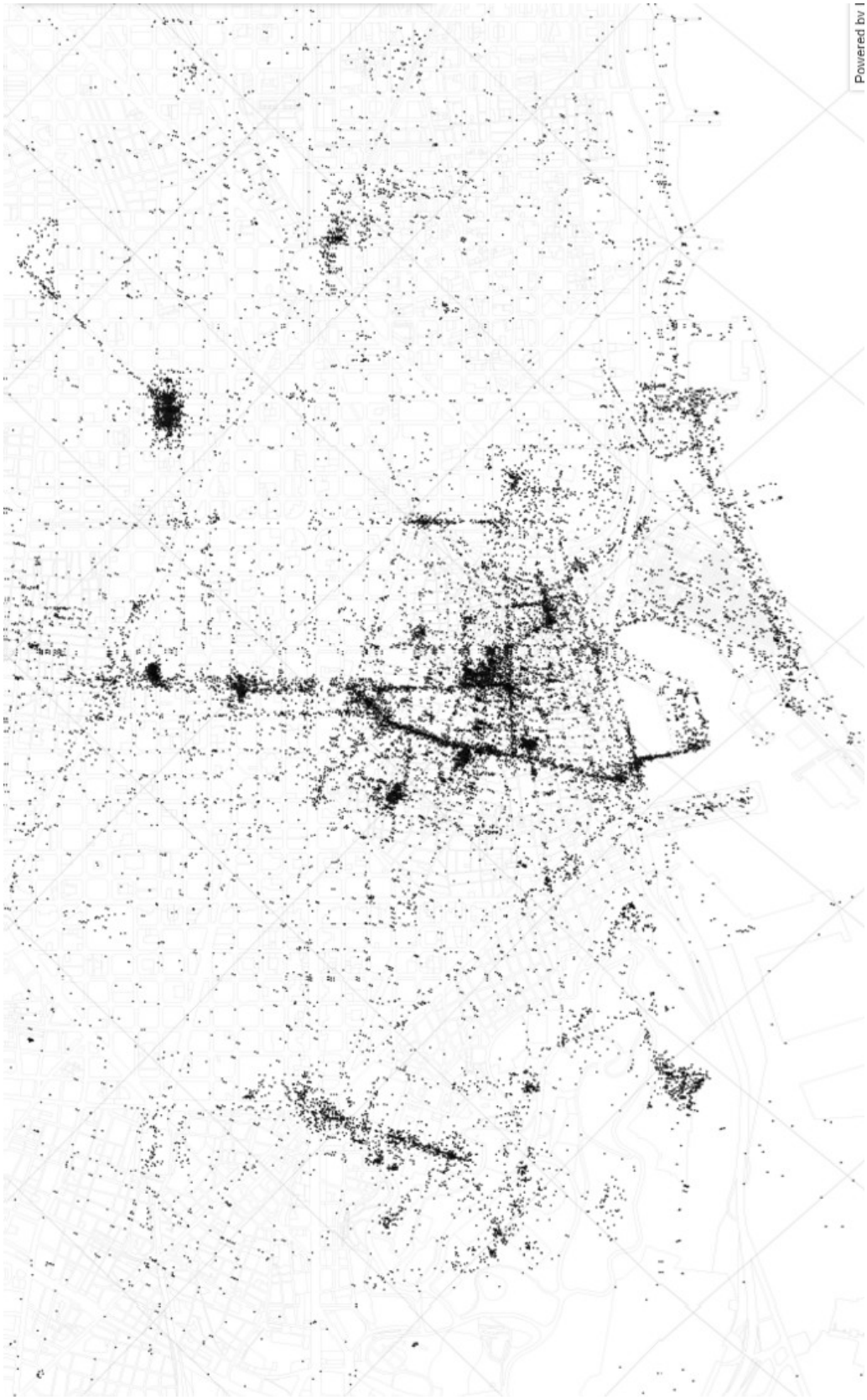


Illustrazione 10: Barcelona is Barcelona,
http://www.atnight.ws/cartographies.php#.VysYHzCLS00

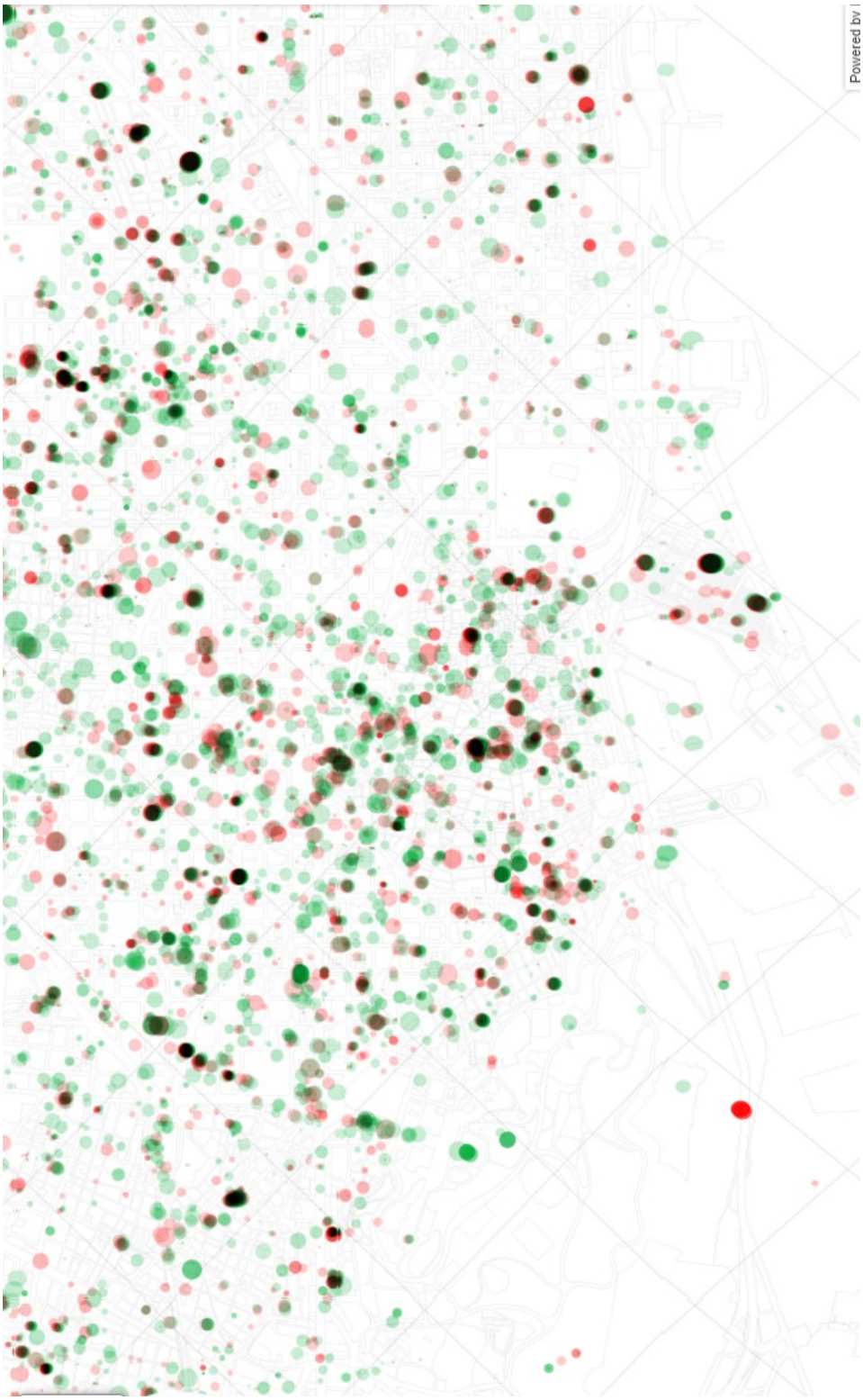


Illustrazione 11: Barcelona Sentiment/Night,
http://www.atnight.ws/cartographies.php#.VwY0oqSLS00

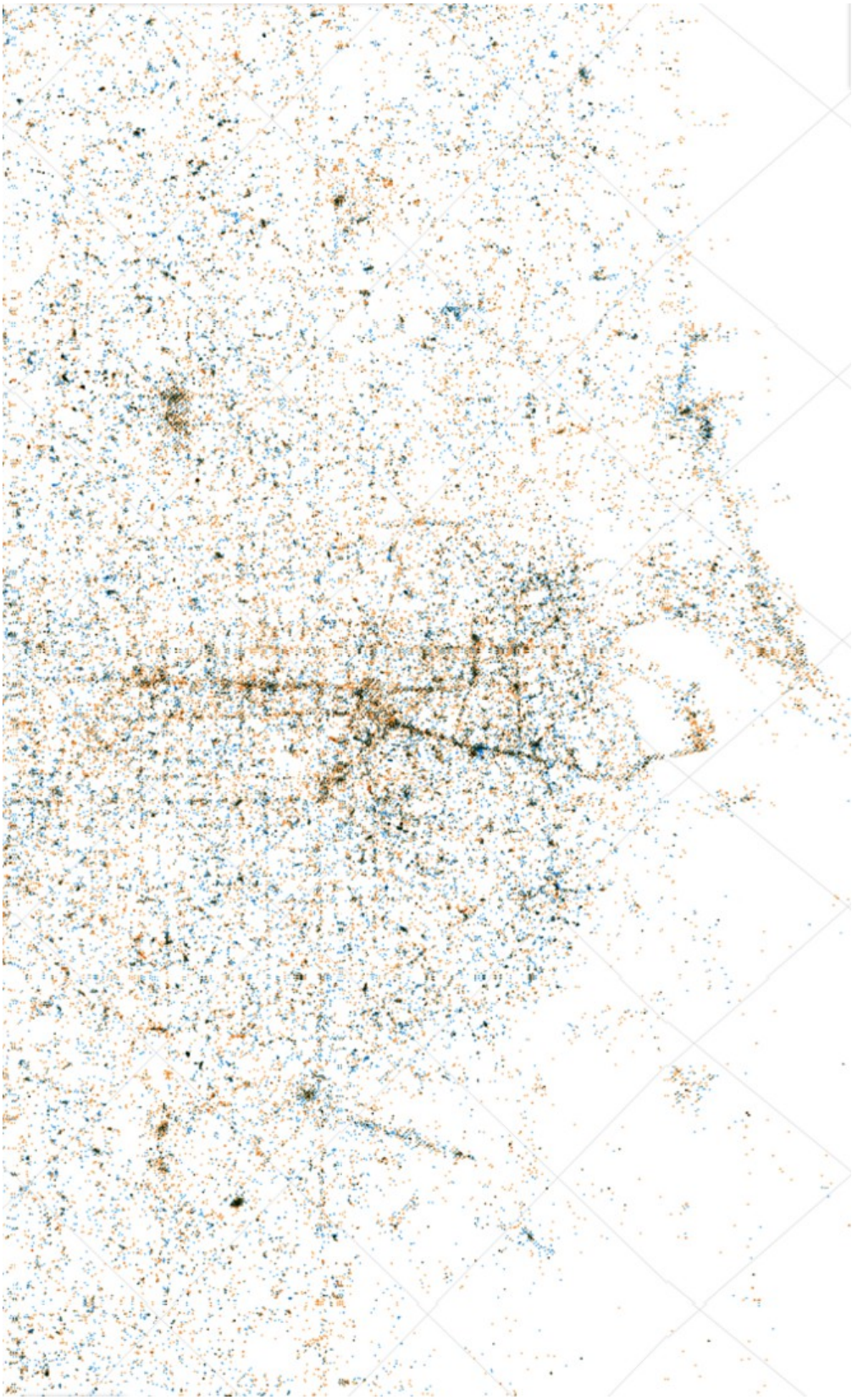


Illustrazione 12: Barcelona Night an Day,
http://www.atnight.ws/cartographies.php#.VwY0oqSLS00

Mobility Patterns

Tweet, accessi e foto. Fino a questo momento i dati utilizzati per rappresentare Barcellona hanno riguardato soltanto fenomeni singoli, verificatisi in un tempo ridotto e uno spazio circoscritto.
La sezione del progetto che si intitola Mobility Patterns vuole invece provare a mostrare la città anche attraverso i suoi flussi, sia che si tratti dei passaggi di taxi che fanno la spola dalla periferia alla Rambla o dei dei tragitti percorsi dalle bici del servizio di Bike Sharing.
I dati raccolti su questi fenomeni, come abbiamo già potuto vedere, vengono poi sovrapposti ad altri e soprattutto viene di nuovo messo in campo il paragone fra notte e giorno.
Fra tutte le mappe pubblicate si può notare anche *Movement vs. Density* in cui un confronto fra gli itinerari geolocalizzati dei taxi e il numero di abitanti per quartiere mette in evidenza quanto alcune delle strade più trafficate passino in aree in cui non vive nessuno. Le zone più battute dai tassisti sono infatti quelle centrali dove però la densità di abitanti è estremamente ridotta.

Activity and Usage

Le ultime due mappe del progetto sci concentrano sulla rappresentazione di come vengono vissute le diverse zone di Barcellona. Nella prima, *Night Time*, sono rappresentati gli accessi a Google durante la notte, evidenziando così le aree dedicate alla vita notturna della città, al riposo e al lavoro. L'ultima visualizzazione sfrutta invece i dati dei social network, isolando questa volta delle triadi di parole inerenti a momenti di riposo. I punti segnati sul tessuto urbano pur concentrandosi nelle vie attorno alla Rambla formano un tessuto continuo che arriva fino in periferia.

Come sarà la Barcellona del futuro?

Tutti questi dati non servono solo a capire i segnali nascosti di una città ma identificano anche le tendenze da seguire per un futuro piano di sviluppo urbanistico. Quelli rappresentate dalle mappe di *atNight* non sono infatti tutti i dati che si possono raccogliere in un ambiente urbano.
Il modello a cui si guarda ora è quello delle Smart City, piene di sensori per misurare ogni tipo di parametro, dal traffico alla velocità. Una marea di dati sta già arrivando sui tavoli degli ingegneri e degli architetti che si occupano di pensare al futuro delle metropoli di tutto il mondo. Una marea di dati che cambierà il modo di pensare all'urbanistica, come sostengono Mar Santamaria-Varas e Pablo Martinez-Diez.

“In conclusion, it is crucial to delve into this line of research. Today, urban planning still relies upon traditional cartographic information (topography, plot division, usage) and neighbourhood-level statistics. These long-established practise are inadequate in comprehending the interaction of citizenship on the urban skin and the territory. The newly and accessible cartographic information will provide an inestimable tool for citizen empowerment, enabling individuals to take collective decisions about the intagible city we actually inhabit”.

Illustrazione 13: Movement Vs. Density,
http://www.atnight.ws/cartographies.php#.VwiHUqSLS00

Illustrazione 14: Fiesta, Food & Safety,
http://www.atnight.ws/cartographies.php#.VwiHUqSLS00

5. Visualizzare per ammirare. Nuove forme di espressione

Forse questo capitolo potrebbe sembrare solo un divertissment. Eppure l'estetica ricopre un ruolo fondamentale nel processo di data visualization. Che si tratti della scelta dei colori, del tipo di forme o della disposizione dei grafici nello spazio, nella costruzione di una dataviz rimangono sempre delle scelte da prendere che necessitano anche di una sensibilità artistica.

Questo aspetto non è preponderante in tutte le visualizzazioni. Basti pensare alla differenza tra quelle create in campo scientifico e quelle in campo giornalistico.
Per quanto riguarda le prime il fattore più importante riguarda la fedeltà dei dati riportati mentre una dataviz stampata sulle pagine di un quotidiano deve anche tenere conto di tutti i meccanismi di graphic design che concorrono a rendere quell'immagine interessante per il lettore e gradevole da fruire.

Ci sono però casi in cui l'elemento estetico prende il sopravvento sugli altri e la data visualization diventa prima di tutto una forma di espressione.

Una dataviz al museo. I pittori astratti dello studio Pitch Interactive

Sono molti i luoghi in cui si possono trovare dei grafici. Ci sono quelli più tradizionali, quasi scontati, come le pagine dei libri di scuola, i report trimestrali delle aziende o gli articoli di giornale. Ma esistono anche dei posti in cui un grafico sembrerebbe quasi fuori contesto come ad esempio le pareti di una galleria d'arte.
Eppure la data visualization è arrivata anche in questo ambiente e non per essere inserita all'interno di un pannello illustrativo su un pittore o un movimento artistico ma per essere esposta al pari di un'opera figurativa.

Dal dicembre 2015 al marzo 2016 si è tenuta a Londra *Big Bang Data*[24], una mostra che ha raccolto visualizzazioni realizzate da giornalisti, designer e società che si occupano di analisi dei dati. E questo non è certo un caso isolato.

Wesley Grubbs è un data designer dello studio Pitch Interactive che ha la sua base a Berkeley, in California. Nel 2013 insieme ai suoi collaboratori ha realizzato un'opera dal titolo *Invisible Montpellier* commissionatagli da La Panacèe, un museo dedicato alle nuove forme di comunicazione che si trova proprio nella città a sud della Francia.

Il suo approccio verso la resa grafica della visualizzazione dei dati è definito dall'articolo *A Process Dedicated to Cognition and Memory*[25], pubblicato su *New Challenges for Data Design* di David Bihanic.

Il paradigma della sfumatura estetica che caratterizza i suoi lavori si può leggere in un sua riflessione in cui accosta la comprensione di un sistema complesso alla meraviglia suggerita da un paesaggio naturale.

“Design is critical in the work we do, and we spend a significant amount of time during the production and postproduction focusing on colors, type, alignment, and other visual elements because they are a crucial part to the comunication aspect of visualization. The human brain is

24 http://bigbangdata.somersethouse.org.uk
25 W. Grubbs, *A Process Dedicated to Cognition and Memory*, in D. Bihanic *New Challenges for Data Design*, 2015

wired to not only process but to attract to intricate imagery, for example, looking at a mountain range or watching waves on a beach. […] When intricate imagery has a story embedded whitin, it only stimulates our interest more".

Per Grubbs le visualizzazioni complesse di grandi quantità di dati non sono totalmente artificiali ma si inquadrano in una serie di esperienze visive già familiari all'uomo, come può essere il continuo infrangersi delle onde sulla sabbia o una catena montuosa. Nei suoi lavori quindi cerca di riprendere proprio questi elementi, cerca di riportare graficamente delle informazioni sfruttando pattern che richiamino il paesaggio naturale. Non si serve di diagrammi che ritiene troppo astratti come i grafici a barre. E questo non si riferisce solo alle forme ma anche alla composizione e alla scelta dei colori.

"What I try to do is create a visual or esthetic that borrows forms or colors from shapes I find in nature whenever possible. For example, my color picking process come from photographs that I take of natural formations: sunset, daiseis and cloud. Nature is far superior at picking and matching colors than I am".

Il suo modo di approcciarsi alla data visualization è stato notato anche dal Moma, il Museum of Modern Art di New York. Nel 2010 lo studio Pitch Interactive aveva realizzato una dataviz che mostrava il flusso giornaliero di chiamate fatte dai cittadini di New York al 311, il numero per segnalare situazioni che richiedono un intervento delle forze dell'ordine per situazioni non urgenti, come la rimozione di un'automobile in divieto di sosta o la caduta di un albero in strada che non ha provocato danni. Questo lavoro è stato prima pubblicato sulla rivista *Wired* e poi esposto al Moma nel 2011 nell'ambito della mostra *Talk to Me*[26].

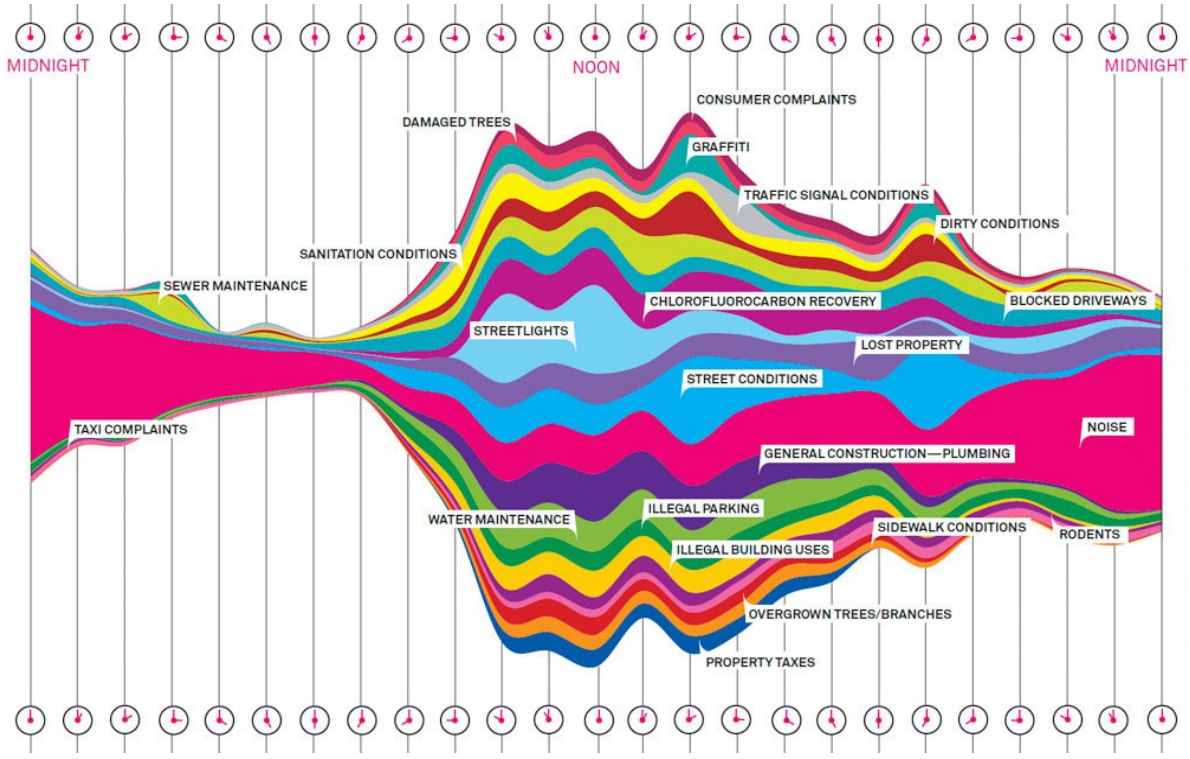


Illustrazione 15: Fonte: http://www.wired.com/2010/11/ff_311_new_york/

26 http://www.moma.org/interactives/exhibitions/2011/talktome

Dopo aver visto il lavoro esposto nel museo di New York, nel 2013 i responsabili dal La Panacèe di Montpellier hanno deciso di commissionare a questo studio di design una dataviz analoga ma basata sui dati della loro città.

Le condizioni di partenza non erano però le stesse. I dati a disposizione per la pubblicazione su Wired erano stati raccolti in due settimane, si riferivano alle chiamate fatte ad un unico numero e per ogni telefonata era stata registrata non solo l'ora ma anche il problema segnalato. Era quindi possibile aggregare questi dati e definire un flusso giornaliero che teneva conto sia del numero di chiamate fatte nei diversi orari della giornata che dei picchi dei diversi tipi di segnalazione. Ad esempio si può vedere che l'argomento *Noise* occupa sopratutto le fasce serali e notturne mentre *Street Condition* è presente solo durante il giorno.

Il dataset a disposizione per il lavoro chiesto su Montpellier era più vario. I dati raccolti infatti si riferivano a segnalazioni non urgenti fatte non ad un unico operatore ma a diversi soggetti e con diversi mezzi di comunicazione, dal telefono alle lettere. Non si poteva così tracciare un flusso giornaliero ma bisognava definire dei criteri entro cui ordinare tutte le segnalazioni. Questi sono quindi le chiavi con cui i dati sono stati classificati.

- ID Type
- Receive Date
- Closed Date
- Method
- Agency Contacted
- Adress of Complaint Source
- Quartiers

Il periodo in cui erano stati raccolti questi dati coincideva poi con tutto il mese di gennaio 2010 ma i numero di record raccolti era parecchio basso rispetto a quelli su cui i designer avevano già lavorato, 1 000 contro 35 000. Gli abitanti di Monpellier infatti sono appena il 2% di quelli di New York. Era impossibile quindi replicare il lavoro su Wired, non c'erano abbastanza dati per definire dei pattern e sopratutto mancava l'indicazione temporale che avrebbe consentito di costruire un flusso giornaliero.

C'era però nei dati francesi un'informazione in più. Ogni segnalazione era geolocalizzata, era indicato il luogo esatto da cui era partita. Il team di Picth Intercative ha così deciso di utilizzare questa come chiave di volta per la sua visualizzazione.
I punti in cui era stato fatto lo stesso tipo di segnalazione sono stati indicati sulla mappa della città e uniti per creare un poligono. In questo modo sono state sovrapposte diverse forme, ognuna delle quali riferita ad un argomento diverso, dalla presenza di graffiti alla spazzatura lasciata in strada. Per distinguerle le une dalle altre è bastato caratterizzarle con colori diversi.

Create le prime bozze, questa squadra di designer ha voluto però orientare il suo lavoro anche in funzione del luogo in cui sarebbe stato esposto.

“We decide to explore more abstract visual rapresentation of the data that we had. After all, this piece was for an exhibit in an art museum. We mixed statiscal analysis with esthetic exploration”.

Dopo aver definito quindi una palet di colori, delle forme che rappresentassero dei dati risultando però esteticamente attraenti e una composizione che riuscisse a mettere insieme tutti questi

elementi, l'ultimo compito rimasto era trovare qualcosa che permettesse alla dataviz di avere un solido ancoraggio alla realtà. Qualcosa che ricordasse subito al visitatore che quell'opera non era solo un quadro astratto. Dopo aver provato con la mappa dei trasporti pubblici, troppo invadente e troppo carica di informazioni, è stato così scelto il fiume Le Lez. Il corso d'acqua scorre infatti accanto alla parte est di Montpellier definisce una collocazione geografica riconoscibile.
Basta il colpo d'occhio per capire quanto questa mappa si diversa da quelle del progetto *atNight*. I dati di partenza sono molto simili, c'è sempre la pianta di una città e ci sono sempre dei dati da sovrapporre. Quello che cambia è proprio lo scopo con cui viene realizzata la visualizzazione. Mentre le mappe di Barcellona erano improntate solo a svelare al pubblico i lati nascosti della metropoli, qua l'obiettivo è creare qualcosa che colpisca esteticamente lo spettatore. Così la dataviz creata assomiglia quasi ad un quadro astratto.

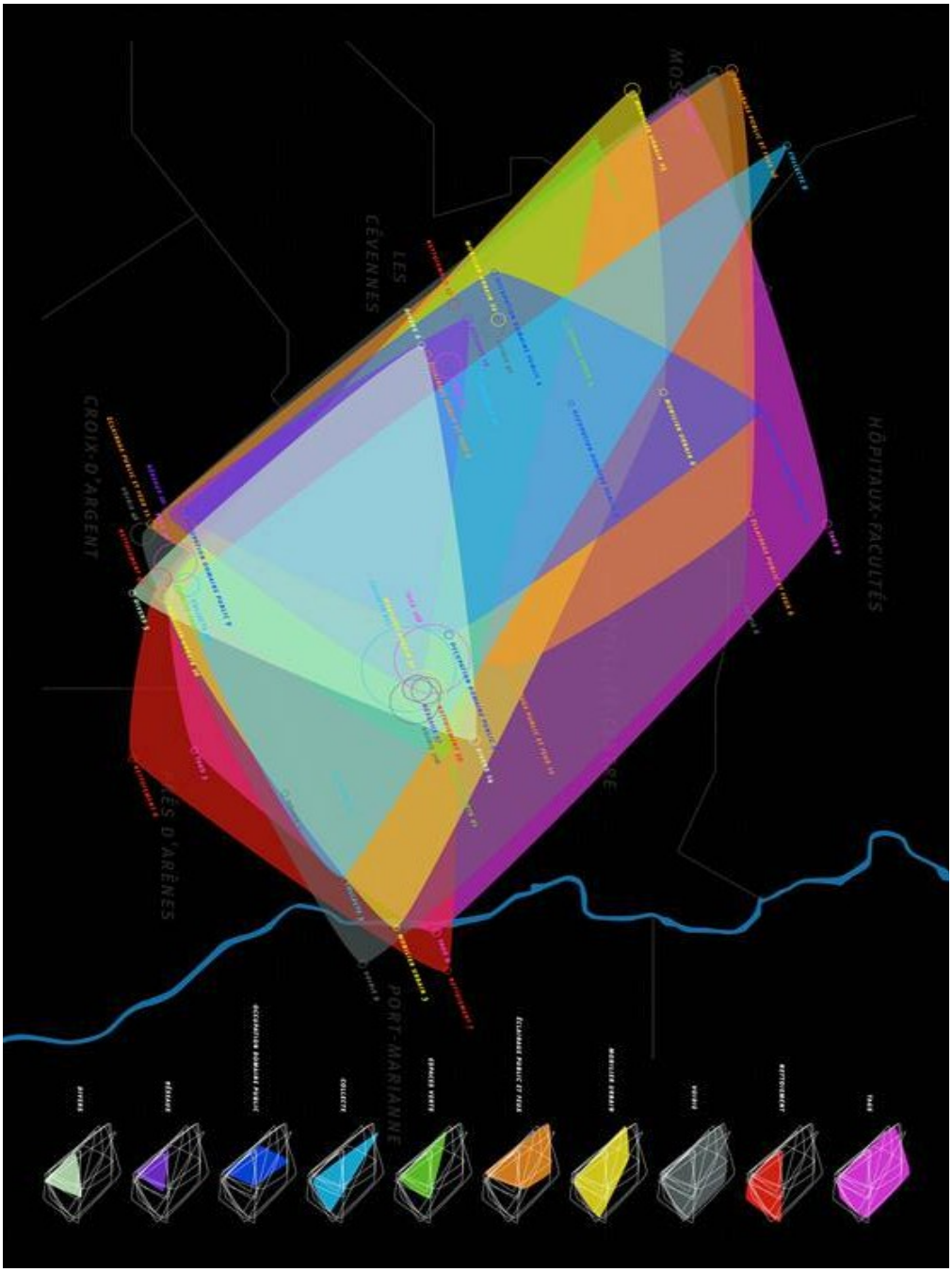


Illustrazione 16: Fonte: http://pitchinteractive.com/work/Montpellier.html

Un diario di grafici. I report di Nicholas Felton

Tra il 2009 e il 2010 Geoff McGhee ha girato un documentario dal titolo *Journalism in The Age of Data*[27]. Nel corso delle riprese ha intervistato i designer, gli informatici e i giornalisti che allora stavano muovendo in primi passi in quella disciplina che ora è ben conosciuta come data journalism.

Fra i professionisti intervistati in questo documentario c'era anche Nicholas Felton[28], un designer di New York che nel 2005 ha cominciato un esperimento di data design. Ha deciso di servirsi dei grafici per raccontare la sua vita e così, invece che scrivere un diario o aprire un blog, ha cominciato a pubblicare dei report annuali con tutti i dati delle sue giornate.

Il primo di questi lavori è nato nel 2006 con lo scopo di riprendere in mano tutti i fatti accaduti l'anno precedente. La base di partenza sono stati i dati che aveva già a disposizione, come gli eventi segnati in calendario o la cronologia di LastFm.com[29], una delle prime web radio dove gli utenti registrati potevano pubblicare le canzoni ascoltate. Un progetto quindi nato per svago, per cominciare a maneggiare un linguaggio che solo negli anni successivi avrebbe rivelato le sue potenzialità. Ecco come Felton commenta infatti il suo lavoro davanti la telecamera di McGhee

"I tought would by interesting to friends and family and turned out to have a much broader appeal".

I suoi lavori sono diventati così sempre più complessi. Ha ampliato il numero di fenomeni da monitorare, arrivando ad esempio a registrare quotidianamente la sua temperatura corporea, e ha approfondito quelli che già lo interessavano. Se del 2005 ad esempio vengono riportati i voli in aereo, già nel 2007 sulle pagine del suo report sono rappresentati tutti i 561 viaggi fatti con la metropolitana della sua città o le 138 corse in taxi.

Questi report, a metà fra arte figurativa e scritto autobiografico, sono stati notati anche da chi si occupa di data visualization da una prospettiva del tutto diversa.

Fernanda Viegas è una delle sviluppatrici di Many Eyes, il software per l'elaborazione grafica dei dati lanciato da Ibm. Ecco come commenta il lavoro di Felton nel quinto capitolo di *Journalism in the Age of Data* intitolato *Life as a Data Stream*[30].

"It is a reflection of how much data visualization is becoming a new expressive language".

Oltre a comporre ogni anno i suoi diari, Felton si è anche impegnato a creare degli strumenti affinché chiunque possa registrare e visualizzare i propri dati. Così ha fondato assieme a Ryan Case un sito che permette attraverso una serie di interfacce visuali di creare dei lavori simili ai suoi report, segnando di giorno in giorno i chilometri percorsi, i caffè bevuti o qualsiasi altro fenomeno si voglia monitorare. Il portale si chiama Daytum[31] e bastano pochi click e un breve tutorial per cominciare a leggere la propria vita attraverso i dati.

La quantità di informazioni a disposizione per questo tipo di lavoro è cresciuta sempre di più, arrivando a segnare un'impennata con la diffusione degli smartphone, vere e proprie sonde portatili

27 http://datajournalism.stanford.edu
28 http://feltron.com
29 http://www.last.fm
30 https://www.youtube.com/watch?v=IV1yk1DFeCY
31 http://daytum.com

in grado di misurare e visualizzare quasi ogni tipo di attività.
Già nell'anno dell'intervista di McGhee era chiaro l'interesse di Felton per questi strumenti che in quel periodo avevano appena iniziato la loro diffusione di massa. Il primo modello di iPhone era stato lanciato solo due anni prima.

Se già nel 2010 questo designer era suggestionato dalle possibilità che avrebbero offerto gli smartphone del futuro per raccogliere informazioni sui loro possessori non è difficile sapere come questa sua intuizione si sia trasformato negli ultimi anni. Nel 2014 ha lanciato Reporter[32], una app nata con lo stesso scopo di Daytum ma con il vantaggio di poter essere aggiornata sempre e con pochi tocchi.

L'interesse di Felton sulla ricerca di modalità diverse di raccontare la sua vita ha incuriosito anche uno dei colossi della comunicazione contemporanea che basa il suo successo proprio sulla condivisione da parte dei suoi utenti di informazioni personali. Ha lavorato infatti insieme a Facebook per la progettazione della funzione *Timeline* che in Italia è arrivata come *Diario*.

L'ultimo suo report risale al 2014 ed è totalmente diverso dagli altri per un aspetto: l'automazione. Come si legge nella nota che accompagna l'edizione ormai il mondo dei dati è cambiato. Se nei primi report doveva utilizzare delle soluzione studiate ad hoc per registrare i suoi spostamenti o il numero di ore spese in determinate attività, ora ci sono app in grado di memorizzare tutto e fornire anche immediatamente soluzioni di visualizzazione.

“This is the tenth and final Feltron Annual Report. The world of personal data has changed considerably since the project began in 2005 and this edition attempts to capture its current state. While previous editions have relied on custom solutions to gather ethereal personal data, this edition is based entirely on commercially available applications and devices. Using an array of products and software, the author’s car, computer, location, environment, media consumption, sleep, activity and physiology were instrumented and logged”[33].

32 http://www.reporter-app.com
33 http://feltron.com/FAR14_07.html

Illustrazione 17: Report 2005, Fonte: http://feltron.com/FAR05.html

Illustrazione 18: Report 2007, Fonte: http://feltron.com/FAR07.html

Illustrazione 19: Report 2014, Fonte: http://feltron.com/far14.html

6. Visualizzare per raccontare. Simon Rogers e le 10 regole del data journalism

Gli esempi riportati nelle pagine precedenti hanno mostrato come si possano visualizzare dei dati per comprenderli meglio o per trasformarli in un'espressione artistica. In questo capitolo invece l'attenzione sarà concentrata su come combinare queste due pratiche per arrivare ad un fine diverso: il racconto.

Il contesto migliore per capire questa forma di data design è il data journalism, il giornalismo che nasce dai dati.

Il titolo di pioniere per questo modo di fare informazione va a Philip Meyer, un giornalista americano che all'inizio degli anni '70 ha pubblicato *Precision Journalism: A Reporter's Introduction to Social Science Methods*[34].

Prima di arrivare a questo libro Meyer aveva maturato una grande esperienza come cronista, tanto che già nel 1968, a 38 anni, era riuscito ad ottenere il premio Pulitzer grazie ad una serie di reportage realizzati assieme ad altri colleghi sulle rivolte della comunità afroamericana che in quegli anni si era riversata per le strade di Detroit.

Dopo il premio decise di prendere un anno sabbatico per frequentare dei corsi alla Harvard University di statistica e sociologia. Da questi spunti nacque l'intuizione per un giornalismo orientato ad utilizzare le tecniche proprie della ricerca scientifica per scovare e approfondire notizie. Le sue idee introdussero nel giornalismo l'uso dei sondaggi, della statistica e dei computer, quando ancora questi non avevano conquistato tutte le scrivanie del pianeta.

La vera rivoluzione è arrivata però all'inizio degli anni 2000, assieme all'esplosione dei Big Data. Le librerie di dati sono diventate sempre più diffuse e facilmente accessibili. Se si vuole capire la portata di questo cambiamento è utile osservare i siti nati negli ultimi anni per garantire la trasparenza della res publica. In tutto il mondo amministrazioni locali e nazionali hanno cominciato a pubblicare on line i loro dati tabelle di numeri relative a diversi settori, dall'istruzione all'inquinamento, dalla sanità alla sicurezza.

I cittadini americani che vogliono approfondire un fenomeno come i crimini commessi nella loro contea possono così utilizzare il portale data.gov[35] e scaricare tutte le informazioni a cui sono interessati. In Italia questo è possibile grazie a diversi progetti.

Su OpenParlamento[36] si può monitorare tutto l'operato dei politici eletti alla Camera e al Senato, analizzando la loro presenza in aula, le mozioni che hanno proposto e persino ogni voto che hanno espresso nella durata della loro carica.

Anche alcuni amministrazioni locali hanno cominciato a lavorare in questo senso. Un ottimo esempio è OpenLombardia[37], il sito di Regione Lombardia che sul modello americano di data.gov pubblica costantemente dati a disposizione di ogni cittadino.

Non ci è voluto molto perché miniere di notizie come quelle qui citate cominciassero a far gola ad

34 P. Meyer, *Precision Journalism: A Reporter's Introduction to Social Science Methods*, 1970
35 https://www.data.gov
36 http://parlamento17.openpolis.it
37 https://www.dati.lombardia.it

un'ampia fetta di giornali. I dati di cui ha cominciato a servirsi il mondo dell'informazione non sono però solo questi. Ci sono anche dati che le redazioni possono riuscire ad ottenere da sole come quelli raccolti nei social network.

Se Philp Meyer ha teorizzato il precison journalism ad averlo riportato nell'era dei Big Data è stato Simon Rogers, fondatore del datablog del *The Guardian* ed ora data editor di Google.

Grazie all'aiuto del suo team di giornalisti, designer e informatici è riuscito a diffondere un modo di approcciarsi all'informazione basato sull'elaborazione e la visualizzazione di grossi dataset. Ormai infatti la definizione di precision journalism ha definitivamente lasciato il posto a quella di data journalism.

Nel suo libro *Fact are Sacred* Rogers spiega i metodi e le fonti adottate da questa nuova forma di informazione partendo proprio della sua esperienza al *The Guardian*. Le prime pagine di questo volume sono dedicate ad un elenco di dieci cose che il lettore apprenderà prima di arrivare alla fine. È difficile non vedere in questa lista un manifesto del data journalism, una serie di principi che lo regolano o soprattutto lo definiscono. Ecco quindi il vademecum lasciato dal giornalista britannico per creare delle notizie partendo dai dati.

It may be trendy but it's not new

La prima riflessione è già stata ampiamente trattata in questo elaborato nel corso della sezione sulla storia della data visualization.
Si riferisce al fatto che informare partendo dai dati non è nulla di nuovo. I primi esempi di questa pratica risalgono addirittura al XIX secolo. Quello che è cambiato è l'accessibilità ai dati. Se infatti negli scorsi secoli questi erano pubblicati in libri tanto voluminosi quanto costosi, ora in molti casi si possono trovare gratuitamente raccolti in semplici tabelle. In questo modo non è più l'uomo a doversi occupare della loro elaborazione ma può affidare il compito ad un computer, dopo averlo istruito con i giusti comandi.

Open data means open journalism

Le statistiche sono diventate alla porta di un click. Basta avere sul proprio computer un qualsiasi pacchetto software in grado di leggere una tabella ed è possibile analizzare e dare forma ai numeri. Certo, non è detto che tutti possano realizzare dei lavori di qualità. Proprio questa semplicità di fruizione rende però i dati in grado di entrare in molte delle notizie di cui si discute all'interno di una redazione. Il data journalism è accessibile a tutti.

Data journalism is sometimes curation

A volte il giornalista è costretto a diventare un vero e proprio curatore di dati. Si deve occupare di selezionare le fonti giuste, pulirli, ordinarli, visualizzarli e presentarli ai suoi lettori. Selezionare il giusto dataset è come selezionare la giusta intervista per un articolo.

We're getting bigger datasets on smaller things

I dataset a disposizione stanno diventando sempre più ampi. Ad esempio i dati di WikiLeaks sulla guerra in Iraq sono arrivati a contare 391 000 records. In questa era abbiamo a disposizione sempre più dati su argomenti sempre più ridotti. Rendere queste informazioni più accessibili anche per un

pubblico non specializzato sta diventando uno dei compiti più importanti a cui sono chiamati i nuovi giornalisti.

It's 80% persipration, 10% inspiration, 10% output

Il data journalism è per la maggior parte sudore, il resto è ispirazione e capacità di presentazione. Spendendo ore a lavorare su un dataset ripulendo i dati o combinandoli insieme è come se il giornalista creasse un ponte tra i dati e le persone che vogliono comprenderli.

It's not all long, complicated investigations

Ci possono essere dataset che necessitano settimane di lavoro per essere analizzati. Alcuni restituiscono la fatica svolta con scoop o reportage interessanti altri invece producono ben poco. Il data journalism però non è solo lavoro a lungo termine ma si sta orientando anche verso forme più brevi, dove è sufficiente analizzare i dati più importante per poi arrivare subito alla notizia.

