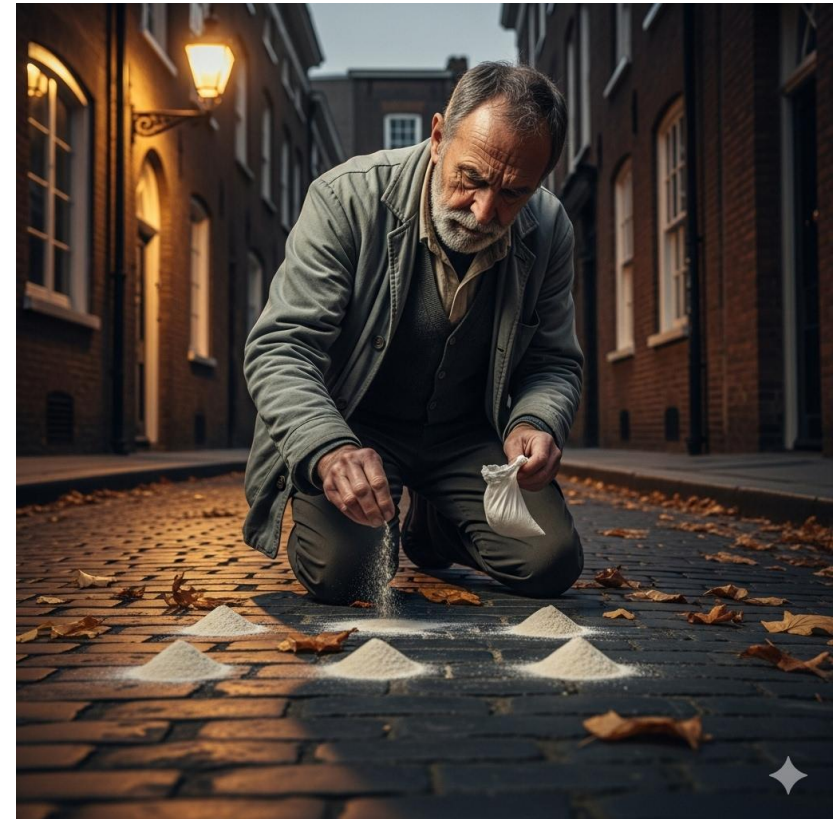


# **I dark data... questi sconosciuti**

*rischi nascosti e valore inespresso*

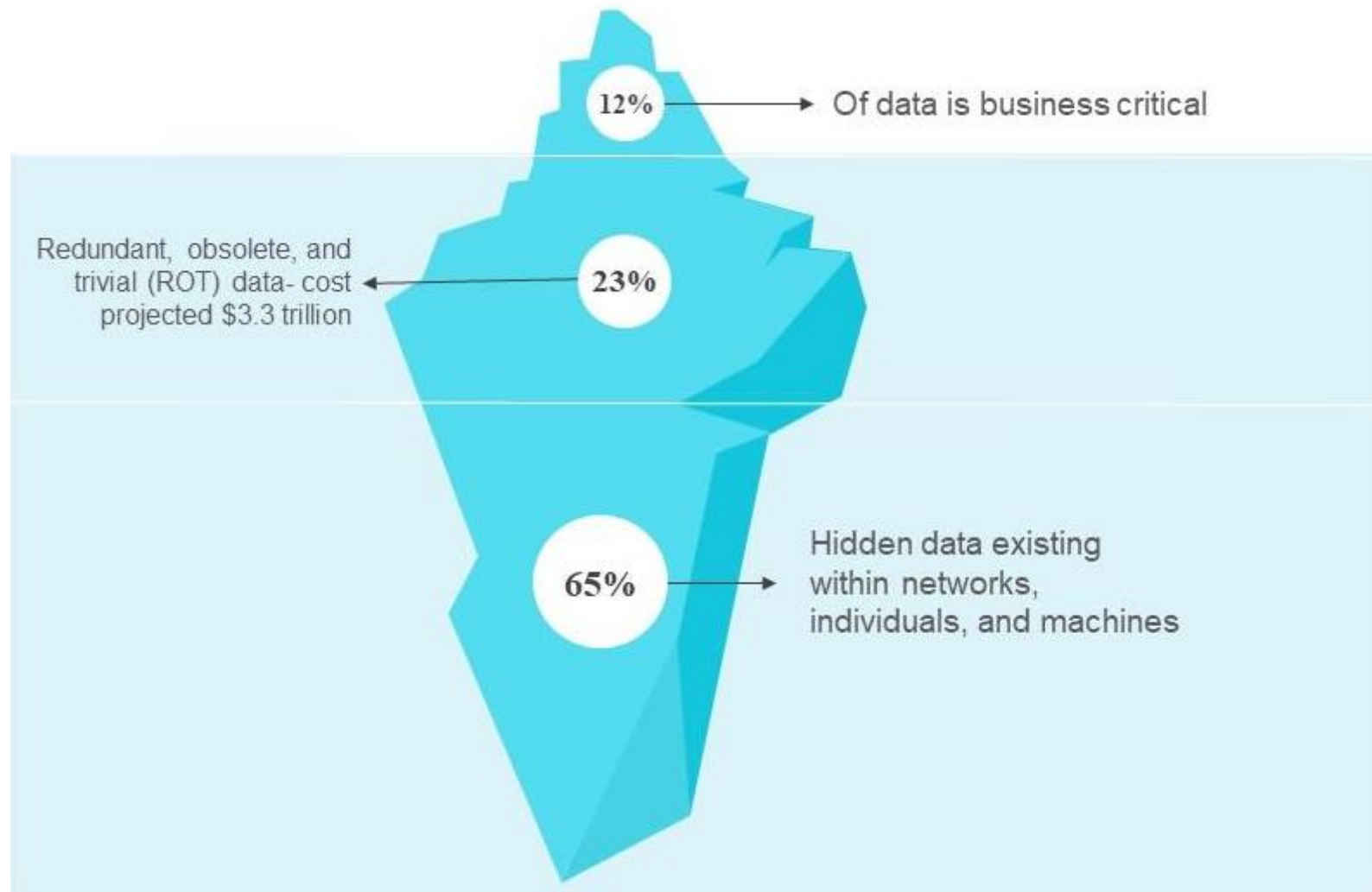
**Prof. Silvano Bari**  
**studiobari@gmail.com**

