Appunti Corso di Data Visualization

Leonardo Alchieri
Dipartimento di Informatica
Università degli Studi di Milano-Bicocca

Indice

1 Informazioni generali

Non si fa pausa e facciamo lezione insieme a un altro corso.

L'esame si basa sulla produzione di un progetto, unico con la parte di Management. Date scritti 16/01 e 03/02. Verranno fatto anche degli orali.

L'esame scritto ha il 40% del peso del voto, in particolare fatto di domande aperte per la parte di *Data Management* e domande a risposta chiusa per la parte di *Data Visualization*. Il restante 60% del voto viene dato dal progetto, che è da fare in gruppi di 3/4 persone.

Il progetto deve essere basato su una storia: dobbiamo raccontare qualcosa con i dati, dopo averli collezionati, processati e quindi descritti. In particolar modo, la visualizzazione deve essere fatta con lo strumento interattivo **Tableau**.

Come parte del progetto dovremo andare da delle persone ignare di quello che abbiamo fatto e fare una valutazione di qualità.

Ci sarà la possibilità di partecipare a 2 **esperimenti**, che riguardano la visualizzazione: come rappresentare l'incertezza. Si potranno ricevere al massimo 1.5 per l'esame finale.

Per avere un buon voto, si devono seguire le indicazione del professore ed evitare gli errori indicati dal professore.

È obbligatoria la lettura del saggio presente online, di titolo **Il Cosa, il Perché e il Come della Visualizzazione Dati.**

Dobbiamo soprattutto puntare a fare un ottimo progetto, dato che spesso le migliori vengono pubblicate su testate giornalistiche.

Le slide sono sufficienti per poter passare l'esame; è inoltre molto consigliabile prendere degli appunti durante la lezione.

Il professore consiglia di comprare uno dei libri che facciano capire come sono importanti i dati nel mondo.

Scrivere al prof Scrivere come oggetto [DATAVITZ RICEVIMENTO 2019] Metti anche matricola e il corso di Laurea. Il suo indirizzo email è

`federico.cabitza@unimib.it`

Lo si può anche scrivere al professore al `0264478888`.

Il professore ha un profilo `Twitter`: si deve scrivergli in Direct se si vuole far parte di un gruppo Twitter, in cui il prof. condividerebbe delle visualizzazioni dei dati.

2 Introduzione

Visualizzazione dei dati La visualizzazione dei dati è essenziale per poter visualizzare l'intero ciclo di vita dei dati, oltre che a quel di saper raccontare la storia dei dati che si presentano.

Una volta preparati dei dati, bisogna andare a esplorarli, per poter capire che tipi di elaborazione sono sensati, per poter tirar fuori delle considerazioni statisticamente sensate. Un'altra parte importante è la valutazione e la comunicazione dei risultati, in cui la **data visualization** permettono di concludere il lavoro che si è fatto e passare al pubblico/utente/committente il progetto.

3 Interazione Uomo-Dato

L'impostazione che prendiamo è quella di vedere come far interagire gli esseri umani con i dati, che sono sempre più.

Ciclo di vita del dato

- Raccolta
- Elaborazione, sia computazionalmente sia a mano con regole e/o procedure
- Analisi
- Condivisione, durante la quale si cerca di visualizzare il dato e renderlo comprensibile anche a persone che non hanno lavorato sul progetto.
- Uso

Vi può essere poi ovviamente un feedback, ovvero vi può essere raccolta dalle persone direttamente. In 3 fasi, la visualizzazione dei dati gioca un ruolo molto importante, ovvero nell'**analisi**, nella **condivisione** e nell'**uso**.

Importanza della visualizzazione Come mai è importante la visualizzazione, sia per la scienza che per il giornalismo della telecomunicazione. Per esempio, per uno scienziato è essenziale far capire quali siano i risultati più importanti raggiunti, e per provare un punto importante.

È molto difficile provare i propri argomenti senza la visualizzazione: essa diventa uno strumento per raccontare il vero. Si noti come, da un punto di vista negativo, questo possa essere usato anche per portare dei dati falsi, delle *fake news*, alla società.