Anyone can do it...

Gli strumenti ci sono e sono a disposizione di tutti. Grazie a Google Fusion tables, Datawrapper, Google Charts, Timetric e tutti gli altri software che nascono ogni giorno tutti possono provare a interrogare i dati.

...but looks can be everything

Il buon design conta davvero. Molti articoli sono resi interessanti non solo dalle notizie trovate ma anche da come sono rappresentate. Un buon designer riesce a cogliere sia l'argomento di cui si sta parlando e che le necessità dei lettori. Spesso questo non può accadere con le visualizzazioni realizzate dalle macchine.

You don't have to be a programmer

Certo, si può diventare degli ottimi informatici per occuparsi di questa materia. La cosa migliore però per chi fa informazione è continuare a pensare ai dati come un giornalista e non come un analista. Quello che conta è trovare le domande giuste da fare. La strada migliore è quindi lasciare che i giornalisti facciano i giornalisti e lavorare sulla parte tecnica fondendo assieme le competenze di persone diverse.

It's (still) all about the stories

Il data journalism non è solo grafici e visualizzazione. Si tratta di raccontare una storia nel miglior modo possibile. A volte questo vuol dire creare una dataviz ma altre volte potrebbero bastare le parole o addirittura un solo numero.
Il data journalism è soprattutto la flessibilità di cercare nuove forme di racconto. Non è nulla di strano, alla fine si tratta solo giornalismo.

Quando i dati diventano notizie. I warlogs di WikiLeaks pubblicati dal *The Guardian*

“You're good with spreadsheets, aren't you?”

Da questa frase è cominciata una delle più grosse inchieste giornalistiche mai pubblicate.
È il 2006 e a rivolgere queste parole a Simon Rogers è un membro della squadra investigative del quotidiano inglese *The Guardian*.

Lo spreadsheet a cui si riferisce conta 92 201 righe, ognuna contente un dettagliato reportage di un singolo evento militare registrato dall'esercito americano in Afghanistan.
Questi documenti provengono tutti da una fonte interna al governo americano e prendono il nome in codice di SIGACTS, i database con tutte le azioni militari significative. È un materiale molto prezioso. Sono racconti di guerra estremamente dettagliati scritti da chi la guerra la sta combattendo sul campo.

I giornalisti che componevano la squadra di questo datablog si erano già confrontati con dataset di tale portata. Tempo prima avevano lavorato sul COINS, il dataset che raccoglieva tutte le spese della tesoreria inglese dove erano registrate le uscite di ogni dipartimento per un periodo di circa due anni. Gli sviluppatori del *The Guardian* crearono per questa occasione il COINS Explorer, un'interfaccia che permetteva ai reporter di navigare in questi dati per cercare delle storie. Per i warlogs di WikiLeaks gli obiettivi erano gli stessi: aiutare i giornalisti ad accedere alle informazioni, analizzare i dati e renderli comprensibili ai lettori.

Premessa. WikiLeaks, Julian Assange e le sue fonti

“WikiLeaks is a giant library of the world’s most persecuted documents. We give asylum to these documents, we analyze them, we promote them and we obtain more” Julian Assange[38]

In *Fact are Sacred*[39] il giornalista Simon Rogers racconta come la redazione del *The Guardian* ha elaborato diversi dataset sulle guerre in Medio Oriente provenienti dal sito WikiLeaks. Prima di seguire Rogers nella descrizione del suo lavoro è necessario però soffermarsi proprio su questa fonte per capire il peso delle informazioni trattate dai giornalisti inglesi.

WikiLeaks è un'organizzazione internazionale senza scopo di lucro, fondata da Julian Assange nel 2006[40]. È specializzata nell'analisi e nella pubblicazione di grandi dataset di materiali sottoposti a censura o qualsiasi altra forma di restrizioni. I documenti su cui lavora riguardano soprattutto guerre, spionaggio e corruzione. Al momento ha raccolto e analizzato più di 10 milioni di file.

Il sito funziona attraverso il sistema dei whisteblower, informatori anonimi che decidono di inviare materiale riservato per far conoscere aspetti nascosti di governi o aziende. Chiunque decida di fidarsi di WikiLeaks e dei sistemi di cifratura del sito può così caricare i file in suo possesso.

Ogni dataset analizzato ha quindi una fonte e una storia diversa.

Il 25 luglio 2010 vengono rilasciati i documenti militari relativi alla guerra in Afghanistan. Questi

38 Dichiarazione proveniente da un'intervista rilasciata al quotidiano tedesco *Der Spiegel* il 20 luglio 2015, http://www.spiegel.de/international/world/spiegel-interview-with-wikileaks-head-julian-assange-a-1044399.html
39 S. Rogers, *Fact are Sacred,* 2013
40 https://wikileaks.org/What-is-Wikileaks.html

file coprono un periodo che va da gennaio 2004 al dicembre 2009. Si tratta di una delle fughe di notizie più estese nella storia del governo americano.
Questo caso è stato fondamentale per la storia del sito fondato da Assange. Data l'importanza dei dati raccolti WikiLeaks ha deciso infatti di non distribuire queste informazioni solo dal suo portale ma di chiedere l'aiuto di tre grandi giornali, oltre al *The Guardian* hanno lavorato a questo progetto anche il *New York Times* e il *Der Spiegel*.

Il 5 aprile 2010 nel corso di una conferenza stampa a Washington, WikiLeaks ha diffuso un video che mostra l'uccisione di 12 civili iracheni, tra cui due giornalisti dell'agenzia stampa Reuters, da parte di una coppia di elicotteri Apache statunitensi. All'origine di questi omicidi pare ci sia stato un errore di valutazione. I soldati avrebbero scambiato la videocamera di un giornalista per un'arma. A maggio dello stesso anno viene arrestato Chelsea Manning, soldato dell'esercito americano accusato di aver diffuso non solo il video ma anche centinaia di migliaia di documenti riservarti sulla guerra in Iraq. Ad ottobre circa 300 000 di questi file verranno divulgati sempre da WikiLeaks.

L'ultimo dataset che verrà qui trattato non riguarda informazioni di guerra ma di carattere diplomatico. Dal 28 novembre 2010 il sito di Assange ha infatti pubblicato notizie riservate sull'operato della diplomazia statunitense. Si tratta di 251 287 documenti confidenziali provenienti da 274 ambasciate americane distribuite in tutto il mondo. Il periodo coperto è molto ampio, si va infatti dal 1966 fino al febbraio 2010. Anche in questa occasione molte testate giornalistiche hanno partecipato alla divulgazione delle informazioni rilasciate.

Parte I. Afghanistan

Quando il team diretto da Simon Rogers cominciò a pensare a quali articoli scrivere partendo da questo primo stock di dati gli obiettivi che si posero furono due: offrire al lettore un quadro generale della guerra in corso e trovare delle storie da raccontare, storie di chi stava ancora combattendo sul fronte.
La prima cosa che divenne chiara è che il database non poteva essere pubblicato per intero. Il rischio di diffondere indizi sugli informatori della Nato era troppo alto e questo avrebbe messo in difficoltà non solo i militari ma anche i civili che stavano collaborando con loro. Senza contare che era importante creare un sistema per cui i giornalisti del team investigativo guidato da David Leigh e Nick Davies potessero navigare in questi dati.

Quello che arrivò in redazione infatti era un enorme file xls di 92 301 righe e molte di queste contenevano dati non strutturati o semi strutturati. La prima operazione consisteva nel rendere accessibili i dati. I reporter hanno quindi creato un database interno, dove si potessero cercare le storie attraverso parole chiave o eventi specifici. Ogni documento a disposizione era così ordinato con parametri precisi: ora, data, descrizione, vittime e soprattutto latitudine e longitudine.

Una delle prime storie emerse da questi dati fu quella degli IED[41], attacchi con bombe artigianali piazzate lungo le strade e molto difficili da prevedere. Dal database emersero circa 7 500 IED registrati fra il 2004 e il 2009. Alcuni erano semplici esplosioni, altri erano accompagnati anche da imboscate in cui i nemici utilizzavano armi di piccolo calibro e granate. Oltre 8 000 ordigni vennero invece trovati e bonificati. La maggior parte di questi incidenti si era verificata nel sud dell'Afghanistan, la regione controllata dalle truppe britanniche e canadesi. In alcuni periodi poi erano stati registrati dei picchi particolarmente acuti. Nel settembre 2010 quando nei tre giorni attorno le elezioni presidenziale scoppiarono sulle strade del Paese più di 100 bombe.

41 Improvised Explosive Device

Da tutte le informazioni a loro disposizione i giornalisti del *The Guardian* hanno quindi estratto una storia, visualizzandola attraverso una serie di cartine e grafici. La narrazione però non si è fermata qui ma è stata completata anche da articoli e fotografie. Una sorta di pacchetto multimediale che ha reso possibile la piena comprensione da parte del lettore delle vicende raccontate[42].

I reporter hanno voluto ampliare il loro lavoro offrendo agli utenti la possibilità di analizzare questi dati e fornendo anche una descrizione accurata del processo da loro seguito per arrivare a questa storia. Chiunque quindi può non solo scaricare i dati ma anche visualizzarli come preferisce e caricare la propria dataviz sull'apposito gruppo Flickr[43].
Questa scelta non ha garantito solo una migliore trasparenza nei confronti dei lettori ma ha anche aperto il giornale ai loro contributi.

Parte II: Iraq

Questi warlogs sono stati rilasciati nell'ottobre 2010 ed erano formati da ben 391 000 record. Secondo Simon Rogers la grande quantità di dati a disposizione ha reso la guerra in Iraq il conflitto la più documentato nella storia.
Sono registrati infatti molti eventi, anche di poca importanza. Fra tutti i dati però quello che emerge con più forza riguarda il numero dei morti, la maggior parte dei quali sono civili.

Come nel caso precedente, anche qui i reporter hanno preferito non pubblicare tutto il dataset a disposizione, sempre per tutelare le fonti che hanno fornito tutte le informazioni qui contenute. È possibile però scaricare uno spreadsheet con tutti i 60 000 incidenti in cui qualcuno ha perso la vita. Questi eventi sono stati tutti geolocalizzati ed inseriti all'interno di una mappa per vedere quali fossero le zone più toccate dal conflitto.

Con un dataset così ampio è possibile cominciare a riflettere anche sul concetto di validation.
La data analyst Katy Börner inserisce spesso nei suo processi di visualizzazione questo passaggio. Si tratta di sottoporre le analisi condotte sui dati ad un esperto di dominio. Qualcuno che sappia bene l'argomento di cui trattano i dati. Questo serve per evitare di trovare false correlazioni fra dati e quindi stabilire causalità inesistenti.

Il team di giornalisti del *The Guardian* si è rivolto così a Jacob Shapiro, professore di Politiche e Affari Internazionali alla Princeton University. Visionando le informazioni a disposizione l'accademico ha così svelato che non sono state raccolte tutte le morti avvenute durante il conflitto in Iraq ma solo quelle registrate dalle Multi-National Forces. In questo elenco di vittime non sono infatti incluse quelle che si sono verificate in eventi in qui erano coinvolte i militari della Coalizione o le Unità Irachene, due realtà così tanto coinvolte in questa guerra da non aver nemmeno il tempo di annotare tutte le morti sul campo.

La situazione quindi è senza dubbio più grave di quanto sia stato rappresentato. Ma questo ha permesso comunque ai reporter di estrarre una serie di dati aggregati che fornissero un immediato colpo d'occhio del quadro generale.

Si scopre così che delle 109 032 morti registrate nel database 66 081 sono state di civili, 15 196 delle forze di sicurezza irachene e 23 984 di ribelli insorti.
Oltre alle classiche visualizzazioni dove gli eventi sono stati geolocalizzati, i giornalisti britannici

42 http://www.theguardian.com/world/datablog/2010/jul/26/wikileaks-afghanistan-ied-attacks
43 https://www.flickr.com/groups/guardiandatastore/pool/page7

hanno anche usato il data design per spiegare ai lettore quanto fosse complesso il database su cui stavano lavorando.
Jonathan Stray e Julian Burgess della Associated Press hanno così creato per *The Guardian* un network graphs prendono solamente gli 11 616 SIGACT collezionati nel mese di dicembre 2006[44]. Ogni evento è indicato da un cerchio e collegato ad altri con linee più o meno spesse in base al numero di parole chiave in comune. Il risultato non ha alcuno scopo informativo sulla materia trattata, infatti è difficile cogliere delle informazioni sul conflitto in Iraq all'interno dei cerchi e delle linee che rappresentano i documenti visualizzati. In compenso è fondamentale per mostrare al lettore come fosse complessa la mole di dati con cui i giornalisti hanno dovuto rapportarsi.

Parte III: Le ambasciate americane nel mondo

L'ultimo stock di dati rilasciati da WikiLeaks e trattati da *The Guardian* riguarda i cablogrammi che raccontano l'andamento della diplomazia americana nel mondo. Siamo nel dicembre 2010 e si tratta di 251 287 dispacci, proveniente da 250 sedi fra consolati e ambasciate.

La quantità di dati è ancora più impressionante dei casi precedenti. Si tratta infatti soprattutto di testi, dati non quantitativi e quindi più complessi da analizzare.

Ogni cablogramma conteneva infatti:

- Fonte: ambasciate o altri organi diplomatici
- Lista di destinatari
- Soggetto: un breve sommario del cablogramma
- Tags: parole chiave contenute all'interno del testo
- Corpo del testo: tutto il contenuto del cablogramma

Anche in questo caso, per motivi di sicurezza, i giornalisti hanno scelto di non riportare il contenuto di tutti i dispacci ma solamente alcune alcune loro parti.

Per divulgare questo dataset è stata predisposta un'interfaccia in grado di guidare la navigazione attraverso mappe e parole chiave, delle infografiche statiche che riportano i dati complessivi per offrire un inquadramento generale e soprattutto è stata offerta ancora la possibilità per i lettori di scaricare tutti i dati, già puliti ed ordinati. Anche qui però, come negli articoli pubblicati sul giornale, è stato eliminato il corpo del testo.

44 http://www.theguardian.com/news/datablog/2010/dec/16/wikileaks-iraq-visualisation

Illustrazione 20: Spreadsheet sugli attacchi IED scaricabile dagli utenti, http://www.theguardian.com/world/datablog/2010/jul/25/wikileaks-afghanistan-data

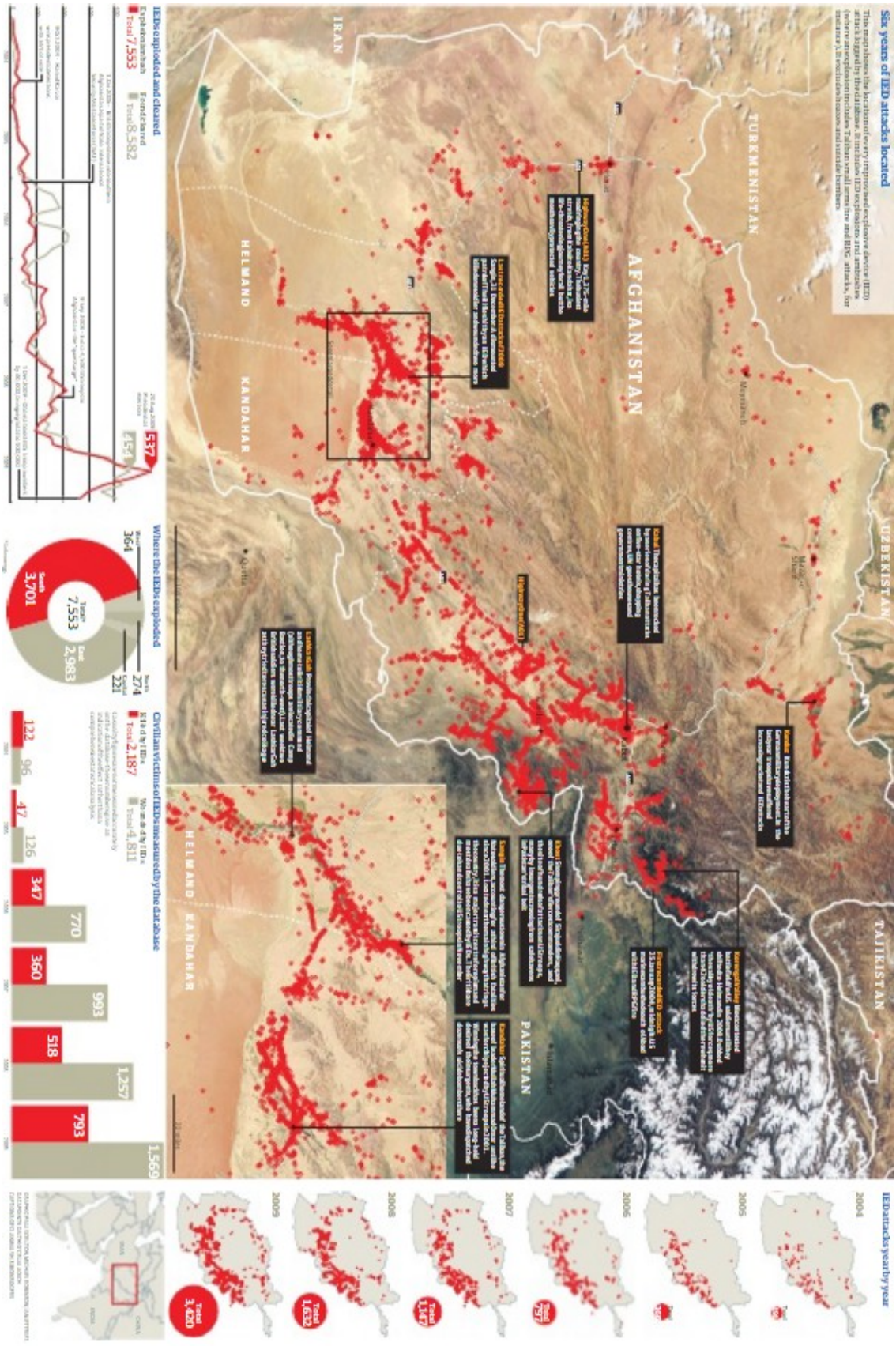


Illustrazione 21: Infografica IED Afghan war logs, https://www.scribd.com/doc/34850058/Afghanistan-IED-attacks-2006-to-2009

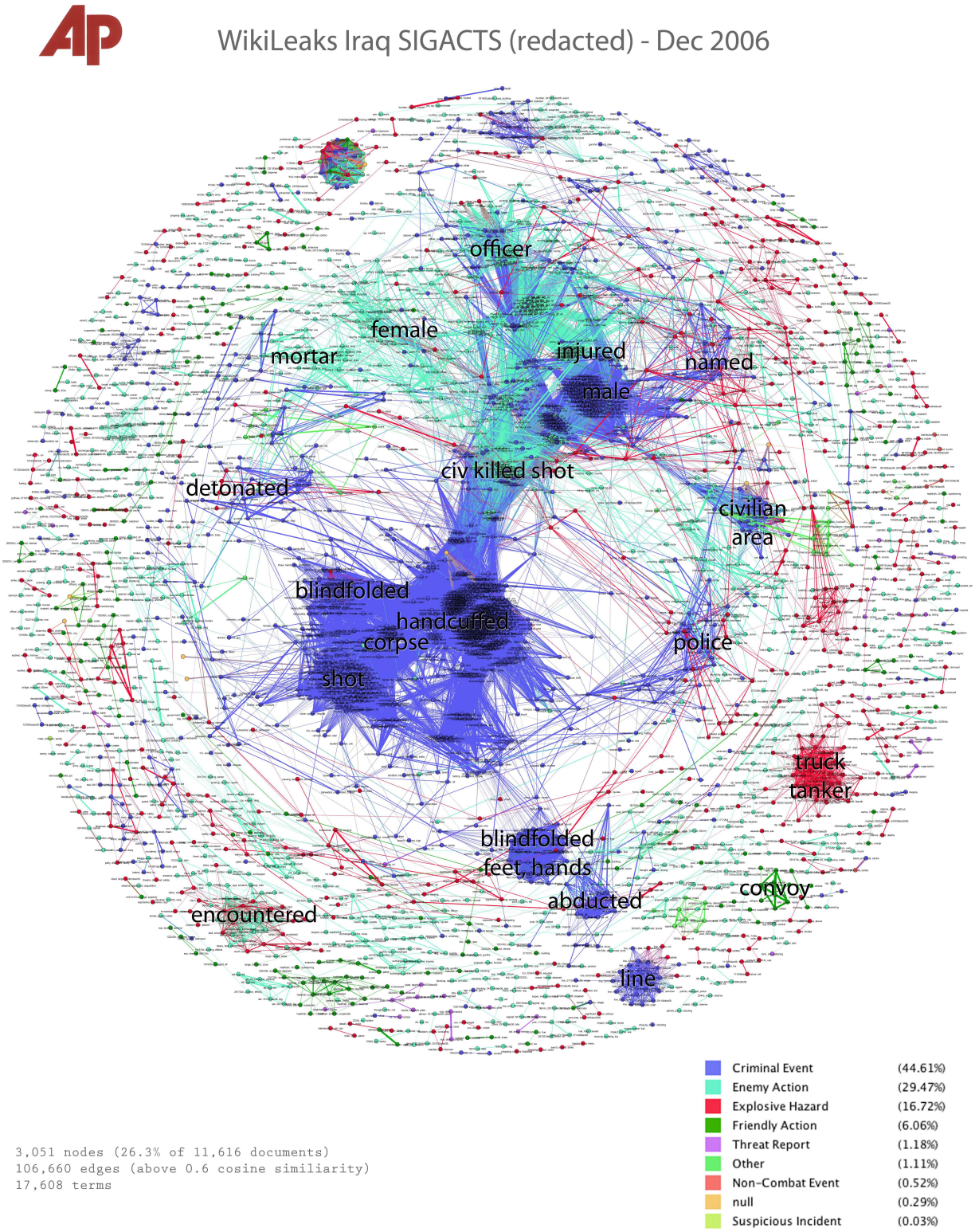


Illustrazione 22: La rete dei war logs sulla guerra in Iraq, http://jonathanstray.com/wp-content/uploads/2010/12/SIGACTS-dec-2006-hi-res2.jpg

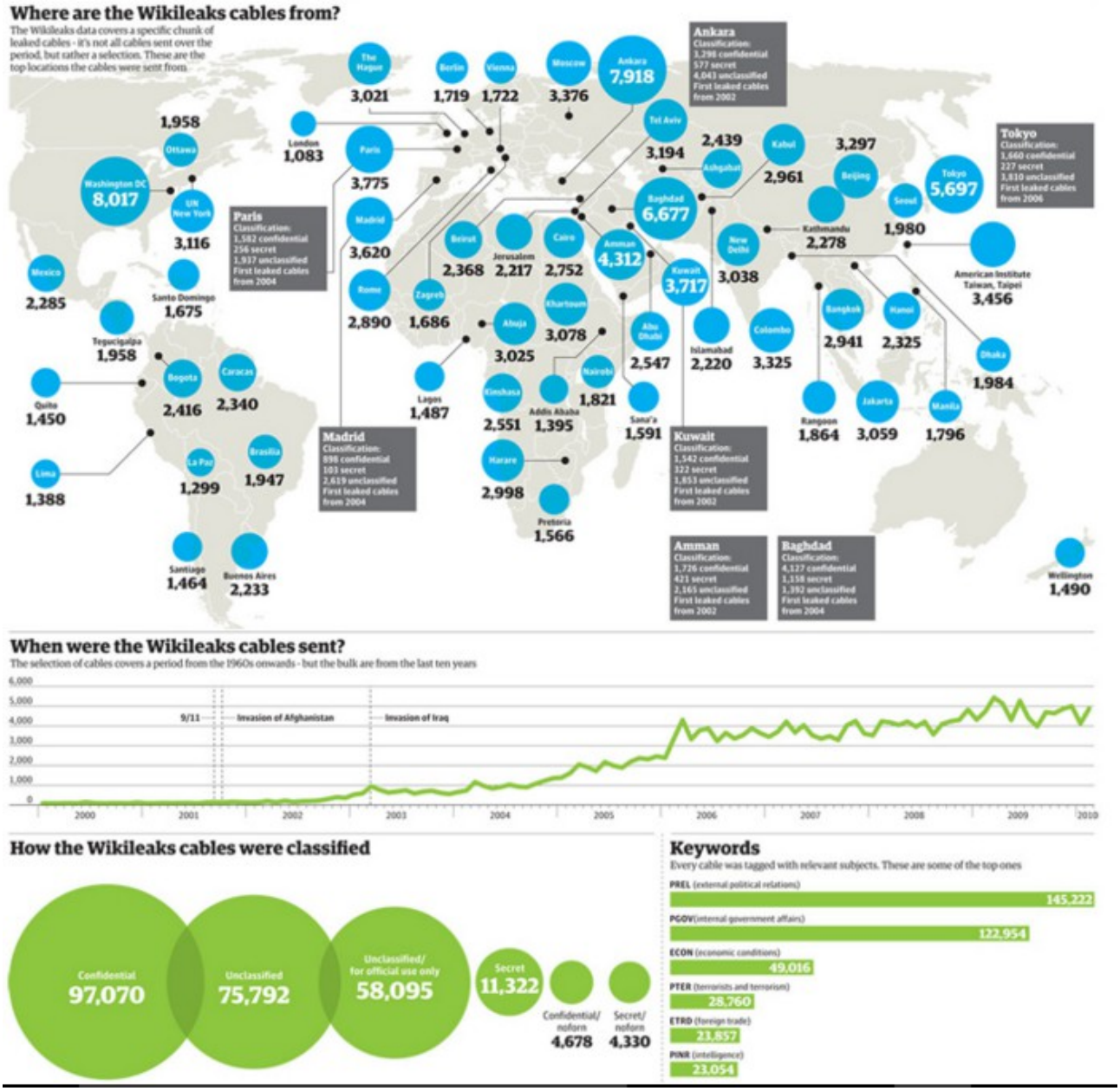


Illustrazione 23: Infografica dei cablogrammi delle ambasciate americane, http://www.theguardian.com/news/datablog/2010/nov/29/wikileaks-cables-data

L'opacità dei dati. Valentina Manchia e gli oggetti del data journalism

Analizzando i passaggi che hanno portato alla visualizzazione dei dati rilasciati da WikiLeaks emergono degli spunti di riflessione interessanti sul data journalism.
Ogni passaggio che è stato fatto, ogni dato che è stato strutturato e poi divulgato è frutto di una scelta. Nel caso degli attacchi IED, ad esempio, il team di giornalisti non ha deciso solo su quali dati concentrarsi ma anche come processarli e come disporli su mappe e grafici.

Il lavoro che appare al lettore è quindi una costruzione. Un'idea che forse va un po' a scardinare tutti quei luoghi comuni per cui i dati, le statistiche e quindi i grafici non sono altro che la rappresentazione più fedele della realtà.

Già Tufte parlava infatti di chartjunk, grafici spazzatura che non riportavano fedelmente i numeri di partenza ma ingannavano il lettore giocando sulle dimensioni e le distanze delle forme, tradendo così la fedeltà ai numeri. Se si pensa quindi che questo processo di mistificazione potrebbe avvenire ancora prima della fase di visualizzazione allora il data journalism perde quell'aurea di totale aderenza alla realtà dei fatti che lo circonda.

Con questa affermazione non si vuole certo sostenere che i reporter del *The Guardian* abbiano voluto deliberatamente snaturare i dati raccontati ai loro lettori. È necessario però riflettere su come le dataviz siano sempre costruzioni che nascono da un autore.

Su questo tema si è espressa Valentina Manchia, una semiologa dell'Università di Siena che nel 2015 ha pubblicato un articolo dal titolo *Il data journalism e la rappresentazione visiva delle informazioni tra trasparenza e opacità. Gli Afghan War Logs di WikiLeaks riletti dal Guardian*[45].

Il primo punto che affronta è quello della rappresentazione.
Rappresentare un oggetto esistente vuol dire scegliere un punto di vista per descriverlo. Un esempio di questo processo possono essere le mappe geografiche.

“La mappa esiste solo grazie al punto di vista che decide da che angolazione restituire allo sguardo il territorio che rappresenta. Una mappa, pertanto, e così anche nel nostro caso, fornisce accesso a un oggetto (a un territorio), e allo stesso tempo è l'ostensione di questo stesso oggetto da parte di un soggetto”.

La sua riflessione per quanto riguarda il data design non si ferma però qui. Le dataviz infatti non solo rappresentano un oggeto ma contribuiscono anche a crearlo. Quelle reti di collegamenti, quegli insieme di dati si riferiscono certo a qualcosa di reale ma la loro rappresentazione è un oggetto del tutto inedito, diverso da qualunque cosa esista.

“Questi oggetti, le visualizzazioni grafiche e visive del *The Guardian*, non sono rappresentazioni in senso classico: non rimandano a un oggetto del mondo, ma a un soggetto costruito che contribuiscono a creare, e che rende conto delle relazioni e delle connessioni tra i dati viste *sotto un certo rispetto*. Sono rappresentazioni opache, che portano volutamente le loro tracce di produzione e di interpretazione sulla scena”.

45 V. Manchia, *Il data journalism e la rappresentazione visiva delle informazioni tra trasparenxa e opacità. Gli AfghanWar Logs di WikiLeaks riletti dal Guardian*, , in M. Serra, O. Gomez, *Transparencia y Secreto*, 2015

La narrazione di questi dati nasce quindi da una costruzione che poi verrà rappresenta da uno strumento arbitrario. Sommando questi due pensieri, l'opacità dei dati diventa doppia dato ce “c'è l'opacità dell'organizzazione del discorso sui dati, e l'opacità del dispositivo che mette in scena questo discorso”.

Proprio per questo motivo l'autrice dell'articolo elogia un'altra scelta del team di reporter che si è occupato di lavorare su questo materiale: la possibilità aperta a tutti di scaricare i dati.

“I dati […] sono offerti ai lettori sia perché sanzionino il corretto lavoro effettuato sui materiali, sia come ostensione dell'oggetto, considerato come base da cui partire, e non come qualcosa da restituire nella sua integrità”.

Questa traduzione porta con sé il marchio del soggetto che la opera, un marchio irrinunciabile per arrivare alla comprensione di dati che altrimenti, anche nella loro forma rielaborata e ripulita, apparirebbero comunque troppo complessi per il pubblico.

In questo tipo di comunicazione la trasparenza si può verificare dunque solo esplicitando i processi seguiti per arrivare al risultato e offrendo magari al lettore l'opportunità di confrontarsi con la materia prima che costituisce il punto di partenza della visualizzazione.

6. I numeri dei libri. Franco Moretti e la data visualization in ambito umanistico

C'è ancora un aspetto del data design che merita di essere toccato in questa rassegna. Secondo la classificazione proposta all'inizio di questo capitolo la sua naturale collocazione dovrebbe essere nella sezione *Visualizzare per capire* ma per le sue caratteristiche merita una sezione a parte.

Associare l'utilizzo delle forme e dei colori alla comprensione di dati in ambito scientifico può apparire quasi naturale per chi si occupa di queste materie. È forse più difficile invece immaginare come questo tipo di processi vengano ripresi negli studi umanistici.

Franco Moretti è professore di letteratura britannica presso la Stanford University e fondatore del *Center for the Study of the Novel*. Nel 2005 ha pubblicato *Graphs, Maps, Trees*[46] un volume in cui propone un approccio quantitativo allo studio della letteratura mutuato da tecniche di visualizzazione dei dati.

Grafici, mappe e alberi. Tre strumenti che secondo questo autore possono permettere di comprendere la letteratura come fenomeno di massa, non fermandosi solo sulle opere più importanti o sulle vite degli autori che le hanno scritte.

“Graphs, maps, and trees place the literary field literally in front of our eyes – and show us how little we sill know about it. It is a double lesson, of humility and euphoria at the same time: humility for what literaly history has accomplished so far (not enough), and euphoria for what still remains to be done (a lot)”.

Graphs

Le prime riflessioni sulle dataviz arrivano da uno spostamento di interesse che passa dagli eventi straordinari a quelli ordinari, dalle vicende legate agli autori più significativi di una determinata epoca a quelle che riguardano la massa dei lettori.

Come punto di partenza Moretti sceglie una domanda.

“What literature would we find, in 'the large mass of facts'?”

Per rispondere comincia una raccolta di dati. Una ricerca attraverso archivi e biblioteche per recuperare informazioni relative alla letteratura di massa. Informazioni che poi vengono presentate e visualizzate attraverso una serie di grafici sulla diffusione del romanzo in diversi paesi del mondo, sull'evoluzione delle differenze di genere fra autori o ancora sulla distribuzione del mondo di questa forma di narrativa.

Uno dei più interessanti fra questi lavori è sicuramente quello che mostra l'evoluzione dei generi del romanzo britannico fra il 1740 e il 1900.

Dopo aver consultato un centinaio di studi sulla letteratura britannica Moretti ha identificato 44 generi di romanzo che si distribuiscono in forme diverse in questo periodo. Un numero eccezionale se si pensa che sui 160 anni analizzati indica la nascita di un nuovo genere ogni 4 anni.

Per definire questo veloce susseguirsi di generi a volte simili a volte completamente opposti,

46 F. Moretti, *Graphs, Maps, Trees*, 2005

ognuno di loro è stato rappresentato da una barra disposta su un asse temporale. La lunghezza della barra coincide con il periodo compreso tra la prima apparizione del genere e la sua scomparsa dal mercato librario.

Illustrazione 24: Fonte: https://ariddell.org/reconstructing-graphs-maps-trees.html

Questa visualizzazione può aprire molte riflessioni. Si possono riconoscere dei picchi di creatività dove in poco tempo compaiono generi nuovi, si possono vedere generi o addirittura gruppi di generi che scompaiono improvvisamente tutti assieme. Ci si può anche chiedere perché il ciclo di vita di

un genere duri in media dai 25 ai 30 anni.

Ai fini di questo elaborato non è significativo riflettere su queste domande ma piuttosto riflettere su quante domande possa aprire un approccio di questo tipo. È bastato allargare lo sguardo dai singoli casi alla collettività per mostrare un'idea diversa di storia del romanzo. Una storia che non è più una percorso lineare tracciato orientandosi solo sulle opere capitali ma piuttosto un flusso estremamente vario formato da generi più rilevanti e sottogeneri. Un flusso che solo preso nella sua complessità può definire cosa è il romanzo per la storia della letteratura.

"For most literary historians, I mean, there is a categorical difference between 'the novel' and the various 'novelistic (sub)genres': the novel is, so to speak, the substance of the form, and deserves a full general theory; subgenres are more like accidents, and their study, however intersting, remains local in character, without real theoretical consequences. The forty-four genres […] suggest a different historical picture, where the novel does not develop as a single entity but by periodically generating a whole set of genres, and then another, and another […] In other words , the novel is *the system of its genres*".

Maps

Dopo i grafici cronologici, la seconda famiglia di strumenti del data design che può migliorare la comprensione del romanzo sono le mappe. E sono due i modi in cui possono essere utilizzate.

Il primo riguarda la geografia della trama. Si tratta qui di utilizzare delle mappe semantiche che identifichino i rapporti spaziale fra elementi diversi all'interno del romanzo, siano questi personaggi, eventi o vicende.

All'inizio del XIX secolo nel Regno Unito erano popolari ad esempio le *Village Stories*, un genere che si basava su racconti sviluppati attorno lo stesso villaggio. Una delle esponenti più importanti di questo genere fu Mary Mitford che tra il 1824 e il 1832 pubblicò una serie di romanzi in cinque volumi dal titolo *Our Village*[47]. Il villaggio al centro dei racconti si chiamava Mile Cross e si trovava in Berkshire, una contea dell'Inghilterra sud-orientale.

Il primo volume di questa serie è composto da 24 storie che Moretti ha disposto su una mappa circolare.
Al centro c'è il nucleo più denso, i 10 racconti ambientati all'interno del villaggio. Disponendo gli altri a seconda delle distanze dal villaggio dei luoghi che fanno loro da scenario ci si accorge di come si venga a creare una sorta di sistema solare nel quale si possono riconoscere due orbite.
Una è quella più vicina, dove vengono posizionati i racconti che si sviluppano entro un miglio dal villaggio e una seconda più lontana, formata da quelli che invece si sviluppano entro due miglia. Solo 3 racconti, ambientati in città ben lontane da Mile Cross si collocano all'esterno di questa mappa.

Lo schema che si viene così a creare mostra perfettamente la struttura di un romanzo appartenente al genere delle *Village Stories.*

47 M. Mitford, *Our Village*, 1824-1832

FIGURE 2: *Mary Mitford*, Our Village, *volume I [1824]*

Illustrazione 25: Fonte: https://digitalpublichistory.wordpress.com/2013/02/09/graphs-maps-and-trees-oh-my/

Mappe di questo tipo possono essere utili per prepararsi all'analisi letteraria. Permettono infatti di ridurre il testo ad un piccolo numero di elementi e creare un oggetto artificiale in grado di aprire nuovi spiragli di discussione mostrando aspetti che non sarebbero emersi tanto facilmente da una semplice lettura.

"You *reduce* the text to a few elements, and *abstarct* them from the narrative flow, and construct a new, *artificial* object like the maps […] With a little luck these maps will be more than the sum of their parts: they will posses 'emerging' qualities, which were no visible at the lower level".

C'è poi un modo ulteriore di utilizzare questo tipo di schema ossia non disporlo solo sul piano astratto ma sovrapporlo ad una vera mappa geografica per studiare le posizioni dei soggetti analizzati rispetto all'ambiente in cui sono inseriti.

Sempre Moretti in *Atlas of the European Novel*[48] propone diverse mappe letterarie di questo tipo fra cui *Protagonist of Parisian novel, and objects of their desire*.
Il punto di partenza per questa visualizzazione è una serie di romanzi scritti nel XIX secolo e ambientati a Parigi incentrati su storie d'amore. L'autore della mappa ha qui voluto disporre sulla cartina della capitale francese non solo i nomi dei protagonisti maschili ma anche gli oggetti del loro desiderio, rappresentati con una stella.

FIGURE 10: *Protagonists of Parisian novels, and objects of their desire*

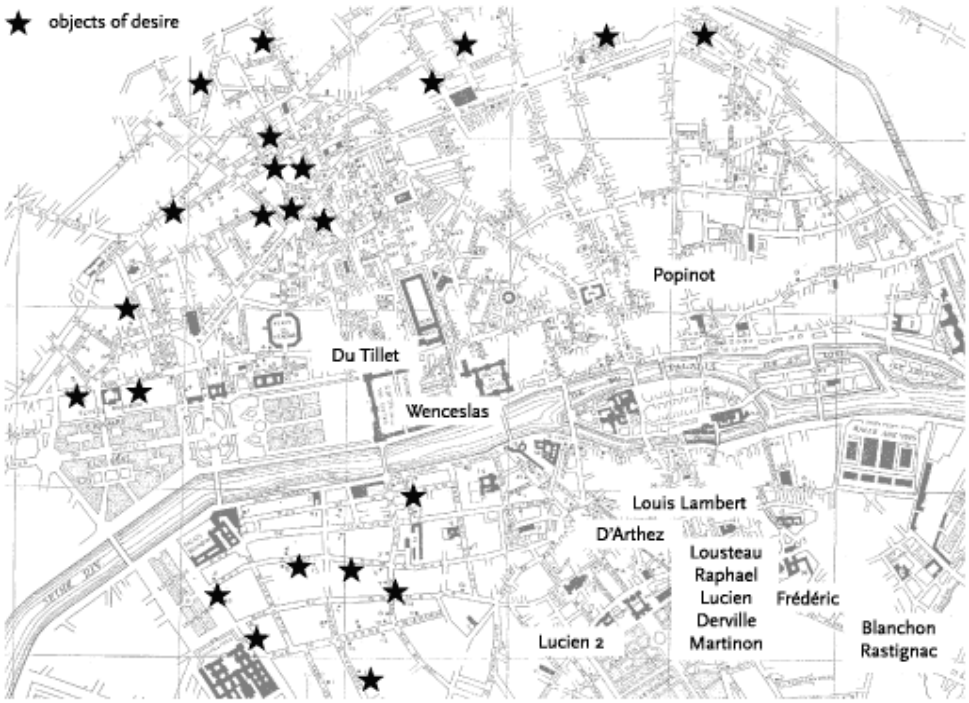


Illustrazione 26: Fonte: https://newleftreview.org/II/26/franco-moretti-graphs-maps-trees-2

Qui si può vedere come fosse molto diffuso topos letterario in base al quale l'amante bramava una donna che abitava sul lato opposto della Senna.

Trees

L'ultimo esempio di visualizzazione dei dati in ambito letterario viene invece dagli studi di Charles Darwin. In *The Origin of Species*[49] il biologo utilizza i tree graphs per spiegare la teoria dell'evoluzione, illustrando come all'interno del mondo animale sopravvivano solo i soggetti che riescono ad adattarsi meglio all'ambiente naturale. Questi soggetti sviluppano infatti caratteristiche peculiari che permettono loro di vivere meglio e di portate avanti il processo evolutivo. I soggetti che invece non riescono a trovare sistemi di adattamento altrettanto validi soccombono.