# I dark data



Ciò che non vediamo  
condiziona il nostro mondo

# I dark data



fonte: [www.slideteam.net](http://www.slideteam.net)

# I dark data

## **1° caso**

***L'organizzazione è consapevole che i dati sono mancanti, incompleti o soggetti a incertezza.***

**I dark data sono i dati raccolti, elaborati o conservati da un'organizzazione ma che non vengono utilizzati attivamente per prendere decisioni o ottenere insight.**

**Sono dati "oscuri" perché non visibili o non analizzati, ma comunque esistenti e potenzialmente utili.**

*Esempi di dark data:*

*Email aziendali archiviate ma mai analizzate.*

*Log di sistema o server.*

*Registrazioni audio delle chiamate con i clienti non trascritte o analizzate.*

*Immagini, video, documenti non strutturati.*

*Feedback raccolti ma non strutturati (es. commenti sui social).*

# I dark data

## **2° caso**

***L'organizzazione non è a conoscenza della mancanza di dati o presume che i dati disponibili siano completi.***

**Ci sono cose ignote che conosciamo; ossia sappiamo che ci sono certe cose che non conosciamo.**

**Ma ci sono anche cose ignote che non conosciamo, ossia quelle che non sappiamo di non conoscere.**

*Donald Rumsfeld (ex segretario alla Difesa degli USA)  
Conferenza stampa del Dipartimento della Difesa, 12 febbraio 2002)*

# I dark data: definizione

## **dark data**

**sono i dati che non abbiamo:  
dati che forse vorremmo avere,  
speravamo di avere, o pensavamo di avere,  
ma che, nondimeno, non abbiamo.**