Parole vs Immagini Ci sono delle ricerche che sostengono che una figura vale circa 60000 parole: secondo questo studio, gli elementi visuali sono elaborati dal nostro apparato cognitivo 60000 volte più velocemente, e ci vuole addirittura il doppio del tempo e leggere parole piuttosto che immagini.

Quando si tratta di *data visualization*, si andrà a vedere quanto efficienti saranno le immagini prodotte.

Data Science vs Data Visualization Bisogna avere un atteggiamento oggettivo e onesto nel confronto dei dati che si hanno a disposizione. La scienza si basa sul *metodo scientifico Galileiano*: analogamente, dobbiamo applicare un pensiero simile al trattamento e alla visualizzazione dei dati.

La data visualization si può organizzare come:

1. Formulare una domanda.
2. Fare un punto esplicito.
3. Sviluppare una procedura per mostrare la tesi.
4. Raccogliere i dati.
5. Visualizzarli e pubblicarli.
6. Vedere la qualità della dimostrazione.

Raccontare una storia L'obiettivo di una buona visualizzazione dei dati non è solamente quello di mostrare dei risultati, ma anche quello di raccontare una storia. L'uso dei dati e la loro rappresentazione tramite dei grafici mostra degli obiettivi particolari: si deve **parlare di qualcosa**. Non basta solamente mostrare solo i dati, ma si devono fornire anche altre informazioni, che arricchiscano la narrativa.

Punti di vista Spesso il punto di vista in cui si presentano i dati cambia quello che si comunica: a seconda di come ci mettiamo cambia l'interpretazione. Quello che si capisce dai dati cambia radialmente a seconda di come vengono visualizzati. Il **field of view** è una chiave di lettura essenziale per l'interpretazione corretta dei dati.

Tipi di grafici Diversi tipi di grafici possono conferire informazioni diverse anche sullo stesso tipo di dati.

Scatterplot In particolare, gli **scatterplot** sono una delle risorse migliore nel caso di mancanza di altre soluzioni: riescono sempre a dare informazioni generali sui dati.

Heatmap Si possono usare delle **heatmaps** in alcune situazioni per mostrare come alcune caratteristiche siano correlate o meno tra di loro. Vengono spesso usate in genomica, per vedere la correlazione tra i geni e le caratteristiche che vengono espresse.

Si noti come, poiché si basano sulla percezione dei colori, possono dare informazioni apparentemente errate: la nostra percezione del colore cambia a seconda dei colori presenti nell'intorno. Quasi tutte le heatmap potrebbero avere degli elementi che potrebbero sembrare associati a un colore maggiore, quando in realtà sono un altro: questo si nota soprattutto quando si eseguono dei confronti con altri fenomeni nel dataset.

Cartogrammi Sono dei plot geografici che modificano l'estensione territoriale in base alla quantità di un fenomeno, e.g. popolazione.

4 miti della *data visualization*

1. Visualizzare i dati è relativamente semplice: *basta inserire un diagramma.*

   Questo ovviamente è un disastro, e i programmi che si usano non sono in grado di prevedere quali siano le migliori scelte. Spesso vengono prodotte le visualizzazioni a minor sforzo cognitivo.

   Attento: *se si commettono uno degli errori presentati dal prof, la visualizzazione verrà rigettata.*

2. La visualizzazione dei dati va di moda, ma presto nessuno la userà più.

   In realtà, la visualizzazione dei dati non è recente e non una moda: la storia degli stati moderni è anche la storia di come gestire in maniera efficiente una grande quantità di dati. La visualizzazione di dati è sempre stato il modo per visualizzare grosse quantità di dati su carta.

3. Visualizzare i dati dovrebbe essere l'ultima cosa da fare.

   In realtà, prima si visualizzano, meglio è: permette una migliore comprensione del problema che si ha davanti.

4. Le visualizzazioni sono belle, ma non portano un reale valore.

   In realtà, la visualizzazione dei dati è proprio la maniera in cui si possono presentare in maniera chiara quello che si vuole trasmettere.