48 F. Moretti, *Atlas of the European Novel*, 1999
49 C. Darwin, *The Origin of Species*, 1859

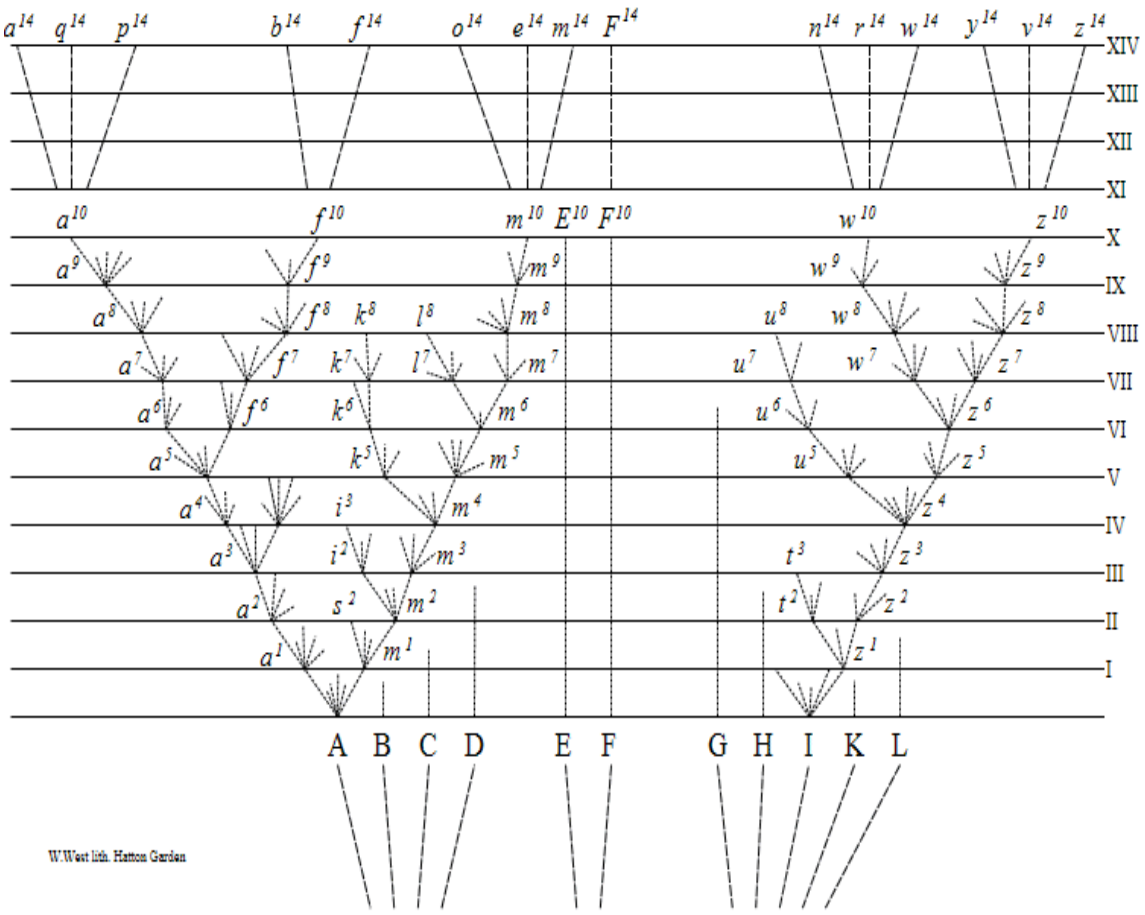


Illustrazione 27: Fonte: http://www.age-of-the-sage.org/evolution/charles_darwin/tree_of_life.html

Nell'ambito della critica letteraria questo modello può avere diversi utilizzi, uno è sicuramente la possibilità di analizzare l'evoluzione di un elemento letterario in opere diverse.
Franco Moretti ha provato a studiare attraverso un tree graph le prime fasi della British Detective Fiction, un genere che ha visto primeggiare fra tutti gli autori della fine del XIX Arthur Conan Doyle, il creatore del personaggio di Sherlock Holmes.

L'elemento scelto per questa analisi la presenza di indizi nelle storie di Conan Doyle. Si tratta di tessere di un mosaico sparse nelle descrizioni o nei racconti di vittime e testimoni che l'investigatore di Baker Street riesce ad assemblare alla fine di ogni caso per trovare la soluzione corretta.

Questo espediente letterario non è usato allo sesso modo da altri autori suoi contemporanei che scrivevano lo stesso tipo di racconti. L'autore dello schema qui rappresentato ha infatti diviso le opere appartenenti al British Detective Fiction di questo periodo in quattro gruppi usando come elemento di distinzione proprio l'utilizzo dell'indizio all'interno della storia.

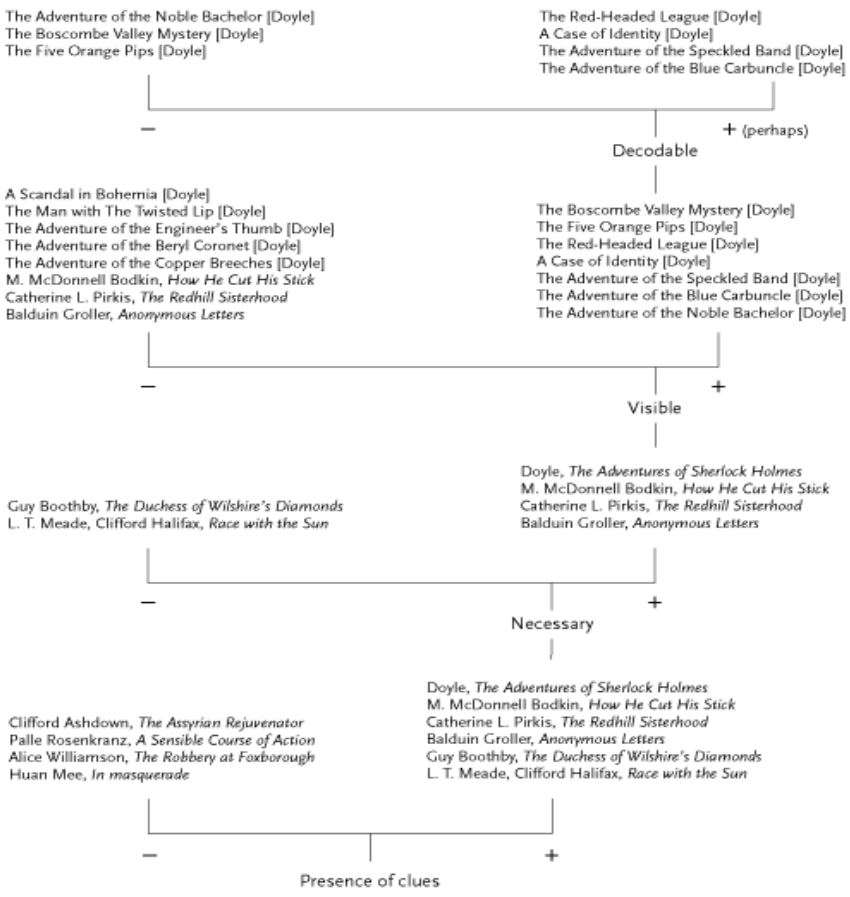
Così il grafico si sviluppa distinguendo diversi gruppi di racconti in cui gli indizi:

1. sono inclusi nella storia
2. sono inclusi ma non sono necessari allo sviluppo del caso
3. sono inclusi, sono necessari allo sviluppo del caso ma non sono visibili

4. sono inclusi, sono necessari allo sviluppo del caso, sono visibili ma non decodificabili

Si può vedere come in questo schema i racconti di Doyle facciano un uso più massiccio degli indizi, mentre quelli di altri autori lo usino di meno. Nell'ultima parte di grafico poi la presenza di altri scrittori scompare e rimane solamente una divisione interna tra i racconti con protagonista Sherlock Holmes.

FIGURE 3: *Presence of clues and the genesis of detective fiction*



From the standpoint of technique, the devices employed by Conan Doyle in his stories are simpler than the devices we find in other English mystery novels. On the other hand, they show greater concentration . . . The most important clues take the form of secondary facts, which are presented in such a way that the reader does not notice them . . . they are intentionally placed in the oblique form of a subordinate clause . . . on which the storyteller does not dwell.

Viktor Shklovsky, *Theory of Prose*

Illustrazione 28: Fonte: https://newleftreview.org/II/28/franco-moretti-graphs-maps-trees-3

Questo grafico, come gli altri affrontati da Moretti, alza il sipario su diverse domande. È stata la

presenza di indizi a garantire la longevità dell'opera di Doyle? Perché i lettori sono più attratti dalle storie che contengono questo elemento?

Theory are nets

Anche qui la data visualization non offre soluzioni immediate ma piuttosto punti di vista differenti che aprono nuovi scenari d'analisi.

La citazione utilizzata da Moretti chiarisce bene questo pensiero. Il poeta Novalis affermava infatti "theory are nets and only he who casts will catch". Ed è proprio in questo senso in cui, secondo il critico italiano, bisogna introdurre la data visualization all'interno dello studio della letteratura. È la nascita di un nuovo metodo di analisi.

"Yes, theory are nets, and we should evaluate them, not as ends in themselves, but for how they *concretely change the way we work*: for how they allow us to enlarge the literary field, and re-design it in a better way, replacing the old, useless distinctions (high and low; canon and archive; this or that national literature...) with new temporal, spatial, and morphological distinctions".

Certo, un approccio del genere mostra anche evidenti limiti di analisi. Predilige infatti la spiegazione delle macrostrutture all'interpretazione dei singoli testi in un ottica che riprende "a materialistc conception of form".

Questi strumenti quindi potrebbero anche presentare dei grossi limiti ma è solo approfondendo il loro uso che si capirà fin dove potranno arrivare.

"Opening new conceptual possibilities seemed more importan than justifying them in every detail".

Capitolo 2
La sintassi dei grafici

1. Tradurre i dati in immagini

Nei manuali di data design c'è una formula che torna spesso: “Making sense of data”.
Sotto queste parole sono raggruppati tutti i processi per passare da un dataset a delle informazioni comprensibili, passando dai numeri a delle figure semplici da comprendere e analizzare.

Come già detto, all'aumentare dei dati a disposizione sono aumentati anche le possibilità per visualizzarli. Pie chart, bar chart, mappe e visualizzazioni dinamiche sono così entrare nelle nostre giornate in ambiti del tutto diversi tra loro.
Nel primo capitolo di questo elaborato è stato dimostrato che esempi di data design si possono trovare dalle riviste scientifiche ai quotidiani, dai musei alle applicazioni per smartphone che misurano i parametri di salute.

Sembra quasi che si stia affermando un nuovo linguaggio, un nuovo modo di comunicare delle informazioni. Un linguaggio che negli ultimi anni è esploso, diventando alla portata di tutti grazie all'ampia di strumenti open source disponibili in rete.

Insieme a questi strumenti sono nati così libri, corsi di studio e cattedre universitarie dedicate proprio a questa materia.

Katy Börner insegna Information Science presso l'Indiana University di Bloomington e negli ultimi anni si è dedicata molto ai processi di visualizzazione dei dati.
Dopo aver pubblicato diversi materiali su questo argomento, come *Visual Interfaces to Digital Libraries*[50] o *Atlas of Science*[51], nel 2013 ha cominciato un MOOC[52], Massive Online Open Course, dedicato proprio al data design.

Le lezioni sono state trasmesse su una piattaforma gratuita creata da Google, Google Course Builder, e sono state seguite da studenti provenienti da tutto il mondo.
Il corso è durato 6 settimane ma si è prolungato ben oltre la sua naturale scadenza. Tutte le lezioni, i tutorial e gli esercizi sono infatti ancora disponibili sul sito dell'università mentre le tematiche affrontate sono stati pubblicati in *Visual Insights, A Pratical Guide to Making Sense of Data*[53].

Oltre allo studio di un metodo di lavoro e allo sviluppo di diversi strumenti di visualizzazione, le lezioni della ricercatrice statunitense sono fondamentali anche per il loro contributo nell'identificare le forme esistenti di dataviz.

Un lavoro che nel suo volume successivo, *Atlas of Knlowledge*[54], è diventato talmente accurato da trasformarsi in una vera e propria grammatica dei grafici. Le riflessioni pubblicate in questo libro arrivano infatti a sciogliere le visualizzazioni in tutti i micro elementi che le costituiscono.

50 K. Börner, *Visual Interfaces to Digital Libraries*, 2003
51 K. Börner, *Atlas of Science*, 2010
52 http://ivmooc.cns.iu.edu
53 K. Börner e David E. Polley, *Visual Insights, A Pratical Guide to Making Sense of Data*, 2014
54 K. Börner, *Atlas of Knowledge*, 2014

2. Alla base della lingua. I morfemi grafici

Nella linguistica descrittiva l'unità minima dotata di senso è il morfema, la più piccola combinazione di caratteri che riesce a produrre un significato.
In questa categoria rientrano le radici dei verbi, i suffissi e le vocali tematiche. Ad esempio la particella *-si* è un morfema perché può identificare la coniugazione riflessiva di un verbo.

Secondo la grammatica dei grafici i morfemi sono gli elementi all'origine di qualsiasi visualizzazione, le basi da cui partire quando si vuole rendere comprensibile una dataset attraverso una serie di forme.
Katy Börner ha distinto tre aree in cui si possono si possono dividere questi segni:

A. Simboli Grafici
B. Simboli Linguistici
C. Simboli Pittorici

A. Simboli Geometrici

In geometria i punti e linee sono elementi privi di superficie, caratterizzati dall'avere rispettivamente zero e una dimensione. Nel mondo dei grafici questi elementi sono molto utilizzati dando però per scontato che nella loro rappresentazione grafica acquisiscano anche una superficie. Anzi, proprio la loro superficie, più o meno ampia o più o meno colorata, può diventare la variabile che permette di visualizzare le diverse caratteristiche dei dati rappresentati.

Punto

L'uso principale dei punti è quello di definire una collocazione precisa all'interno di una visualizzazione, una caratteristica che permette loro anche di chiarire la densità di un fenomeno visualizzato.

Linea

Le linee possono essere usate per definire elementi del paesaggio in cui la dimensione della lunghezza prevale sulle altre, come fiumi o strade, oppure il percorso seguito da un oggetto, sia esso un'automobile o un tornado. Le linee vengono utilizzate anche per definire le relazioni fra elementi diversi, in questi casi spesso si trasformano in frecce.

Area

L'uso più classifico delle aree è quello di definire il territorio occupato da un organismo amministrativo, sia questa un città, una regione o una provincia. I poligoni possono anche indicare lo spazio occupato da punti dotati dello stesso valore.

Superficie

Con le superfici entriamo nell'ambito degli elementi tridimensionali. Visualizzazioni con questo tipo di forme non sono molto diffuse ma possono essere utilizzate in diversi ambiti. Nella mappa qui riportata ad esempio le variazioni di altezza della superficie servono per indicare le differenze fra i prezzi delle case in Inghilterra rispetto alla loro posizione.

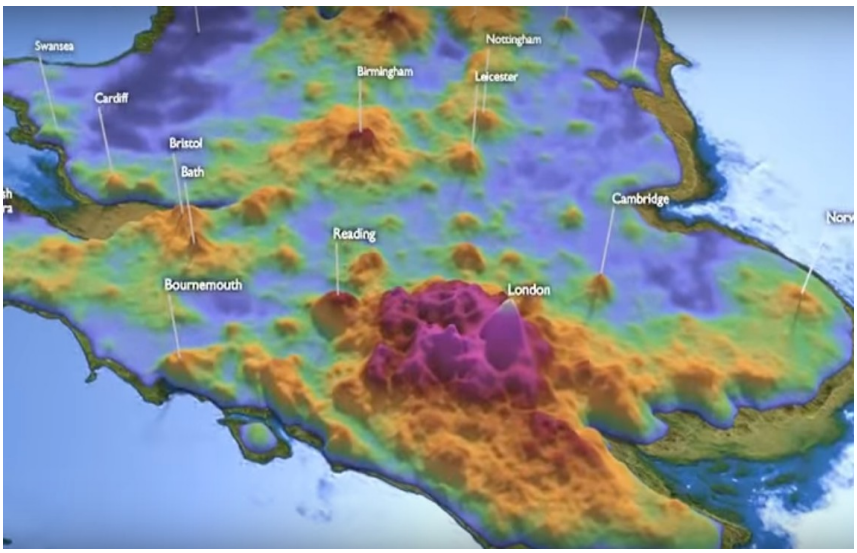


Illustrazione 29: House Price of England and Wales, fonte: https://www.youtube.com/watch?v=1q2e62VV1pg

Volume

Anche il volume è un elemento tridimensionale. Può essere usato per dare rendere più chiari gli elementi di una visualizzazione.

B. Simboli Linguistici

Nel data design i simboli linguistici come lettere, numeri e segni di interpunzioni non compaiono solo nei titoli o nelle didascalie ma possono diventare simboli grafici dotati di senso. Uno degli esempi più classici è quello delle *word cloud*, nuvole dove ogni parola ha una dimensione diversa in base ad un valore definito. La visualizzazione qui sotto riporta le parole più utilizzate nel *Moby Dick* di Herman Melville.

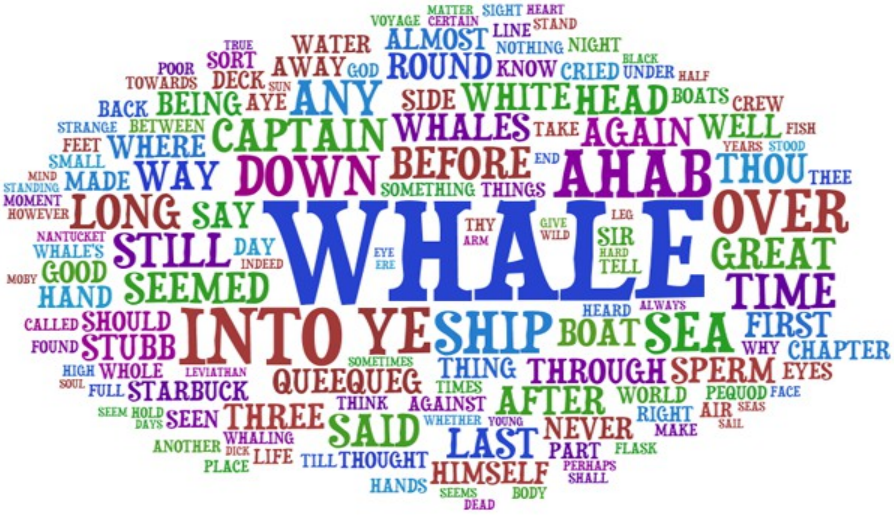


Illustrazione 30: Fonte: http://www.robweir.com/blog/2008/06/beautiful-word-clouds.html

C. Simboli Pittorici

Un simbolo pittorico è un segno convenzionale per rappresentar nozioni complesse, come quantità, qualità o relazioni.

Immagini e Icone

Le immagini sono riproduzioni fedeli degli oggetti che rappresentano, come ad esempio le nuvole disposte sulle cartine delle previsioni meteo, mentre le icone sono simboli astratti che solitamente hanno bisogno di una legenda per essere compresi.

Glifi Statistici

I glifi statistici prendono le loro forme dalle visualizzazioni tipiche della statistica, come bar graph o i line graph. La differenza qui è che non hanno griglie o valori di riferimento.
L'uso più classico per questo tipo di figure sono le sparkline, grafici densi di informazioni ma grandi quanto lo spazio occupato da una striscia di parole.

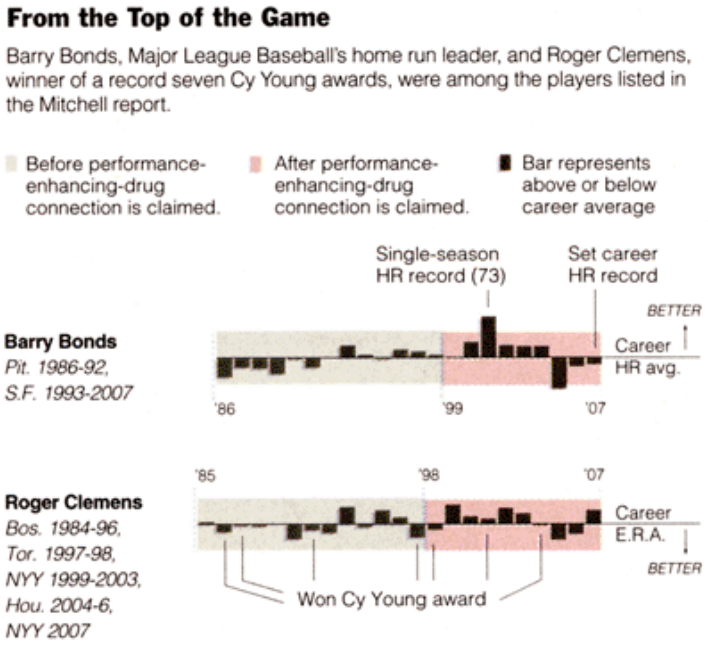


Illustrazione 31: Le Sparklines delle statistiche di due giocatori di baseball, fonte: http://www.edwardtufte.com/

3. Cambiare aspetto per cambiare significato. Le variabili grafiche

I simboli grafici non hanno un significato intrinseco. Un punto, una linea o un'area non vogliono dire nulla se non sono inseriti in un contesto spaziale, se non hanno un colore o una forma collegati ad un determinato valore oppure ancora se non sono confrontati con altri simboli grafici.
Per completare questi morfemi del data design è quindi necessario introdurre altri aspetti, altri elementi che siano in grado di rendere più completa l'informazione offerta.

È qui che entrano in gioco le variabili grafiche, le variabili che definiscono il significato dei simboli.

La prima distinzione operata da Börner è quella fra una variabile spaziale e una retinica.

La variabile spaziale definisce la posizione di un oggetto nello spazio mentre quella retinica identifica delle caratteristiche proprie dell'oggetto stesso. Per questo secondo caso si apre un ventaglio molto ampio di possibilità che merita una trattazione più approfondita.

Le variabili retiniche coinvolgono infatti diversi aspetti: forma, colore, texture, proprietà ottiche e movimento.

Forma

Le forma con cui un simbolo grafico ci appare è definita dalla combinazione di diversi elementi:

- **Dimensione**. Spesso alla dimensione di un simbolo grafico viene associata una variabile di tipo quantitativo. Più la dimensione aumenta, più quel simbolo rappresenta un valore maggiore.

- **Figura**. La figura di un oggetto si può dividere in tre tipi: geometrica, naturale o astratta. Le figure geometriche sono i poligoni o i solidi che siamo abituati a conoscere, come triangoli o cubi, quelle naturali riprendono le forme dell'oggetto a cui si riferiscono, come uomini o animali, mentre quelle astratte sono simboli come glifi o icone.

- **Rotazione**. Si riferisce all'inclinazione dell'oggetto e può trasmettere informazioni sia quantitative che qualitative.

- **Curvatura**. Sono i gradi rispetto ai quali un simbolo grafico è curvato

- **Angolo**. È l'angolo entro cui avviene l'intersezione fra due simboli grafici.

- **Chiusura**. È la distanza che separa due simboli grafici.

Colore

Il colore è una delle variabili grafiche più usate nell'ambito della data visualization. Può servire semplicemente per distinguere un oggetto dall'altro oppure può rappresentare una variabile rispetto ad una scala che va da una tonalità all'altra o da un grado di saturazione all'altro.
All'interno del colore si possono distinguere altre tre variabili.

- **Tono**. Si può definire anche tinta e viene utilizzato solitamente per caratterizzare dati qualitativi. Ad esempio in una dataviz che si riferisce alla geografia di un territorio le zone boscose possono essere rappresentate in verde mentre quelle sabbiosi in giallo.

- **Luminosità**. È la quantità di luce che viene riflessa da un oggetto. I due estremi di questa scala sono il bianco e il nero.

- **Saturazione**. Indica quanto pigmento è contenuto all'interno di un colore. Gli oggetti con una saturazione maggiore appaiono in primo piano mentre quelli con una saturazione minore vengono solitamente usati come sfondo.

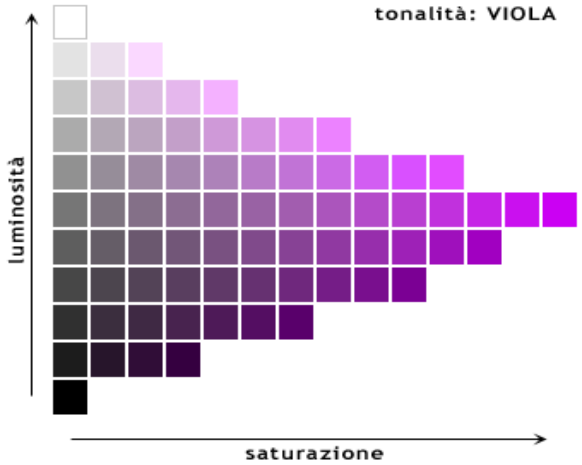


Illustrazione 32: Differenza fra Saturazione e Luminosità rispetto alla tonalità "Viola", fonte: http://web.mclink.it

Texture

La texture non è una variabile intrinseca del simbolo grafico ma si riferisce piuttosto alle relazione che più simboli grafici creano fra di loro nello stesso spazio.

- **Spacing**. È la densità e si riferisce alla quantità di spazio che separa gli elementi grafici.

- **Granularity**. In questo caso parliamo di una scala che indica la dimensione dei simboli grafici rispetto allo stesso spazio.

- **Pattern**. Sono i modelli, le trame create dalla ripetizione di un simbolo grafico.

- **Orientation**. Anche in questo caso si parla di rotazione ma se nelle variabili relative alla forma questa era riferita ad un singolo oggetto, ora riguarda tutti gli oggetti presenti in un determinato spazio.

- **Gradient**. Indica l'aumento o la diminuzione di una delle proprietà precedenti, spesso viene utilizzato per dare un'idea di prospettiva.

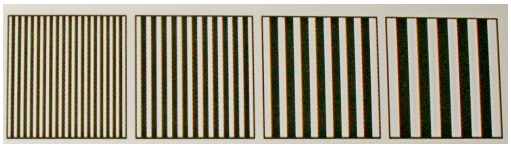


Illustrazione 34: Esempio Granularity, fonte: K. Börner, Atlas of Knowledge

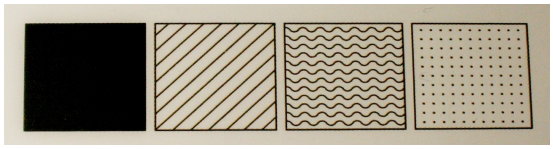


Illustrazione 33: Esempio Pattern, fonte: K. Börner, Atlas of Knowledge

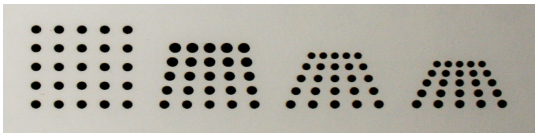


Illustrazione 35: Esempio Gradient, fonte: K. Börner, Atlas of Knowledge

Proprietà Ottiche

Le proprietà ottiche di un soggetto possono essere utilizzate per definire una scala di importanza fra elementi simili.

- **Sfocatura**. Una sfocatura maggiore indica solitamente incertezza mentre una definizione più alta corrisponde ad una sicurezza maggiore.

- **Trasparenza**. Anche in questo caso la trasparenza indica una scala di importanza rispetto ad elementi simili. Quando più oggetti si sovrappongono quelli più trasparenti vengono coperti.

- **Ombreggiatura**. L'ombreggiatura di un oggetto può essere utilizzata sia per finalità estetiche, come per dare un aspetto tridimensionale, oppure per indicare valori qualitativi.

- **Profondità Stereoscopica**. È uno stratagemma ottico che suggerisce un'idea di profondità accostando due immagini

Movimento

Il movimento è una variabile grafica che si riferisce solamente a tutte quelle dataviz che utilizzano dei sistemi di animazione. Katy Börner ha provato ad identificare tre valori che si possono trovare in questo tipo di visualizzazioni.

- **Speed**. Si riferisce alla velocità con cui un elemento si muove ma non alla sua direzione.

- **Velocity**. Questa variabile si riferisce non solo alla velocità con cui un oggetto si muove ma anche alla sua direzione.

- **Rhythm**. È il modello con cui si definisce il numero di volte in cui un gruppo di oggetti cambia nel tempo.

4. Simboli e variabili grafiche. La mappa delle combinazioni

La fusione di simboli e variabili grafiche ha permesso a Börner di creare una tabella in cui vengono affrontate tutte le possibili combinazione tra simboli e variabili. Gli spazi bianchi sono le intersezioni dove non è stato possibile trovare un esempio efficace.

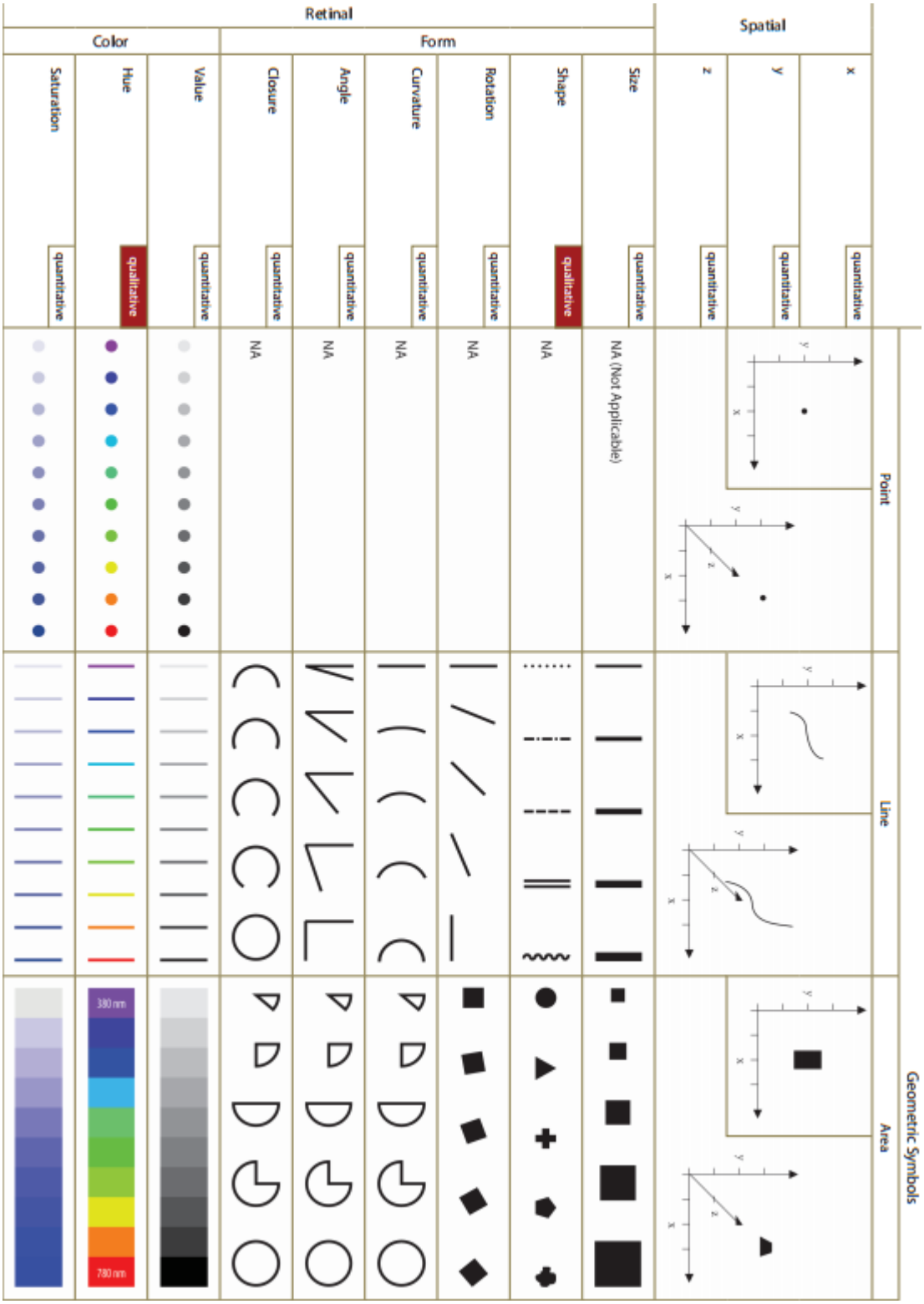


Illustrazione 36: Fonte: http://scimaps.org/

Illustrazione 37: Fonte: http://scimaps.org/

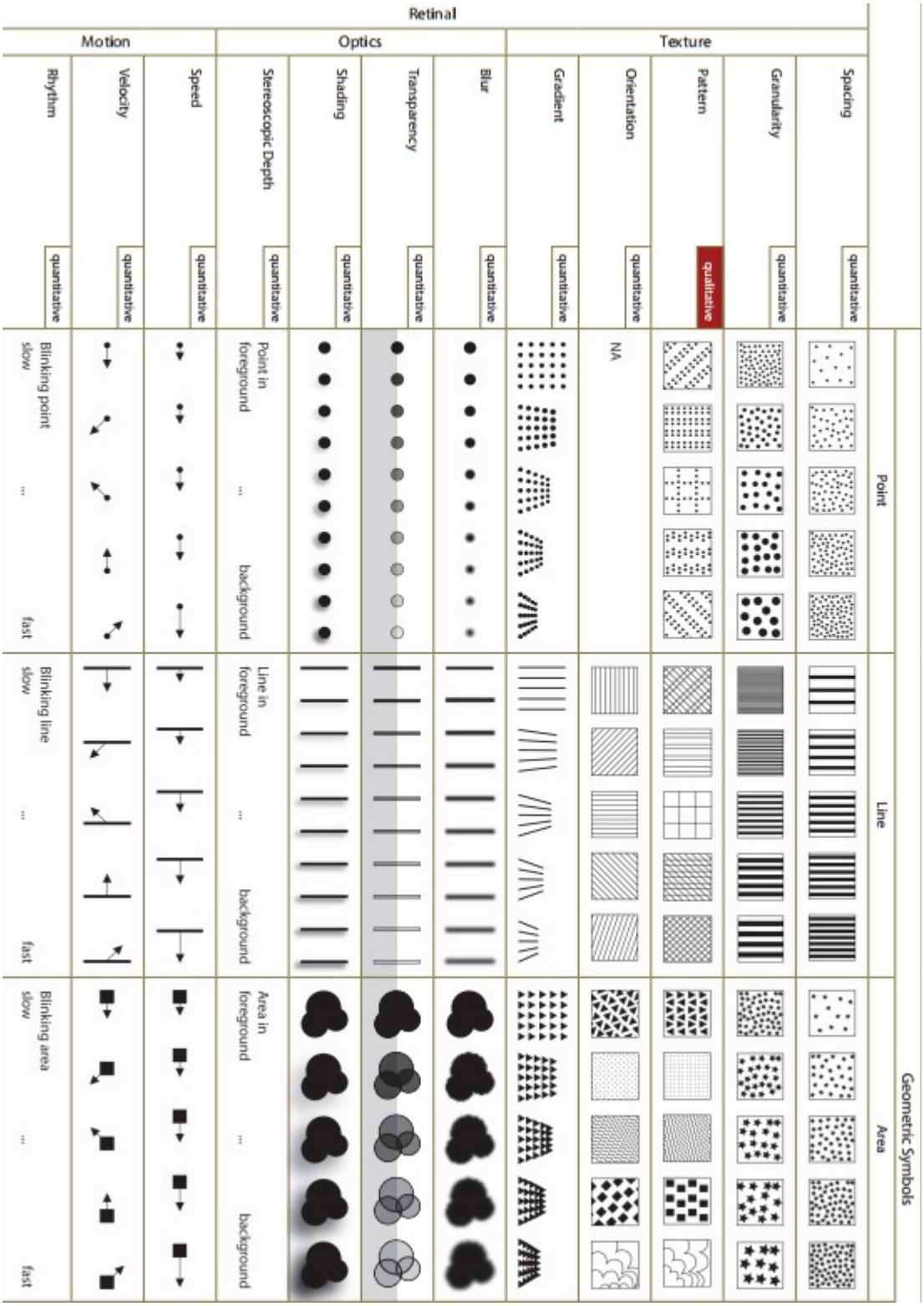


Illustrazione 38: Fonte: http://scimaps.org/

Illustrazione 39: Fonte: http://scimaps.org/

5. L'unione dei morfemi. Il lessico dei grafici

Rimanendo sul paragone con la linguistica descrittiva, l'unione di diversi morfemi crea le parole, il lessico che permette la costruzione di ogni frase. Se quindi per morfemi nel mondo del data design intendiamo i simboli grafici declinati in tutte le loro possibili variabili, per lessico vengono presi in considerazione invece quei modelli di grafici che elaborati e combinati fra di loro possono dare vita a qualsiasi dataviz.

Per questione linguistiche è preferibile utilizzare le definizioni inglesi di questi modelli. Uno di essi infatti prende il nome di "Graphs", una parola che nella lingua italiana si potrebbe tradurre solamente come "Grafico". Questo termine però in italiano può essere benissimo utilizzato per indicare qualsiasi forma di data design.

I modelli che combinano i morfemi grafici trattati negli scorsi paragrafi sono cinque: Tables, Chart, Graphs, Maps e Network Layouts.

Questa distinzione non è inedita nella storia di questa materia ma si tratta dell'ultima versione di un sistema che il cartografo francese Jacques Bertin aveva già cominciato ad elaborare nel 1967 con il suo *Semiologie Grafique*[55].

Tables

Le *tables* sono la forma più primordiale di dataviz. Quella che compare ovunque e quella con cui lavorano sia i grafici e gli analisti. È la materia grezza, il primo passaggio in cui i dati vengono strutturati.

Ogni tabella è formata da righe e colonne. Nel linguaggio informatico di gestione dei database le righe vengono definite "tuple" o "records" mentre le colonne "attributi" perché definiscono una caratteristica specifica dei dati. Le intersezioni fra righe e colonne prendono il nome di "celle". Nelle celle le informazioni vengono codificate attraverso simboli numerici e linguistici.

Solitamente la prima riga della tabella serve solo per indicare i nomi degli attributi. Le celle possono essere raggruppate per tipologie e per distinguerle si utilizzano anche variabili come colori o dimensioni.

	A	B	C	D	E	F	G
1	State	County	Voting	# Households	% Households	# low income	% smokers
2	Alabama	Autauga	R	602	3,7	9274	20,87
3	Alabama	Baldwin	R	1707	3,07	30657	21,74
4	Alabama	Barbour	R	726	6,97	9505	32,72
5	Alabama	Bibb	R	596	7,94	7087	33,7
6	Alabama	Blount	R	844	4,38	14841	29,11
7	Alabama	Bullock	D	473	11,96	4459	38,29
8	Alabama	Butler	R	541	6,46	6788	31,8
9	Alabama	Calhoun	R	1281	2,82	20814	18,51
10	Alabama	Chambers	R	828	5,71	9317	25,47
11	Alabama	Cherokee	R	512	5,25	8294	34,5
12	Alabama	Chilton	R	819	5,38	11941	30,31
13	Alabama	Choctaw	R	620	9,76	6986	43,97

Illustrazione 40: Dateset sulle condizioni di salute della popolazione degli U.S.A.

55 J. Bertin, *Semiologie Grafique*, 1967

Addirittura alcune tabelle possono contenere al loro interno degli attributi grafici simile alle sparklines.

	Total		Low	High	6/1/2008
All Books	$ 325,926		$8,021	$13,466	$ 9,161
Arts & Photography	$ 42,844		$776	$1,995	$ 996
Children's Books	$ 43,595		$786	$1,951	$ 1,368
Computers & Internet	$ 43,187		$799	$1,995	$ 1,358
History	$ 25,722		$350	$1,407	$ 1,021
Mystery & Thrillers	$ 44,833		$758	$1,984	$ 1,471
Romance	$ 42,365		$791	$1,988	$ 1,129
Science Fiction & Fantasy	$ 41,377		$789	$1,936	$ 859
Sports	$ 41,871		$756	$1,937	$ 959

Illustrazione 41: Dateset relativo alle vendite di una libreria

Charts

Le *chart* rappresentano dati quantitativi e qualitativi senza un preciso sistema di riferimento. Più che fornire la rappresentazione di un dato sono orientate infatti a definire il confronto fra dati diversi. Gli esempi più famosi di questo tipo di dataviz sono le *pie chart* e le *doughnut chart*.

Le *pie chart* sono una delle forme più conosciute di data design nonché uno dei modi più semplici per definire la differenza fra percentuali. Tuttavia sono state sollevate parecchie critiche sui principi alla base di questo tipo di *chart* perché l'occhio umano farebbe fatica a decifrare precisamente la differenza di grandezze fra aree mentre riuscirebbe a confrontare meglio delle lunghezze.
Uno dei più noti detrattori delle *pie chart* è Edward Tufte che ha dichiarato:

"The only worse design than a pie chart is several of them, for then the viewer is asked to compare quantities located in spatial disarray both within and between pies"[56] .

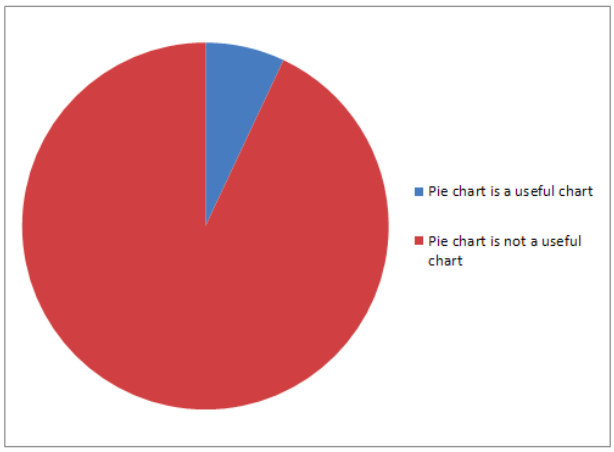


Illustrazione 43: Pie Chart apparsa in un articolo sull'utilità di questo strumento pubblicato sul sito http://priceonomics.com/

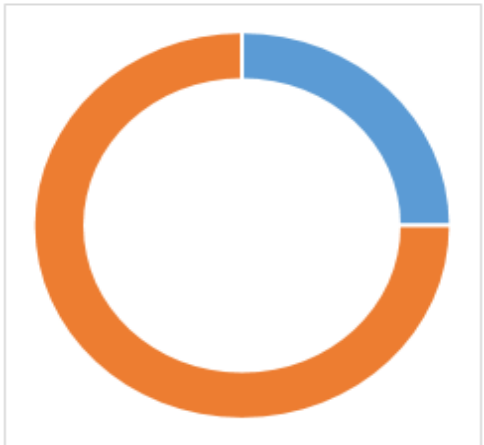


Illustrazione 42: Esempio di Doughnut Chart

56 E. Tufte, *The Visual Display for Quantitative Information*, 1983

Gli altri due modelli che appartengono a questo gruppo sono le *bubble chart* e i *tag cloud* o *word cloud*.
Nelle *bubble chart* dati diversi sono rappresentati da cerchi diversi disposti nello spazio. Sono molto utili per rappresentare dati quantitativi perché ad ogni elemento può essere assegnata una grandezza diversa.

