In Python i dataset si esportano come ***dataframe***.

Spesso in alcuni dataset non ci sono delle variabili significative.

Librerie per fare visualizzazioni in Python:

• Altair

• Matplotlib → piuttosto grezzo ma simile alle visualizzazioni che si fanno in MATLAB

• Ggplot → in R si può importare in Python ( <https://realpython.com/ggplot-python/> )

Slide 1

*Information design*

*Information architecture*

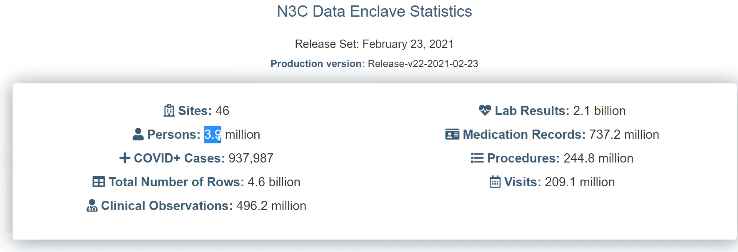
Slide 2

La capacità di ***processing*** sta aumentando sempre di più, come anche la quantità di dati disponibili.

Tutti i diversi plot nascono dalla volontà di visualizzare un particolare tipo di dato ma poi vengono **generalizzati** a più tipi di dati.

**Wurman** è specializzato nella psicologia dell’informazione e ha definito *l’****information anxiety*** ossia l’ansia che si ha quando si hanno troppe informazioni.

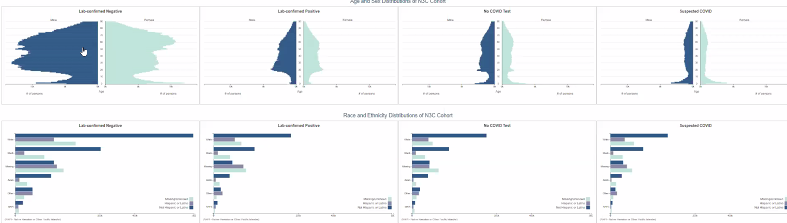
Slide 3

**N3C** è un enclave che ha unito dapprima ricercatori degli Stati Uniti, poi man mano sta unendo diversi ricercatori di tutto il mondo.

Si tratta di un ampio dataset.

Come si fa ad arrivare ad una visualizzazione che dia l’idea del tipo di dati presenti?

È una schermata che con dei plot vuole dare idea di come sia fatto il dataset.



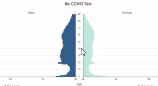
↑ Questi sono degli **istogrammi non normalizzati**.

Vuole essere una **dashboard** di descrizione del dataset e funziona bene: cerca di di dare una distribuzione dei pazienti.

Gli istogrammi vanno normalizzati.

Questo è un esempio di grosso dataset, dove ci sono queste visualizzazioni:

**Grouped bar chart** 

**Istogramma verticale** o **simmetrico** 

Hanno cercato di dare una descrizione visuale del dataset.

Abbiamo visto quindi cosa si può intendere per ***big data***.

Come big data si possono intendere anche le reti/grafi sociali che hanno milioni di edge che collegano i diversi nodi del grafo.

Slide 4

Come si pianifica una dashboard (come quella della slide 3) che dia una descrizione semplice dei dati e che, per uno che non sa magari l’argomento, risulti abbastanza chiara?

Bisogna seguire i 6 principi di visualizzazione dello studioso Edward Tufte:

* Mostra le comparazioni
* Mostra cause ed effetti nelle visualizzazioni (se è il caso)
* Usare dati multivariati (se si hanno)
* Integrare tutti i tipi di modalità
* Stabilire una credibilità dei dati
* Focalizzarsi sul contenuto

Slide 5

Minard descrive con un’immagine la campagna di Russia.

È stato il primo a fornire una descrizione visuale che mostra bene le *comparazioni*.

1mm di larghezza di banda rappresenta un certo numero di soldati (in rosso a SX).

In nero è indicato il rientro.