Un'importante consapevolezza è che la *Data Visualization* viene percepita in maniera diversa a seconda della persona che la fruisce. Gli artefatti permettono un'interpretazione dei dati che altrimenti non sarebbe possibili, ma è anche in grado di condizionare la nostra capacità cognitiva.

4 Semiotica

Introduzione Come base, si utilizza il libro *Semiotics* di *Daniel Chandler*. Con **semiotica** si intende lo *studio dei segni*, ovvero tutto quello che ha a che fare con la comunicazione.

La definizione di **segno** è quello di qualsiasi cosa che possegga una qualche forma di significato. Essi sono quindi delle cose percepite che significano qualcosa in un certo contesto sociale, ovvero in un insieme di interpreti competenti.

Il linguista *Ferdinand de Sasseure* è considerato il padre fondatore della disciplina, e l'Americano *Charles Sanders Peirce*riconobbe il bisogno dell'analisi dei segni, e che addirittura da essi dipendano tutti i nostri metodi di comunicazione della conoscenza.

Segno Secondo *Peirce*, qualunque cosa può essere un segno se qualcuno lo interpreta come tale, ovvero che sia in grado di significare qualcosa.

Questa definizione spesso viene assegnata a quella di *simbolo*, sebbene essi siano solamente una parte dei segni.

Secondo *de Sasseure*, un segno è l'unione due interfacce, un *significato* e un *significante*. Esso non è ciò che lega le due interfacce, ma tutto l'insieme. Si ha quindi che le interfacce sono due facce dello stesso foglio.

Secondi *Peirce*, un segno è qualcosa che sta per qualcuno per qualcos'altro riguardo a qualcos'altro, in qualche modo o capacità. Ovvero si ha un modello triadico, *rappresentazione*, *oggetto* e *significato*.

I due modelli non sono in competizione: il modello di *Peirce* è più un'estensione di quello del francese.

Per avere un segno bisogna avere *oggetto*, *come è interpretato* e il *significato*.

Pensare in maniera *logico-simbolica* viene definita come **semiosi**, ovvero il processo con cui interpretiamo segni. La metafora del **triangolo** è essenziale e ricorrente nella semiotica.

Una delle idee di come rappresentare il segno è quella di **tripode**, ovvero l'oggetto che deve essere tenuto in piedi da tutte e tre le cose: **oggetto, significato, interpretante**. Si noti come in realtà questi tre oggetti hanno avuto differenti nomi nella letteratura.

Simbolo Peirce ha creato una sorta di tassonomia dei segni, ovvero *simboli*, *indici* e *oggetti*. In realtà, questa divisione non è correttissima: si parlerebbe di una modalità con cui si presenta un segno.

In particolare, il **simbolo** è la modalità in cui il significante non somiglia al significato, ma vi è associamento per arbitrarietà.

Tipi di segni In qualunque scena che ci troviamo davanti, vi sono diversi tipi di segni, sia indicale che simbolico, che interagiscono tra di loro per dare maggiore significato a quello che si vede. Per esempio, un mappa racchiude in sé molti segni. La stessa $\mu\omega\rho\varphi\eta$ può indicare più cose diverse.

Legame con la Data Visualization Poiché la visualizzazione dei dati si basa molto su indici e icone, ovvero si cerca di creare segni attraverso dati. A partire da delle "tabelle" pieni di numeri e simboli. L'uso di grafici permette di trasmettere i significati in maniera più efficace, tramite l'uso di tutti gli strumenti della semiotica.

Nelle visualizzazione dei dati si vuole "asciugare" l'elemento simbolico e aumentare la parte **iconica** e/o **indicale**, che agisce su dei meccanismi innati dell'essere umano.

4.1 *Un approccio semiotico alla data visualizzazione*

Introduzione Lezioni da parte di **Agata Meneghelli, PhD** in Semiotica, che lavora oggi nel settore privato.

Che cosa è la Semiotica? La semiotica è la scienza che studia, analizza e cerca di spiegare tutti i fenomeni di significazione, che siano segni, fenomeni, risorse, meccanismi, etc.