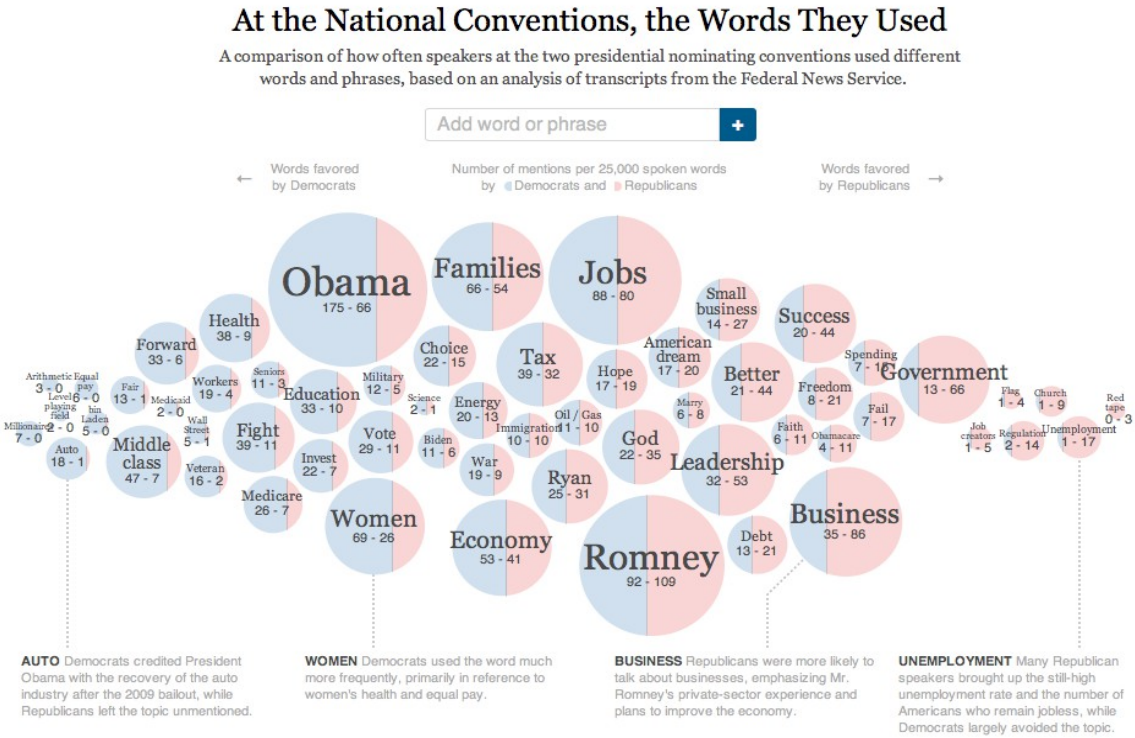


Illustrazione 44: Buble Chart interattiva delle parole più usate alle Convention Nazionali dei due principali partiti politici degli U.S.A., fonte: http://www.nytimes.com/interactive/2012/09/06/us/politics/convention-word-counts.html?_r=0

I *tag cloud* sono invece nuvole di parole in cui, come visto nell'esempio già riportato nelle pagine precedenti, ogni termine ha una dimensione diversa dagli altri in base ad un valore definitivo, come il suo numero di ricorrenze in un testo.

Graphs

I *graphs* si differenziano dalle *chart* perché il sistema di riferimento per rappresentare i dati è ben definito, infatti solitamente i *graphs* si basano sul modello del piano cartesiano che dispone sull'asse delle ascisse e quello delle ordinate i valori su cui si vuole costruire la rappresentazione.

Esistono molte varianti di questo tipo di visualizzazioni, come i *bar graphs*, i classici grafici a barre o i *line graphs*, in cui invece i dati rappresentati sul piano sono uniti da una linea. Negli *scatter plots* ogni dato viene rappresentato semplicemente come un punto mentre i *bubble graphs* possono essere simili alle *bubble chart* ma nel loro caso le misure dei cerchi sono stabilite sulla base di un sistema che permette non solo di confrontare i valori rappresentati ma anche di misurarli.

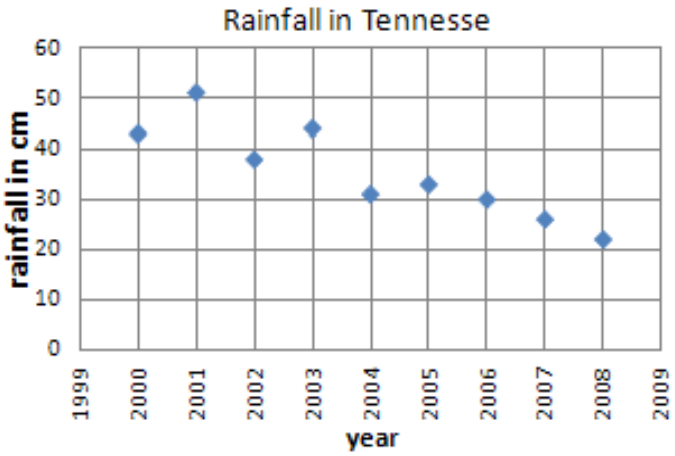


Illustrazione 45: Scatter Plot che rappresenta il livello delle precipitazione annue in Tennesse, fonte: http://www.mytestbook.com/

L'ultimo esempio per questa branca di dataviz sono i *parallel coordinate graphs* in cui vengono accostati non due ma diversi valori sullo stesso piano utilizzando più assi. In questo caso i punti rappresentati sono collegati da linee che definiscono l'appartenenza ad una stessa registrazione.

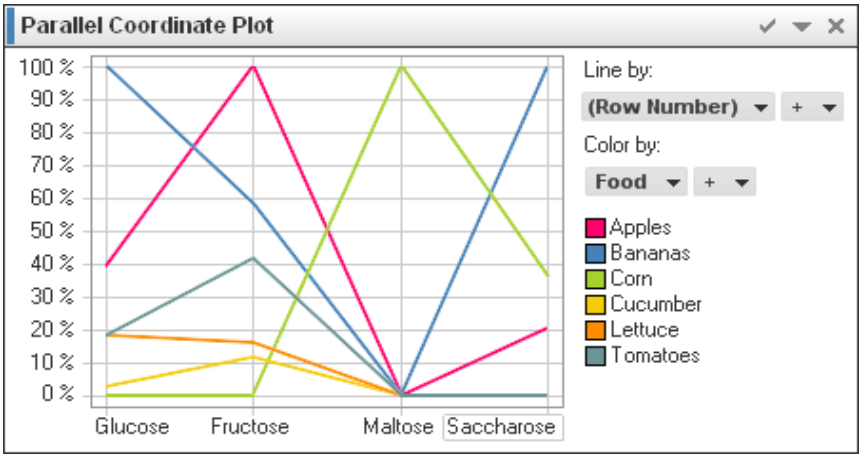


Illustrazione 46: Parallel Coordinate Graph sugli zuccheri contenuti in diversi alimenti, fonte: http://stn.spotfire.com/

Maps

Quando si parla di *maps* non ci si riferisce solo a quelle geografiche ma anche a tutte quelle visualizzazioni che rappresentano i dati secondo la loro relazione spaziale. Spesso in questo tipo di dataviz alla locazione geografica vengono sovrapposte altre informazione attraverso variabili grafiche come la forma o il colore.

Gli esempi più importanti di *maps* sono:

- **Cartograms**. Le aree geografiche rappresentate subiscono delle distorsioni in base ai dati che rappresentano.

- **Choropleth Map**. Qui i diversi valori sono rappresentati tramite l'uso del colore. Queste mappe sfruttano la divisione del territorio in aree, come ad esempio quelle amministative.

- **Isochrone Map**. A differenza delle *Choropleth Map* i dati non sono rappresentati in base alle aree ma in base a dove si sviluppa il fenomeno. Sono le mappe utilizzate per spiegare gli eventi atmosferici.

- **Proportional Symbol Map**. In questo caso i dati vengono rappresentati attraverso dei simboli distribuiti sulla mappa in base all'area geografica a cui si riferiscono.

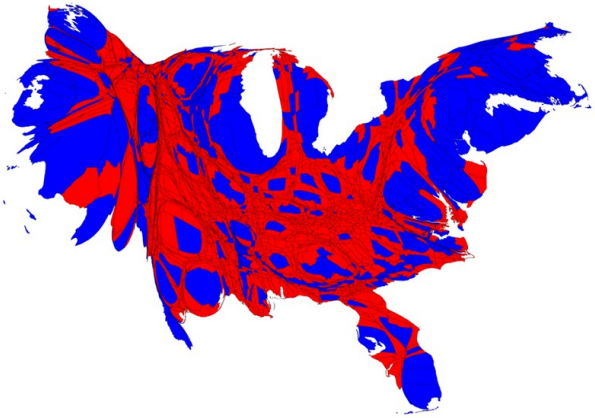


Illustrazione 47: Cartogram che rappresenta il territorio americano in base alla popolazione di ogni contea e al voto deciso alle elezioni presidenziali del 2012. Il rosso è per i democratici, il blu per i repubblicani.

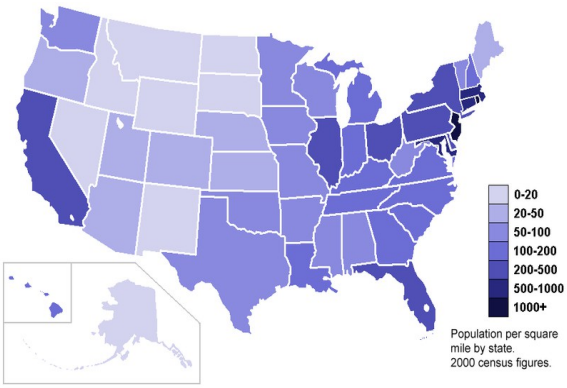


Illustrazione 48: Choropleth Map che rappresenta la densità di popolazione per miglio quadrato in ogni Stato.

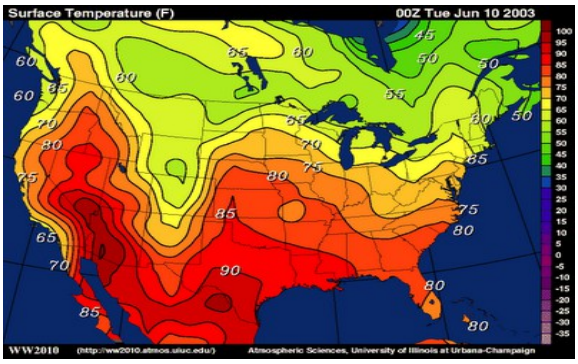


Illustrazione 49: Isochrone Map sulla temperatura media in Nord America in una giornata di giugno del 2003

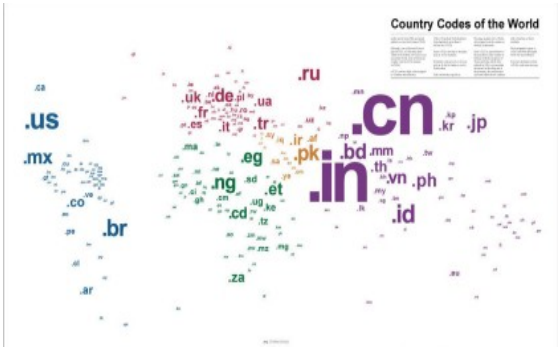


Illustrazione 50: Distribuzione dei domini internet nazionali nel mondo. Alla grandezza della sigla corrisponde il numero di domini registrati.

Network Layouts

Quella dei *network layout* è una famiglia di grafici che serve per rappresentare i collegamenti fra elementi diversi. Solitamente qui vengono utilizzati dei nodi per indicare i soggetti analizzati e delle linee per esprimere i collegamenti fra questi nodi.

Linee e nodi possono essere elaborati tramite la sovrapposizione di diverse variabili. I nodi possono

essere posizionati nello spazio secondo un criterio definito o un algoritmo causale e possono essere di colori, forme e dimensioni diverse. Lo stesso accade per linee che solitamente sono più o meno spesse a seconda dell'importanza del collegamento che rappresentano.

Le relazioni definite da una dataviz si dividono a loro volta in due tipi: *tree* e *network*.

Le visualizzazione che si riferiscono al modello *tree* sono quelle basate su gerarchie, dove ci sono elementi principali da cui si sviluppano elementi secondari. Quelle che invece seguono il modello *network* indicano elementi che sono connessi attraverso una rete non gerarchica.

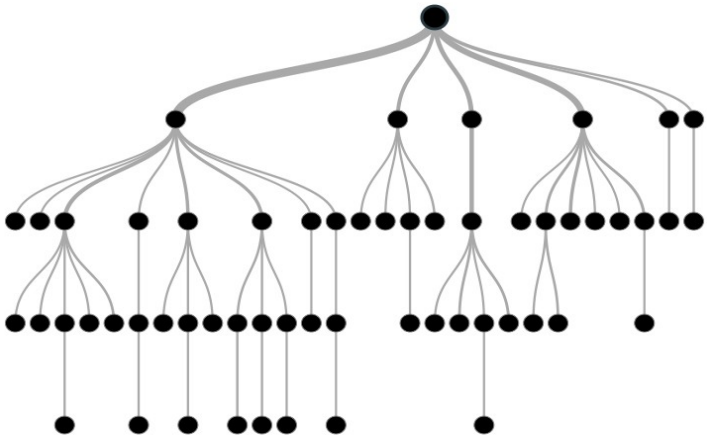


Illustrazione 51: Esempio Tree

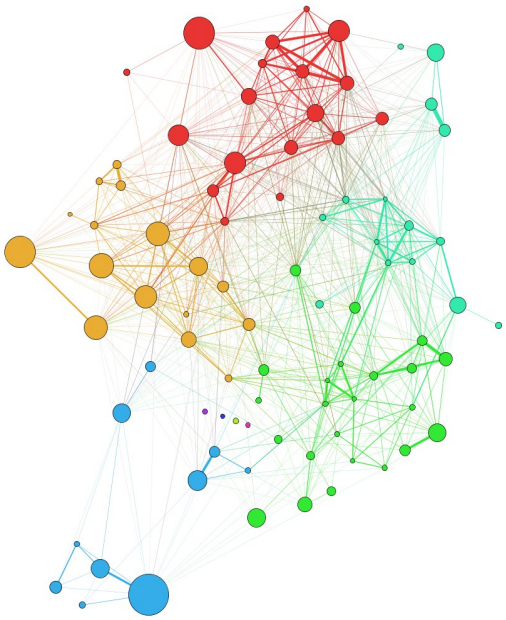


Illustrazione 52: Esempio Network

6. La sintassi complessa di Edward Segel e Jeffry Heer

Fino a questo punto le dataviz sono state considerate come singoli prodotti di una traduzione semantica che trasforma in numeri in forme.

Questo ragionamento può valere forse quando si parla di infografiche statiche, come quelle che si trovano sulle pagine di libri o giornali. Tuttavia già in questi media è possibile imbattersi in visualizzazioni formate dalle combinazione di grafici differenti che prendono il nome di *dashboard*.

Nell'era di internet le combinazioni di grafici differenti possono diventare però molto complesse. Ora sono state create infatti forme interattive di dataviz dove non solo è possibile scegliere i dati da rappresentare con un grafico ma anche navigare attraverso grafici diversi.

Hanno iniziato quindi ad affermarsi delle visualizzazioni esplorabili in cui, dopo una presentazione iniziale dei dati a disposizione, è l'utente stesso a costruirsi il grafico che più lo interessa.

Un caso interessante è quello creato da Amanda Cox una giornalista laureata in statistica che si occupa di data journalism al *New York Times*. Nel 2009 ha pubblicato *The Jobless Rate for People Like You*[57].

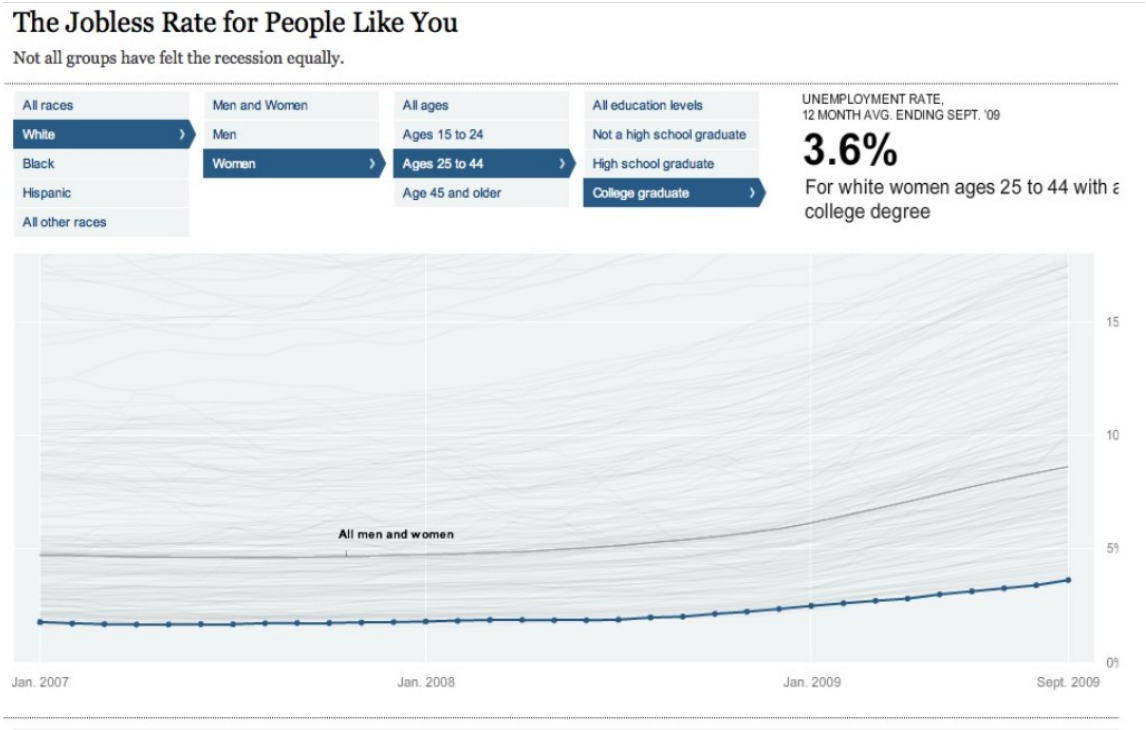


Illustrazione 53: Jobeless Rate for People Like You, fonte: http://www.nytimes.com/interactive/2009/11/06/business/economy/unemployment-lines.html?_r=0

I filtri posti sopra al grafico permettono al lettore di selezionare il profilo che più gli interessa per scoprire a quale tasso di disoccupazione è associato. Il titolo invita infatti a personalizzare questa dataviz non limitandosi ad un'esplorazione casuale ma cercando proprio la combinazione di filtri che rispecchia la propria situazione.

57 http://www.nytimes.com/interactive/2009/11/06/business/economy/unemployment-lines.html?_r=0

Esistono migliaia di esempi simili a questo, più o meno riusciti e più o meno semplici da navigare. Nella dataviz di Amanda Cox infatti i filtri a disposizione permettono di agire solo su un grafico e su un determinato dataset mentre non è difficile trovare ora soluzioni dove sono analizzati dataset diversi sfruttando varie tipologie di grafici.

Se quindi per comprendere gli elementi alla base della data visualization è stato creato un parallelo con la linguistica descrittiva, ora è il momento di spostarsi sul fronte dell'analisi del periodo, capendo come il lessico dei grafici possa creare un discorso complesso.

L'accostamento di grafici differenti e la possibilità di esplorarli come fossero capitoli di un libro permette infatti di arrivare alla costruzione di un racconto. Una storia che si sviluppa passando da una *bar chart* ad una mappa, da una linea temporale ad un grafico di relazione.

Edward Segel e Jeffrey Heer sono due informatici e ricercatori della Standford University che hanno deciso di affrontare questa tematica cercando di capire quando una serie di data visualization possa acquisire lo status di narrazione. Nel 2010 hanno pubblicato un articolo dal titolo *Narrative Visualization: Telling Stories with Data*[58] che definisce tre modelli attraverso cui è possibile creare una storia partendo da una serie di dataviz.

Prima di affrontare i modelli proposti da Segel e Heer è necessario capire cosa i due autori intendano per storia.
La definizione che utilizzano è quella riportata da Jonathan Harris, un artista e computer scientist che definisce se stesso uno storyteller.

“I define 'story quite loosely. To me, a story can be small as a gesture or as large as a life. But the basic elements of a story can probably be summed up with the well-worn Who/What/Where/When/Why/How”.

I criteri usati per definire una storia da Harris sono quindi simili a quelli adottati nel giornalismo secondo l'arcinota regola delle 5 W. Una distinzione che ritorna anche nei MOOC organizzati dall'Università dell'Indiana dove le lezioni erano divise in quattro blocchi, ognuno dei quali affrontava una specifica rea tematica: When, Where, What, Whit Whom.

Una dataviz quindi per raccontare una storia dovrebbe rispettare gli stessi criteri di un articolo di giornale. Cercando di definire una notizia attraverso tutti i suoi aspetti.

Una storia poi dovrebbe avere uno filo cronologico e quindi un inizio, uno sviluppo e una fine. Prendendo come riferimento sempre il giornalismo si potrebbe qui introdurre il meccanismo della *top* down. Questo modello di scrittura di un articolo prevede l'inquadramento di una notizia nelle prime righe, in modo tale da fornire subito al lettore le coordinate principali per poi procedere con l'approfondimento dei dettagli. Il fenomeno rappresentato dai grafici verrebbe quindi prima riportato in modo sommario e poi esplorato con passaggi più dettagliati. L'ultima tappa di questo percorso sarebbe dunque una conclusione che termini e renda organico tutto il lavoro.

Da questa premessa Segel e Heer hanno cominciato a catalogare una serie di dataviz interattive prese da diverse testate giornalistiche: *New York Times*, *The Guardian, The Financial Times*, *The Washington Post* e *Slate*.

58 E. Segel e J. Heer, *Narrative Visualization: Telling Stories with Data*, 2010

Dopo una prima analisi del materiale raccolto, hanno stilato una lista di caratteristiche proprie di queste dataviz che andavano dalla presenza di testo all'uso di bottoni di navigazione fino alla possibilità di zoomare dei punti specifici.
Ogni infografica è stata quindi inserita in una tabella e catalogata in base questa lista, segnando semplicemente con un + o un – l'assenza o la presenza di una determinata caratteristica.

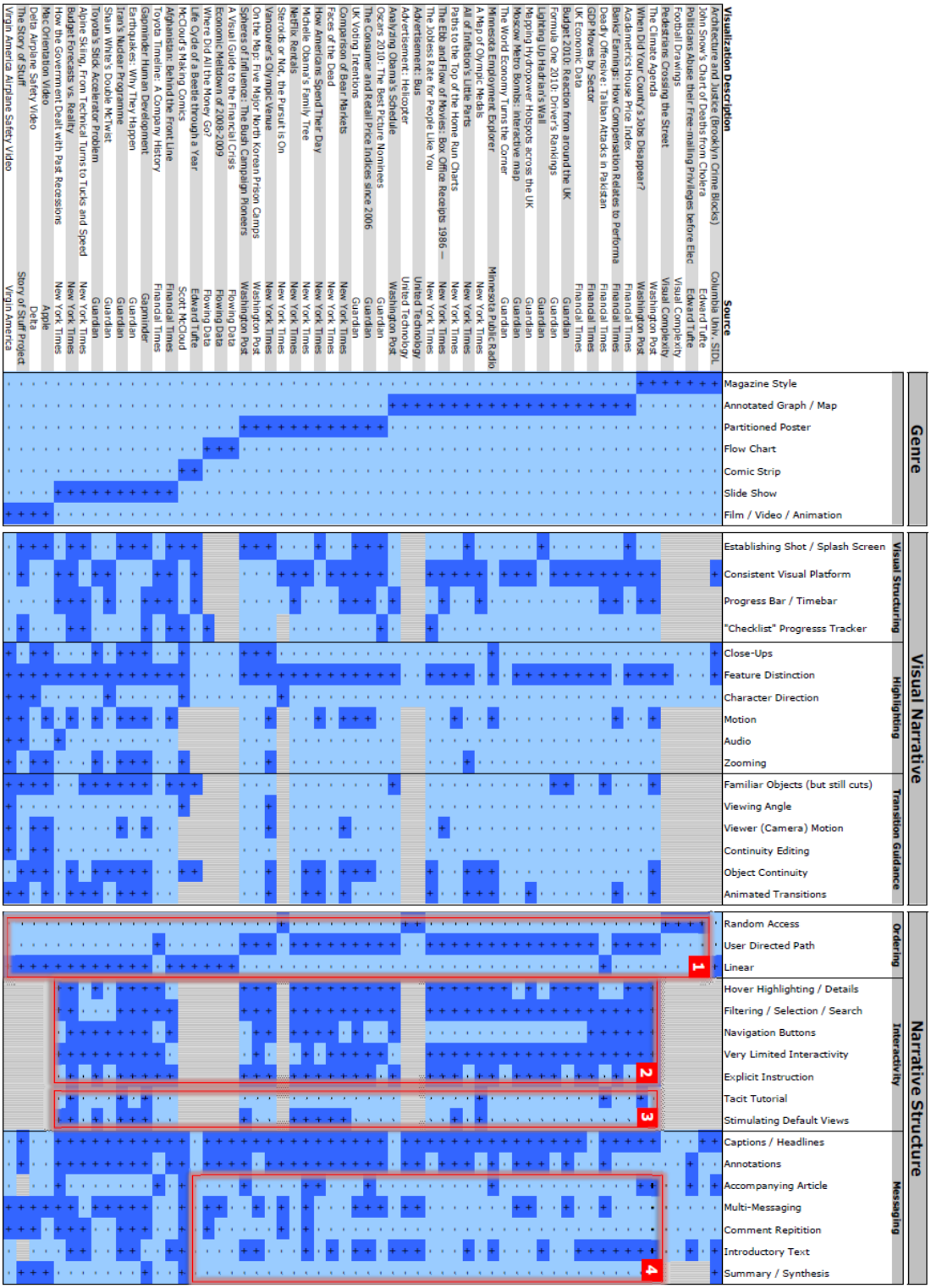


Illustrazione 54: Tabella di Segel-Heer

Tre modelli, tre strumenti per realizzarli. Drill-Down Story, Interactive Slideshow e Martini Glass Structure.

Dall'elaborazione dei dati ricavati dopo la raccolta di 58 dataviz Segel e Heer hanno teorizzato tre modelli diversi per la costruzione di una infografica interattiva. Questi modelli sono pensati in base alla relazione autore-lettore, e riportati seguendo una scala che va da un approccio guidato dall'autore, *author-driven*, ad uno invece guidato dal lettore, *reader-driven*.

Un approccio guidato dall'autore prevede una trama lineare, la concentrazione su un messaggio e nessuna possibilità di interazione da parte del lettore. Questo tipo di costruzione funziona bene quando l'obiettivo è proprio quello di creare un racconto ben definito, con un inizio e una fine chiari.

Un approccio guidato dal lettore identifica invece un ordine dei contenuti non gerarchico e un alto grado di interattività. Attraverso questo sistema è possibile di solito accedere ad una serie di strumenti per analizzare a fondo i dati a disposizione.

Le dataviz che seguono perfettamente un approccio guidato dall'autore o uno guidato dal lettore sono ovviamente poche. La maggior parte di loro si colloca nel mezzo, affrontando tre diversi sistemi che possono essere così identificati: Drill-Down Story, Interactive Slideshow e Martini Glass Structure.

Nelle pagine seguenti verrà accostato ad ogni modello un esempio e un software per realizzarlo.

La parola al lettore. Drill-Down Story

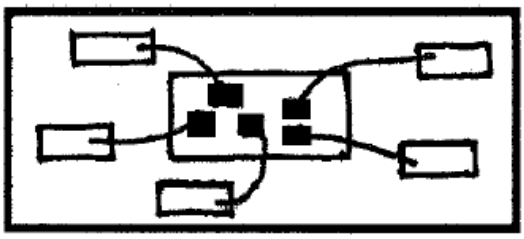


Illustrazione 55: Modello Drill-Drown Story

In questo tipo di visualizzazioni è il lettore a scegliere l'ordine con cui scoprire il racconto.
Una Drill-Down Story presenta infatti un tema generale e poi lascia la possibilità di analizzare le parti ritenute più interessanti. L'utente diventa quindi una sorta di archeologo che davanti ad un sito da esplorare comincia a scavare dove pensa di trovare i reperti più importanti.

Esempi del genere si possono riscontrare soprattutto nelle mappe dove vengono geolocalizzati elementi che poi, spostando sopra il cursore del mouse o toccandoli su un touch screen, possono essere ulteriormente esplorati.

Uno di questi casi è la visualizzazione pubblicata sul *Washington Post* dal titolo *On the Map: Five Major North Korean Prison Camps*[59]. In questa Drill-Down Story che rappresenta i campi di prigionia in North Korea l'analisi avviene su due livelli. In una prima mappa vengono rappresentati tutti i luoghi su cui si hanno delle informazioni. Selezionando un luogo si apre una finestra con una breve descrizione e cliccando ulteriormente si accede ad una nuova mappa in cui il luogo scelto viene rappresentato più da vicino ed è replicato il sistema a finestre informative adottato nella mappa iniziale.

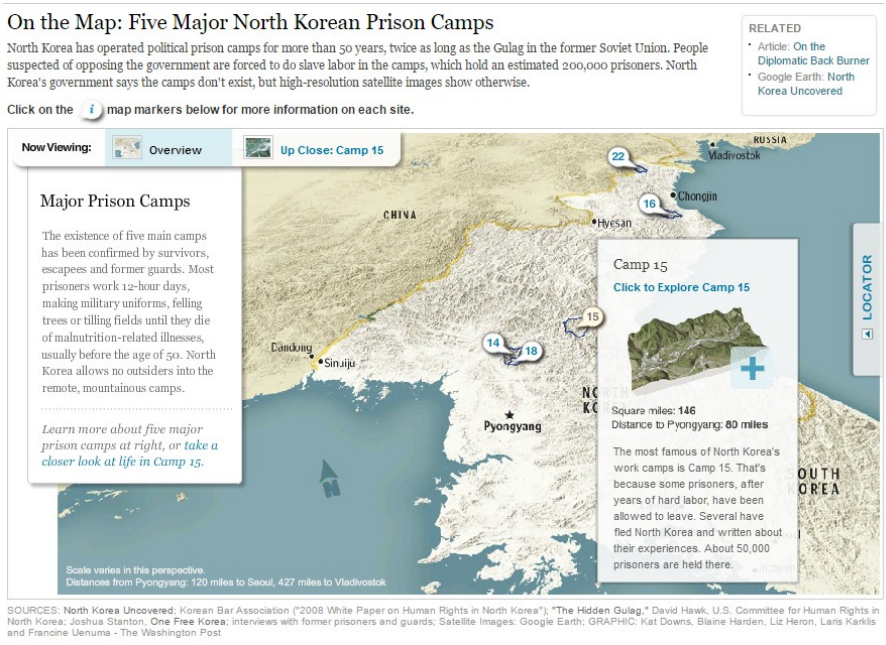


Illustrazione 56: Primo livello: Overview

59 http://www.washingtonpost.com/wp-srv/special/world/north-korean-prison-camps-2009/

Illustrazione 57: Secondo Livello: Camp 15

L'informazione attraverso le mappa, StoryMap JS

StoryMap Js[60] è uno strumento on line a cui si può accedere in maniera gratuita e senza bisogno di scaricare un applicazione desktop.

Le visualizzazioni che permette di creare sono divise solitamente in due parti. La metà a sinistra è occupata dalla mappa con dei simboli che indicano dove è possibile interagire. Nella metà di destra invece compaiono le informazioni associate ad un determinato simbolo, disposte come una scheda informativa. Oltre ai testi e alle fotografie, ad ogni punto è possibile associare anche video e audio.

Il software offre la possibilità di navigare anche tramite slideshow. Definendo la sequenza dei punti da osservare è possibile quindi passare da una posizione all'altra. Questa possibilità è utile per realizzare dei racconti basati su un preciso ordine cronologico. In questo caso la Drill-Down Story si trasformerebbe però in un altro modello: l'Interactive Slideshow.

Qui di seguito è riportata una dataviz creata dal team Yahoo India in cui vengono rappresentati tutte le squadre di calcio che hanno partecipato al mondiale tenutosi a Rio de Janiero nel 2014.

60 https://storymap.knightlab.com/

Illustrazione 58: Fonte: https://cdn.knightlab.com/libs/storymapjs/latest/embed/?url=https%3A%2F%2Fwww.googledrive.com%2Fhost%2F0B45EeRGUhSinVmVqaWIxakhjcDQ%2Fpublished.json

Un compresso fra lettore e autore. Interactive Slideshow

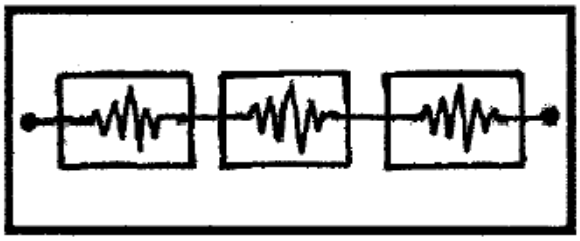


Illustrazione 59: Modello Interactive Slideshow

L'Interactive Slideshow si presenta come una serie di slide su cui però il lettore è libero di fermarsi, tornare indietro oppure proseguire.
In questo modo l'autore può dividere la storia che vuole raccontare in diversi capitoli mentre l'utente è libero di approfondirne alcuni e di sorvolare su altri. Si tratta quindi di un ottimo compromesso fra lettore e autore.

Un sistema di questo tipo è utilizzato ad esempio in *Valar Morghulis*[61], una dataviz creata sempre dal *Washington Post* su *Game of Thrones*, una serie televisiva in onda dal 2011 sul canale HBO tratta dalla saga *A Song of Ice and Fire* di George R.R. Martin.

La serie di ambientazione fantasy racconta la storia di diverse casate nobiliari che in un mondo di cavalieri, dame e magia si contendono il trono reale.
In questo contesto le morti di protagonisti, comparse o interi eserciti giocano un ruolo fondamentale. I giornalisti del *Washington Post* hanno così deciso di raccoglierle tutte con un'infografica basata sul modello Interactive Slideshow. Per ogni stagione della serie è possibile accedere ad una sezione che riepiloga tutti i personaggi e addirittura gli animali spirati nei diversi episodi. I defunti sono poi ordinati in ordine di importanza, dai più noti a quelli che hanno meritato solo pochi fotogrammi. Portando il cursore su ognuno di loro è poi possibile accedere anche ad una finestra informativa che ne riassume la storia.

Illustrazione 60: Overview, selezionado una stagione della serie è possibile accedere alle informazioni sulle morti.

61 https://www.washingtonpost.com/graphics/entertainment/game-of-thrones/#season-one

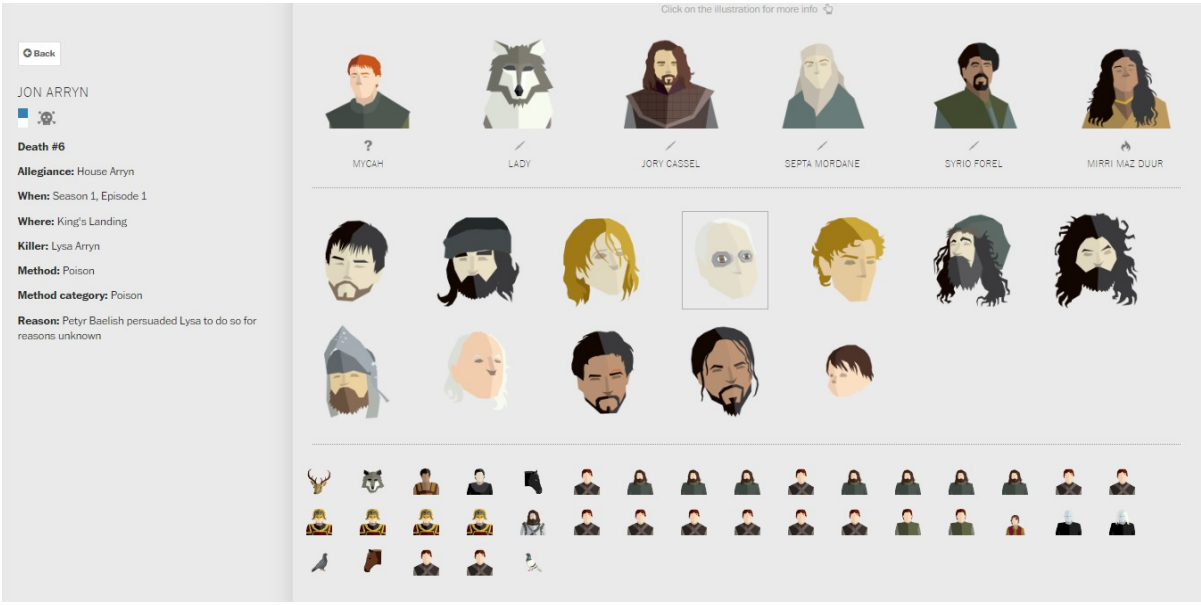


Illustrazione 61: Season 1, dai personaggi principali si passa poi a quelli secondari

Presentazioni su Mappe Concettuali, Prezi

Uno degli strumenti che più si accostano a questo tipo di modello è Prezi[62]. Si tratta di un sito che permette di realizzare delle presentazioni basate sul concetto di mappa tematica. Quando si inizia a costruire uno slideshow con Prezi la prima operazione da fare infatti è creare una mappa o una figura che si vuole esplorare. Successivamente si potranno comporre delle slide posizionate in punti diversi che permetteranno all'utente di muoversi all'interno di questo spazio.

Prezi presenta però anche due grossi limiti. Se infatti il lettore ha la possibilità di muoversi avanti e indietro nelle slide non può però scegliere di esplorare una sezione precisa della mappa creata ma deve seguire il percorso stabilito. In secondo luogo le slide create con questo strumento non sono interattive ma statiche.

Nella pagina successiva sono riportate la mappa generale e una slide di un progetto dal titolo *Animais da Amazonia*[63]. Questo lavoro è valso al suo creatore Guilherme Criscuolo un Prezi Award nel 2014, un premio assegnato ogni anno dagli sviluppatori di questo strumento agli utenti che hanno realizzato le visualizzazioni migliori. La categoria in ha trionfato era intitolata “Best Educational”.

62 https://prezi.com/

63 http://blog.prezi.com/latest/2015/1/5/the-best-prezis-of-2014?utm_source=Facebook&utm_medium=Social&utm_content=Blog&utm_campaign=PreziAwards

Illustrazione 62: Overview

Illustrazione 63: Slide 1

La guida dell'autore e la libertà di navigazione. Martini Glass

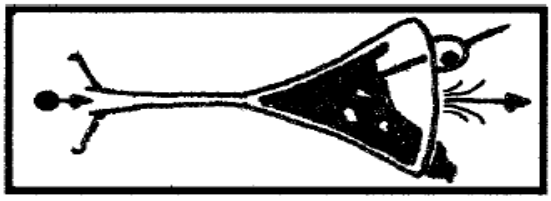


Illustrazione 64: Modello Martini Glass

L'ultimo modello di data visualization teorizzato da Shegel e Heer è quello più utilizzato e forse anche più complesso.
Per rendere graficamente l'idea di questo tipo di dataviz i due ricercatori hanno usato un bicchiere di Martini rovesciato, così da definire due fasi ben distinte.

Nella prima il lettore viene accompagnato dall'autore alla scoperta del dataset. Gli vengono indicati i punti principali su cui focalizzare l'attenzione e gli viene spiegato come sono strutturate le informazioni che ha davanti. Questo passaggio può avvenire attraverso un testo, un video di presentazione o direttamente dei grafici.

Nella seconda fase invece è il lettore, forte delle competenze acquisite grazie al contributo dell'autore, a poter navigare liberamente nei dati messi a disposizione.
Ecco spiegata la correlazione con il bicchieri di Martini rovesciato. La prima fase è rappresentata dallo stelo del calice, un percorso definito in cui l'autore accompagna il lettore mentre il secondo passaggio è la parte più ampia, in cui i lettore è libero di esplorare.

Questo modello è lo stesso che è stato seguito per il caso studio *Terra Malata*, esposto in questo elaborato a pagina 136.

Liberi di analizzare i dati, Tableau Public

A differenza degli strumenti presentati nelle pagine precedenti, Tableu Public[64] è un software che può essere scaricato gratuitamente e utilizzato anche offline. Il suo miglior pregio è la possibilità di gestire ampi dataset svolgendo sia processi di analisi che di visualizzazione.
I dati possono essere rappresentati attraverso mappe o grafici e allo stesso tempo elaborati, filtrati e aggregati. Grafici diversi sono assemblati insieme attraverso la funzione *Dashboard* e *Dashboard* diverse sono posizionate in sequenza mediante la funzione *Story*. Molto importante è qui l'uso dei filtri, pannelli che permettono all'utente di selezionare le informazioni a cui è interessato.

È proprio grazie a questa opzione che è possibile replicare il modello del Martini Glass. Si può infatti cominciare la propria *Story* con una serie di grafici che inquadrino l'argomento trattato in un dataset, riportando magari le informazioni principali e definendo il contesto in cui si inseriscono i dati. Dopo aver fornito le coordinate necessarie ad orientarsi si può passare a visualizzazioni più complesse che lasciano al lettore maggiore libertà di analisi.

64 https://public.tableau.com/s/

Capitolo 3
Data storytelling. La collaborazione con il Crisp

1. Tre definizioni per passare dalla data visualization al data storytelling

Davanti alla classificazione degli ambiti di utilizzo della data visualization e degli elementi che compongono la sua grammatica emerge chiaramente un dato: in questa disciplina non esiste un'unica figura professionale.

I progetti mostrati fino a questo momento hanno infatti messo in luce che la visualizzazione dei dati è un campo ibrido nel quale rientrano competenze diverse. Fra gli autori dei casi analizzati nelle pagine precedenti compaiono infatti informatici, designer, analisti, giornalisti, medici, storici e critici letterari.

Sono professionisti che provengono da ambiti di studio molto vari, da quelli scientifici a quelli umanistici. È naturale quindi chiedersi quali competenze siano indispensabili creare un qualsiasi tipo di infografica.

A questa domanda è possibile rispondere con almeno due definizioni di data visualization.

La prima è quella più circoscritta ed è riferita esclusivamente alla composizione. Questa risposta prevede che i dati di partenza siano già pronti, raccolti e ordinati in un dataset al quale si può accedere senza problemi. La creazione di una dataviz viene qui intesa esclusivamente sul piano del design e quindi la figura professionale più qualificata per lavorare in questo ambito sarebbe quella del designer. Per poter trasformare dei dati in figure è necessario possedere conoscenze grafiche, che si tratti di saper utilizzare pennini e righelli o software dedicati.