Di ogni punto viene data la posizione geografica.

Scrive anche il numero totale della truppa in quello step, in quel punto geografico.

Questa visualizzazione è bella perché consente di comparare la partenza (prima) rispetto al rientro (dopo).

Fa vedere anche le truppe che sono andate e ritornate indietro subito…fa vedere un sacco di cose.

Si possono comparare diverse cose nelle visualizzazioni, ad esempio nei tweet si possono compare i fake tweet da quelli reali.

Quindi le comparazioni esistono ovunque nelle visualizzazioni.

Slide 6

*Show casuality* vuol dire dove il grafico deve poter mettere in evidenza una conseguenza di cose e fare in modo che ciò sia visibile.

Questo è il primo grafico che rappresenta un **radial plot**.

In questo plot Nightingale (la fondatrice delle crocerossine) mise in evidenza come le morti non fossero dovute tanto alle guerre ma alle malattie all’interno dell’esercito.

Nightingale misura dall’interno, dal centro (il totale dell’area blu parte dal centro – il totale di ogni area parte dal centro):

- l’area blu: rappresenta i morti per una malattia che non dipende dalle guerre

- l’area rossa: rappresenta coloro che sono morti per guerra

- l’area nera per altre cause

Con questo grafico voleva mostrare l’effetto della cattiva igiene nell’esercito.

E c’è riuscita tanto che la regina fece una campagna vaccinale (per risolvere il problema delle morti).

Ha utilizzato il radial plot per far notare il decorrere degli anni: a Gennaio fa più freddo e ci sono più morti.

Ecco perché non ha usato un **bar chart**/**bar plot**.

Slide 7

Un **line chart** del numero di soldati all’andata ed il numero di soldati al ritorno.

Questi dati va bene unirli perché si tratta di dati temporali.

Utilizza un unico grafico e fa vedere un line chart (il quale mostra insieme alle date le temparature) che permette un confronto tra le truppe che sono riuscite ad andare e quelle che sono riuscite a ritornare, fornisce una visualizzazione geografica e fa vedere anche le truppe che si sono staccate e che si sono riunite.

Oggi chiameremo questa visualizzazione, come visualizzazione multimodale → ossia visualizzazione con dati multimodali, ovvero dati provenienti da più modalità (da una conta, da dati geografici e dalla temperatura).

Slide 9

Come si applica ai nostri giorni la visualizzazione di Minard?

Tante volte noi umani tendiamo ad usare due sole variabili, come *salary* e *yrs.since.phd (numero di anni passati dalla laurea)*; ma se si volesse considerare anche la *carica*, quindi un’altra variabile?

Posso usare i colori.

Si hanno quindi tre dimensioni: *yrs.since.phd* è intero, *salario* è continuo, ,mentre la terza dimensione (il *rank*, la *carica*) è categorica.

Si può usare uno **scatter plot** in cui si vede che tutti i professori prendono tendenzialmente un salario più alto e lavorano dopo anni dalla laurea.

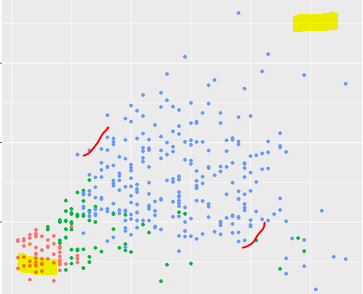
Gli assistenti hanno tendenzialemente un salario più basso e lavorano subito dopo la laurea.

Questo scatter plot, in unico grafico mette in evidenza le *comparazioni*, i *dati multivariati* e mostra una *casualità* (ossia appena ti sei laureato è più probabile che tu sia assistente).

Mette in evidenza tuitte le informaizoni che si vogliono mettere in evidenza.

Se non si fosse considerata la terza dimensione, non si sarebbero considerate delle informazioni.

Ci sono anche degli outlier (valori anomali) ossia dei professori che nonostante tali hanno un salario basso.

In giallo si ha una **corrispondenza lineare** tra le variabili, per dire tanti anni per laurearsi tanto salario.