La visualizzazione dei dati è un oggetto semiotico, il cui obiettivo è quello di trasmettere del senso. Si noti come in realtà anche i **dati** stessi sono oggetti semiotici, poiché anch'essi costruiti. Entrambi non sono "dati" a monte, ma sono entrambi dei processi costruiti (anche i dati).

Interpretazione Secondo quanto detto da *Peirce*, si ha uno schema triadico **segno - oggetto - interpretante**, anch'esso un altro segno che traduco il primo segno per esplicitarne il significato. Si possono quindi avere delle catene di interpretanti, in quanto sono essi stressi segni.

Interpretanti Ogni interpretante successivo aggiunge significato a quello precedente, un fenomeno prettamente semiotico.

In particolare, si può identificare come *interpretante* un grafico che visualizza dei dati, visti come *segni*: tramite questo modo, si riesce ad arrivare a un

tipo particolare di interpretazione, e.g. il prezzo delle case è effettivamente aumentato.

Testo Secondo **Umberto Eco**, con *testo* si intende la catena di enunciati legati tra loro da vincoli di coerenza, sia gruppi di enunciati emessi contemporaneamente sulla base di più sistemi semiotici.

Quindi, dalla definizione di Eco, si può dedurre che un testo è qualsiasi cosa che può essere interpretato da qualcuno, con però delle caratteristiche, per esempio è più ampio di un semplice segno.

Visualizzazione dei dati Come detto, i dati sono frutto di un processo semiotico, frutto di creazione dei segni, e tra la visualizzazione come testo e i segni (dati) è un processo di interpretazione dei dati. Non si deve quindi pensare in termini di segni isolati, ma in termini di *testi*, ovvero segni che interagiscono tra di loro.

Sia nella progettazione che nell'esecuzione, è importante tenere presente la *storia* che si vuole raccontare attraverso quello che si fa.

Semiotica Dizionario Secondo la *semiotica dizionario*, il significato di un segno può essere descritto come un insieme di unità minime di significato. Quando si opera con la definizione da dizionario, si possono creare degli schemi ad albero. Questo tipo di rappresentazione è che rimane troppo limitata: risulta statica, non ammette cambiamento e rimane disconnesso dal mondo.

Semiotica Enciclopedia Nel caso di enciclopedia, si hanno legami tra differenti segni, ovvero vi sono un insieme infinito di interpretazioni, dipendenti dalle culture. Essa è posseduta un maniera diverse dalle singole persone, sebbene sia comunque un costrutto globale.

Una enciclopedia può essere una fonte enorme di segni e interpretazioni dati per la visualizzazione dei dati. Sebbene sia sicuramente un vincolo, è una forte risorsa da cui si può attingere.

Cooperazione testuale In un testo, ci sono molte cose che rimangono lasciate esplicite: *un testo è un meccanismo pigro, che vive sul plusvalore introdottovi dal destinatario.* Esso per funzionare ha quindi sempre bisogno di qualcuno che vada a riempire gli spazi vuoti.

4.2 *A semiotic analysis of some exemplary cases*

Recap Come detto, qualsiasi visualizzazione dati è costruita, e anche i dati sono essi stessi dei significanti. In particolare, abbiamo visto che la visualizzazione dati è un **testo** che racconta una storia. Abbiamo poi parlato delle enciclopedie come fonti di conoscenza comune, che guidano e arricchiscono l'interpretazione del mondo.

Ogni testo presuppone un **lettore modello** di quella visualizzazione: esso non è reale, ma è solamente chi lo scrittore si presuppone leggerà.

Lettore modello In un testo, ci sono due figure che interagiscono: lo scrittore e il lettore. Esistono però una serie di persone immaginate, come il **lettore modello**, che interagiscono nella mente delle persone fisiche.

Questo lettore immaginario deve essere proiettato sul testo, e lo si deve aiutare per capire come muoversi. Se però esso è pensato e/o proiettato male, si entra in **decodifica aberrante**, ovvero si è raccontata un'altra storia, fuori dalle nostre intenzioni.