La seconda risposta invece cerca di allargare il campo. Per arrivare ad un'infografica infatti non è necessario solo sapere come rappresentare i dati ma anche analizzarli e raccoglierli. Per data visualization a questo punto si potrebbe intendere tutto il processo che porta alla realizzazione di una dataviz ossia: raccolta dei dati, analisi e visualizzazione.
A questo punto però non è più appannaggio di soli designer ma si presenta la necessità di coinvolgere altre figure. Idealmente potremmo quindi pensare ad un informatico che raccolga i dati ed li ordini all'interno di un database per poi passarli ad uno statistico che li analizzi.

In questo processo la data visualization si presenterebbe due volte. Una prima volta assolverebbe quella funzione di *Visualizzare per capire* analizzata nel primo capitolo, diventando uno strumento per riuscire a comprendere i dati da analizzare. La seconda volta invece entrerebbe in gioco dal punto di vista della comunicazione e servirebbe per divulgare le informazioni raccolte.

Da questa definizione è però interessante sviluppare un'altra domanda.
Se l'obiettivo finale del processo di data visualization è quello di trasmettere delle informazioni che nascono dai dati, i grafici sono sempre il modo migliore per farlo?
È possibile infatti che non tutte le visualizzazioni siano sufficienti a questo scopo o addirittura che non siano neanche necessarie. A volte i grafici devono essere collegati e spiegati da parti di testo e altre volte il risultato più importante di un'analisi su un dataset è semplicemente un numero.

A questo punto è possibile pensare ad un processo di raccolta, analisi e comunicazione dei dati in cui la data visualization sia solo un passaggio e non un punto di arrivo obbligatorio.
È proprio qui che si inserisce la definizione di data storytelling, il racconto delle storie che nascono dai dati. Un racconto che prevede necessariamente la combinazione di diversi linguaggi, sopratutto testo e grafici.

Il data sorytelling non si presenta solo nel giornalismo, come nel caso degli Afghan War Logs pubblicati sul *The Guardian*, ma potrebbe essere utilizzato anche in un contesto scientifico, diventando un modo diverso per intendere la comunicazione dei dati. Alle competenze necessarie definite finora se ne andrebbe quindi ad aggiungere un'altra: la capacità di raccontare. Ed è proprio in questo senso che una figura con una formazione umanistica potrebbe mettere in gioco le sue capacità.

2. Prove di data storytelling. Il Crisp

Il Crisp, Centro di Ricerca Interuniversitario per i Servizi di Pubblica Utilità alla Persona è una rete accademica interdisciplinare che ha lo scopo di condurre ricerche nel mondo dei servizi.
È stato fondato nel 1997 e ha sede nell'Università degli Studi di Milano-Bicocca.

La squadra di ricercatori è formata di circa 20 persone provenienti da ambiti differenti, dall'informatica alla statistica fino al design. Il direttore è il professor Mario Mezzanzanica che si occupa di Sistemi di Elaborazione delle Informazioni.

I progetti di questo centro riguardano ambiti differenti, dalla scuola alla sanità fino al mondo del lavoro. Ad accomunare gli studi svolti in questi contesti c'è però un unico fattore: l'utilizzo dei dati.
Il Crisp si è infatti specializzato nella raccolta, analisi e visualizzazione delle informazioni contenute nei dati.

Uno dei progetti seguiti da questo centro di ricerca è *Il Quadrante del lavoro*[65], un portale creato in collaborazione con Regione Lombardia in cui vengono raccolti i dati relativi al mondo del lavoro nel territorio regionale

Grazie all'interessamento del professor Mario Mezzanzanica e all'aiuto del project manager del Crisp Matteo Fontana nel febbraio 2016 è cominciata una breve collaborazione volta ad applicare le tecniche di data visualization in un contesto operativo. In questo periodo di lavoro è stato quindi possibile non solo applicare le nozioni teoriche presentate nei primi due capitoli di questo elaborato ma anche approfondire e sviluppare meglio il concetto di data storytelling.

65 http://daslombardia.crisp.unimib.it/pentaho/content/pentaho-cdf-dd/Render?solution=dasLombardia&path=&file=metroHomepage.wcdf

3. Il *Quadrante del lavoro*. Analisi del sito e dei livelli di comunicazione

L'obiettivo del *Quadrante del lavoro* è offrire una visione d'insieme del mercato del lavoro in una delle regioni più produttive d'Europa.
Il target di riferimento non è un pubblico generalista ma un'utenza formata da operatori del settore come analisti di mercato, economisti, imprenditori, giornalisti o aziende.

Fonti

I dati raccolti nel *Quadrante del lavoro* provengono principalmente da cinque fonti. Ognuna fornice una prospettiva diversa e complementare alle altre.

Comunicazioni Obbligatorie

Le Comunicazioni Obbligatorie[66] sono informazioni che tutti i datori di lavoro sono tenuti a trasmettere al Ministero del Lavoro in caso di avviamento, proroga, trasformazione e cessazione di un contratto.
Da questo tipo di dati si raccolgono quindi diverse informazioni che si possono distinguere in generali, modalità di lavoro e forme contrattuali.

Generali

- **Avviamento**: instaurazione di un rapporto di lavoro.
- **Cessazione**: termine di un rapporto di lavoro.
- **Saldo Avviamenti e Cessazioni**: differenza tra avviamenti e cessazioni.
- **Variazione Tendenziale**: variazione in termini percentuali rispetto allo stesso periodo dell'anno precedente
- **Trasformazione**: trasformazione legale del contratto che lega lavoratore e datore di lavoro. Questo cambiamento riguarda il passaggio da una tipologia di contratto ad un altra.
- **Skill Level**: livello di istruzione formale necessario allo svolgimento della professione.

Modalità di Lavoro

- **Tempo parziale orizzontale**: il lavoro si distribuisce su tutti i giorni ma ad orario ridotto
- **Tempo parziale verticale**: il lavoro è a tempo pieno ma solo per alcuni giorni della settimana, del mese o dell'anno.
- **Tempo parziale misto**: si tratta della combinazione fra le due modalità di lavoro.

66 https://www.co.lavoro.gov.it/

Forme Contrattuali

- **Apprendistato**: rapporto di lavoro con cui un datore si impegna a formare un apprendista fino a farlo diventare un lavoratore qualificato. L'apprendistato è un tipo di contratto considerato permanente. Al termine del periodo di formazione dovrebbe infatti trasformarsi in un tempo indeterminato.

- **Contratto di Somministrazione**: accordo commerciale fra due soggetti, uno denominato *utilizzatore* e l'altro *somministratore*. Il somministratore assume i lavoratori e li mette a disposizione per esigenze professionali di tipo carattere periodico o limitato nel tempo. È l'evoluzione del contratto di lavoro interinale.

- **Lavoro a progetto**: rapporto di collaborazione coordinata continuativa, si tratta del vecchio co.co.co.

- **Tempo determinato**: contratto di lavoro in cui la data di scadenza viene stabilita nel momento della siglatura.

- **Tempo Indeterminato**: contratto di lavoro che dopo un periodo di prova si trasforma in una assunzione senza scadenza.

Istat

I dati dell'Istat[67] sono la principale fonte di informazione statistica per il mercato del lavoro italiano. A differenza delle comunicazioni obbligatorie, si tratta di dati campionari. Vengono quindi selezionati dei campioni per poi proiettare sulla popolazione nazionale i risultati ottenuti.

Le informazioni rilasciate dall'Istat sono molto utilizzate nel giornalismo e per questo i suoi indicatori sono ormai ben conosciuti anche da un pubblico non specializzato.

- **Forza Lavoro**: persone occupate dai 15 anni in su e disoccupati dai 15 ai 74 anni.

- **Tasso di Attività**: rapporto tra le persone appartenenti alle forze di lavoro e la corrispondente popolazione di riferimento.

- **Tasso di Occupazione**: rapporto tra gli occupati e la corrispondente popolazione di riferimento.

- **Tasso di Disoccupazione**: rapporto tra le persone in cerca di occupazione e la corrispondente forza lavoro.

- **Posizione nella professione**: livello acquisito dal lavoratore nella professione esercitata. Le definizioni cambiano tra lavoratori dipendenti e indipendenti**.**

67 http://www.istat.it/

Eurostat

I dati Eurostat[68] permettono di riportare gli indicatori dell'Istat sullo senario europeo. Si concentrano soprattutto sui *Quattro Motori per l'Europa*, le quattro regioni considerate economicamente trainati per il vecchio continente: la Lombardia, il Baden-Wurttemberg, la Catalogna e il Rodano-Alpi.

Inps

Dall'Istituto Nazionale di Previdenza Sociale[69] vengono presi i dati relativi agli ammortizzatori sociali, le misure di sostegno al reddito riservate a determinate categorie di lavoratori. Fra le più note ci sono qui la Cassa Integrazione Guadagni e la Cassa Integrazione Straordinaria.

Unioncamere Movimprese

Unioncamere[70] è l'Unione Italiana delle Camere di Commercio, Industria, Artigianato e Agricoltura. Ogni tre mesi pubblica il dossier Movimprese, un'analisi statistica sulla nata-mortalità delle imprese italiane. Questa ricerca viene commissionata a InfoCamere[71], la struttura per la gestione del patrimonio informativo e dei servizi del sistema camerale.

I due indicatori più importanti che questa fonte riesce a fornire sono:

- **Natalità**: il tasso di natalità viene calcolato come rapporto tra le imprese iscritte ad Unioncamere e le imprese attive. Possono esserci infatti imprese registrate ma inattive.

- **Mortalità**: il tasso di mortalità viene calcolato come rapporto tra le imprese cessate e quelle attive.

Oltre a queste due indicatori da qui si possono estrarre anche tutti i dati relativi alle forme giuridiche con cui le imprese si costituiscono, dalle società di capitale a quelle di persone fino alle ditte individuali.

Architettura del sito

La home page si presenta come una composizione di quadri.
Su quelli statici scorrono delle informazioni mentre su quelli navigabili compare il titolo dell'argomento trattato nella sezione del sito a cui rimandano.

Esistono cinque sezioni fisse che vengono aggiornate ogni tre mesi:

- **Dinamiche Lavorative**: l'andamento del mercato del lavoro viene descritto attraverso i dati delle Comunicazioni Obbligatorie.

- **Indicatori Occupazionali**: qui sono rappresentati principalmente i dati provenienti dall'Istat e Eurostat.

68 http://ec.europa.eu/eurostat
69 https://www.inps.it/portale/default.aspx
70 http://www.unioncamere.gov.it/
71 http://www.infocamere.it/

- **Crisi Aziendali**: le difficoltà registrate dalle imprese sono raccontate attraverso i dati dell'Inps.

- **Dinamiche delle Imprese**: ad essere protagonisti sono i numeri forniti da Movimprese.

- **Indicatori di Sintesi**: è la sezione che offre il panorama più ampio sul mondo del lavoro. Sono riportati gli indicatori più rilevanti di tutte le altre sezioni.

Nella schermata si possono poi trovare altre quattro sezioni che trattano argomenti specifici, raccogliendo dati da diverse fonti.

- **Giovani e Lavoro**

- **Donne e Lavoro**

- **Marcroregione Alpina**

- **Expo Milano 2015**: quest'ultima è una sezione temporanea dove sono stati analizzati i dati per capire l'impatto dell'esposizione universale sul mondo del lavoro.

Illustrazione 65: Home Page Quadrante del Lavoro

Sulla home page del sito sono poi presenti altre sei sezioni che non contengono dati.

- **Commenti di Sintesi**: vengono pubblicati dei commenti trimestrali per ogni indicatore presente nel sito.

- **Ultime Notizie**: si possono trovare notizie riguardanti e dati caricati sul portale o su altri siti inerenti al mondo del lavoro.

- **Link Esterni**: sono collegamenti alle fonti utilizzate per aggregare le informazioni.

- **Documentazione**: contiene il glossario degli indicatori utilizzati e una spiegazione della metodologia adottata per la realizzazione del sito.

- **Fonti e Aggiornamento**: elenco delle fonti utilizzate. Per ogni fonte è specificata anche la data dell'ultimo aggiornamento.

- **Contatore**: sono visualizzati gli accessi totali al sito e quelli totalizzati nell'ultimo mese.

I tre livelli di comunicazione

Le informazioni che si possono recuperare da questo sito sul mondo del lavoro possono essere disposte su tre livelli, da quello più generico a quello più approfondito.

A. Commenti di Sintesi

Il primo livello è costituito dai Commenti di Sintesi. Si tratta di documenti pdf pubblicati regolarmente che sintetizzano le informazioni principali in brevi frasi. L'obiettivo qui non è riportare delle riflessioni ma fornire un dato accurato per poi permettere all'utente di analizzarlo.

I Commenti di Sintesi riflettono le sezioni del sito in cui vengono riportati gli indicatori e si dividono quindi in:

- **Commento Generale**

- **Dinamiche Lavorative**

- **Indicatori Occupazionali**

- **Crisi Aziendali**

- **Dinamiche delle Imprese**

In tutti è commenti è presente anche una breve sezione intitolata *Sommario* in cui sono raccolti i dati più importanti.

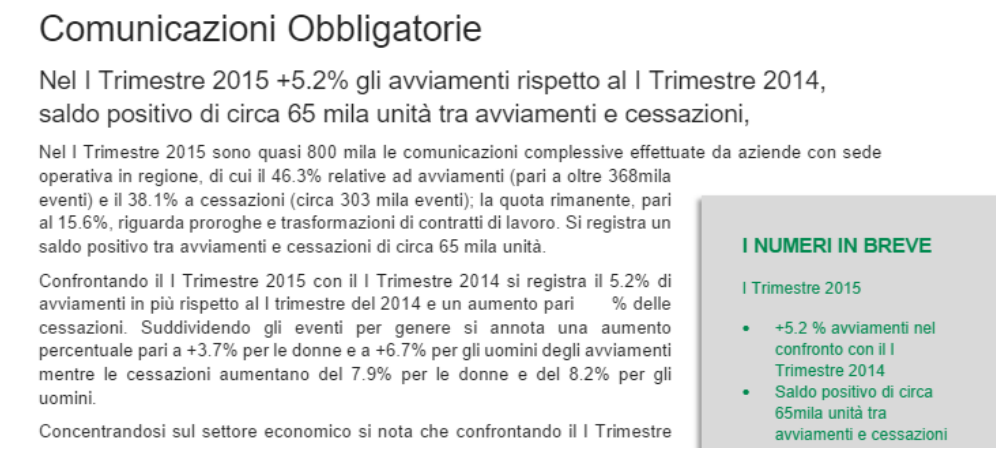


Illustrazione 66: Commento di Sintesi Dinamiche Lavorative I Trimestre 2015

B. Visualizzazioni

Il secondo livello di comunicazione si base invece sulla data visualization.

Accedendo infatti nelle sezioni del sito i dati possono essere esplorati attraverso una serie di grafici, soprattutto *line graphs* e *bar graphs*. Tramite un sistema di filtri è possibile selezionare non solo i dati relativi all'indicatore che si vuole approfondire ma anche scegliere il periodo di tempo o la fascia d'età in cui estrarre i dati.

Per ogni sezione c'è sempre una prima schermata in cui vengono presentati tutti gli indicatori e cliccando sul bottone *Dettagli* si accede ad un'altra pagina in cui vengono approfonditi i dati di ogni indicatore.

Le visualizzazioni che si creano in questo modo non sono statiche. Spostandosi sulla legenda si possono selezionare ulteriormente i valori da visualizzare così che il dato possa essere rappresentato con più chiarezza.

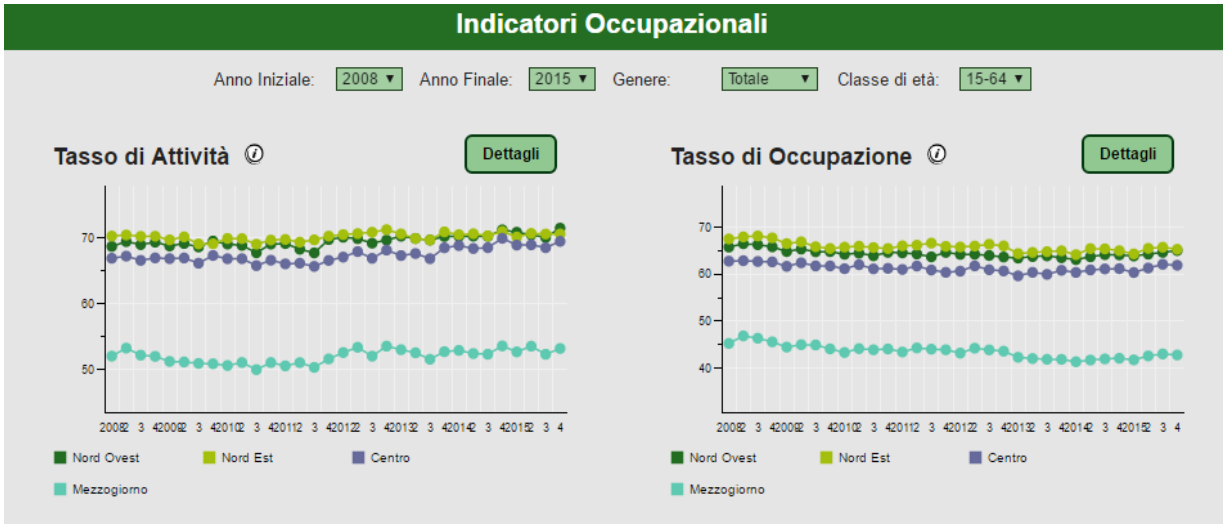


Illustrazione 67: Indicatori Occupazionali, pannello principale.

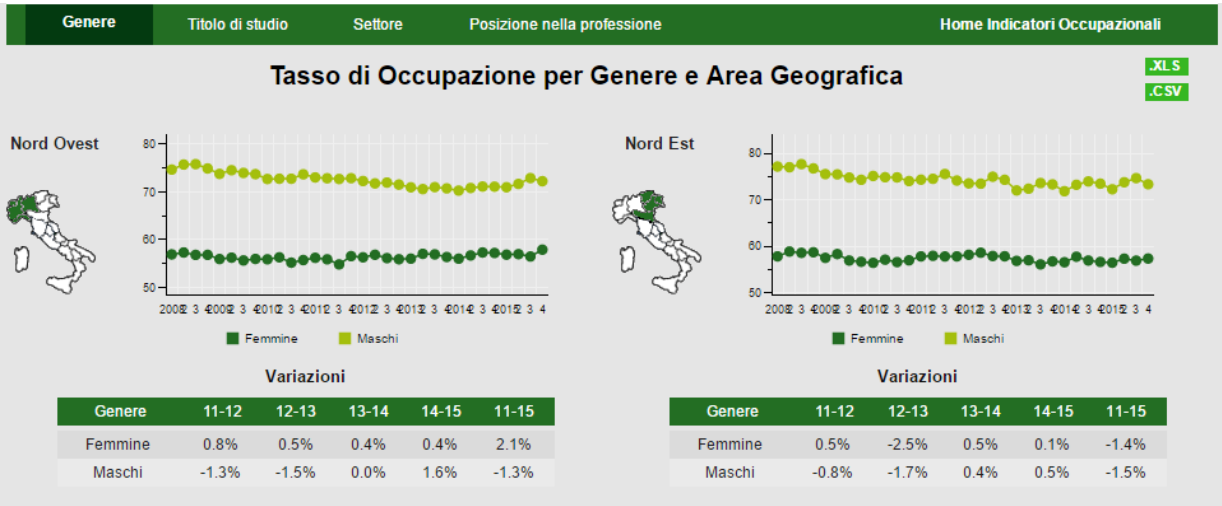


Illustrazione 68: Indicatori Occupazionali, pannello di approfondimento Tasso di Occupazione per Genere e Area Geografica.

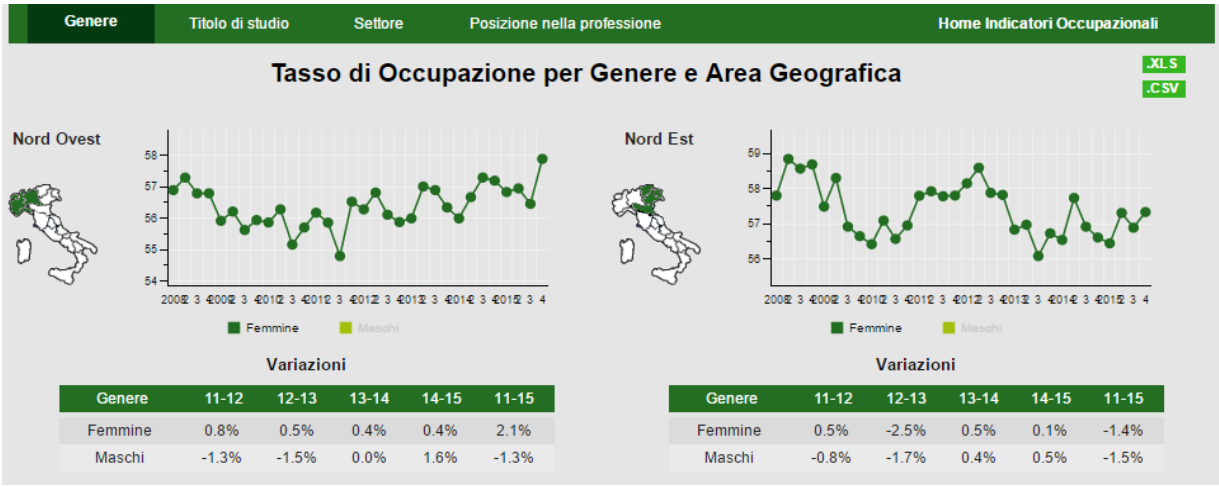


Illustrazione 69: Indicatori Occupazionali, pannello di approfondimento Tasso di Occupazione per Genere e Area Geografica, selezione valore 'Femmine'

C. Tabelle

Il terzo livello di comunicazione è quello più approfondito e quindi quello in cui l'utente è più libero di cercare i dati che gli interessano.

Nelle sottosezioni a cui si accede dopo le pagine principali dei diversi indicatori è possibile scaricare un file csv o xls in cui sono contenuti tutti i dati già presentati nelle visualizzazioni. In questo modo si ha in possesso un formato facilmente utilizzabile per proporre nuove visualizzazioni e nuove analisi.

L'unico elemento di cui bisogna tenere conto è che i dati presenti nelle tabelle, così come quelli delle sottosezioni, vengono selezionati in basi ai filtri scelti nella schermata iniziale di ogni indicatore.

4. Dai commenti alle infografiche. I Report Trimestrali

Il primo progetto nato dalla collaborazione con il Crisp ha coinvolto i Commenti di Sintesi.

Lo scopo di questa sezione è di creare un primo livello di divulgazione, dove l'utente possa cominciare a capire attraverso dei dati aggregati l'andamento del mercato del lavoro. I Commenti di Sintesi dovrebbero essere pubblicati quindi ogni trimestre per restare sempre in aggiornati rispetto alle fonti.
Questo modo di comunicare i dati presenta però almeno due problematiche.

- **Forma**. Pur essendo pochi i dati da riportare, utilizzare uno scritto non sembra essere il modo migliore perché il risultato è un un lungo elenco di dati accostati senza causalità e senza correlazione. Il rischio quindi è che l'informazione non risulti chiara all'utente che vuole analizzare il documento.

- **Pubblico**. Il target di riferimento del *Quadrante del lavoro* resta sempre un pubblico specializzato, abituato ad occuparsi di questo contesto. Dato però che i Commenti di Sintesi costituiscono il primo livello di comunicazione, si potrebbe trovare una soluzione per renderli più comprensibili anche ad un'utenza generalista.

L'obiettivo di questo progetto era dunque trovare un modo per divulgare le stesse informazioni attraverso un'infografica.

I numeri dei download

Prima pensare quali tecniche di data visualization fossero adatte a questo scopo sono stati raccolti i dati dei download avvenuti da novembre 2013 a gennaio 2016.

I download totali dei Commenti di Sintesi sono 2060 e la quota maggiore è quella degli Indicatori Occupazionali che hanno registrato 706 contatti. Spostandosi sui dati cronologici si può vedere come ci siano dei picchi mensili molto alti, dovuti a corsi di formazione interni a Regione di Lombardia in cui i ricercatori del Crisp hanno mostrato a chi si occupa di lavoro negli uffici regionali come utilizzare questo sito. Togliendo questi dati straordinari, normalmente il numero di download si attesta da un minimo di 30 ad un massimo di 100 al mese.

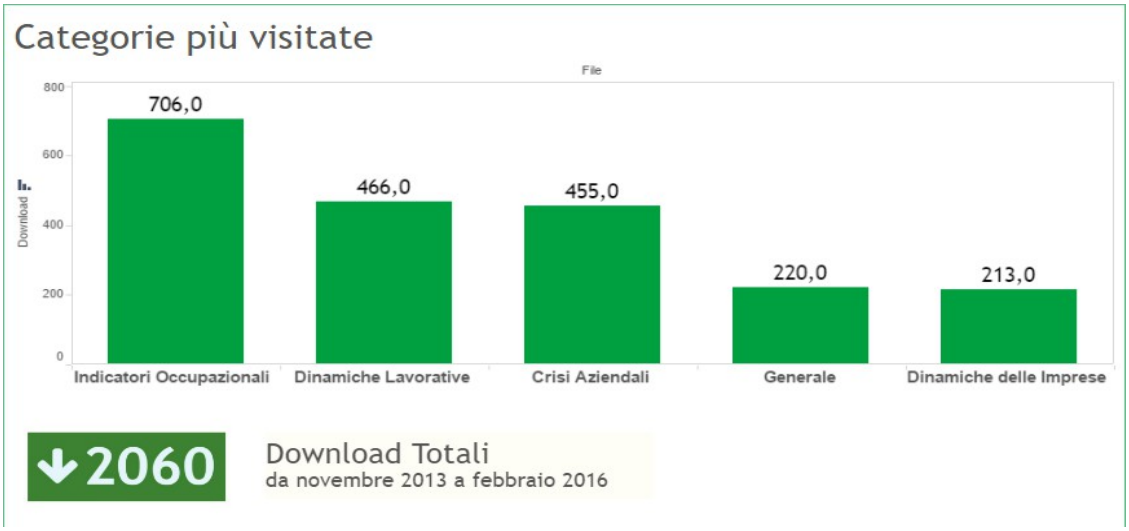


Illustrazione 70: Download Totali Commenti di Sintesi

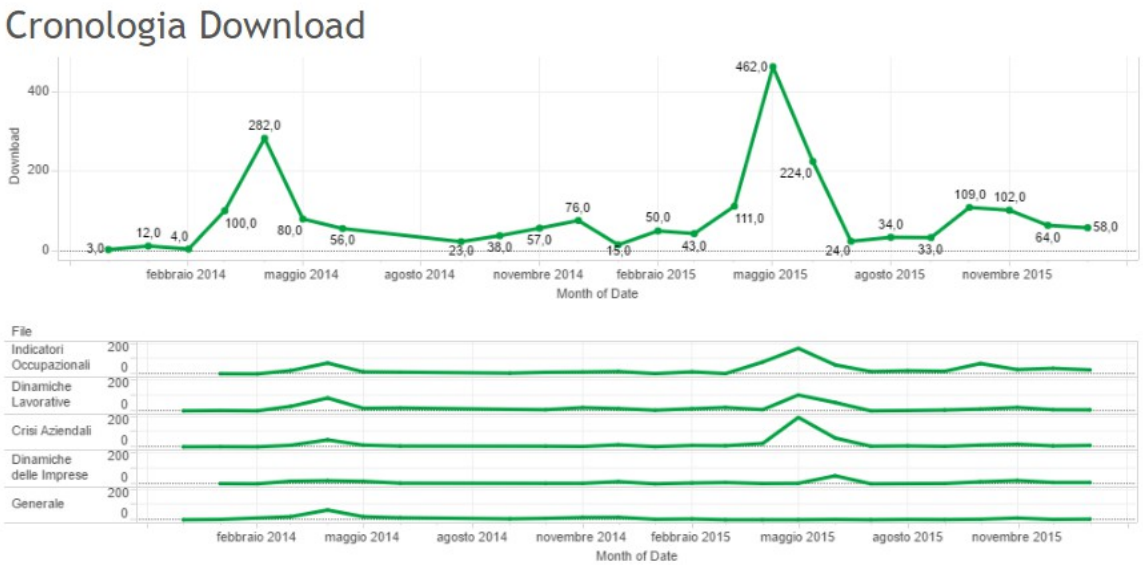


Illustrazione 71: Cronologia Download Commenti di Sintesi

Costruire un'infografica

La sezione scelta come prototipo per questa operazione è stata quella che raccoglie gli Indicatori Occupazionali, dato che si tratta di quella con più download all'attivo nella storia del sito.
Il periodo temporale analizzato è stato l'ultimo trimestre dell'anno 2015.

Una volta definito l'ambito di analisi è cominciata la fase di raccolta dei dati, quelli più interessanti sono stati estratti dall'Istat e disposti all'interno di una tabella.

	A	B	C
1		**Tasso di Occupazione**	**Tasso di Disoccupazione**
2	**Italia**	56.06%	11.9%
3	**Nord-Ovest**	65.1%	8.8%
4	**Lombardia**	65.6%	8.4%
5			
6	**1° 2014**	64.2%	8.8%
7	**2° 2014**	65.0%	7.9%
8	**3° 2014**	65.0%	7.5%
9	**4° 2014**	65.1%	8.5%
10	**1° 2015**	64.6%	8.6%
11	**2° 2015**	65.1%	7.7%
12	**3° 2015**	65.3%	6.7%
13	**4° 2015**	65.6%	8.4%
14			
15	**Uomini**	73.0%	7.6%
16	**Donne**	58.1%	9.4%
17			
18	**G 2014**	37.9%	20.3%
19	**G 2015**	37.0%	20.8%
20			

Illustrazione 72: Tabella dei dati riportati nell'infografica

Struttura del dataset

Le informazioni raccolte sono quindi il risultato di una selezione. Come indicatori sono stati scelti infatti solo il Tasso di Occupazione e il Tasso di Disoccupazione e per entrambi sono state definite le variabili più importanti.

Il punto di partenza è il dato per aree geografiche che evidenzia le differenze della Lombardia rispetto al territorio nazionale e alle regioni del nord ovest.

Gli stessi dati del trimestre sono disposti quindi su un piano cronologico e confrontati con sia con i trimestri del 2015 che con quelli dell'anno precedente.

Il terzo passaggio affronta la differenza di occupazione o disoccupazione fra uomini e donne.

L'ultimo dato che viene presentato è poi differente rispetto agli altri. Le cifre sull'occupazione e sulla disoccupazione giovanile (15-29 anni) non vengono infatti forniti dall'Istat per ogni trimestre ma per ogni anno. In questo report sono stati quindi inseriti proprio perché si trattava del quarto trimestre e quindi erano a disposizione anche i dati annuali.

Strumento

Una volta definiti e strutturati i dati è cominciata la scelta dello strumento grafico per disporre le informazioni. La necessità di creare un file facilmente caricabile sul sito ha portato alla decisione di scegliere un software in grado di comporre un'infografica statistica che potesse essere salvata in formato pdf, png o jpeg.

Dopo alcuni tentativi con Infogr.am[72], Venngage[73] e Visme[74] la scelta è caduta su Piktochart[75] dato che si trattava del software combinava meglio la creazione di dataviz accurate con la possibilità di realizzare una veste grafica interessante.

Grafica

Il layout adottato è semplice per non compromettere la chiarezza delle informazioni trasmesse e il panel di colori riprendere la bandiera della Lomabrdia. Lo schema grafico proposto può quindi essere replicato, sia su altri indicatori che su altri trimestri.

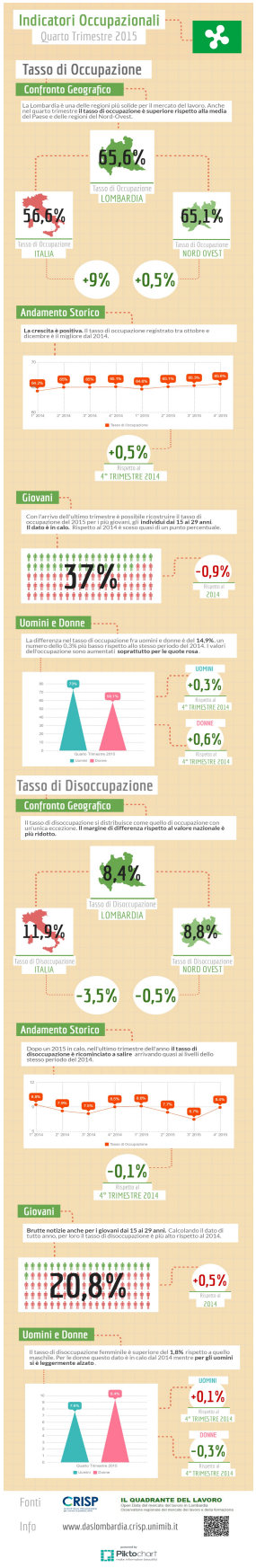
La struttura dei grafici segue l'ordine dei dati raccolti ed è identica sia per il tasso di occupazione che per quello di disoccupazione. Brevi commenti di testo a carattere esplicativo accompagnano ogni sezione e degli specchietti sottolineano le variazioni tendenziali rispetto lo stesso periodo del 2014.

In fondo all'infografica è stato poi aggiunto oltre alle fonti un QR Code che rimanda l'utente alla sezione del *Quadrante del lavoro* sugli Indicatori Occupazionali. La schermata che troverà dopo aver effettuato il collegamento avrà già i filtri temporali impostati per riprendere le informazione contenute all'interno del report.

72 https://infogr.am/
73 https://venngage.com/
74 http://www.visme.co/
75 https://piktochart.com/

Il risultato finale è un'immagine png dal peso di 1,58 MB e dalle dimensioni di 1200x8648 pixel.

Nelle pagine successive questa dataviz verrà riportata un due modi diversi: prima intera e poi divisa in blocchi.


Indicatori Occupazionali
Quarto Trimestre 2015
Tasso di Occupazione
Confronto Geografico
La Lombardia è una delle regioni più solide per il mercato del lavoro. Anche nel quarto trimestre il tasso di occupazione è superiore rispetto alla media del Paese e delle regioni del Nord Ovest.
65,6%
Tasso di Occupazione
LOMBARDIA
56,6%
Tasso di Occupazione
ITALIA
65,1%
Tasso di Occupazione
NORD OVEST
+9%
+0,5%
Andamento Storico
La crescita è positiva. Il tasso di occupazione registrato tra ottobre e dicembre è il migliore dal 2014.
+0,5%
Rispetto al
4° TRIMESTRE 2014
Giovani
Con l'arrivo dell'ultimo trimestre è possibile ricostruire il tasso di occupazione del 2015 per i più giovani, gli individui dai 15 ai 29 anni. Il dato è in calo. Rispetto al 2014 è sceso quasi di un punto percentuale.
37%
-0,9%
Rispetto al
2014
Uomini e Donne
La differenza nel tasso di occupazione fra uomini e donne è del 14,9%, un numero dello 0,3% più basso rispetto allo stesso periodo del 2014. I valori dell'occupazione sono aumentati soprattutto per le quote rosa.
UOMINI
+0,3%
Rispetto al
4° TRIMESTRE 2014
DONNE
+0,6%
Rispetto al
4° TRIMESTRE 2014
Tasso di Disoccupazione
Confronto Geografico
Il tasso di disoccupazione si distribuisce come quello di occupazione con un'unica eccezione. Il margine di differenza rispetto al valore nazionale è più ridotto.
8,4%
Tasso di Disoccupazione
LOMBARDIA
11,9%
Tasso di Disoccupazione
ITALIA
8,8%
Tasso di Disoccupazione
NORD OVEST
-3,5%
-0,5%
Andamento Storico
Dopo un 2015 in calo, nell'ultimo trimestre dell'anno il tasso di disoccupazione è ricominciato a salire arrivando quasi ai livelli dello stesso periodo del 2014.
-0,1%
Rispetto al
4° TRIMESTRE 2014
Giovani
Brutte notizie anche per i giovani dai 15 ai 29 anni. Calcolando il dato di tutto anno, per loro il tasso di disoccupazione è più alto rispetto al 2014.
20,8%
+0,5%
Rispetto al
2014
Uomini e Donne
Il tasso di disoccupazione femminile è superiore del 1,8% rispetto a quello maschile. Per le donne questo dato è in calo dal 2014 mentre per gli uomini si è leggermente alzato.
UOMINI
+0,1%
Rispetto al
4° TRIMESTRE 2014
DONNE
-0,3%
Rispetto al
4° TRIMESTRE 2014
Fonti
CRISP
IL QUADRANTE DEL LAVORO
Info
www.daslombardia.crisp.unimib.it
Powered by
Piktochart

Indicatori Occupazionali

Quarto Trimestre 2015

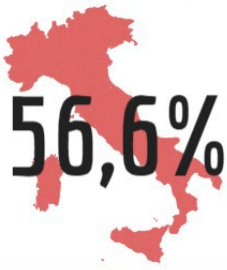
Tasso di Occupazione

Confronto Geografico

La Lombardia è una delle regioni più solide per il mercato del lavoro. Anche nel quarto trimestre **il tasso di occupazione è superiore rispetto alla media** del Paese e delle regioni del Nord-Ovest.

65,6%

Tasso di Occupazione
LOMBARDIA



56,6%

Tasso di Occupazione
ITALIA

65,1%

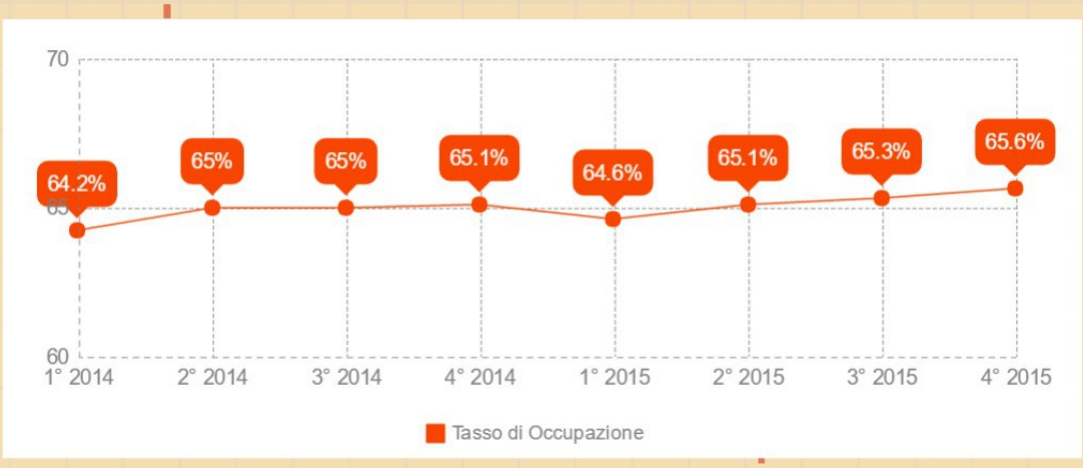
Tasso di Occupazione
NORD OVEST

+9%

+0,5%

Andamento Storico

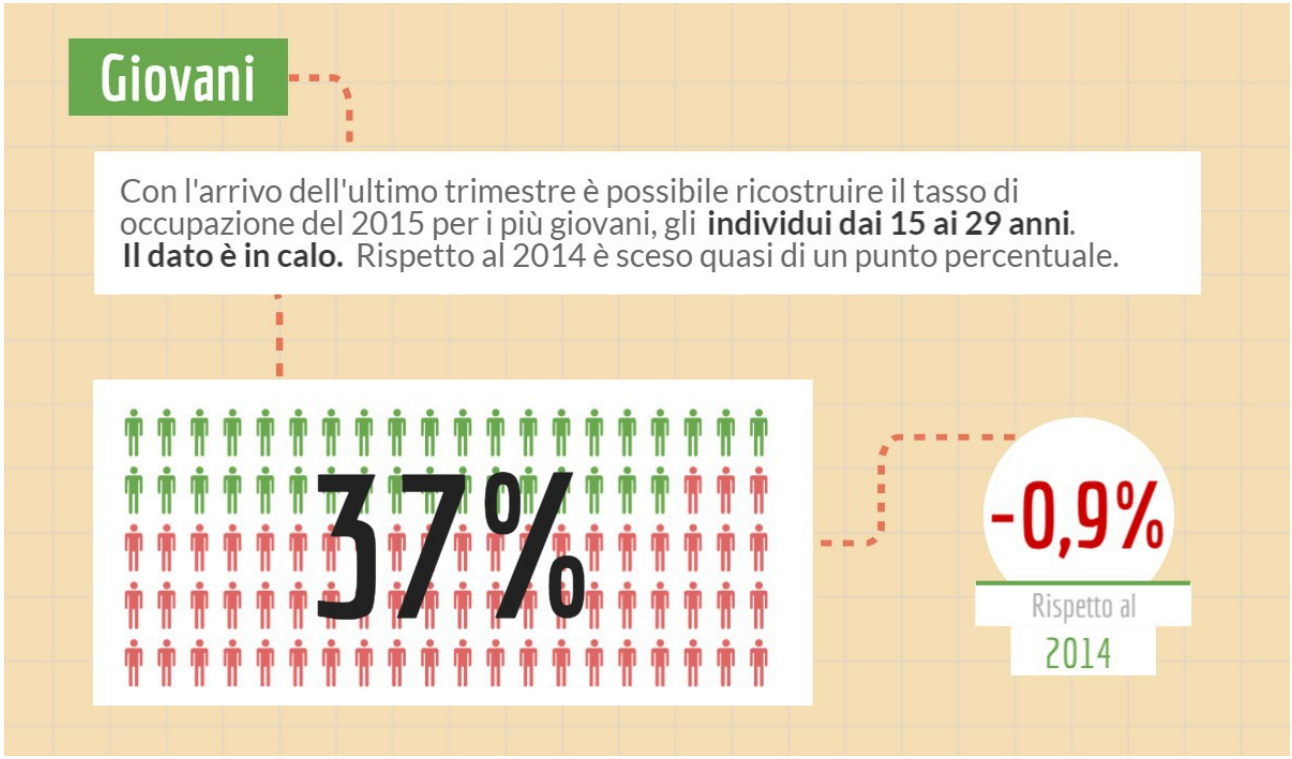
La crescita è positiva. Il tasso di occupazione registrato tra ottobre e dicembre è il migliore dal 2014.