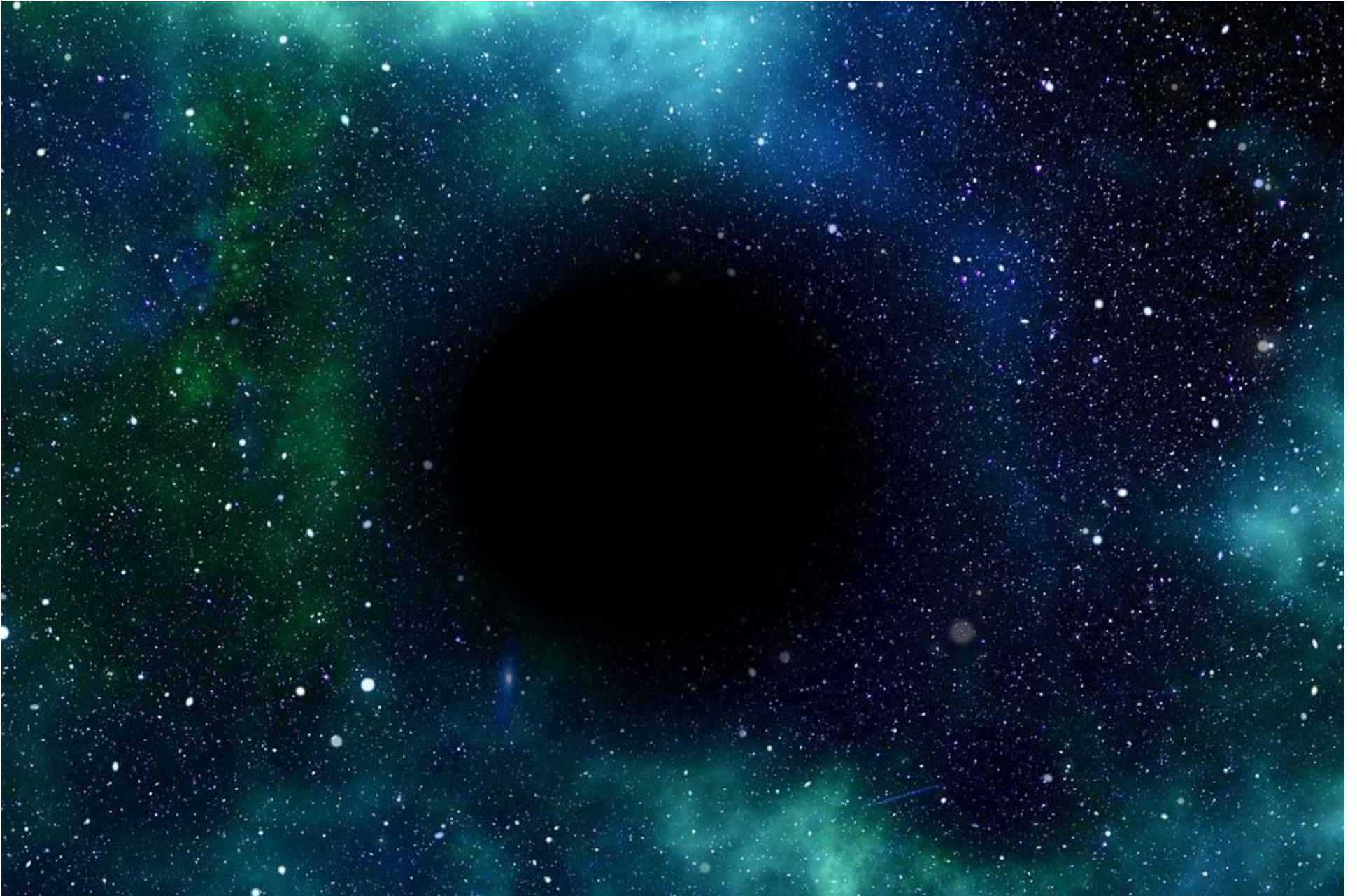
**tutti quei dati che esistono o potrebbero esistere  
ma che non vediamo, non raccogliamo, non utilizziamo o trascuriamo,  
e che tuttavia possono influenzare in modo significativo  
le nostre analisi, decisioni o conclusioni.**

*David J. Hand, Dark Data: Why What You Don't Know Matters*

# I dark data: tipologie

1. Dati di cui conosciamo la mancanza
2. Dati di cui ignoriamo la mancanza
3. Dati che considerano solo alcuni casi
4. Dati di autoselezione
5. Dati ignorati o scartati come irrilevanti
6. Dati che avremmo potuto avere
7. Dati cambiati nel tempo
8. Dati catalogati con definizioni diverse
9. Dati riassunti
10. Dati soggetti a errori di misurazione e margine di incertezza
11. Dati nascosti in formati non strutturati o semi-strutturati
12. Dati gestiti da persone diverse
13. Dati intenzionalmente oscurati
14. Dati fabbricati e sintetici
15. Dati derivanti da estrapolazioni

# I dark data: la materia oscura





# Il «fuoco di sant'antonio»

## HERPES ZOSTER



### COS'È

È un virus che provoca una dolorosa eruzione cutanea, conosciuta anche come **fuoco di Sant'Antonio**. Si tratta dello stesso virus che causa la varicella, ovvero il Varicella-Zoster-Virus.