Mentre in rosso c’è **corrispondenza inversa** tra le variabili.

[ I dati sono stati presi dalla Minessota University, ma in slide non è stata indicata l’origine, quindi la credibilità di questi dati potrebbe essere bassa. ]

[ Mentre nella slide seguente la credibilià è alta poiché è indicata l’origine dei dati stessi. ]

Slide 10

Se ho tante informazioni da dare, non devo preoccuparmi di non riempire il grafico, almeno all’inizio.

In quanto posso usare i testi, posso integrare immagini, audio che sto usando via web.

Questa visualizzazione è la visualizzazione dallo 0 a 300 giorni dopo il centesimo caso Covid confermato nel mondo (in Cina).

Mostra com’è cresciuto il rate di positivi, ossia il rapporto tra il numero di positivi rispetto alla popolazione complessiva.

Usiamo tutte le tipologie di testi che abbiamo a disposizione senza problemi (in modo da spiegare meglio).

Slide 11

Non plottare i dati (ossia nel rappresentarli) senza dire da dove provengono, ma anzi bisogna plottare indicando le sorgenti di questi dati.

Dicendo come sono stati processati tali dati.

Un grafico deve essere massimamente informativo, per essere capibile; non deve mostare cose di cui l’utente non capirebbe il significato.

[ LWBS → le persone che se ne vanno prima di essere visitate (magari perché si scocciano) ].

Questo grafico:

* ha la credibilità poiché è stato detto da dove vengono i dati
* si dà il periodo di raccolta dati
* si dice che sia il *data wrangling* che la visualizzazione sono state fatte con MATLAB.

*Data wrangling* è la fase di raccolta e pulizia dei dati.

[ Pulire → rimuovere dati che non servono

*Data preprocessing* ci si riferisce ad una serie di tecniche di data mining che consistono nel trasformare i dati in un formato comprensibile e adatto alle nostre necessità. ]

Questo grafico compara il 2020 rispetto al 2019.

Il 2019 è indicato in blu, il blu esprime calma.

Il 2020 è indicato in rosso, il rosso esprime pericolo.

Insieme al grafico dettagliato giorno per giorno è visualizzata una linea di trend  per ovviare alla randomicità: infatti si può vedere che ci sono dei momenti nel 2019 in cui ci sono dei picchi.

Si sta facendo una media su trend consecutivi.

La linea tratteggiata dà il trend generale del periodo. 

**Bar plot** e linee di regressione/di trend permettono una *comparazione*, una *casualità* (anche perché sappiamo cosa c’è dietro a questi periodi), usa testi, colori e immagini per visualizzare tre variabili che sono il tempo, il numero di accessi e con i colori si utilizza la terza variabile che è l’anno, ossia 2019 oppure 2020.

Non si rischia di avere dei grafici troppo dettagliati, tanto che l’utente non li capisce?

Ma poi il visualizzatore si abitua. Una volta che gli si spiega la visualizzazione e gli si dà tempo per vedere tutte le visualizzazioni, riescono a capire come funzionano e cosa vogliono rappresentare.

In rosso dà il concetto di quantità, di impatto.

La linea tratteggiata continuata.

La linea tratteggiata permette di vedere se la cosa è cresciuta oppure decresciuta nel periodo.

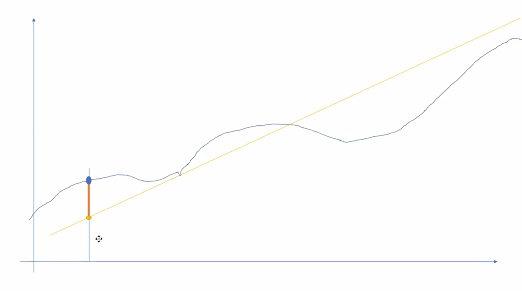
C’è un forte impatto visivo con questi grafici.