+0,5%

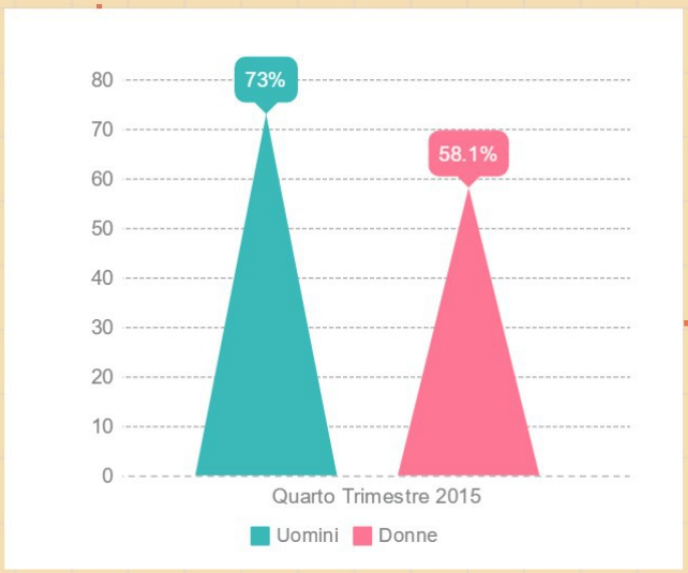
Rispetto al

4° TRIMESTRE 2014


Giovani
Con l'arrivo dell'ultimo trimestre è possibile ricostruire il tasso di occupazione del 2015 per i più giovani, gli individui dai 15 ai 29 anni.
Il dato è in calo. Rispetto al 2014 è sceso quasi di un punto percentuale.
37%
-0,9%
Rispetto al
2014

Uomini e Donne

La differenza nel tasso di occupazione fra uomini e donne è del **14,9%**, un numero dello 0,3% più basso rispetto allo stesso periodo del 2014. I valori dell'occupazione sono aumentati **soprattutto per le quote rosa**.



UOMINI

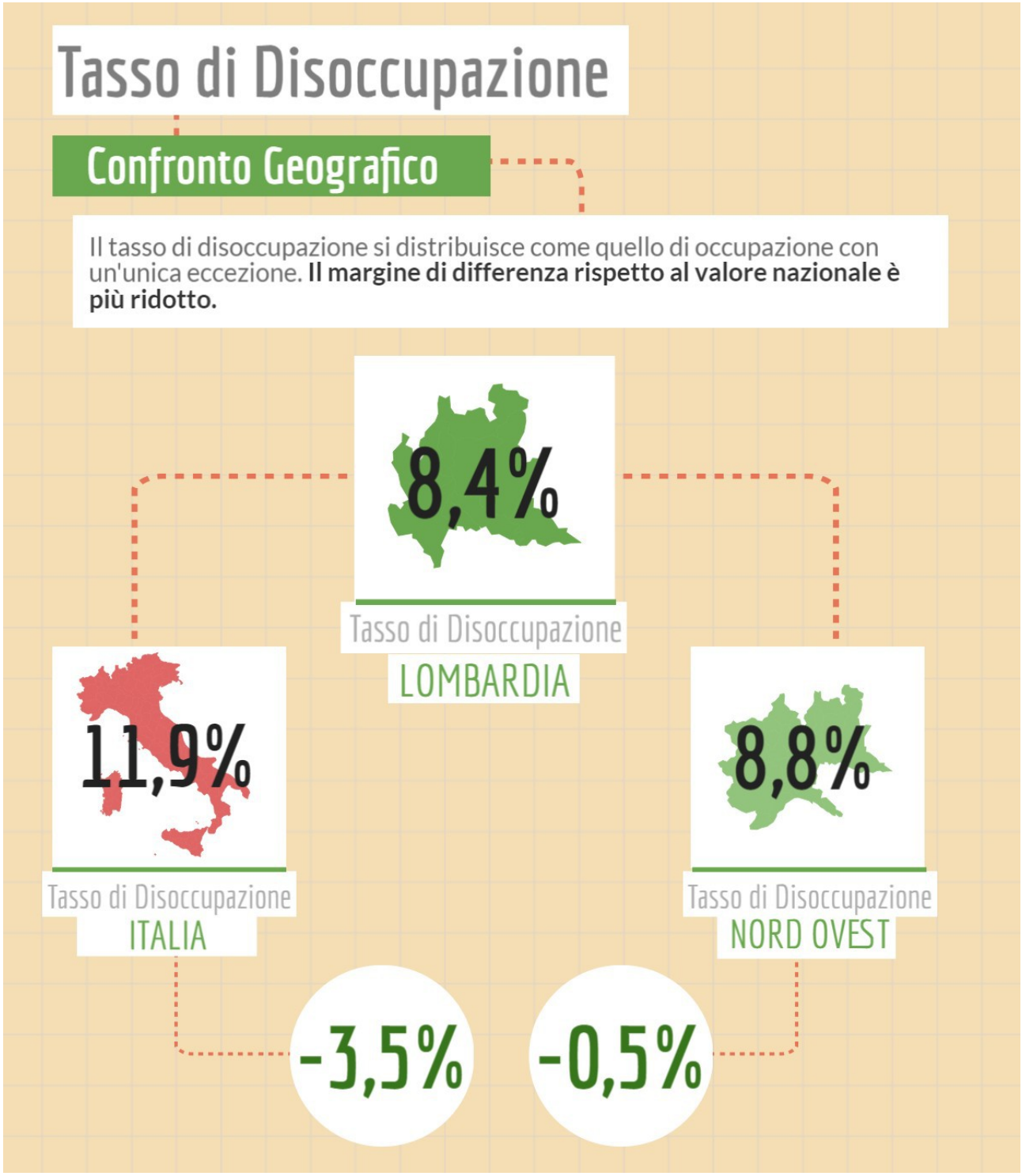
+0,3%

Rispetto al 4° TRIMESTRE 2014

DONNE

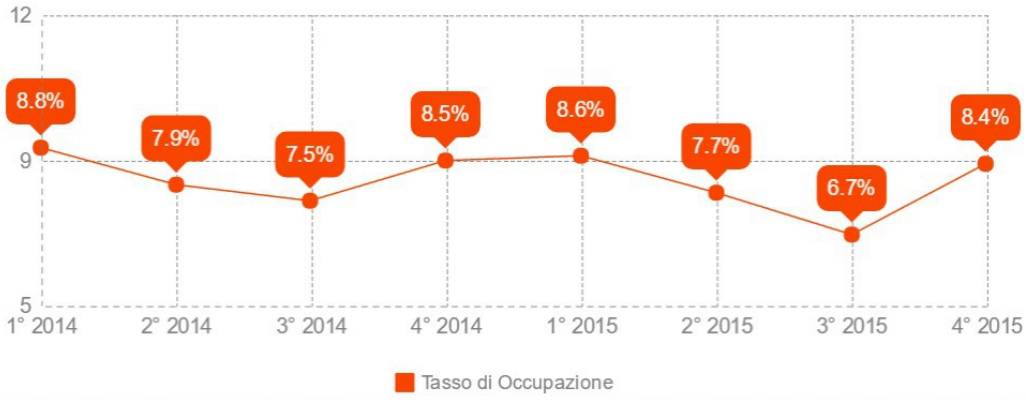
+0,6%

Rispetto al 4° TRIMESTRE 2014


Tasso di Disoccupazione
Confronto Geografico
Il tasso di disoccupazione si distribuisce come quello di occupazione con un'unica eccezione. Il margine di differenza rispetto al valore nazionale è più ridotto.
8,4%
Tasso di Disoccupazione
LOMBARDIA
11,9%
Tasso di Disoccupazione
ITALIA
8,8%
Tasso di Disoccupazione
NORD OVEST
-3,5%
-0,5%

Andamento Storico

Dopo un 2015 in calo, nell'ultimo trimestre dell'anno **il tasso di disoccupazione è ricominciato a salire** arrivando quasi ai livelli dello stesso periodo del 2014.



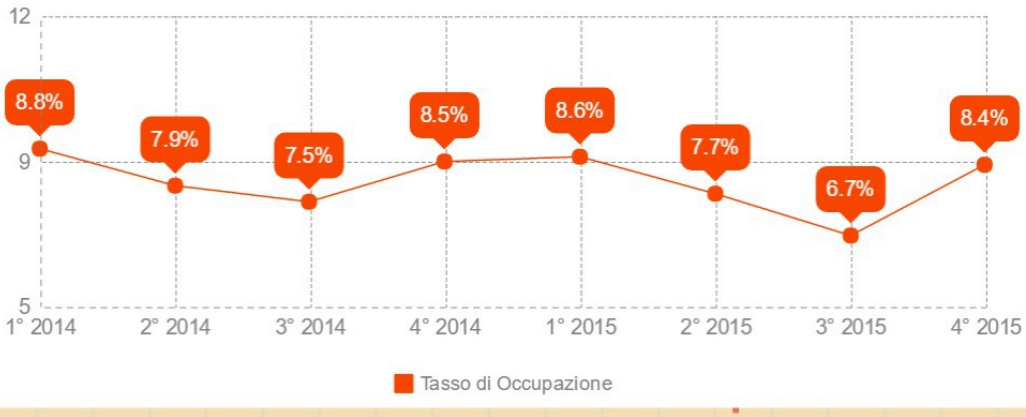
-0,1%

Rispetto al

4° TRIMESTRE 2014

Andamento Storico

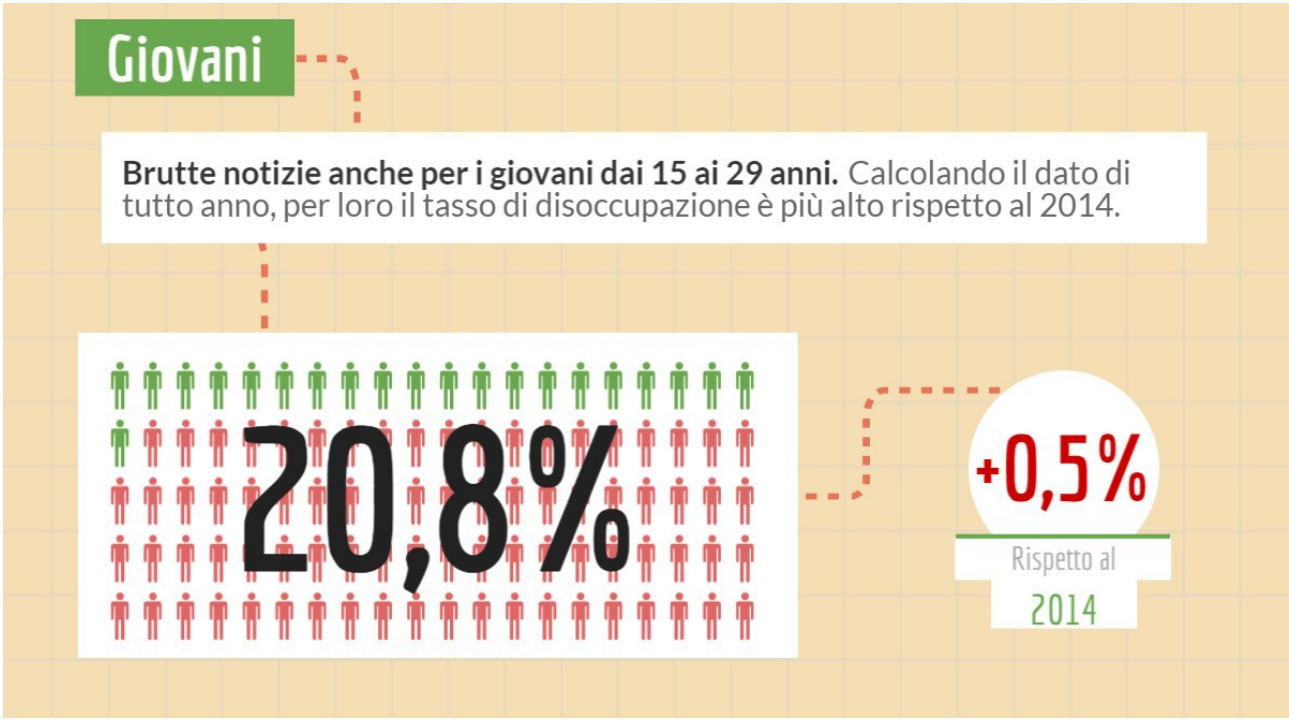
Dopo un 2015 in calo, nell'ultimo trimestre dell'anno **il tasso di disoccupazione è ricominciato a salire** arrivando quasi ai livelli dello stesso periodo del 2014.



-0,1%

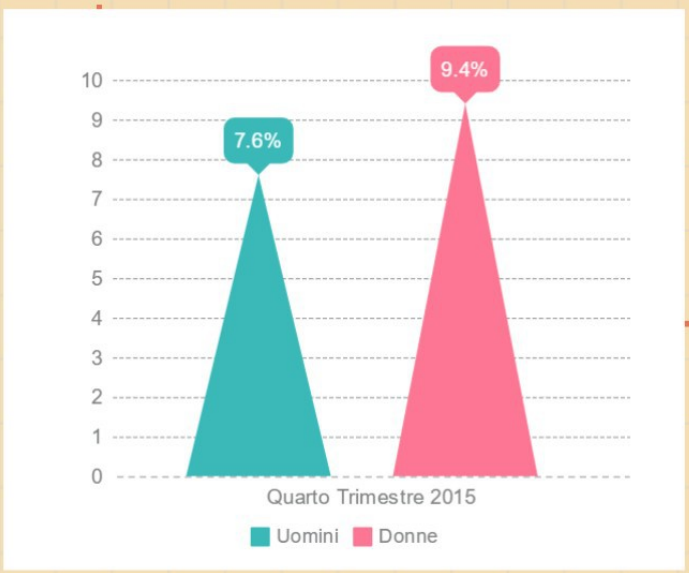
Rispetto al

4° TRIMESTRE 2014


Giovani
Brutte notizie anche per i giovani dai 15 ai 29 anni. Calcolando il dato di tutto anno, per loro il tasso di disoccupazione è più alto rispetto al 2014.
20,8%
+0,5%
Rispetto al
2014

Uomini e Donne

Il tasso di disoccupazione femminile è superiore del **1,8%** rispetto a quello maschile. Per le donne questo dato è in calo dal 2014 mentre **per gli uomini si è leggermente alzato**.



UOMINI

+0,1%

Rispetto al 4° TRIMESTRE 2014

DONNE

-0,3%

Rispetto al 4° TRIMESTRE 2014

Fonti

CRISP
centro di ricerca interuniversitario per i servizi di pubblica utilità

IL QUADRANTE DEL LAVORO
Open Data del mercato del lavoro in Lombardia
Osservatorio regionale del mercato del lavoro e della formazione

Info www.daslombardia.crisp.unimib.it

5. Una storia che nasce dai dati. La scossa del Jobs Act

La prima fase di questo progetto di collaborazione era così volta a creare una sorta di alternativa grafica ad una sezione che prima faceva uso esclusivamente del testo. È però nella seconda fase che è stato possibile applicare il data storytelling, un processo cioè che non mira esclusivamente alla visualizzazione dei dati ma che anzi utilizza queste tecniche prima per comprendere e poi per rappresentare le informazioni all'interno di un racconto.

L'idea alla base di questo nuovo progetto nasce da una delle sezioni già esistenti. Come spiegato nell'analisi del *Quadrante del lavoro*, accanto alle sezioni fisse aggiornate ogni trimestre sono presenti anche delle sezioni focalizzate su temi specifici. Alcuni riguardano aspetti del mondo del lavoro sempre attuali, come *Giovani e lavoro* o *Donne e lavoro* ma altri, come *Expo Milano 2015*, si concentrano su un evento specifico e limitato nel tempo.

Partendo da questo spunto è nata così l'idea di scegliere una tematica e svilupparla attraverso il data storytelling. Osservando come è cambiato di recente il mercato del lavoro, la decisione non è stata affatto difficile. In questo ambito il 2015 è stato segnato infatti dall'arrivo del Jobs Act, quella serie di provvedimenti volti a rendere più flessibili e più diffusi i contratti a tempo indeterminato.

Scelto l'argomento la sfida è stata quella di capire prima di tutto la modalità con cui si intendeva raccontare questa storia.
La scelta è caduta ancora sull'utilizzo del software Piktochart, questa volta per due motivi.
Il primo è quello di garantire una sorta di coerenza grafica con i report trimestrali mentre il secondo è la possibilità offerta da questo strumento di fondere grafici e testo.

L'obiettivo qui è stato quello di creare un racconto nato dai dati accompagnando il lettore attraverso una storia che avesse un inizio, uno sviluppo e una conclusione.

Documentazione e scelta delle fonti

La prima tappa di questo percorso è stata l'acquisizione di dominio. Prima di parlare di questa riforma era infatti necessario capire come fosse strutturata e quali aspetti del mondo del lavoro coinvolgesse.
Dopo la lettura di dossier, articoli di giornale, leggi e commenti di imprenditori ed economisti, è stato così possibile definire una serie di tappe in cui la riforma si è sviluppata nel corso del 2015.

- **Gennaio**. Entra in vigore la nuova legge di stabilità. Le aziende che assumono dipendenti a tempo indeterminato non versano i contributi previdenziali per i tre anni successivi alla firma del contratto. Il tetto fissato per questo esonero è di 8 060 euro all'anno.

- **Marzo**. Nasce il Jobs Act, arrivano i contratti a tempo indeterminato con tutele crescenti. Rispetto ai vecchi contratti a tempo indeterminato si caratterizzano per una maggiore flessibilità in uscita, licenziare diventa più semplice. Viene penalizzato l'utilizzo dei contratti a progetto.

- **Dicembre**. In questo mese comincia a circolare la voce che gli sgravi fiscali per assumere a tempo indeterminato verranno dimezzati nel 2016. Si registra un picco di avviamenti. Da gennaio 2016 il tetto per l'esonero dai contributi previdenziali è stato effettivamente ridotto a 3 250 euro all'anno per i primi tre anni.

Una volta definito il contesto è emerso quindi che i dati più importanti da prendere in considerazione per capire l'impatto di questa riforma riguardavano la tipologia di contratti firmati nel corso del 2015.

La sezione principale da cui sono state prese le informazioni è stata così *Dinamiche Lavorative*, quella che utilizza i dati provenienti dalle Comunicazioni Obbligatorie. Leggendo queste cifre si è capito se il numero di contratti a tempo indeterminato fosse effettivamente aumentato e se in generale il numero di avviamenti fosse superiore rispetto a quello degli altri anni.

Per cercare di fornire un'idea chiara del rapporto fra avviamenti e cessazioni inizialmente il valore su cui basare tutte le visualizzazioni è stato il saldo, il risultato del totale degli avviamenti meno il totale delle cessazioni. Su consiglio però dei ricercatori del Crisp questo valore è stato sostituito dal numero di avviamenti.

Questo saldo infatti rischiava di essere letto come un saldo occupazionale mentre il suo significato è leggermente diverso. Non si tratta di un saldo di stock, che definisce l'esatto numero di lavoratori presenti in un dato periodo ma di un saldo di flusso, un indicatore che piuttosto riporta una tendenza generale dell'andamento del mercato.

Così dal saldo si è passati a prendere in considerazione esclusivamente gli avviamenti. Questi infatti ricalcano gli stessi trend dei dati delle cessazioni che invece non sono stati inseriti nel lavoro finale per non appesantire il contenuto informativo.

Per quanto riguarda poi l'arco di tempo coperto dalle informazioni, i dati del 2015 sono stati confrontati con quelli del 2014 e in alcuni casi anche degli anni precedenti. I dati più recenti invece arrivano fino a gennaio 2016 e servono a capire come sono cambiati i numeri degli avviamenti a tempo indeterminato una volta terminato il primo stock di incentivi statali.

Raccolta dei dati e prime visualizzazioni

Una volta scelte le fonti su cui concentrarsi è cominciata la fase di raccolta dei dati. In alcuni casi è stato sufficiente utilizzare le tabelle che si possono acquisire dalle varie sezioni del *Quadrante del lavoro* ma in altri volte è stato necessario approfondire i dati che la piattaforma metteva a disposizione.

Questo è capitato in due occasioni. Prima si è rivelato necessario per capire il fenomeno delle trasformazioni. Per avere un quadro completo della situazione occorreva infatti sapere che tipo di contratto avessero i lavoratori che nel 2015 sono passati ad un tempo indeterminato.

Una seconda volta è stato necessario per comprendere le dinamiche dei settori lavorativi. Con il *Quadrante del lavoro* è infatti possibile andare ad analizzare come si è sviluppato il numero di contratti a tempo indeterminato solo nei macro settori: primario, costruzioni, industria e terziario. Nella dataviz realizzata invece vengono presi in considerazione i micro settori, dalla ristorazione fino al trasporto marittimo.

Tutti i dati ritenuti rilevanti sono stati poi sistemati all'interno di tabelle xls e in alcune di queste sono stati aggiunti degli attributi per calcolare percentuali, somme o differenze.

Per capire a questo punto quali fossero gli andamenti, i picchi o le proporzioni fra un dato e l'altro si

è rivelato indispensabile procedere con una prima serie di visualizzazioni con cui trovare le informazioni più interessanti.

Il software scelto per questa operazione è stato Tableau Public. Questo strumento permette non solo di visualizzare grandi quantità di dati ma anche di svolgere alcune operazioni. È possibile infatti sommare fra di loro le cifre contenute in determinate celle oppure filtrare i dati in base ad tuple o attributi.

Le visualizzazioni di questa fase non sono avvenute senza un ordine preciso ma sono la risposta a delle domande. Il modello è lo stesso di un'intervista. Quando un giornalista vuole conoscere la storia o le idee di una persona non gli lascia il microfono o il registratore davanti permettendogli di parlare liberamente. Cerca piuttosto di porre delle domande per stimolare e ordinare il discorso, seguendo un filo logico che poi permetterà anche al lettore di comprendere meglio quello che l'intervistato vuole dire.

Allo stesso modo quando si creano delle visualizzazioni per conoscere i dati che si hanno a disposizione è importante conoscere le domande che si vogliono fare, capire su quale aspetto di un fenomeno interrogare i dati. In questo caso ad esempio il focus dell'intervista era basato come fosse cambiato il numero di avviamenti di contratti a tempo indeterminato rispetto allo scorso anno.

Sono stati realizzati così diverse dashboard che hanno permesso di capire quali informazioni inserire all'interno della storia.

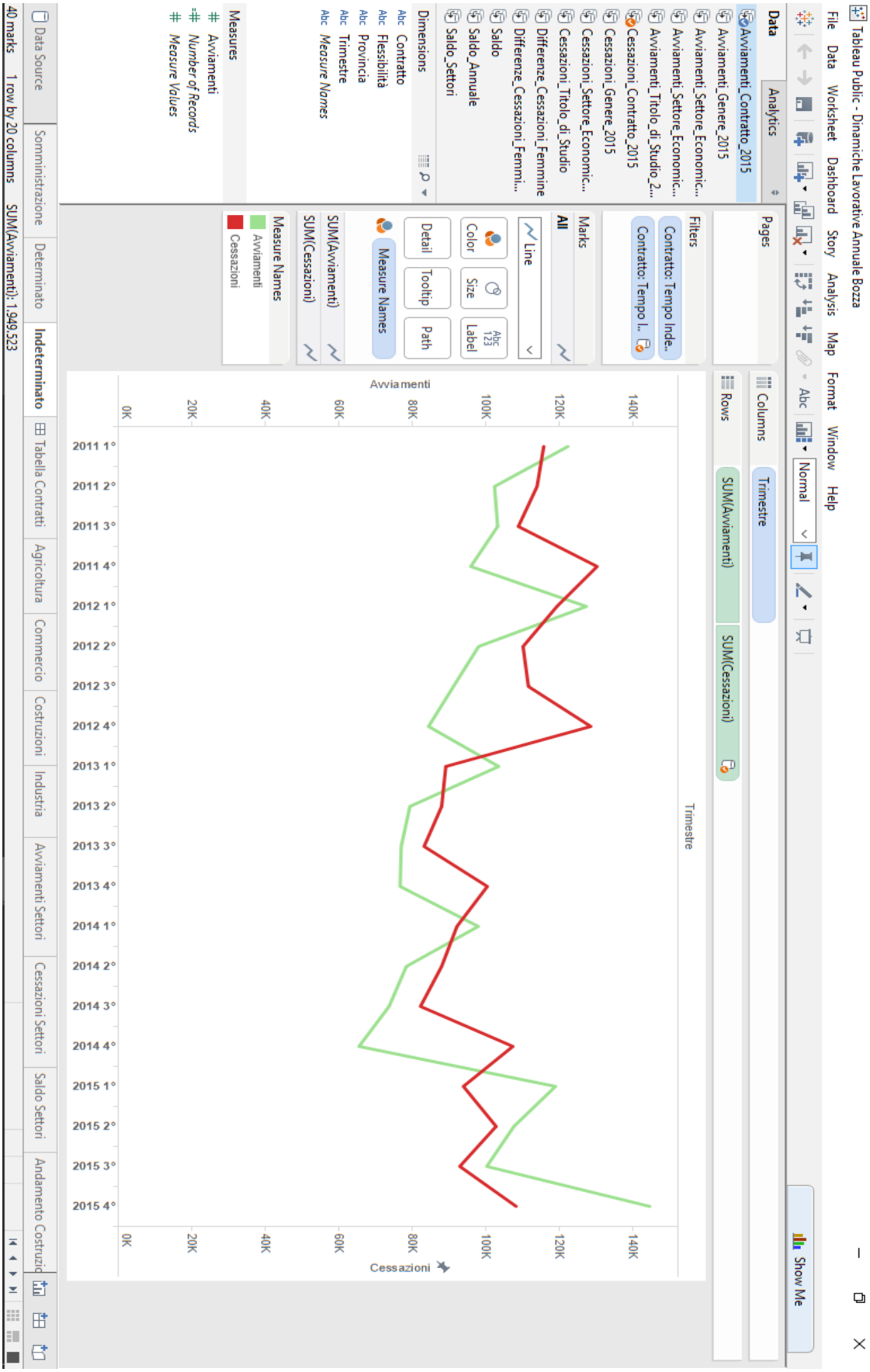


Illustrazione 73: Andamento Avviamenti e Cessazioni per i Contratti a Tempo Indeterminato dal 2011 al 2015

Illustrazione 74: Numero di Contratti a Tempo Indeterminato firmati nel settore Terziario nel 2014 e nel 2015

Estrazione delle Informazioni e Struttura della Storia

Dopo aver visualizzato tutti i dati presenti nelle fonti è stato possibile definire le informazioni da riportare ed organizzarle all'interno di una struttura definita prendendo come modello un articolo di giornale. I primi paragrafi corrispondono quindi alla *top down* le prime righe di un articolo in cui vengono presentati gli elementi più importanti della notizia. Scendendo con la visualizzazione queste prime indicazioni vengono sviluppate e approfondite fino ad arrivare agli ultimi grafici in cui la storia si conclude. Nella sezione finale prima si fa riferimento agli ultimi dati a disposizione per poi proiettarsi su quali potrebbero essere gli sviluppi nel 2016.

- **Meglio del 2014**. In questo primo paragrafo, oltre a definire i criteri di analisi, viene introdotto l'argomento trattato e inserito il confronto fra gli avviamenti conteggiati nel 2014 e nel 2015.

- **Le tappe della riforma**. Elenco degli eventi principali che hanno caratterizzato questa riforma: legge di stabilità, Jobs Act e esoneri.

- **Contratti più Stabili**. Confronto tra gli avviamenti di contratti a tempo indeterminato nel 2014 e nel 2015.

- **L'Indeterminato Conviene**. Confronto tra la percentuale di contratti a tempo indeterminato sul totale degli avviamenti firmati nel 2015 e nel 2014. Andamento storico di altri tipi di contratto come quello a progetto o quello che regola il lavoro a somministrazione.

- **Contratti Trasformati**. Percentuali delle tipologie di contratto che avevano i lavoratori passati nel 2015 ad un tempo indeterminato.

- **Posto Fisso in Fabbrica e in Cantiere**. Totale degli avviamenti a tempo indeterminato nei singoli settori economici e confronto con i dati del 2014.

- **Gennaio Freddo**. Primi dati del 2016 rispetto agli avviamenti per ogni tipo di contratto.

- **La Scossa Continua**. Confronto fra gli esoneri concessi nel 2015 e quelli previsti per il 2016. Riflessioni conclusive sul futuro disegnato dalla Riforma del Lavoro.

Il titolo scelto per questa storia è *La Scossa del Jobs* Act. Dai dati analizzati emerge infatti che nel 2015 c'è stato un drastico cambiamento nello scenario del tempo indeterminato, una forte scossa appunto. La metafora del titolo viene ripresa anche nell'ultimo paragrafo e serve per sottolineare che la sfida per il 2016 è capire se questa scossa sia stata un fenomeno circoscritto o se abbia sbloccato meccanismi in grado di cambiare le dinamiche del mondo del lavoro anche nei prossimi anni.

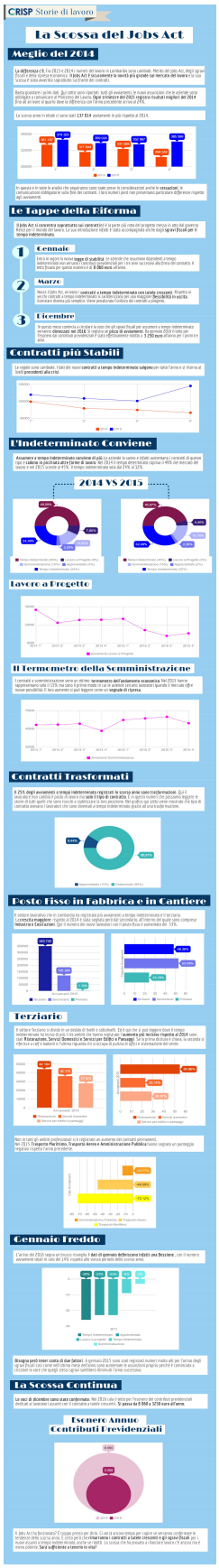
Il layout scelto per questa visualizzazione richiama quello adottato per il report mostrato nelle pagine precedenti ma si differenzia nella scelta dei colori. In questo modo è possibile capire la provenienza comune e allo stesso tempo distinguere le due tipologie di infografica.

Anche in questo caso, fra le fonti poste in fondo al lavoro è stato inserito un QR Code che rimanda alla home page de *Il Quadrante del Lavoro*.

Storie di Lavoro. Proposta di una rubrica

All'inizio di questa dataviz compare una breve intestazione: *Crisp, Storie di Lavoro*. La proposta che viene formulata è infatti che questa visualizzazione non costituisca un evento isolato ma che sia una sorta di puntata pilota per una rubrica periodica, da pubblicare magari in coincidenza con l'uscita dei report trimestrali.

Di volta in volta si potrebbe scegliere infatti un tema legato al mondo del lavoro da analizzare e raccontare.


CRISP Storie di lavoro
La Scossa del Jobs Act
Meglio del 2014
Le Tappe della Riforma
1 Gennaio
2 Marzo
3 Dicembre
Contratti più Stabili
L'Indeterminato Conviene
2014 VS 2015
Lavoro a Progetto
Il Termometro della Somministrazione
Contratti Trasformati
Posto Fisso in Fabbrica e in Cantiere
Terziario
Gennaio Freddo
La Scossa Continua
Esonero Annuo Contributi Previdenziali
Fonti
CRISP
IL QUADRANTE DEL LAVORO
Info
www.daslombardia.crisp.unimib.it
Powered by Piktochart

CRISP Storie di lavoro

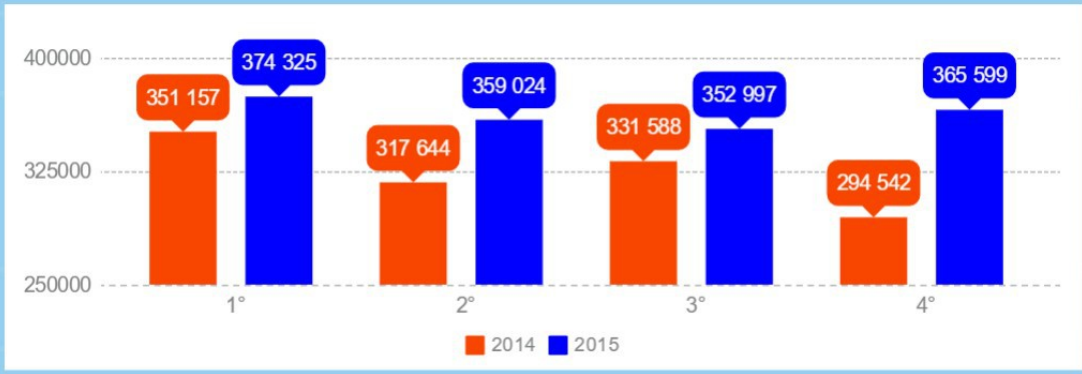
La Scossa del Jobs Act

Meglio del 2014

La differenza c'è. Tra 2015 e 2014 i numeri del lavoro in Lombardia sono cambiati. Merito del Jobs Act, degli sgravi fiscali e della ripresa economica. **Il Jobs Act è sicuramente la novità più grande sul mercato del lavoro** e la sua scossa è stata avvertita soprattutto sul fronte dei contratti.

Basta guardare i primi dati. Qui sotto sono riportati tutti gli avviamenti, le nuovi assunzioni che le aziende sono obbligate a comunicare al Ministero del Lavoro. **Ogni trimestre del 2015 registra risultati migliori del 2014** fino ad arrivare al quarto dove la differenza con l'anno precedente arriva al 24%.

Lo scorso anno in totale ci sono stati **137 014** avviamenti in più rispetto al 2014.



In questa e in tutte le analisi che seguiranno sono state prese in considerazioni anche le **cessazioni**, le comunicazioni obbligatorie sulla fine dei contratti. I loro numeri però non presentano particolare differenze rispetto agli avviamenti.

Le Tappe della Riforma

Il Jobs Act si concentra soprattutto sui contratti ed è la parte più nota del progetto messo in atto dal governo Renzi per il mondo del lavoro. La sua introduzione infatti è stata accompagnata anche dagli **sgravi fiscali per il tempo indeterminato**.

1 Gennaio

Entra in vigore la nuova **legge di stabilità**. Le aziende che assumono dipendenti a tempo indeterminato non versano i contributi previdenziali per i tre anni successivi alla firma del contratto. Il tetto fissato per questo esonero è di **8 060 euro** all'anno.

2 Marzo

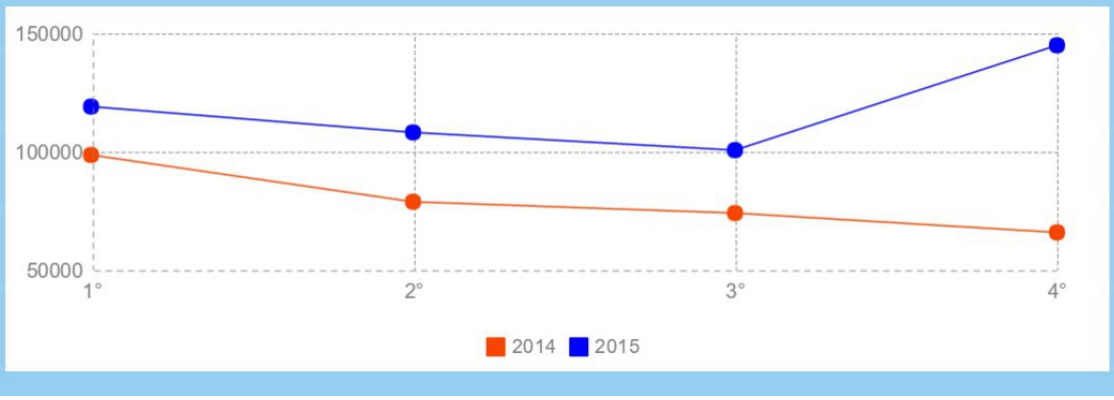
Nasce il Jobs Act, arrivano i **contratti a tempo indeterminato con tutele crescenti**. Rispetto ai vecchi contratti a tempo indeterminato si caratterizzano per una maggiore **flessibilità in uscita**, licenziare diventa più semplice. Viene penalizzato l'utilizzo dei contratti a progetto.

3 Dicembre

In questo mese comincia a circolare la voce che gli sgravi fiscali per assumere a tempo indeterminato verranno **dimezzati nel 2016**. Si registra un **picco di avviamenti**. Da gennaio 2016 il tetto per l'esonero dai contributi previdenziali è stato effettivamente ridotto a **3 250 euro** all'anno per i primi tre anni.

Contratti più Stabili

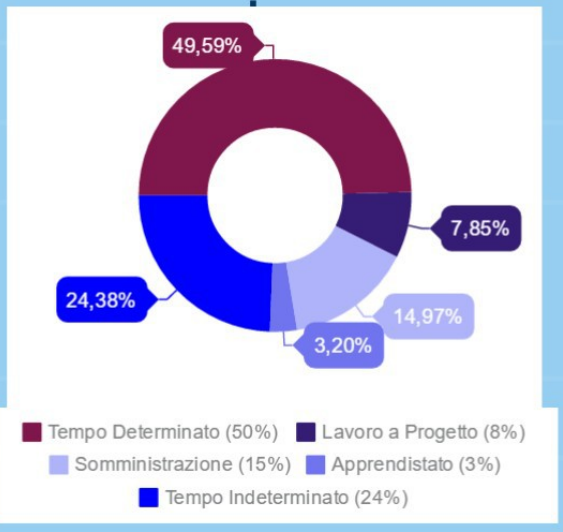
Le regole sono cambiate. I dati dei nuovi **contratti a tempo indeterminato salgono** per tutto l'anno e si ritorna ai livelli **precedenti alla crisi**.

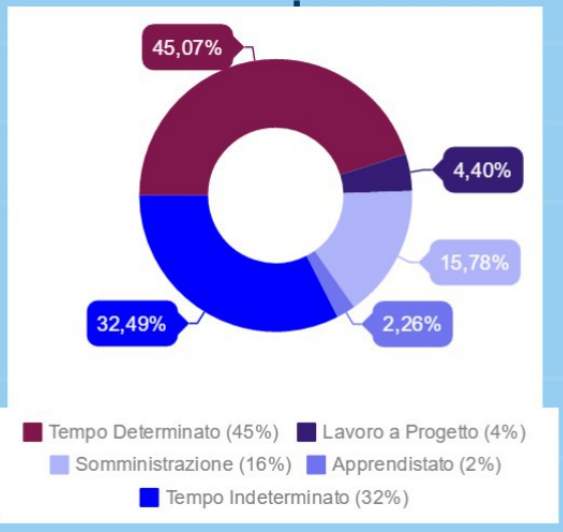


L'Indeterminato Conviene

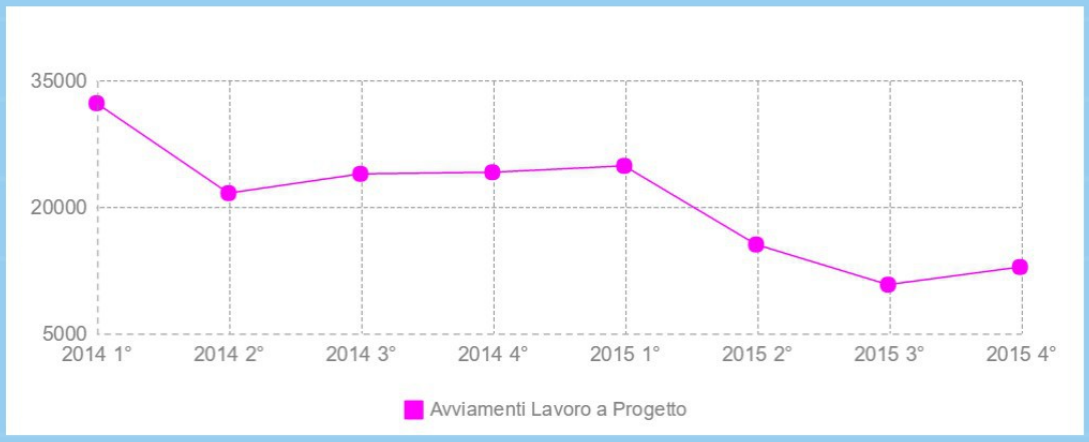
Assumere a tempo indeterminato conviene di più. Le aziende lo sanno e infatti aumentano i contratti di questo tipo e **cadono in picchiata altre forme di lavoro**. Nel 2014 il tempo determinato copriva il 49% del mercato del lavoro e nel 2015 scende al 45%. Il tempo indeterminato vola dal 24% al 32%.

2014 VS 2015



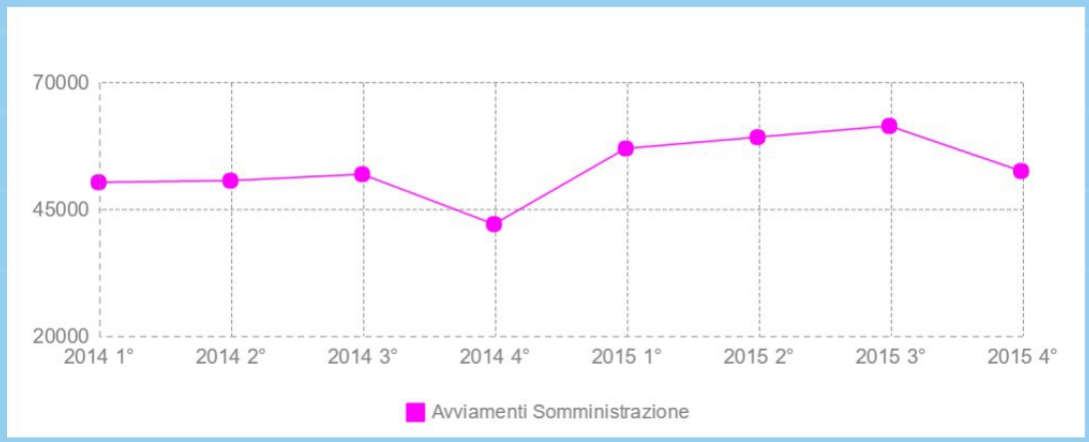


Lavoro a Progetto



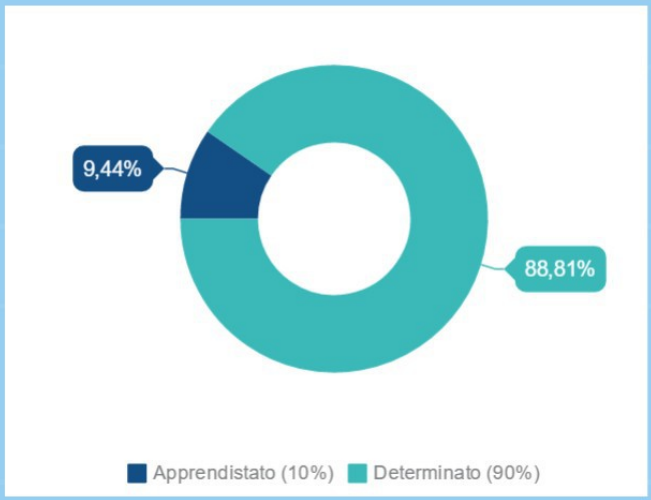
Il Termometro della Somministrazione

I contratti a somministrazione sono un ottimo **termometro dell'andamento economico**. Nel 2015 hanno rappresentano solo il 15% ma sono il primo modo in cui le aziende cercano lavoratori quando il mercato offre nuove possibilità. Il loro aumento si può leggere come un **segnale di ripresa**.