### I SINTOMI

Possano comprendere:

- sviluppo di **lesioni vescicolari eritematose e pruriginose**
- febbre
- **prurito**
- mal di testa
- **brividi**
- dolori di stomaco
- **spossatezza.**





### I RIMEDI

Di solito l'infezione da Herpes Zoster **guarisce spontaneamente**, come nel caso della varicella. Tuttavia, considerata la fastidiosità dei disturbi o la concomitanza di malattie debilitanti, **si possono impiegare dei trattamenti locali e sistemici per ridurre i sintomi.**

### LA PREVENZIONE

Trattandosi di un'infezione virale è buona norma **non avere contatti diretti con pazienti colpiti dall'infezione** o con i loro indumenti.

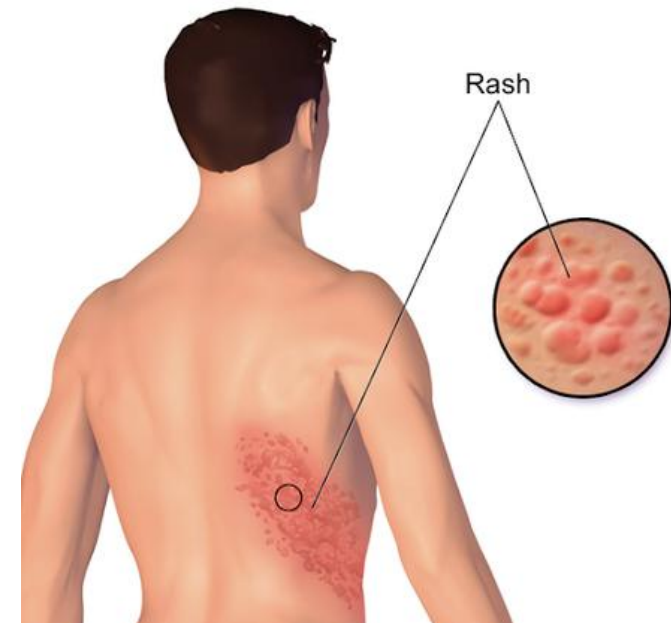




### IL VIRUS PUÒ RIATTIVARSI?

Il sistema immunitario produce anticorpi sufficienti a bloccare la malattia, ma il virus **può restare inattivo nel tessuto nervoso per anni o anche per tutta la vita.**

## HUMANITAS



**160.000 nuovi casi ogni anno** (su persone che hanno contratto la varicella in passato).

**6,3 casi ogni 1.000 persone-anno** (al di sopra dei 50 anni)

**circa due terzi dei casi** si manifestano in persone che hanno **più di 50 anni.**

L'Herpes Zoster è una patologia che può essere altamente invalidante e il rischio di svilupparla nel corso della vita per un individuo adulto è circa **1 su 3**

## TARN (Trauma Audit & Research Network)

### ➤ 1988: Report by RCS: The Management of patients with Major Injuries

*"Serious deficiencies in the management of severely injured patients"*

- Enhancing pre-hospital care, ensuring appropriate medical intervention
- Rapid transfer to the best local facility
- Assessing the use of helicopters
- Adopting ATLS principles
- Integrating trauma services within and between hospitals
- Investing in rehabilitation services

#### ■ System of Auditing & Research



### ➤ TARN Established in 1989: Based Salford Royal Hospital



# Dati di cui conosciamo la mancanza



**Evgeny Mirkes**

**«Handling missing data in large healthcare dataset: a case study of unknown trauma outcomes»**

**Studio su 165.559 casi di trauma, di cui 19.289 casi con esito sconosciuto.**



# Dati di cui conosciamo la mancanza



**165.559 casi di trauma, di cui 19.289 casi (11,6%) con esito sconosciuto**

**Nella ricerca traumatologica, conoscere l'esito significa sapere se il paziente è sopravvissuto o no per almeno trenta giorni dopo il trauma**

**Che fine hanno fatto i 19.289 casi?**

## **1° caso**

tutte le 146.270 persone con esiti noti sono sopravvissute e ristabilite senza alcun trattamento

(i 19.289 esiti ignoti sono morti entro due giorni dall'ammissione ma lo ignoriamo)