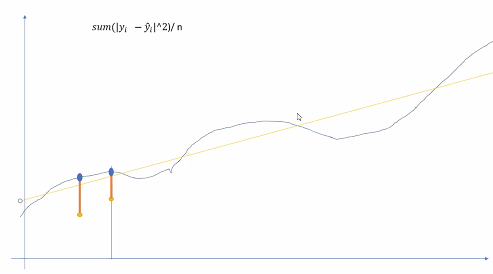
*Spiegazione regressione lineare*

La regressione lineare è la retta che minimizza l’errore quadratico medio.

Si cerca la retta (gialla) per cui la sostituzione di tutti i punti della retta a tutti i punti della funzione (in blu)

[ c’è sempre bisogno degli assi] si avrebbe un errore dato dalla lunghezza del segmento rosso (tra i due punti):



Se ci si sposta bisogna fare la stessa cosa:

**Yi** → i punti della retta blu (i punti azzurri)

**Yi** “cappuccio” → i punti della retta gialla (i punti gialli)

Si sostituiscono i punti blu con quelli della retta giallla.

Se si calcola la distanza e si eleva al quadrato, si ha l’**errore quadratico**.

Se lo si divide per n punti considerati→ si ha l’**errore quadratico medio**.

In sostanza la regressione lineare è la retta che si avvicina di più alla mia funzione.

La retta è descritta da un **m** e da un **q**. Bisogna ottimizzare.

La regressione lineare serve per fare le rette di trend.

*Establish credibility*: indicare sempre l’origine dei dati.

Slide 12

*Focalizzarsi sul contenuto*

Questi grafici sono stati fatti come grafici di denuncia, sono stati messi in una rivista scientifica e si voleva dare una classificazione diversa dei vari consumi.

 è un **donut chart**.

Slide 13

È un’immagine bella colorata, ma ci sono dei problemi.

Ad esempio sull’asse y, a volte è indicato grammi a volte kilogrammi.

Problemi: lo stesso concetto poteva essere reso in un modo più semplice.

Non c’è scritto sul **bar chart** che si tratta di una scala logaritmica.

Meglio evitare di usare scale diverse, non si capisce cosa si sta visulizzndo, c’è confusione.

Slide 14

*Moiré effect* → degli effetti di background che creano solo rumore

Quando si sovrappongono più linee a diverse orientamenti, il cervello crea dei pattern che si incrociano e danno fastidio, creano confusione.

I donut chart sono poco chiari.

Per mettere in evidenza la popolazone degli Stati Uniti rispetto a quella mondiale, si poteva usare un **mosaic chart** anziché una visualizzazione tridimensionale.

Slide 15

Una nuova versione renderebbe tutto più chiaro e più semplice.

Sull’asse x è stata considerata quindi la scala logaritmica.

Slide 17

Il giornale The Guardian addirittura ha messo sul suo sito questi plot e ha chiesto, per chi riesce a capirle, di fornire la soluzione e si otterrà un premio.

Attenzione quando si usano i **pie chart**.

Slide 18

Attenzione a fare cose di questo tipo.

Questa è stata considerata una delle peggiori **infografiche**.

Slide 19

[ La *dashboard* in un’unica vista descrive un dataset permettendo di fare visulizzazioni. ]

Come sta l’**information architecture** all’**information design** alla **data visualization**.

**IA**: disciplina/ambito di ricerca che si occupa della strutturazione dei dati e delle informazioni in un modo sostenibile (ossia in un modo che possa essere supportato da evidenza) ed effettivo.

Scopo della IA: vuole aiutare a chi ha mole di dati enormi a capirle.

Slide 20

La **data visualization** si pone all’interno dell’IA.

L’IA comprende tutte le scienze che si occupa del rendere le informazioni visibili, disponibili e trattabili da parte degli utenti.

La data visualization si occupa di rendere visualizzabili, di far sì che si vedano dei dati.

I dati sono delle informazioni; fino a poco tempo fa si trattava solo di dati numerici perché erano comprensibili dal computer ed era l’unico tipo di dati che riusciva a trattare, ma al giorno d’oggi si parla anche di tweet per dire.