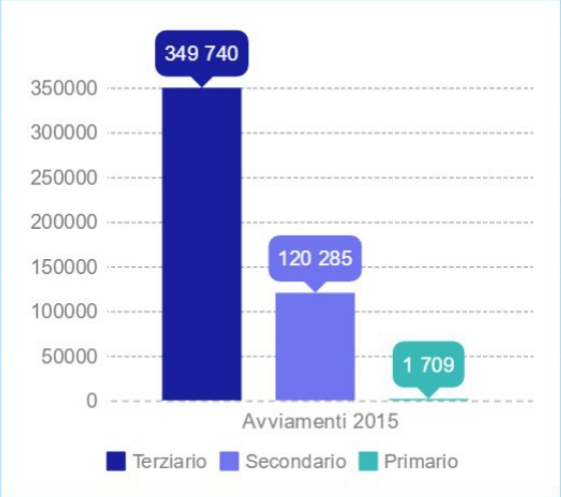
Contratti Trasformati

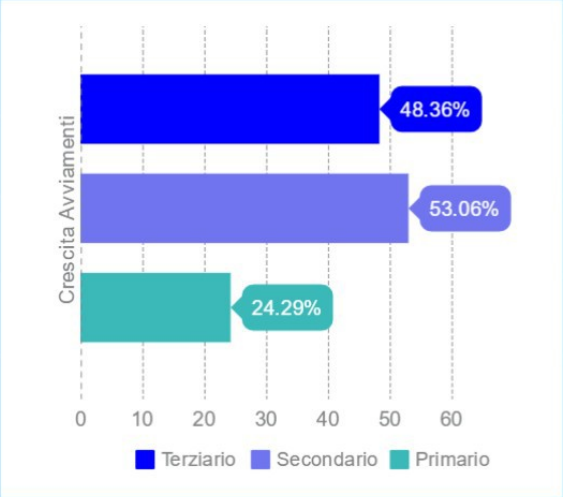
Il 25% degli avviamenti a tempo indeterminato registrati lo scorso anno sono trasformazioni. Qui il lavoratore non cambia il posto di lavoro ma **solo il tipo di contratto**. È in questi numeri che possiamo leggere le storie di tutti quelli che sono riusciti a stabilizzare la loro posizione. Nel grafico qui sotto viene mostrato che tipo di contratto avevano i lavoratori che sono diventati a tempo indeterminato grazie ad una trasformazione.



Posto Fisso in Fabbrica e in Cantiere

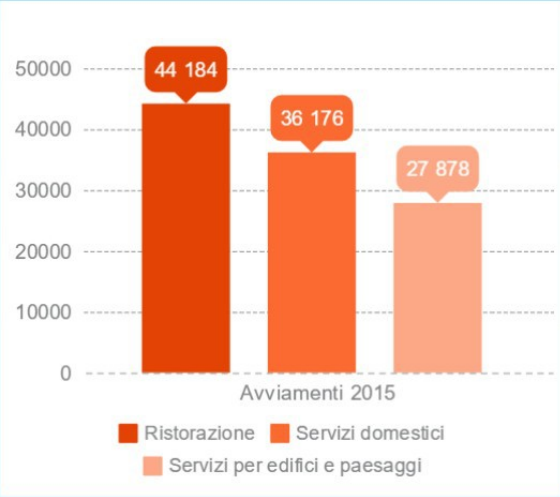
Il settore lavorativo che in Lombardia ha registrato più avviamenti a tempo indeterminato è il terziario. La **crescita maggiore** rispetto al 2014 è stata segnata però dal secondario, all'interno del quale sono comprese **Industria e Costruzioni**. Qui il numero dei nuovi lavoratori con il posto fisso è aumentato del 53%.

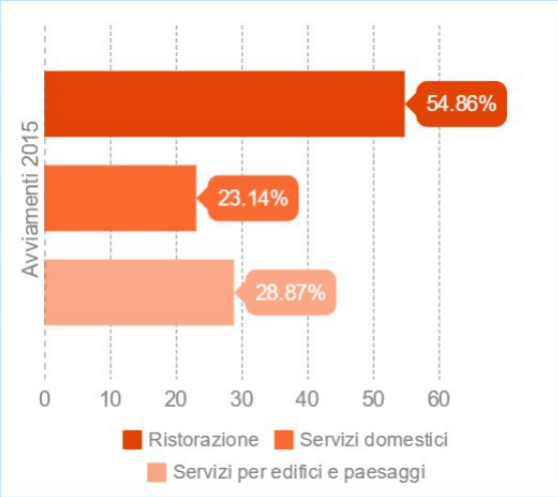




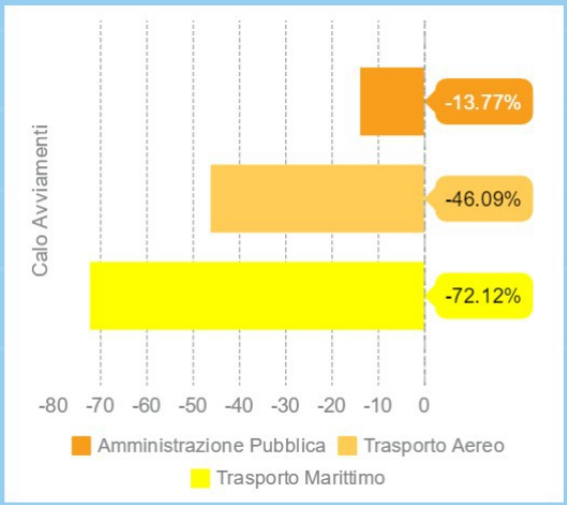
Terziario

Il settore Terziario si divide in un dedalo di livelli e sottolivelli. Ed è qui che si può leggere dove il tempo indeterminato ha inciso di più. I tre ambiti che hanno registrato l'**aumento più incisivo rispetto al 2014** sono stati **Ristorazione, Servizi Domestici e Servizi per Edifici e Paesaggi**. Se la prima dicitura è chiara, la seconda si riferisce a colf e badanti e l'ultima riguarda chi si occupa di pulizia di uffici e sistemazione del verde.



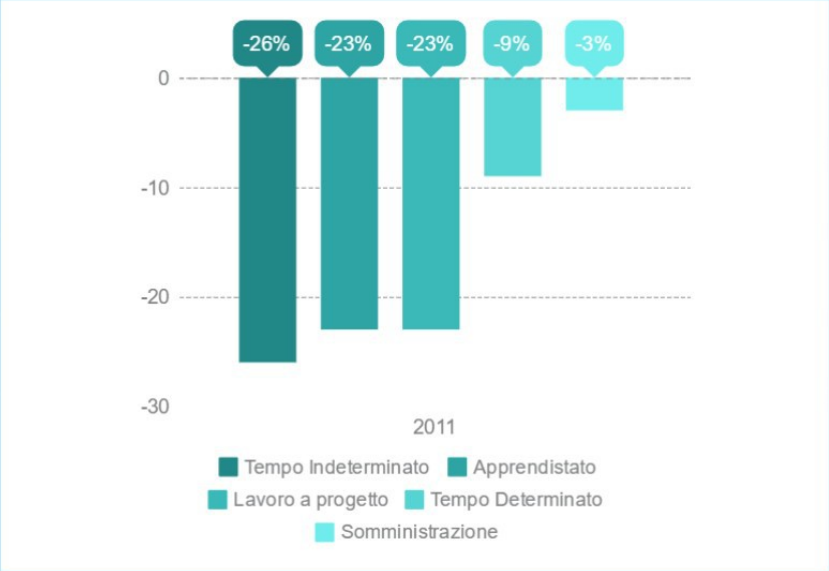


Non in tutti gli ambiti professionali si è registrato un aumento dei contratti permanenti.
Nel 2015 **Trasporto Marittimo, Trasporto Aereo e Amministrazione Pubblica** hanno segnato un punteggio negativo rispetto l'anno precedente.



Gennaio Freddo

L'arrivo del 2016 segna un brusco risveglio. **I dati di gennaio definiscono infatti una flessione** , con il numero avviamenti totali in calo del 14% rispetto allo stesso periodo dello scorso anno.

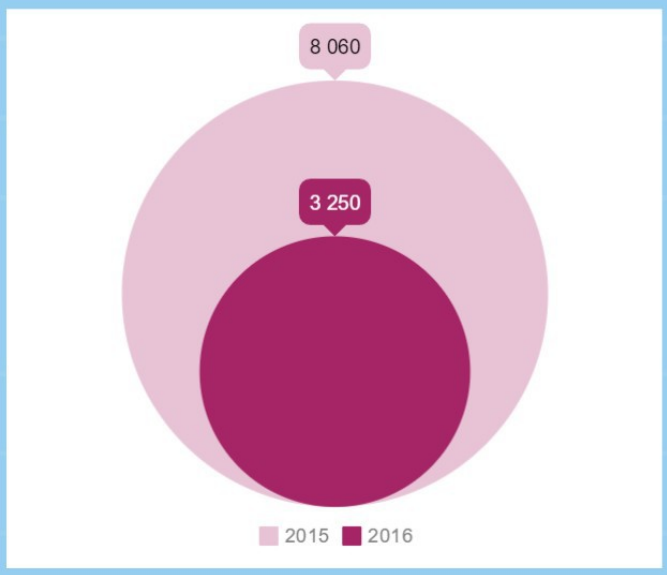


Bisogna però tener conto di due fattori. A gennaio 2015 sono stati registrati numeri molto alti per l'arrivo degli sgravi fiscali così come nell'ultimo mese dell'anno sono aumentate le assunzioni proprio perché è cominciata a circolare la voce che quegli stessi sgravi sarebbero diminuiti l'anno successivo.

La Scossa Continua

Le voci di dicembre sono state confermate. Nel 2016 cala il tetto per l'esonero dei contributi previdenziali dedicati ai lavoratori assunti con il contratto a tutele crescenti. **Si passa da 8 060 a 3250 euro all'anno.**

Esonero Annuo Contributi Previdenziali



Il Jobs Act ha funzionato? È troppo presto per dirlo. Ci vorrà ancora tempo per capire se verranno confermate le tendenze dello scorso anno. È certo però che **rimarranno i contratti a tutele crescenti e gli sgravi fiscali** per i nuovi assunti a tempo indeterminato, anche se ridotti. La scossa che ha provato a rilanciare lavoro c'è ancora ma è meno potente. **Sarà sufficiente a tenerlo in vita?**

Fonti

CRISP
centro di ricerca interuniversitario per i servizi di pubblica utilità

IL QUADRANTE DEL LAVORO
Open Data del mercato del lavoro in Lombardia
Osservatorio regionale del mercato del lavoro e della formazione

Info www.daslombardia.crisp.unimib.it

6. Le correlazioni spurie di Tyler Vigen e i tre errori di Alberto Cairo

Cosa c'entra il numero dei suicidi per strangolamento e la spesa degli Stati Uniti d'America per le scienze, la tecnologia e la ricerca spaziale?

Apparentemente nulla ma i dati dicono il contrario, come si può leggere da questo *line graphs* basato sulle informazioni provenienti dal U.S. Office of Management and Budget e dai Centres for Disease Control & Prevention.

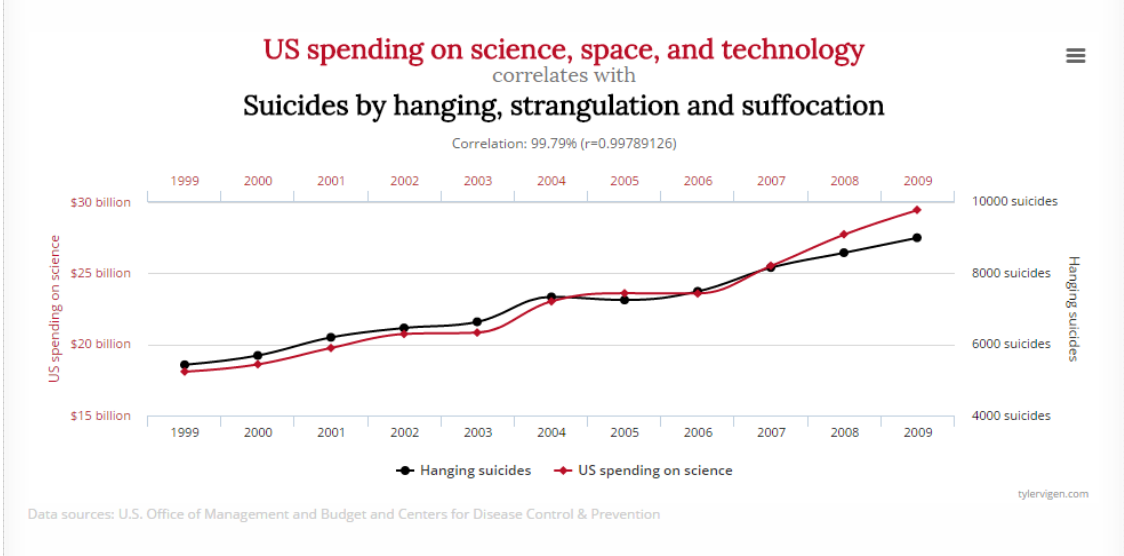


Illustrazione 75: Fonte: http://tylervigen.com/spurious-correlations

La correlazione in questo caso è addirittura del 99,97% e guardando la linea del grafico è impressionante vedere come i due fenomeni vadano di pari passo.

Nella realtà però non esiste alcuna causalità fra questi due fenomeni. La visualizzazione qui riportata infatti è stata creata da Tyler Vigen[76] uno studente di giurisprudenze della Harvard University molto interessato alle correlazioni spurie, le correlazioni cioè che avvicinano due fenomeni senza che questi abbiamo alcun rapporto di casualità.

Questo è uno dei rischi più grossi che può correre una persona che si avvicina al mondo della data visualization. E non si tratta di un problema proprio solo di chi ha una formazione umanistica. Anche uno statistico che non ha alcuna competenza del contesto in cui sta analizzando dei dati può cadere alla tentazione di stabilire cause ed effetti solo perché due fenomeni seguono gli stessi andamenti.

La possibilità di commettere errori nella creazione di una dataviz è un tema di cui si è interessato anche Alberto Cairo, professore di Visual Journalism alla School of Comunicatione dell'University of Miami.

Nel 2016 ha pubblicato infatti *The Truthful Art: Data, Charts, and Maps for Communication*[77], una sorta di guida per chi intende utilizzare il data design nell'ambito della comunicazione.

76 http://tylervigen.com/spurious-correlations
77 A. Cairo, *The Truthful Art: Data, Charts, and Maps for Communication*, 2016

In questo prontuario per data journalist parla anche dei meccanismi attraverso cui la mente può essere ingannata dai grafici.

Questi problemi nascono da quello che Cairo identifica come un modello di pensiero scritto sulle parti più antiche della nostra corteccia celebrare, un pensiero che si basa sull'intuizione. La capacità che ha l'uomo di arrivare subito ad una conclusione partendo da pochi elementi. Un modello che rispecchia un processo primitivo basati su “Ci sono rumori tra i cespugli>potrebbe essere un predatore>corri o stai pronto a difenderti”.

Questa intuizione primordiale, questa tendenza a giudicare a prescindere dalla conoscenza di un fenomeno si ripresenta anche nel mondo della data visualization e in particolare sotto tre errori del pensiero concatenati fra loro: Patternicity Bug, Storytelling Bug e Confirmation Bug.

Il Patternicity Bug riguarda la tendenza da parte dell'occhio e della mente umana di trovare dei pattern, delle trame anche dove queste non ci sono, una tendenza che nel mondo statistico prende il nome di apofenia. Nel suo libro Cairo mostra una serie di grafici che dichiara rappresentare il tasso di disoccupazione in nove nazionale tra il 2010 e il 2015. Invita l'utente a soffermarsi su questi tabelle, provando a trovare dei modelli, suggerendo la ripetizione di alcuni valori negli stessi periodi dell'anno oppure la presenza di *line graphs* molto simili fra di loro.

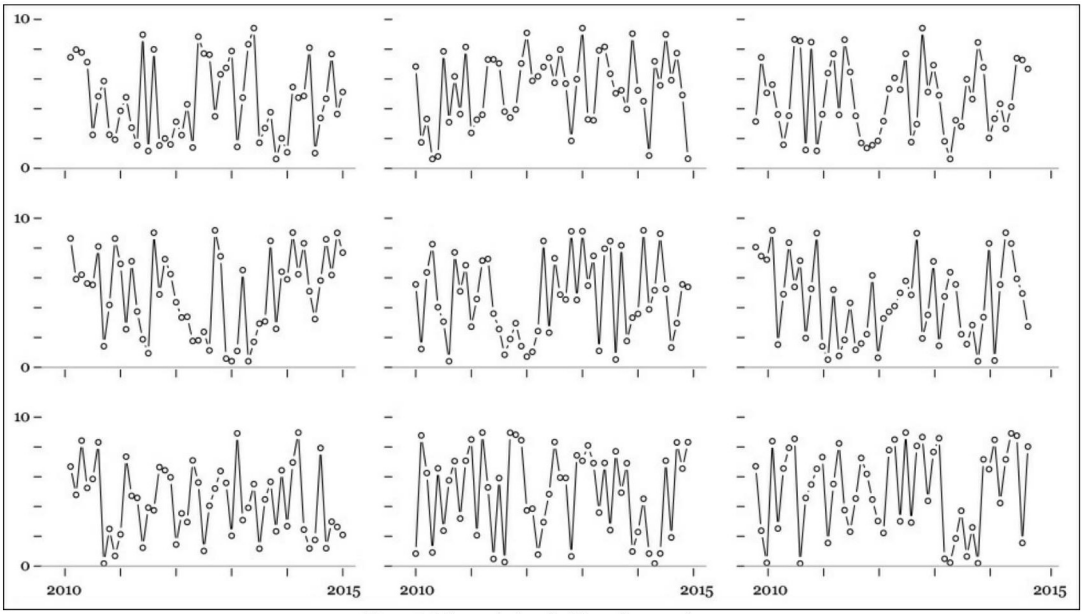


Illustrazione 76: Serie di nove grafici proposta da Alberto Cairo in The Truthful Art

Dopo questo breve esperimento, l'autore rivela che si tratta di valori inesistenti, generati in modo del tutto casuale. Ammette però che pur conoscendo l'origine di questi grafici anche lui guardandoli cade ancora nella tentazione di cercare dei modelli che si ripetono.

Il secondo di questi errori di percorso è una conseguenza del primo ed è lo Storytelling Bug, la tendenza cioè a stabilire dei nessi causa-effetto.

Una volta che infatti si scopre un modello o si trovano dei dati simili tra loro, l'uomo cerca naturalmente di associarli e di trovare delle ragioni a queste correlazioni. L'errore iniziale prende dunque corpo e quello che era una semplice deduzione erronea arriva a diventare una storia completamente sbagliata.

A questo secondo errore ne deriva quindi subito un terzo: il Confirmation Bug.

Nel 2003 lo psicologo Geoffey Cohen ha presentato una serie di linee guida per il welfare ad un gruppo di liberali e ad uno di conservatori. Le persone che avevano dichiarato un'appartenenza politica liberale si mostravano in accordo con le politiche conservatrici se veniva detto che queste erano promosse dal Partito Democratico. Lo stesso meccanismo avveniva poi in forma speculare anche per i sedicenti conservatori. Quando veniva chiesto loro perché appoggiassero una politica piuttosto che l'altra, questi rispondevano di basarsi solo su un'analisi attenta delle informazioni contenute nelle proposte.

L'esperimento di Cohen non vuole certo dimostrare che le politiche dei due principali partiti statunitensi siano perfettamente scambiali ma piuttosto che l'uomo tende a recepire e riproporre solo i dati che lo intessano, o meglio, che sono necessari alla conferma di un modello e di una storia da lui trovati.

Per cercare di contrastare questi fenomeni, Cairo suggerisce di accrescere sempre il valore delle proprie visualizzazioni in due diverse direzioni.
Da una parte bisogna cercare di sviluppare la profondità delle ricerche sui dati, creando modelli con differenti livelli di analisi. Dall'altra non ci si deve concentrare solo su un singolo fenomeno ma coinvolgere nella propria ricerca anche fenomeni diversi così da poter sempre avere un metro di paragone.

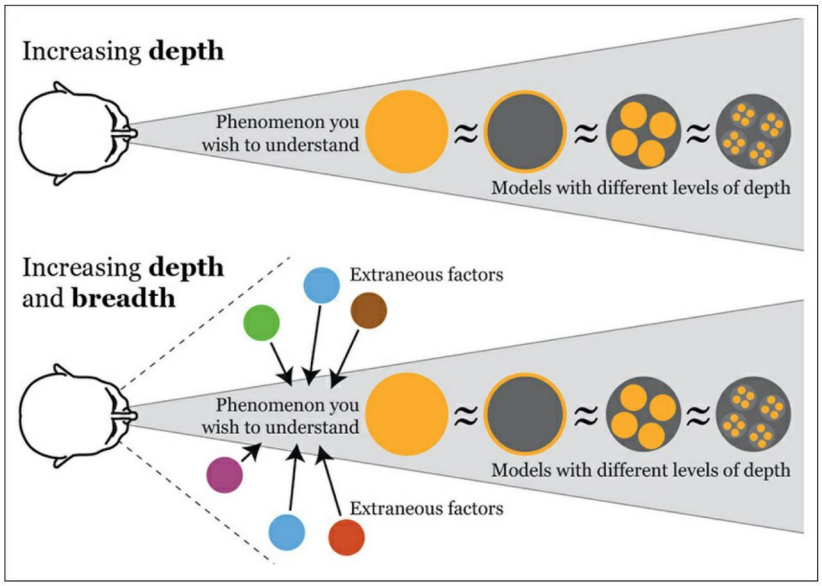


Illustrazione 77: Milgiorare Profondità e Ampiezza, lo schema proposto da Cairo

Anche se l'intuizione quindi è ancora un meccanismo ben radicato nella mente dell'uomo, l'invito che Cairo fa nel suo libro è quello di cercare di combatterla. Di non rincorrere subito la storia intravista nei dati ma di fermarsi e pensare, approfondire ed ampliare, solo dopo si potrà decidere quale storia raccontare.

“Don't rush to write a headline or an entire story or to design a visualization immediately after you find an interesting pattern, data point, or fact. Stop and think. Look for other sources and for people who can help you escape from tunnel vision and confirmation bias. Explore your information at

multiple levels of depth and breadth, looking for extraneous factors that may help explain your finding. Only *then* can you make a decision about what to say, and how to say it, and about what amount of detail you need to show to be true to the data".

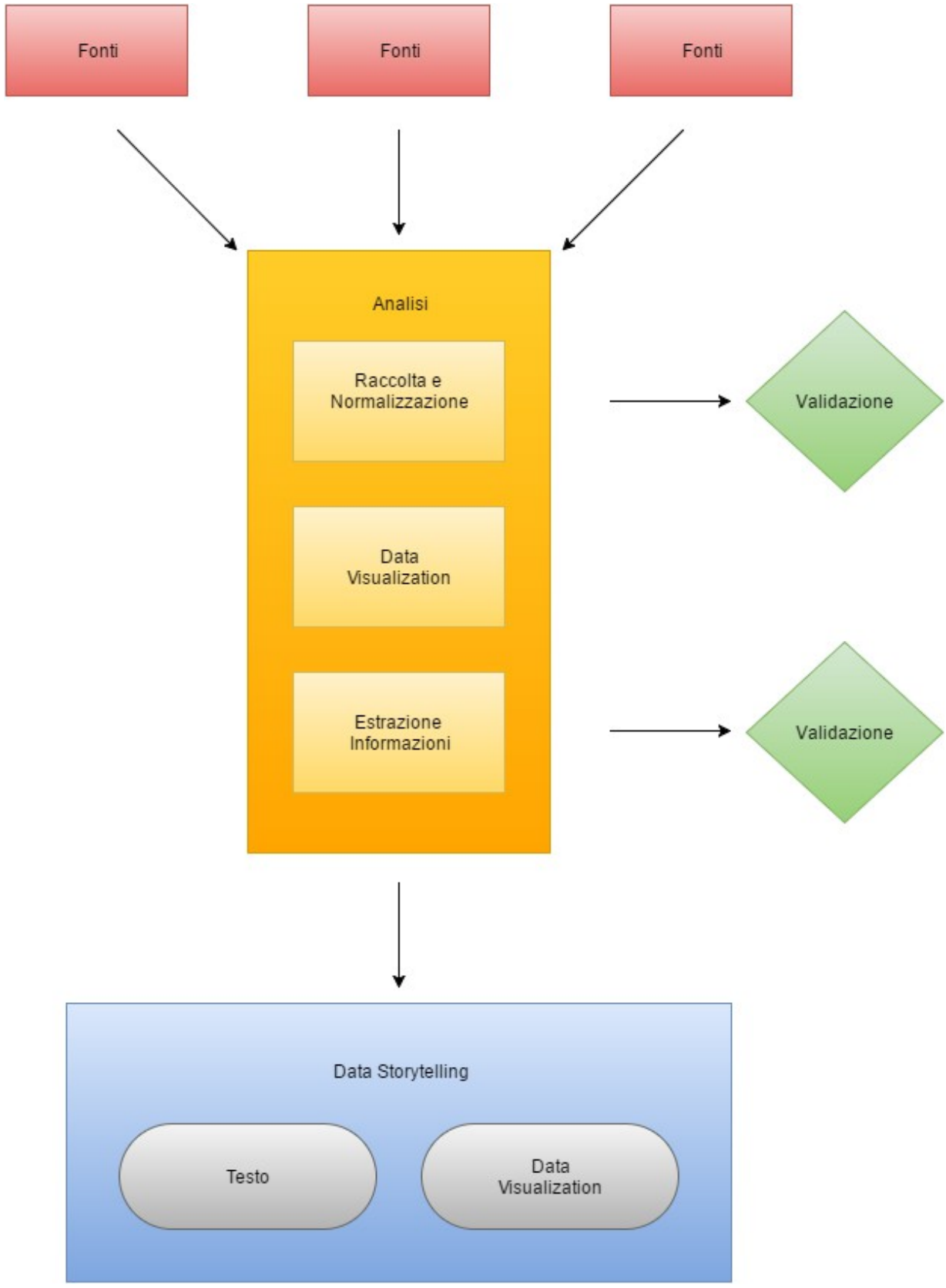
Dato che il campo della data visualization è caratterizzato dalla presenza di competenze diverse è necessario quindi cercare di ricorre sempre ad una validazione, al supporto di qualche figura esperta del dominio trattato che possa convalidare le deduzioni a cui si è arrivati. Per chi arrivano da studi umanistici la necessità di questa validazione coinvolge anche la metodologia applicata per l'analisi o la raccolta dei dati.

Questa riflessione si riferisce però solo alle occasioni in cui un'unica persona si debba occupare di tutto il processo di data storytelling. Se lo stesso processo avvenisse all'interno di un team composto da professionisti diversi la validazione non sarebbe più necessaria perché ognuno si occuperebbe del passaggio di sua competenza.

Nel corso del lavoro su *La Scossa del Jobs Act* la necessità di una validazione è arrivata in due momenti. Prima nel momento in cui si è reso necessario stabilire un indicatore su cui basare il data storytelling. La proposta iniziale era quella di utilizzare il saldo fra avviamenti e cessazioni ma la scelta che si è rivelata essere migliore è stata utilizzare solo i numeri degli avviamenti. Nel secondo caso invece è servita per verificare delle informazioni estratte dall'analisi dei dati.

7. Dai dati al data storytelling. Uno schema di lavoro

Al termine di questo progetto è stato possibile definire un flusso di lavoro per il data storytelling.



Il percorso qui suggerito si divide in tre parti: scelta delle fonti, analisi dei dati e data storytelling.

La scelta delle fonti è il primo passo da fare per poter cominciare questo percorso. È una scelta che non riguarda il formato con cui si possono trovare dei dati ma semplicemente la loro origine.
Esistono diversi tipi di fonti, da quelle che forniscono dati strutturati, come il caso delle tabelle del *Quadrante del lavoro* a quelle che invece forniscono dati non strutturati che poi andranno raccolti e ordinati attraverso appositi database.

Il concetto di fonte cambia in base alle competenze tecniche che si possiedono. Le tabelle sul lavoro in Lombardia utilizzate ne *La Scossa del Jobs Act* ad esempio non sono dati originali ma hanno già attraversato un processo di normalizzazione e aggregazione che ha permesso all'utente di poterle gestire con Excel. Questo metodo sarà più adatto per un giornalista mentre un informatico interessato allo stesso tipo di lavoro potrebbe sviluppare un sistema in grado di registrare autonomamente dati provenienti da diverse fonti.

Anche l'analisi dei dati varia a seconda della figura professionale che la sta conducendo. Una volta definito un dataset per chi ha una formazione umanistica lo strumento della data visualization potrà essere utile per cercare nei dati delle risposte alle domande che si hanno in mente. In questo momento è possibile anche cominciare a capire quali forme di dataviz possono essere più efficaci. L'ultimo passaggio dell'analisi è rappresentato dall'estrazione delle informazioni. È in questo momento che vengono selezionate solo le informazioni che si ritengono più importanti, quelle che poi saranno riprese nella fase finale.

Il data storytelling è il racconto che chiude il processo. Qui vengono disposte in sequenza le informazioni estratte e vengono scelti i media attraverso cui comunicarle. Questa è una decisione che riguarda sia il tipo di dataviz da utilizzare che i testi da inserire. Tutto dipende dai dati estratti attraverso l'analisi. È possibile infatti che alla fine di tutto questo processo non occorra neanche un grafico per comunicare l'informazione estratta dai dati.

La validazione si inserisce in due fasi. Prima è necessario controllare che la raccolta e normalizzazione dei dati sia avvenuta correttamente in modo tale da avere un dataset solido da cui partire. Alla fine del processo di analisi è poi importante controllare con un esperto di dominio che le informazioni ottenute poggino su basi attendibili per evitare di cadere nelle trappole indicate da Tyler Vigen e Alberto Cairo.

Conclusioni

La data visualization fornisce gli strumenti per raccontare delle storie complesse ma non sempre questi strumenti sono sufficienti a trasmettere un'informazione chiara.

Dopo aver analizzato le strutture che regolano la composizione dei grafici e i loro ambiti di utilizzo è possibile affermare che le infografiche sono in grado di sviluppare delle narrazioni, caratterizzate da una struttura, un ritmo e uno sviluppo cronologico. E questa potenzialità è espressa in maniera ancora più evidente quando vengono accompagnate anche da un testo. Da sola una visualizzazione non sempre è esauriente. Il lettore che si approccia a questo linguaggio rischia infatti di perdersi nella complessità dei dati o trovare delle informazioni che si rivelano distanti dalla realtà.

Dalla data visualization si passa quindi al data storytelling, il racconto che nasce dai dati. Ed è qui, nella capacità di creare e sviluppare una storia, che si possono collocare delle competenze forse ancora inedite in questo ambito: quelle provenienti da un ambito umanistico.

Racconti di questo tipo possono essere strumenti di comunicazione efficaci in diversi ambiti, dal data journalism alla divulgazione scientifica, ma la loro creazione è una sfida non priva di insidie.
Chi si trova a scrivere una storia nata dai dati deve infatti camminare su un terreno dove spesso sono richieste diverse competenze tecniche.

È per questo che durante il progetto di creazione di un'infografica illustrato nel terzo capitolo è stato necessario definire un flusso di lavoro diviso in tre fasi.

Il punto di partenza è la scelta delle fonti, da qui si passa alla raccolta e all'analisi dei dati per poi arrivare al data storytelling, il momento in cui vengono definite le strategie migliori per trasmettere le informazioni raccolte.
In questo percorso le tecniche di data visualization vengono utilizzate due volte, prima nella fase di analisi, per comprendere i dati che si hanno davanti, e poi alla fine, per comunicarli.

È giusto però segnalare almeno due rischi a cui si può andare incontro in questo processo.

Il primo riguarda chi non possiede competenze di informatica o statistica. Raccogliere e analizzare dati può essere relativamente semplice quando ci si concentra su piccole cifre. Se però la materia prima aumenta e gli strumenti si fanno più sofisticati è necessario ricorrere all'aiuto di esperti che possano garantire la correttezza dei metodi utilizzati e delle informazioni estratte. Il data storytelling si conferma così un campo multidisciplinare dove è necessaria la collaborazione di figure professionali differenti.

Il secondo rischio è invece legato al processo creativo. Una dataviz e ancora di più una storia che nasce dai dati non sono la semplice trasposizione di una serie di numeri. Si tratta sempre di costruzioni, di oggetti nati da un dataset trasformato dalle mani di uno o più autori. Questo è un meccanismo inevitabile che però può essere reso trasparente. Al fine di creare una narrazione onesta sarà quindi necessario non solo informare il lettore delle procedure utilizzate ma anche permettergli di accedere direttamente alle fonti.

Appendice 1
Caso Studio *Terra Malata*

Terra Malata

Il progetto Terra Malata nasce per studiare le potenzialità del software Tableau Public e comprendere allo stesso tempo tutti i passaggi necessari per trasformare un corpus grezzo di dati in una visualizzazione interattiva.

Open Data Lombardia

Il punto di partenza per cominciare questo percorso è stato reperire una fonte attendibile. La scelta è caduta su Open Data Lombardia[78], il portale di Regione Lombardia che raccoglie dataset su diversi temi, dalla sanità alla sicurezza, passando per energia, ambiente e istruzione. Il sito nasce per promuovere la trasparenza della pubblica amministrazione e quindi le informazioni che si possono trovare sono da considerarsi tutte ufficiali. Ogni dataset è accompagnato da una breve descrizione che spiega a cosa si riferiscono i dati, come sono stati raccolti e a quando risale l'ultimo aggiornamento. Tutti i file possono essere scaricati in diversi formati. Dopo una veloce rassegna dei dataset archiviati nel catalogo, quello selezionato per questa esercitazione è stato "Elenco_dei_siti_contaminati_sul_territorio_lombardo.xls".

Struttura del dataset e normalizzazione

L'argomento affrontato da questa raccolta di dati è la presenza in Lombardia di terreni contaminati da sostanze inquinanti.
La tabella è formata da 889 righe e da 8 colonne, ognuna delle quali contiene dati di carattere soprattutto geografico.
Ecco infatti quali informazioni si possono trovare per ogni terreno:

- Provincia
- Numero
- Comune
- ID_Anagrafe
- Denominazione_Sito
- Indirizzo_Località
- N_Civico
- Classificazione

La prima operazione condotta su questo dataset è stata quella di tagliare le parti superflue.
È stata eliminata la colonna che conteneva il numero delle righe e quella con il codice dell'anagrafe.
É stata tolta anche la colonna nominata "Classificazione". Qui infatti non erano contenute informazioni rilevanti ma in ogni cella ritornava sempre la stessa voce: "Contaminato". L'ultima colonna che non ha superato la fase di sfoltimento è quella contenente i numeri civici, nella maggior parte dei casi infatti questo dato non era stato registrato, risultando quindi nullo.

Dopo aver reso più maneggevole il dataset, il secondo passaggio di questo lavoro ha previsto la normalizzazione dei caratteri. Molte voci erano costituite da nomi di luoghi, da quelli delle province fino alle città. Non tutti però utilizzavano gli stessi criteri di scrittura. Alcuni si presentavano in maiuscolo, altri in minuscolo, alcune parole erano divise da linee altre semplicemente da spazi. Le province poi non comparivano con il nome completo ma solo con la loro sigla, un tipo scrittura che

78 https://www.dati.lombardia.it/

impedisce il riconoscimento come dato geografico a Tableau Pubblic.
Le sigle delle province sono quindi state sciolte e i nomi di città riscritti con la prima lettera maiuscola e le altre minuscole. Le denominazioni dei singoli siti e i loro indirizzi sono invece stati trasformati tutti in maiuscolo, in modo tale da garantire chiarezza e omogeneità.

Il processo di normalizzazione si è quindi concluso riempendo le celle prive di qualsiasi voce con il segno grafico "_".

Intervista ai dati

Prima di avviare il software e cominciare il processo di visualizzazione, bisogna formulare le domande da sottoporre ai dati, per guidare non solo la scelta dei grafici ma anche la sequenza in cui presentarli. Ecco quelle scelte per l'esercitazione

- Dove si trovano i terreni contaminati?
- Quali sono le zone più inquinate?
- Il mio comune ha delle aree da bonificare?

Il modello narrativo utilizzato è stato quello del Martini Glass, teorizzato da Edward Segel e Jeffrey Herr. Secondo questo schema la visualizzazione si sviluppa in due fasi. Nella prima l'autore accompagna l'utente mostrandogli le informazioni più importanti del dataset, nella seconda invece è l'utente stesso ad esplorare liberamente le risorse messe a disposizione.

Visualizzazione

Davanti a queste esigenze, la struttura della visualizzazione è stata definita attraverso la modalità Story di Tableau Public. Grazie a questa opzione è infatti possibile creare una sequenza formata da più *Sheet*, i fogli di lavoro su cui si costruiscono i grafici, o *Dashboard*, pannelli su cui si possono unire più *Sheet*.

Dove si trovano i terreni contaminati?

Per rispondere al primo quesito è stata creata una mappa della regione che geolocalizza tutte le aree registrate come inquinate.

Sulle mappe utilizzate da Tableau Public non compaiono tutti i comuni, spesso quelli più piccoli vengono omessi. Per questo motivo delle 351 città contenute nel dataset, ben 248 erano registrate sotto la voce "unknow". Una condizione che ha reso necessario utilizzare la funzione di geolocalizzazione manuale, inserendo le coordinate decimali di tutte le voci non riconosciute.

Oltre alla geolocalizzazione sono stati inserite anche due variabili grafiche per evidenziare il numero di terreni inquinati presenti in ogni comune. Tutte le città sono infatti contraddistinte da un cerchio che varia in grandezza, dal più piccolo al più grande, e in colore, su una scala cromatica che va dal rosso al nero.

Per rendere esplorabile questa visualizzazione sono stati collegati due diversi filtri. Il primo permette di selezionare e confrontare fra loro le provincie, il secondo invece esegue le stesse operazioni a livello comunale.

Quali zone sono più inquinate?

Alla seconda domanda si è risposto utilizzando una combinazione di due visualizzazioni, una statica e l'altra dinamica.

La prima visualizzazione è una *bubble chart* statica, un grafico in cui la differenza fra gli elementi coinvolti è marcata dalle diverse dimensioni dei cerchi con cui sono rappresentati. In questo modo si può vedere chiaramente la sproporzione fra i terreni inquinati presenti nella provincia di Milano, che ne conta 414, e quelli presenti in provincie meno urbanizzate come Sondrio dove ne vengono registrati solo 3.

La seconda visualizzazione è invece una *bar chart* dinamica che filtra i dati per provincia. In questo modo è possibile ottenere una classifica per vedere quali sono i comuni più inquinati all'interno di aree sostanzialmente simili.

Il mio comune ha delle aree da bonificare?

Dopo aver mostrato le informazioni più rilevanti del dataset, il lettore viene lasciato libero di navigare al suo interno, scoprendo i dati che potrebbero interessarlo di più ossia quelli relativi al suo comune di appartenenza. Qui le informazioni vengono presentate in modo più completo, per la prima volta infatti compare anche il nome del sito contaminato e il suo indirizzo. L'utente non è lasciato solo in questa esplorazione ma può servirsi di un filtro per cercare tutti i dati del comune a cui è interessato.

Pubblicazione

Una volta completata la Story con il titolo e le didascalie necessarie per spiegare il funzionamento dei diversi filtri, la visualizzazione è stata pubblicata sulla piattaforma riservata agli utenti di public.tableau.com[79].

Sviluppi

Il progetto Terra Malata potrebbe evolversi in almeno tre direzioni, potenzialmente complementari fra loro.

Questo studio può essere arricchito con delle informazioni che raccontino in modo ancora più dettagliato la situazione dei terreni da bonificare in Lombardia. Si potrebbe quindi inserire dei dati circa l'ampiezza delle aree in oggetto oppure riguardo i diversi tipi inquinanti riscontrati.

La stessa modalità di analisi potrebbe espandersi ad un territorio più ampio, utilizzando il metodo illustrato per coprire tutte le regioni italiane.

La terza strada possibile è quella della multimedialità. La data visualization in oggetto potrebbe essere dunque il punto di partenza per un lavoro di divulgazione più esteso, arricchito con articoli di approfondimento, fotografie e videointerviste.

79 https://public.tableau.com/profile/valerio.berra#!/

Illustrazione 78: Visualizzazione 1. Dove si trovano i terreni contaminati?

Illustrazione 79: Visualizzazione 2. Quali zone sono più inquinate?

Illustrazione 80: Visualizzazione 3. Il mio comune ha delle aree da bonificare?

Appendice 2
Interviste

1. Un giro di nera tra i dati. Daniele Bellasio

Caporedattore de *Il Sole 24 Ore*, studia i Big Data e conduce Radiotube su Radio24 dove si occupa di tv e social network.

Cos'è il data journalism?

Prima della definizione comincerei con una premessa. Io mi sono avvicinato a questo mondo perché avevo la sensazione che il giornalismo interpretasse le potenzialità dei dati soltanto in una direzione. Il processo che esisteva all'interno dei giornali poteva infatti essere riassunto così. Ci sono dei dati, qualcuno li studia, qualcuno li mette graficamente in modo comprensibile per il lettore e alla fine arriva il giornalista e scrive un articolo.