Esso rimane al gioco, seguendo le regole che sono state delineate. Sebbene a volte ci siano errori da parte dello scrittore, può capitare che il lettore reale decida di non seguire i percorsi interpretativi pensati.

Discorso Quando si ha un testo, un discorso cerca di giungere a degli obiettivi, ovvero si hanno dei costrutti. Queste tecniche vengono usato per immergere le persone all'interno dei videogiochi: le interfacce cercano di oggettificare il mondo del gioco.

Strategie oggettivanti Alcune strategie semiotiche, sebbene dicano la verità, presentano concetti come se fossero sempre presenti, nascondendo come essi siano in realtà costruiti.

Diagrammi Secondo Umberto Eco, i diagrammi, e come estensione potremmo dire la data visualizzazione in generale, sono degli strumenti in cui esiste connessione punto-punto tra espressione e contenuti. Queste forme testuali hanno il potere di essere euristiche.

Non detto Quando in un contesto di testo, e.g. giornalismo, si fa riferimento a dei dati/fatti, ci possono essere una serie di cose non-dette, che però tendono a essere più pensieri piuttosto che fatti. Se queste inferenze sono fatte male, possono giungere dei significati non veri e/o erronei. Spesso gli stereotipi possono emerge da un testo, e sono pericolosissimi: vanno a

generalizzare e annullano molti dettagli, riducendoli a un insieme ristretto di caratteristiche. Questi sono così sedimentati nell'enciclopedia da collegarsi anche in maniera involontaria.

5 DOs and Don'ts

Il professore fa vedere alcuni grafici per mostrare come non siano tutti intuitivi.

La visualizzazione dei dati si possono sia raccontare verità, sia mentire, e.g. una semplice modifica degli assi di una grafico. La visualizzazione, sebbene sia giusta, può veicolare un determinato obiettivo. Davanti a una visualizzazione, si deve cercare di ricostruire l'intenzione.

Durante una visualizzazione dati, è sempre importante porre accanto a una media la **deviazione standard**.

Si può parlare di **fattore di bugia**, o *lie factor*, relativo a una grafico: esso è identificato come la dimensione dell'effetto come mostrato nel grafico e quella che è presente nei dati.

$$\texttt{lie factor} = \frac{\texttt{size of effect shown in graph}}{\texttt{size of effect in data}}$$

dove il *fattore di scala* è dato da:

$$\texttt{size of effect} = \frac{|\texttt{first value} - \texttt{second value}|}{\texttt{first value}}$$

In pratica, il fattore dovrebbe essere contenuto, ovvero tra 0.95 e 1.05.

Secondo grafico Per evitare il problema della scala y, ovvero quando la scala potrebbe raccontare delle bugie, si può usare lo stratagemma di due visualizzazioni associate, in cui una è generale e una seconda che fornisce dati più dettagliati. Spesso si usano degli espedienti per mostrare, per esempio, che si sta eseguendo uno *zoom*.

Assi troncati Per quanto riguarda i grafici a barre, gli esperti sostengono che non si dovrebbe mai troncare l'asse, ovvero che dovrebbe essere fatta partire da 0. Per quanto riguarda i *lineplots*, non si ha consenso in merito, anche se comunque viene lasciato al caso.

In particolare, `Microsoft Excell` tende a troncare gli assi senza eseguire riferimenti.

Limited scope In una visualizzazione dati, è molto importante avere presente quale sia un asse adeguato al fenomeno di riferimento. Nel caso in cui l'asse delle x sia troppo piccolo rispetto al fenomeno di riferimento, si parla di *limited scope*.

Grafici a doppia scala Si noti come i grafici a doppia scala possono presentare delle problematicità: essi sono infatti proni a rappresentare una **correlazione**. Quando si hanno due grafici insieme, è ovvio che si ritenga la presenza di un nesso casuale tra i due.

Come detto, con gli stessi identici dati si possono veicolare informazioni legate a una particolare tesi. In particolare, nel loro uso si potrebbe andare in contro a delle *correlazioni spurie*. Si ricordi che **correlazione non implica causazione**.