Slide 22

Jennifer dà una formalizzazione di cosa vuol dire, dalla parte dell’umano, trattare i dati che provengono dal mondo esterno; quindi definisce la **DIKW**.

Slide 26

Dal nostro punto di vista, questa piramide formalizza ciò che succede ai dati esterni, che sono non strutturati (profumi, colori, stimoli…).

Qualcuno (anche all’interno degli ospedali) prova a collezionarli, a filtrarli e farne una prima codifica.

Da ciò nascono dei dataset un po’ sconclusionati che rappresentano i dati a noi disponibili.

I **comunicatori** (i computer scientists, le aziende come Google, WikiLeaks – aziende a cui i dati piacciono molto) prendono questi dati, applicano i processi di *data wrangling* (ossia pulizia e organizzazione dei dati) in modo da strutturare l’informazione/i dati che a loro interessano e metterli a disposizione dell’utente.

I dati possono essere analizzati per poi essere visualizzati o li si possono visualizzare per poi analizzarli.

*Crawling dei dati* ossia recupero dei dati dal web ed è stato permesso il trattamento di tali dati solo a scopi di ricerca.

Per **strutturazione** si intende anche visualizzazione, darne una forma visuale.

Tramite la visualizzazione, l’utente vede e crea **conoscenza**; ciò porta poi alla **wisdom**, ossia alla saggezza (derivante dalla conoscenza) → il momento esatto in cui l’informazione è fatta propria in modo da poterla usare in momenti futuri.

Slide 27

Tutte queste teorie sono state definite all’interno del campo dell’IA.

Soprattutto quando questo campo si è unito, grazie alla piramide di conoscenza, al campo della psicologia e all’analisi psicologica.

La **scientific/data visualization** è a tutti gli effetti un sottoinsieme dell’**information visualization**, che è un sottoinsieme dell’**information design**, che è un sottoinsieme dell’**IA**.

• Nel campo della *scientific visualization* vediamo grafici, plot, bar chart, radial plot … tutti i plot che ci permettono di vedere e visualizzare i dati.

• *Information visualization* significa **infografica** → ossia unire ai plot testo e immagini per formare come dei mega poster che rappresentano un tutt’uno dell’informazione, non solo le quantità numeriche

(quella di Minard è una sorta di prima infografica, quella della campagna di Russia).

Slide 29

L’**ID** è tutta l’analisi di tutti quei passi che servono ad analizzare i bisogni dell’utente, selezionare gli stili, i metodi più effettivi per comunicare con l’utente, …

Per fare una visualizzazione dobbiamo capire:

* con chi stiamo parlando
* perché vogliamo fare la visualizzazione
* perché quella persona ha bisogno della nostra visualizzazione e cosa è in grado di capire

(ad esempio per un bambino si utilizzano delle visualizzazioni più semplici)

Slide 31

L’**IV** è quel ramo che si occupa di visualizzare delle informazioni astratte, ossia quelle informazioni per cui non si hanno dei numeri.

Le informazioni raffigurate in slide non possono essere espresse con dei numeri.

Slide 32

L’infografica mette insieme sia la visualizzazione delle informazioni che la data visualization che rappresenta numeri.

Un grafico può essere proiettato in uno spazio multidimensionale → i nodi di un grafo possono essere rappresentati in numeri, possono essere proiettati in uno spazio n-dimensionale.

Qui si parla di *scientific visualization* perché si tratta di visualizzazioni che servono nel campo scientifico.

Slide 33

InfoVis, Infographic, DataVis e SciVis si incontrano nell’esplorazione, nell’analisi e nella presentazione.

Ovvero mentre l’InfoVis è fatta per la presentazione delle informazioni, la DataVis e la SciVis sono tendenzialmente fatte sia per la presentazione delle informazioni, ma soprattutto per l’esporazione dei dati (per capire qualcosa di tali dati).

Librerie di visualizzazione di Python:

Seaborn Python

Altair Python

Plotnine