Secondo me questo approccio ai dati è sbagliato. Prima di tutto il giornalista deve essere presente già all'origine della ricerca, per fare delle domande e interrogare i dati a prescindere da quello che un analista può ritenere importante. È il giornalista infatti che decide la storia che vuole raccontare e quindi si pone delle domande e allo stesso tempo le pone ai dati.
La seconda ragione per cui mi sono avvicinato a questo mondo è perché mi sembrava che nelle redazioni tutto questo ambiente fosse interpretato come qualcosa di competenza dei grafici. Questo è un errore madornale. C'è un po' nel mondo dei giornali l'idea che i dati non riguardino i giornalisti che invece si occupano solo di scrivere gli articoli.

La mia definizione di data journalism quindi è doppia. Non è soltanto la riproposizione in forma grafica di storie raccontate attraverso i dati ma è anche la ricerca nei dati di storie da raccontare, magari soltanto con un articolo, senza neanche un'infografica. Bisogna bloccare quel meccanismo, forse un po' troppo facile, che dai numeri porta subito a creare un grafico. Occuparsi di data journalism quindi vuol dire fare giornalismo attraverso lo studio, l'analisi e la riproposizione in forma comprensibile al lettore di storie nate e vissute attraverso lo studio dei dati ma vuol dire anche utilizzare i dati come fonte di ricerca delle notizie. Trovare nuove storie da raccontare in un mondo in cui ormai le notizie sono delle commodity, sono ovunque. I dati sono miniere di notizie, miniere che a volte vengono create inconsapevolmente, come i dati che provengono dai social network, mentre altre volte invece sono il frutto del lavoro di chi li raccoglie, li immagazzina e li pulisce.

C'è poi una sottodefinizione che secondo me è importante. Il giornalista da sempre deve avere una naturale tendenza allo scetticismo nei confronti della forma con cui le istituzioni propongono al pubblico i loro dati. C'è bisogno che sia un po' segugio, che interroghi anche lui i dati, magari aiutato da una squadra di analisti. Anche se facendo così non ribalterà i risultati, potrà sempre trovare nuove sfumature su cui riflettere.

Questo atteggiamento può essere rivolto anche alle metodologie di raccolta. Se io conosco i dati, posso fare domande su come vengono ottenuti.

Esatto

Passiamo ad un altro argomento. A quali nuove domande possono rispondere tutte queste miniere di dati?

È una questione complessa. Sono tre gli approcci che i giornalisti possono avere nei confronti dei

dati. Il primo è quello di verificare con i dati delle teorie, delle riflessioni formulate da chi scrive l'articolo o semplicemente diffuse fra le persone. Magari posso realizzare una serie di interviste a dei passanti su un argomento e poi verificare quanto quelle opinioni rispondano ai fatti.

Il secondo tipo di approccio invece è quello di cui parlavamo prima: capire cosa c'è nei dati, non fermarsi soltanto all'interpretazione ufficiale.

L'ultimo modello per il data journalism parte invece da un'altra premessa, proviamo a fare un esempio. Io voglio raccontare qualcosa di nuovo sull'amministrazione Obama. Qualcosa che non è ancora stato racconto. Decido allora di lavorare insieme a informatici e analisti e quindi comincio a raccogliere dati e cercare di rispondere a delle domande, magari anche banali e non collegate tra loro. Guardo ad esempio le attività dei social network, i picchi delle ricerche in rete oppure dove ci sono state più richieste di sostegno per la raccolta fondi. Una sorta di pesca a strascico. Facendo così comincio a navigare nei dati, comincio a vederli e allora scopro anche qualche traccia interessante da seguire. La storia in questo caso non nasce da una domanda precisa ma da un'esplorazione. È una sorta di giro di nera. Il giro che i giornalisti sono abituati a fare tra le loro fonti per capire se c'è stato qualche evento di cronaca di cui scrivere. Chiaramente una ricerca del genere deve essere ineccepibile nella forma per essere valida. Bisogna darsi dei parametri da seguire e bisogna essere onesti.

In questo caso non c'è forse il rischio di trovare correlazioni fra elementi che non hanno rapporti tra loro? Ad esempio, poniamo il caso che durante la presidenza di Obama sia aumentato il tasso di mortalità in diverse colonie di pinguini in Antartide. Magari la corrispondenza è reale ma la casualità no.

Ed è proprio per questo motivo che i giornali devono cambiare le loro strutture.

In che modo dovrebbero farlo?

Il giornalista per natura si appassiona alla sua tesi. Se è bravo preferisce che venga smentita, così da poter raccontare un'altra storia. Se è pigro invece non vede l'ora di trovare una conferma da parte dei dati. All'interno della redazione c'è quindi bisogno di una figura esperta di statistica per poter riuscire a portare a termine le operazioni di cui abbiamo parlato. C'è bisogno di qualcuno che posso vigilare sul processo di raccolta e analisi dei dati e che possa correggere i risultati senza valore matematico. Oltre a lui sarà necessario però consultare anche un esperto di dominio, qualcuno che conosca a fondo l'argomento.

Il lavoro del giornalista sarà mettere assieme tutte queste competenze in un racconto fruibile agilmente per il lettore. E qui entra in campo lo storytelling. Una quantità di informazioni così complesse avrà bisogno di un lavoro di semplificazione maggiore rispetto ad un comune articolo, un lavoro per cui a volte solo le parole potrebbero non bastare. Occorrerà magari usare un'infografica, una conversazione social interattiva oppure ci potrebbe essere bisogno di continuare ad aggiornare l'articolo in tempo reale. La storia potrebbe anche arrivare a vivere di vita propria.

Possono essere un esempio le data visualization dove i dati vengono aggiornati automaticamente.

Molto spesso i siti che fanno degli open linked data un vanto sono fantastici ma risultano comprensibili soprattutto agli analisti. Proprio lì invece, dove c'è l'aggiornamento in tempo reale, c'è

bisogno del giornalista che spieghi, che analizzi e racconti meglio quello che si vede. È anche un indice di come sia cambiato questo lavoro. Se tu dieci anni fa mi avessi chiesto con chi parlavo di più durante la giornata, ti avrei risposto con il direttore e con i miei collaboratori. Ora invece ti direi con il responsabile della sezione poligrafica.

Simon Roger, fondatore del datablog del *The Guardian*, comincia il suo libro *Facts are Sacred*[80] affermando che il data journalism magari ora è di moda ma non è nulla di nuovo nel mondo dell'informazione. Questo tipo di giornalismo avrà vita lunga o è una bolla destinata a scoppiare?

Se lo leghiamo solo alle infografiche, allora è una moda ed è collegata ad una esigenza di mercato molto precisa. La pubblicità sui siti e sulle app costa poco perché viene interpretata dall'utente come molto invasiva, molto disturbante e perché può essere facilmente esclusa, chiudendo i pop up o installando appositi programmi. Gli editori come provano a risolvere questo problema? Stanno cercando di vendere pubblicità non basandosi su quante persone riescono a raggiungere ma su quanto tempo riescono a far restare il lettore su un contenuto. Lo storytelling e l'utilizzo dei dati in forma dinamica e interattiva nascono quindi per spingere l'utente a dedicare più tempo possibile ad una notizia, lasciandogli anche la possibilità di esplorala in autonomia.

Questa però è una moda, perché magari in futuro esploderà il sistema di pagamento per i giornali online, l'editoria riuscirà a finanziarsi senza le pubblicità ma con le sottoscrizioni e allora le infografiche torneranno ad essere un fenomeno di nicchia. Se i dati verranno usati per raccontare storie che non si trovano altrove allora potranno avere ancora un mercato. I giornali anzi potrebbero specializzarsi solo su questo e fornire un servizio esclusivo simile a quello che ora viene promosso da una piattaforma come Netflix. Le infografiche sono molto legate poi allo strumento che si utilizza, ci sono dei lavori bellissimi che si possono vedere solo su computer ma la fruizione dell'informazione è sempre più orientata al settore del mobile. Mi spiace che quando si parla di dati nel giornalismo ci si orienti sempre sulle infografiche, dovremmo iniziare a concentrarci anche sull'altro fronte, quello del giro di nera, quello delle storie dei dati.

80 S. Rogers, *Facts are Sacred*, 2013

2. Raccontare il mondo con i numeri. Davide Casati e Andrea Marinelli

Giornalisti del *Corriere della Sera* e autori di Datablog, la sezione del Corriere.it nata nell'ottobre 2015 che si occupa di data journalsim.

> Confessiamolo: i dati spaventano. Numeri, cifre, grafici sembrano avere, su molti di noi, un potere di intimidazione. Eppure quelle tabelle sanno raccontare il mondo che viviamo. Offrono angoli di visuale nuovi sulle nostre vite, spiegano ciò che ci interessa e ci emoziona, descrivono quello che appassiona o incuriosisce. Nascoste dietro un volto freddo, insomma, fioriscono storie.
>
> Per questo nasce questa nuova sezione del *Corriere della sera*: il **Datablog**.
>
> L'obiettivo è lanciarci, con voi, in una «caccia al tesoro» per sorprendere - dietro ai dati - direttrici, tendenze e volti della società, dell'economia, dello sport, della salute.
>
> […]
>
> Tra i dati, le firme del *Corriere* hanno già scovato decine di storie. Ma i veri protagonisti di questo blog siete voi: potrete giocare con dati e infografiche e - perché no - suggerirci nuovi percorsi.
>
> Buona lettura e buon divertimento!

Testo 1: Introduzione Datablog Corriere della Sera[81]

C. Davide Casati
M. Andrea Marinelli

Come descrivereste il data journalism?

M. È complesso dare una definizione di Data Journalism. Potremmo dire che è tutto quello che nasce dai dati. Daniele Bellasio de *Il Sole 24 Ore* ne ha dato una definizione perfetta: è il nuovo giro di nera per i giornali. In un momento in cui le testate pubblicano più o meno le stesse notizie, si possono cercare storie che partono dai numeri, da intuizioni che mostrano dei pattern o dei trend a cui non si poteva pensare senza avere a disposizione tutti i dati.

C. Io aggiungere che per i giornalisti non è un'invasione di campo ma la scoperta di un mondo che per moltissimo tempo è stato trascurato. Storicamente il giornalismo italiano si è sempre percepito molto più letterario che numerico, più basato sulla bella scrittura che sull'analisi del dato. Ci stiamo accorgendo di quante storie stanno nascoste dietro i dati.

Come si raccontano queste storie nascoste nei dati, quali competenze servono per poterlo fare?

C. Sicuramente sono necessarie molte più competenze, anche solo per trovare dei database che siano utilizzabili. Spesso quando si parla di data journalism ci si dimentica di quanto sia difficile trovare delle fonti. In Italia poi è davvero complesso. Molte volte i dati sono sepolti in un file pdf e sono quindi difficili da utilizzare. La prima competenza necessarie è la capacità di individuare la fonte, anche se è un tipo di fonte diverso da quelle a cui eravamo abituati.

81 Datablog Corriere.it, http://www.corriere.it/datablog/i-numeri-che-mangiamo/latte/scheda-1.shtml

Senza imbattersi in file pdf, spesso anche quelli che già si trovano in formati più comodi hanno bisogno di un processo di normalizzazione.

La normalizzazione dei database a volte è diventa praticamente un secondo lavoro. È un problema tanto diffuso che proprio l'altro giorno stavamo notando che addirittura i ministri fanno fatica a trovare dei database che siano normalizzati. Carlo Cottarelli[82] ad esempio denunciava non solo che i dati non gli arrivavano ma che quelli a sua disposizione erano imparagonabili. Un'altra competenza fondamentale è poi la capacità di visualizzarsi prima la visualizzazione. È un passaggio complesso. Bisogna infatti tentare di trovare una modalità di comunicare il dato che non sia semplicemente bella ma che sia anche funzionale, che possa guidare il lettore attraverso i dati. Può sembrare banale ma fino a poco tempo fa si chiedeva al grafico di disegnare una pagina, non quale fosse il modello di visualizzazione migliore per trasmettere un messaggio.

Proprio a questo proposito il giornalista inglese David McCandless ha aperto informationisbeautiful.net[83], un blog in cui mostra come si può creare delle infografiche senza avere competenze specifiche nel web design.

M. Un punto interessante è proprio questo. Parlare delle visualizzazioni dal punto di vista del design e dal punto di vista dell'informazione. Spesso le infografiche create dai designer sono bellissime ma la notizia si perde nella loro bellezza, quelle invece meno belle ma più chiare aiutano molto il lettore. Secondo noi non sempre la bellezza si muove di pari passo con l'utilità. Quello che è difficile poi è riuscire a mettere insieme tutti i vari pezzi. Ovviamente sono rare le persone che possono trovare il database, scrivere l'articolo, inserire i dati su Excel e anche visualizzarli. Ci sono diversi software che possono aiutare in questa ultima fase del processo, come ad esempio Tableau, ma non sempre si possono utilizzare.

C. La capacità di lavorare in squadra infatti è un'altra competenza che serve. Tentare di capire dove sta andando il lavoro o anche come lasciarsi guidare dal lavoro. A volte si parte con un'intuizione e poi magari in corso d'opera ci si accorge che questa non è supportata dai dati, si scopre che la notizia sepolta nel database va cercata in tutt'altra direzione. La capacità di visualizzare i dati invece diventerà sempre di più una competenza richiesta. Per noi ora si tratta di una scoperta in un campo del tutto nuovo e io credo che sarà necessaria anche un'educazione all'arte della visualizzazione. Adesso ci stiamo ubriacando della bellezza e delle possibilità che offrono le tecniche a nostra disposizione. Si possono realizzare lavori stupendi che una volta messi in pagina sembrano quasi dei quadri. Piano piano si metterà in moto un'opera di pulizia per rendere quella visualizzazione più funzionale, per creare una distanza simile a quella che esiste tra letteratura e giornalismo. È chiaro ad esempio che si può inserire negli articoli delle suggestioni letterarie ma non si possono certo riportare le notizie in versi. Nel giornalismo questo processo avviene da tanto tempo. Siamo abituati a capire quando stiamo esagerando con i voli pindarici e quando invece in un pezzo è interessante inserire un elemento in più di colore. Con le visualizzazioni è ancora tutto nuovo per noi.

«Confessiamolo: i dati spaventano» è la frase con cui avete iniziato l'introduzione del Datablog. Perché i dati provocano questa reazione, o meglio, perché voi siete partiti da questa premessa?

C. La mia sensazione è che siamo un Paese che forse ha più cultura letteraria che cultura scientifica. C'è l'impressione che i numeri siano aridi, che sia sempre meglio scrivere un tema che svolgere un

82 Commissario Straordinario per la Revisione della Spesa Pubblica nominato dal Governo Letta
83 http://www.informationisbeautiful.net/

compito di matematica.

“La matematica non sarà mai il mio mestiere”.

C. Esatto. Questa tendenza è stata riconosciuta universalmente, tanto da finire anche nelle canzoni. I dati poi sono associati ai numeri e i numeri si collegano spesso alle seccature, sono quelli che si trovano sulla bolletta o sulla divisione in millesimi dei palazzi. Allo stesso tempo però i numeri sono qualcosa che gli scienziati inseriscono in uno scenario di bellezza. Fabiola Gianotti ad esempio lavora al Cern ed è la responsabile dell'esperimento che ha portato alla scoperta del Bosone di Higgs. Per lei la fisica è bellezza. Ha frequentato il liceo classico e poi è finita a studiare fisica. Forse è una visione poco intuitiva ma sono convinto della sua validità. Noi però non cerchiamo la bellezza che si trova nei dati, cerchiamo le loro storie. In questo senso il reperimento di un trend si associa subito con una possibile notizia. Trovare un trend però non basta, bisogna anche capire come comunicarlo. Se metti una tabella sul giornale, questo non aiuta il lettore a capire meglio la notizia anzi, il giornale viene percepito come ancora più difficile.

Quali software avete scelto per costruire le vostre visualizzazioni?

M. Sono state le circostanze a dettare le nostre decisioni. Quando ci siamo avvicinati al progetto, all'idea di fare data journalism, abbiamo scelto gli strumenti che per noi erano più semplici da utilizzare. La base ovviamente è Excel, per motivi evidenti, poi abbiamo usato Tableau perché ci permetteva delle visualizzazioni standard ma allo stesso tempo complesse e visivamente gradevoli e ora siamo passati a strumenti interni alla nostra redazione. Non siamo fissi su un solo software ma cerchiamo di sperimentare fino a trovare quello giusto. Ovviamente si tratta di una ricerca continua ma è questa la base del Datablog.

Passiamo ora all'architettura di questa sezione del Corriere.it. Voi siete partiti con un argomento, il cibo, e ne avete parlato per cinque settimane. Ogni settimana veniva trattato un cibo diverso, dal cacao alla farina, e oltre alle infografiche venivano pubblicati anche articoli scritti da altri giornalisti. Ora invece vi state orientando verso lavori più circoscritti.

M. Tenere lo stesso argomento per più tempo era una richiesta che ci è stata fatta dalla redazione. Ora noi cerchiamo di alternare storie fredde, come quella del cibo, e storie che invece siano più in linea con il tema del giorno o della settimana. Proviamo anche ad alternare sia le tipologie di dati che le visualizzazioni, non vogliamo tenere uno schema fisso per non annoiare chi legge e chi scrive.

C. Su questo tema ci stiamo ancora prendendo un po' la mano. Noi sappiamo di certo che ci saranno notizie molto guidate dai dati, altre volte invece cercheremo dei dati in base alla notizia del giorno. Ci saranno comunque ancora progetti più lunghi focalizzati sullo stesso argomento. Forse quello che verrà più spontaneo ad un giornale sarà partire da un evento e poi cercare una serie di dati. Il progetto sul cibo ad esempio lo abbiamo lanciato perché avevamo un gancio con Expo.

M. Oppure ancora ancora ci saranno focus tematici in cui un articolo partirà dai dati e altri verranno collegati per argomento. Ad esempio il 25 novembre in occasione della Giornata Mondiale Contro la Violenza sulle Donne siamo partiti da un articolo che nasceva dai dati a nostra disposizione a cui abbiamo collegato un articolo dal taglio giuridico e due storie. A me piacerebbe che il Datablog sia anche una hub multimediale per il giornale in cui possano trovare spazio vari tipi di formati.

Perché avete deciso di interessarvi al data journalism?

C. La spinta è arrivata fondamentalmente dal riconoscimento dell'esistenza di una filone di notizie che non stavamo prendendo. Forse è sbagliato parlare di filone di notizie, sarebbe meglio dire un filone di racconto delle notizie che non stavamo prendendo. Ed era un filone interessante, non si trattava solo di una meteora. Il processo è stato simile all'introduzione del long form in Italia. È una cosa che vedi prima dall'esterno, soprattutto dai siti stranieri e quindi cominci a fruirne. A un certo punto ti rendi conto che in quelle visualizzazioni c'è qualcosa che non sparisce, che esiste un pozzo che non si sta ancora esaurendo e quindi in qualche modo ti viene voglia di provare. All'inizio è complicato perché, come dicevamo prima, si tratta di un processo che coinvolge delle competenze diverse. Implica anche un cambiamento nella struttura redazionale che non è certo banale. Però tutto parte dal desiderio di fare anche tu qualcosa che hai apprezzato da altre parti. La data visualization permette di trattare una notizia in un modo che altrimenti è impossibile.

Da quali fonti attingente per i database?

M. La prima risposta è Google. Chiaramente cerchiamo di utilizzare il più possibile i database ufficiali. Per qualsiasi articolo poi specifichiamo anche la fonte.

Cosa si può scoprire con la data visualization? Oltre a dare soluzioni a vecchi problemi, come ad esempio la possibilità di mostrare insieme dati che altrimenti sarebbero stati solo elencati, quali campi di esplorazione si aprono con questo strumento?

C. Sono letteralmente una valanga. Secondo me ora è anche difficile dirli tutti perché è la quantità di dati a disposizione che è cambiata. Noi siamo davanti ad una diffusione di dati che prima era impensabile, ogni cosa in questo momento è digitalizzata. Siamo diventati delle sonde ambulanti perché giriamo con aggeggi che indicano dove siamo, cosa stiamo facendo. Sono tutti dati che prima semplicemente non esistevano. Io non so quali campi si possano aprire. I primi che mi vengono in mente sono gli ambiti in cui i dati c'erano già ed eravamo abituati ad averli ma gli orizzonti più interessanti saranno quelli in cui adesso i dati non si possono neanche immaginare. Come ad esempio potrebbe accadere nel campo delle digital humanities.

Il fatto che tu abbia un dato e che tu possa lavorarci non vuol dire però che tu ci possa trovare un'informazione significativa. Se io mi mettessi a studiare il rapporto con mia moglie attraverso un'analisi dei dati, forse non ci troverei nulla di significativo. Questo non sarebbe però un problema dello strumento ma un problema di scelta dello strumento. Abbiamo in mano ora una risorsa enorme e ora proveremo a sfruttarla in tutti i campi possibili, a giocarci come bambini. È un mondo enorme di possibilità, quasi vertiginoso.

C'è anche un po la ricerca del Santo Graal in questi Big Data...

C. Sì, una delle tentazioni da evitare è la convinzione che dentro lì ci sia la riposta a tutto. Sicuramente il libro del mondo è scritto anche con i caratteri matematici. Che questi però lo esauriscano è tutt'altro che certo.

3. Il data scientist nell'era dei Big Data. Mario Mezzanzanica

Professore associato di Information System presso l'Università degli Studi di Milano Bicocca. Dal 2005 è direttore scientifico del CRISP, il Centro di Ricerca Interuniversitario per Servizi di Pubblica Utilità alla Persona. Dal 2010 è direttore e presidente del comitato di coordinamento del in Business Intelligence and Big Data Analytics promosso dall'Università degli Studi di Milano-Bicocca.

Cosa sono i Big Data e come sono strutturati?

In termini di definizione parlare di Big Data oggi significa parlare di capacità di organizzare infrastrutture nuove e complesse che mettano a disposizione grandi quantità di dati in tempo reale. All'interno di questi enormi insiemi ci sono dati strutturati, semistrutturati e non strutturati.
Il termine Big Data nasce per un problema di metrica. Si era arrivati ad un livello di disponibilità del dato tale per cui una metrica tradizionale non riusciva più a rappresentarlo in modo significativo.

Cosa si intende per metrica?

Per metrica intendo la tradizionale scala di classificazione dei dati, di cui fanno parte ad esempio megabyte e gigabyte. Riferendoci ai Big Data useremmo numeri così elevanti che sarebbe difficili renderli comprensibili a tutti. Un'unità di misura non è utile solo per fissare degli standard ma anche perché diventa uno strumento di comunicazione. Il termine Big Data è nato proprio perché c'era un problema di comunicazione rispetto alla quantità di dati oggi disponibili.

Perché questi Data stanno diventando sempre più Big?

Sono in corso diversi cambiamenti dal punto di vista economico e sociale. Noi siamo nella cosiddetta era dell'informazione, in cui i dati stanno diventano un fattore strategico fondamentale per le aziende e per il sistema sociale. Da questo punto di vista la capacità di utilizzare, fruire e valorizzare la conoscenza che può derivare dai dati è un fattore importante. Se prima per le aziende il vero patrimonio era il capitale a disposizione, ora la risorsa principale sta diventando l'informazione, soprattutto in certi settori del mercato.

In secondo luogo c'è stato un aumento dei dati a disposizione per il sistema dei servizi o sistema sociale. Attraverso i canali del web c'è una possibilità di espansione molto rapida e i dati diventano sempre più importanti, anche per scopi personali. È sufficiente pensare che quando si cerca un ristorante ci si può servire di una quantità enorme di informazioni, come ad esempio le recensioni degli utenti che ci sono già stati. Il dato e l'informazione che se ne ricava possono diventare quindi un fattore interessante non solo per le imprese ma anche per le persone e quindi per il sistema socio economico nel suo complesso. È qui che i Big Data diventano fondamentali.

Come si inserisce il Crisp in questo panorama?

Il Crisp è un centro di ricerca che si occupa di studi sui servizi di pubblica utilità con un approccio multidisciplinare. Ci sono competenze legate alla statistica, alla tecnologia dell'informazione, all'economia e alla giurisprudenza. Nel corso della sua storia il Crisp ha sempre affrontato i fenomeni di cui si è occupato studiandoli con un'analisi quali-quantitativa, cercando di utilizzare dati sempre più rilevanti. Le nostre fonti spaziano infatti dai dati statistici ai dati amministrativi,

fino ai dati provenienti dalle aziende o dal web. Noi cerchiamo sempre di entrare nella struttura più complessa del dato.

Perché è nato il master in Business Intelligence and Big Data Analytics?

Questo master è nato cinque anni fa per due motivi. Da una parte si stavano evolvendo gli studi che portavamo avanti nel nostro centro di ricerca, dall'altra invece molte aziende con cui noi eravamo abituati a dialogare stavano cercando figure professionali che fossero capaci di utilizzare i dati nel contesto lavorativo, persone quindi in grado di seguire tutto il ciclo dei dati. Originariamente la nostra offerta era rivolta a neolaureati provenienti da ambiti tecnico-scientifici ma in questi anni abbiamo ricevuto anche parecchie richieste di persone che avevano già cominciato da qualche anno a lavorare in azienda. In questo caso si tratta spesso di individui collocabili entro i primi cinque anni della loro attività lavorativa che vedono i temi che affrontiamo come un'opportunità di crescita.

Recentemente si sono affacciate anche persone che vengono da un panorama più umanistico. Noi le abbiamo accolte comunque per il grande interesse dimostrato verso gli aspetti più legati alla comprensione dei dati, in modo da poter meglio occuparsi della parte interpretativa e di racconto. Abbiamo scoperto con estremo piacere che questi soggetti hanno fatto una bella fatica sotto il profilo tecnico ma i riscontri sono stati estremamente positivi, hanno potuto infatti acquisire le informazioni necessarie per cogliere quali sono i fattori in gioco nel trattamento dei dati e allo stesso tempo valorizzare tutta la parte di interpretazione, storia e rappresentazione del dato.

Nei documenti di presentazione del master viene tratteggiata la figura professionale del data scientist. Quali sono le sue competenze?

Ci sono ancora molti dubbi riguardo la definizione reale. Da quello che ho avuto modo di imparare, il data scientist è una figura professionale che sa mettere insieme competenze multidisciplinari finalizzate al trattamento del dato. Per trattamento del dato intendo la capacità di definire i dati per la loro significatività e quindi la capacità di rappresentare una popolazione.

Esistono quindi due aspetti legati a questo compito. Il primo è un problema più statistico, ossia quello di trattare tecnicamente il dato. Il secondo invece riguarda più la qualità dei dati e la possibilità di rendere il dato significativo per rappresentare un contenuto.

Tutto questo processo definisce il ruolo del data scientist nell'azienda. Potremmo dire ancora che questa è un persona in grado di valorizzare con un approccio multidisciplinare un patrimonio di informazioni per rendere un servizio ai diversi dipartimenti aziendali. Per far sì quindi che i dati diventino realmente un valore per le attività dell'azienda. In questi passaggi entrano in gioco competenze di natura statistica, informatica, grafica e inerenti anche all'ambito della comunicazione.

Cosa può dare al mondo dei Big Data chi proviene dall'ambito umanistico?

Secondo mi ci sono due fattori in gioco molto interessanti. Prima di tutto il concetto che abbiamo utilizzato prima di competenza multidisciplinare è sostanziale, non formale. Questo vuol dire che oggi per fare determinati tipi di lavori occorre mettere insieme diverse competenze, una dinamica che può verificarsi solo se si crea uno strato comune di linguaggio. Bisogna avere dei concetti comuni, delle conoscenze sopra le quali si può costruire l'interdisciplinarità.

Senza scomodare Ford, in questo ambito il concetto della catena di montaggio è abbastanza superato?

Esattamente, questo fatto oggi è evidentissimo nel trattamento di qualsiasi dato in un sistema complesso. Questo è un fattore importante. Avere una multidisciplinarietà dei discenti crea una coscienza di questo strato comune. In secondo luogo c'è poi l'apporto della specificità. Facciamo un esempio. L'apporto della specificità fa capire ai tecnici che non basta essere bravi a trattare i dati, bisogna capire a quali domande rispondere per far sì che il trattamento dei dati sia davvero un supporto alla conoscenza. Dalla parte opposta invece, chi gestisce la conoscenza deve avere il supporto di chi può garantire un dato di qualità, l'informazione generata sarebbe altrimenti forviante.

Nuovi strumenti aprono nuove prospettive. Quali possibilità ha la data visualization di fornire un punto di vista del mondo ancora inedito?

C'è un problema di fondo di natura culturale da affrontare. In gioco c'è sempre di più la possibilità di capire cosa fa nascere l'informazione e cosa crea conoscenza. Quello che fa nascere informazione è la disponibilità del dato ma ciò che rende informativo il dato è la domanda che si ha. Da un punto di vista culturale non si può pensare che una dato sia direttamente collegato al verificarsi di un fenomeno.

Questa interpretazione diventa estremamente interessante perché rinforza un paradigma di approccio tipico della statistica, ossia che è la domanda che pongo a dirmi quali informazioni sto cercando e quali dati possono aiutarmi. L'altro passaggio molto delicato è capire come raccontare le risposte a queste domande. Questo è il tema delle infografiche, uno strumento che sicuramente ha ancora molta strada da fare per quanto riguarda la fruizione da parte dell'utente. Anche qui però le domande che si pongo sono importanti. Scorrendo le infografiche che vengono utilizzate adesso, si capisce come certe nascano per rispondere a delle domande ma non ad altre. Se non viene chiarita la domanda a cui si vuole rispondere è poi difficile raccontare una storia. Certo un'infografica deve essere anche bella e intuitiva ma l'intuizione è reale quando è chiara la domanda di partenza, non il contrario. La strada da seguire è proprio in questa direzione.

Il punto di partenza quindi è che le infografiche non siano rappresentazioni visive di tabelle Excel ma storie che nascono da domande.

Dovrebbe essere così, però molto spesso ora non è chiaro a quali domande risponde l'infografica. É come se statisticamente si prendessero dei dati costruiti per rispondere a delle domande e li si utilizzassero per rispondere ad altre. Questo secondo me è il rischio del data journalism oggi. Si tratta di una sfida culturale molto interessante che mette in gioco tecnica, tecnologia e modelli di analisi.

4. Dal visibile all'invisibile. Valentina Manchia

Membro del gruppo di ricerca Omar Calabrese dell'Università di Siena, studia le dinamiche di interazione tra verbale e visivo, negli ultimi anni si è occupata di information design.

Partiamo con una definizione. Cos'è la data visualization?

In semiotica visualizzazione dei dati significa rendere visibile quello che visibile non è. Questa è una delle definizioni che arriva dall'ambito della data visualization pura. Mi sto riferendo a Edward Tufte che nelle sue opere affronta questo rapporto tra visibile e invisibile.

Nella mia ricerca cerco quindi di mostrare le strategie di visibilità che di volta in volta vengono messe in atto dai soggetti che usano i dati. Chi generalmente si occupa di data visualization costruisce una narrazione a partire da dati quantitativi che cerca di convogliare per produrre un senso, una significazione. All'interno del mio lavoro io ho cercato di analizzare diversi casi in cui questo avviene. Ho preso in considerazione designer puri, team giornalistici al lavoro con i dati e anche diversi artisti visivi che hanno agito in questo senso.

La procedura a mio parere parte sempre dall'individuazione di un campo di attenzione, chiamiamolo un dataset. Una volta che i dati sono a disposizione, il lavoro da fare è quello di porre una domanda a questo insieme di dati. Una domanda che è già un orientamento e quindi produce non una risposta, ma un taglio, uno sguardo, un punto di vista su questi dati. Il risultato finale deriva da questo punto di vista ma è allo stesso tempo una messa in forma. Qui infatti entra in gioco la parte di design che a mio parere è importantissima ed è il momento in cui si costruisce la narrazione. Da domande e da scelte di narrazione diverse nascono visualizzazioni diverse.

Perché occuparsi di questo tema dal punto di vista semiotico?

Capire come si passa dai dati puri ad un discorso sui dati non è banale e per questo l'interesse della semiotica è altissimo. Dove c'è traduzione, dove c'è costruzione, dove c'è narrazione non c'è chiaramente corrispondenza tra il dato e quello che salta fuori. Questo meccanismo è affascinante per chi studia la mia materia. Non troverai mai un semiologo che ritenga un oggetto talmente banale da non dover essere analizzato. Gli oggetti trasparenti, quelli che non hanno nulla da rivelare, non esistono ed è particolarmente interessante riflettere su questo aspetto quando si parla di visualizzazione dei dati.

La retorica che spesso gira intorno a questi argomenti infatti è che il giornalismo basato sui dati sia un giornalismo obiettivo che non ha bisogno di traduzioni o di articoli di accompagnamento. La famosa frasetta secondo cui i dati parlano da soli è a mio parere da un lato pericolosa, perché di fatto non è così, dall'altro è semioticamente interessante. Ed è proprio quando ci troviamo davanti ad un'affermazione del genere che mostrare i meccanismi retorici diventa stimolante. Io lavoro su questo modo di costruire narrazioni, un modo che sarà sempre più diffuso visto la direzione che stanno prendendo i media.

La tua ricerca è un'eccezione o esistono contributi di altri semiologi sulla data visualization?

Quando io ho cominciato a scrivere di questi temi era il 2010 e, almeno in Italia, non ho trovato altri colleghi che si occupassero esattamente di questi oggetti dal punto di vista che mi interessava. Devo però sottolineare che da metà degli anni duemila ci sono diversi studiosi che hanno cominciato ad

occuparsi di semiotica del web, soprattutto per quanto riguarda la costruzione del discorso e del layout collegato alla presentazione di informazioni. Giovanna Cosenza, Isabella Pizzini e Leonardo Romei ad esempio stanno lavorando molto in questo ambito.

Come riesce un'immagine a raccontare una storia?

Anzitutto le data visualization sono immagini come le altre, quindi sono sottoposte anche loro alle stesse leggi. Sono oggetti bidimensionali, hanno dei colori, hanno delle forme e hanno delle strutture che possono essere lette come in ogni altro prodotto visivo. Anche una fotografia o un quadro si confrontano con regole di composizione e attirano l'attenzione in modi molto simili. Queste visualizzazioni hanno quindi tutti elementi che la semiotica classica delle immagini è già abituata a prendere in considerazione.

Questa era la risposta più generale, ora scendiamo nello specifico. Per quanto riguarda le dataviz, la capacità di raccontare una storia è strettamente legata all'interazione fra testo e immagini. È estremamente difficile avere una buona dataviz di sole immagini e niente testo. A mio parere la ricchezza di questo tipo di comunicazione è proprio la capacità di far dialogare assieme immagini e testo. Il testo però non può essere solo una didascalia dell'immagine, così come l'immagine non può essere solo accessoria al testo. È piuttosto un meccanismo composto da ingranaggi che si compenetrano.

Poi chiaramente possono anche essere analizzati tutti gli elementi visivi che fanno parte di una dataviz. Si può studiare il loro impatto, le loro caratteristiche, la texture, le derivazioni iconografiche o il rapporto fra le diverse figure. Con questi elementi le visualizzazioni statiche riescono a costruire una narrazione a strati, mentre per le visualizzazioni dinamiche bisognerebbe affrontare altri modelli e il discorso diventerebbe ancora più complicato. Per riuscire a raggiungere l'obiettivo di creare una narrazione su un'immagine statica è importante secondo me guardare anche verso altri linguaggi, come ad esempio quello del fumetto. Ci sono autori come Gianni De Luca che hanno lavorato molto sulla possibilità di rappresentare una narrazione complessa in una sola pagina.

Nei tuoi saggi parli spesso di narrazioni oneste, cosa intendi con questa parola?

Una visualizzazione è onesta quando dichiara il più possibile i meccanismi e gli intenti che le stanno dietro. Ad esempio il quotidiano inglese *The Guardian* è stato estremamente attento nel mostrare tutti i passaggi che hanno portato dall'acquisizione del data set di WikiLeaks alle visualizzazioni che sono state pubblicate. Onestà ancora più conclamata perché nel loro lavoro finale c'erano link dappertutto dove era possibile scaricare gli spreadsheet dei dati di cui si stava parlando. L'idea che sta alla base di questa scelta non è solo quella di essere chiari e trasparenti nei confronti dei propri lettori ma anche quella di renderli consapevoli che la visualizzazione delle informazioni è una costruzione complessa. Nel momento in cui ci sono queste premesse, c'è anche onestà.

Il *The Guardian* sapeva però anche un'altra cosa quando ha deciso di mettere a disposizione i dati su cui ha lavorato. Sapeva di non rischiare niente nel mettere on line le sue fonti perché nessuno avrebbe trovato immediatamente gli stessi risultati. Pubblicare gli spreadsheet in loro possesso non avrebbe creato un doppione perché un quantitativo così esteso di materiale avrebbe potuto essere analizzato solo da uno specialista. Quando si presenta una visualizzazione molto creativa, tipica di alcuni quotidiani come USA Today, non avviene niente di tutto ciò. In questi casi ci sono grossi disegni, dimensioni che non sono collegate ad una quantità precisa e allo stesso tempo manca la

possibilità di accedere ai dati di partenza. Ecco, qui non ho delle buone visualizzazioni. Edward Tufte parlava addirittura di chart junk, grafici spazzatura.

Glossario

- **Dashboard** Pannello su cui sono accostati più grafici.
- **Data Design** Branca del design che si occupa di trovare soluzioni per la visualizzazione dei dati. Questa formula può indicare anche tutto il processo di visualizzazione dei dati.
- **Data Visualization** Trasposizione grafica di un dataset. Questa formula può indicare anche tutto il processo di visualizzazione dei dati.
- **Dataset** Collezione di dati solitamente strutturati.
- **Dataviz** Abbreviazione di *Data Visualization.* Si riferisce ad una forma grafica di rappresentazione dei dati.
- **Dato Non Strutturato** Dati non inseriti all'interno di schemi e tabelle. Sono esempi di *Dati Non Strutturati* le fotografie e i testi.
- **Dato Strutturato** Dati conservati in database organizzati secondo schemi e tabelle.
- **Infografica** Deriva dall'unione tra i termini *Information* e *Graphic*. Sinonimo di *dataviz* è una forma grafica di rappresentazione dei dati.
- **Normalizzazione** Nel linguaggio di gestione dei database è l'eliminazione della ridondanza informativa e del rischio di incoerenza del database.
- **Record** Si riferisce alla singola registrazione di un dato. All'interno di una tabella coincide solitamente con una riga.
- **SpreadSheet** Foglio elettronico usato per raccogliere, aggregare e svolgere le prime analisi su un dataset.

Bibliografia

- A. Cairo, *The Functional Art: an Introduction to Information Graphics and Visualization*, 2012
- A. Cairo, *The Truthful Art: Data, Charts, and Maps for Communication*, 2016
- B. Marr, Big Data: *Using Smart Big Data, Analytics and Metrics To Make Better Decision and Improve Performance*, 2015
- D. Bihanic, *New Challenge for Data Design*, 2015
- E. R. Tufte, *The Visual Display of Quantitative Information*, 2001
- E. Segel e J. Heer, *Narrative Visualization: Telling Stories with Data*, 2010
- F. Franchi, *Designing News*, 2013
- F. Moretti, *Graphs, Maps, Trees*, 2005
- F. Tissoni*, Social network, Comunicazione e marketing*, 2014
- K. Börner, D. E. Polley, *Visual Insights, A Pratical Guide to Making Sense of Data*, 2014
- K. Börner, *Atlas of Knowledge*, 2014
- K. Börner, *Atlas of Science, Visualizing What We Know,* 2010
- K. Börner, *Visual Interfaces to Digital Libraries*, 2003
- M. Santamaria-Varas e P. Martinez-Diez, *atNight: Nocturnal Landscape and Invisible Networks*, 2015
- N. Yau, *Visualize This*, 2011
- P. Meyer, *Precision Journalism: A Reporter's Introduction to Social Science Methods*, 1970
- S. Rogers, *Facts are Sacred*, 2013
- V. Manchia, *Il data journalism e la rappresentazione visiva delle informazioni tra trasparenxa e opacità. Gli AfghanWar Logs di WikiLeaks riletti dal Guardian*, in M. Serra, O. Gomez, *Transparencia y Secreto*, 2015

Sitografia

- http://datajournalism.stanford.edu/
- https://www.academia.edu/
- http://www.atnight.ws/
- http://bigbangdata.somersethouse.org.uk/
- http://www.co.lavoro.gov.it/
- http://www.corriere.it/
- http://www.crisp-org.it/
- http://daslombardia.crisp.unimib.it/
- https://www.dati.lombardia.it/
- http://daytum.com/
- http://www.densitydesign.org/
- http://ec.europa.eu/eurostat
- https://www.edwardtufte.com/tufte/
- http://ella.slis.indiana.edu/
- http://feltron.com/
- https://www.flickr.com/
- http://flowingdata.com/
- http://www.gapminder.org/
- https://gephi.org/
- http://www.infocamere.it/
- http://www.infodata.ilsole24ore.com/
- https://infogr.am/
- http://www.informationisbeautiful.net/
- http://www.inps.it/portale/default.aspx
- http://www.internetlivestats.com/
- http://www.istat.it/it/
- https://it.pinterest.com/
- http://www.last.fm/

- http://www.lastampa.it/
- http://www.mapofscience.com/
- http://moma.org/
- http://www.nytimes.com/
- http://parlamento17.openpolis.it/
- https://piktochart.com/
- https://piktochart.com/
- http://pitchinteractive.com/
- https://prezi.com/
- https://public.tableau.com/s/
- http://www.reporter-app.com/
- http://www.robweir.com/blog/
- http://scimaps.org/
- http://www.spiegel.de/
- https://storymap.knightlab.com/
- http://www.ted.com/
- http://www.theguardian.com/international
- http://tylervigen.com/spurious-correlations
- http://www.unioncamere.gov.it/
- https://venngage.com/
- http://www.visme.co/
- https://www.washingtonpost.com/regional/
- https://wikileaks.org/
- https://www.youtube.com/

Videografia

- G. McGhee, *Journalism in the Age of Data,* 2010. Il documentario completo si può trovare su www.youtube.com
- H. Rosling, *How not to be ignorant about the world*, 2014. La lezione completa si può trovare sul sito del progetto TED, www.ted.com.
- K. Börner e D. E. Polley, Data Visualization MOOC, 2013. Le lezioni complete si possono trovare su www.ivmooc.cns.iu.edu