Granularità dell'asse x Come detto, la definizione dell'asse delle ordinate, in particolar modo nei *barplot*, possono comunicare informazioni differenti.

Pie chart L'uso di aerogrammi è fortemente sconsigliato, a meno di casi particolari, e.g. in cui ci sono solamente poche informazioni e ben distribuite.

3D graph Come regola buona, non si devono usare grafici in 3 dimensioni se non strettamente necessari; ovvero rappresentare solamente le dimensioni presenti nei dati.

6 Infografiche

Mi sono perso la lezione precedente. Si è detto nella lezione precedente che bisogna stare attenti alle percentuali. Nelle visualizzazioni è spesso più utile usare le **frequenza naturali**.

Intervalli di confidenza Nella visualizzazione dati è molto importante identificare la presenza di errori. La più semplice di queste visualizzazioni consiste con l'uso di **barre di errore**, che però risulta una modalità poco comprensibile da persone non specialiste, oltre che fuorviante su che cosa si voglia rappresentare – può essere usato per diversi intervalli di confidenza e/o misure.

Alcune tecniche si basano sull'uso di sfumature intorno al valor medio, in modo tale da avere delle stime visuali più dirette.

Altre visualizzazioni si possono basare sul **boxplot notch**.

Vedi sul sito web `https://www.data-to-viz.com/caveats.html` per avere un'idea degli errori che non si devono fare nella visualizzazione dei dati.

Non si deve essere specialisti, ma il più inclusivi possibile, ovvero per permettere a un maggior pubblico di comprendere l'analisi.

7 The Light Side

In una visualizzazione dati, non si deve aver paura ad avere una **tesi**, ovvero le cose devono essere mostrate in maniera chiara e onesta. La valutazione di un'infografica non è puramente soggettiva, e ci sono alcuni criteri che possono essere importanti. Tra questi:

- Efficienza.
- Effettività.
- Soddisfazione.

Ci si potrebbe inventare anche una nuova visualizzazione.

Vediamo i giusti ingredienti per creare una buona visualizzazione.

- Scale fatte bene.
- Sistemi di coordinate.
- Indizi visuali.
- Contesto.

È importante integrare il testo con il grafico. Anche se la visualizzazione è interattiva, si deve dare una chiave di lettura tramite delle spiegazioni/testo. Per esempio, in presenza di serie temporali può essere importante apporre delle annotazioni, per rendere più intuitivo usare del testo.

Se non si hanno delle annotazioni, la tesi che si propone è più debole.

8 Assessment

Ci sono spesso alcuni pattern che si cerca di spiegare con l'uso di una *data visualization*.

Una **visualizzazione giusta** è una che non porta delle distorsioni ai dati. Una **visualizzazione eccellente** è invece una che riesce a trasmettere in pieno in certo significato. In alcuni casi una visualizzazione giusta può essere la soluzione migliore.

Usare dinamicità è importante, e l'uso di **boxplot** può essere utile per rappresentare indicatori di tendenza centrale, e.g. mediana, quartili etc. Come alternativa, si possono usare altre tecniche, tipo i **violin plot**.

Valutazioni Ci sono sia metodi qualitativi che quantitativi. Tra i primi, quelli qualitativi potrebbero essere di tipo **euristico**. Alcuni quantitativi sono per esempio **user test**. Qualcosa a metà strada tra entrambi sono invece i **questionari psicometrici**, ovvero in cui si cerca di misurare delle valutazioni/opinioni degli utenti relativamente alla visualizzazione.

Un protocollo, noto come **think aloud protocol**, un tipo di valutazione euristica. Questo si dovrebbe eseguire prima di fare una valutazione con più persone.

Euristica Le *euristiche* sono dei suggerimenti/modi furbi di fare le cose. In questo caso, si ha il mantra di Shneiderman.

User test In questo caso, si cercano due o tre domande sulla visualizzazione, per vedere quante persone riescano a rispondere correttamente. Si va inoltre ad analizzare quanto le persone dicano le risposte giuste. In questo modo si vede l'efficacia e l'efficienza.