# Dati di cui conosciamo la mancanza



**165.559 casi di trauma, di cui 19.289 casi (11,6%) con esito sconosciuto**

**Nella ricerca traumatologica, conoscere l'esito significa sapere se il paziente è sopravvissuto o no per almeno trenta giorni dopo il trauma**

**Che fine hanno fatto i 19.289 casi?**

## **1° caso**

tutte le 146.270 persone con esiti noti sono sopravvissute e ristabilite senza alcun trattamento

(i 19.289 esiti ignoti sono morti entro due giorni dall'ammissione ma lo ignoriamo)

**Risultato: non c'è bisogno di alcuna cura per ristabilirsi  
(ma almeno l'11,6% dei pazienti muore)**

# Dati di cui conosciamo la mancanza



**165.559 casi di trauma, di cui 19.289 casi (11,6%) con esito sconosciuto**

**Nella ricerca traumatologica, conoscere l'esito significa sapere se il paziente è sopravvissuto o no per almeno trenta giorni dopo il trauma**

**Che fine hanno fatto i 19.289 casi?**

## **1° caso**

tutte le 146.270 persone con esiti noti sono sopravvissute e ristabilite senza alcun trattamento

(i 19.289 esiti ignoti sono morti entro due giorni dall'ammissione ma lo ignoriamo)

**Risultato: non c'è bisogno di alcuna cura per ristabilirsi  
(ma almeno l'11,6% dei pazienti muore)**

## **2° caso**

gli esiti di 19.289 pazienti non sono stati registrati perché sono morti tutti nei primi trenta giorni: se la valutazione dell'esito si basa sul contattare i pazienti trenta giorni dopo l'ammissione per sincerarsi delle loro condizioni, nessuno dei deceduti potrebbe rispondere alle domande

# Dati di cui conosciamo la mancanza



**165.559 casi di trauma, di cui 19.289 casi (11,6%) con esito sconosciuto**

**Nella ricerca traumatologica, conoscere l'esito significa sapere se il paziente è sopravvissuto o no per almeno trenta giorni dopo il trauma**

**Che fine hanno fatto i 19.289 casi?**

## 1° caso

tutte le 146.270 persone con esiti noti sono sopravvissute e ristabilite senza alcun trattamento

(i 19.289 esiti ignoti sono morti entro due giorni dall'ammissione ma lo ignoriamo)

**Risultato: non c'è bisogno di alcuna cura per ristabilirsi  
(ma almeno l'11,6% dei pazienti muore)**

## 2° caso

gli esiti di 19.289 pazienti non sono stati registrati perché sono morti tutti nei primi trenta giorni: se la valutazione dell'esito si basa sul contattare i pazienti trenta giorni dopo l'ammissione per sincerarsi delle loro condizioni, nessuno dei deceduti potrebbe rispondere alle domande

**Risultato: potremmo non registrare mai la morte di molti pazienti**

# Dati di cui conosciamo la mancanza



**165.559 casi di trauma, di cui 19.289 casi (11,6%) con esito sconosciuto**

**Nella ricerca traumatologica, conoscere l'esito significa sapere se il paziente è sopravvissuto o no per almeno trenta giorni dopo il trauma**

**Che fine hanno fatto i 19.289 casi?**

## 1° caso

tutte le 146.270 persone con esiti noti sono sopravvissute e ristabilite senza alcun trattamento

(i 19.289 esiti ignoti sono morti entro due giorni dall'ammissione ma lo ignoriamo)

**Risultato: non c'è bisogno di alcuna cura per ristabilirsi (ma almeno l'11,6% dei pazienti muore)**

## 2° caso

gli esiti di 19.289 pazienti non sono stati registrati perché sono morti tutti nei primi trenta giorni: se la valutazione dell'esito si basa sul contattare i pazienti trenta giorni dopo l'ammissione per sincerarsi delle loro condizioni, nessuno dei deceduti potrebbe rispondere alle domande

**Risultato: potremmo non registrare mai la morte di molti pazienti**

## altri casi

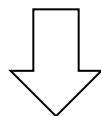
esiti non registrati per pazienti trasferiti in altri ospedali o altri reparti o non presentati alle visite successive

# Dati di cui conosciamo la mancanza

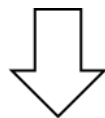


**conclusione finale dello studio:**

**gli esiti ignoti non sono mancanti «in modo casuale»  
non si può semplicemente ignorarli**



**errori significativi e distorsioni**

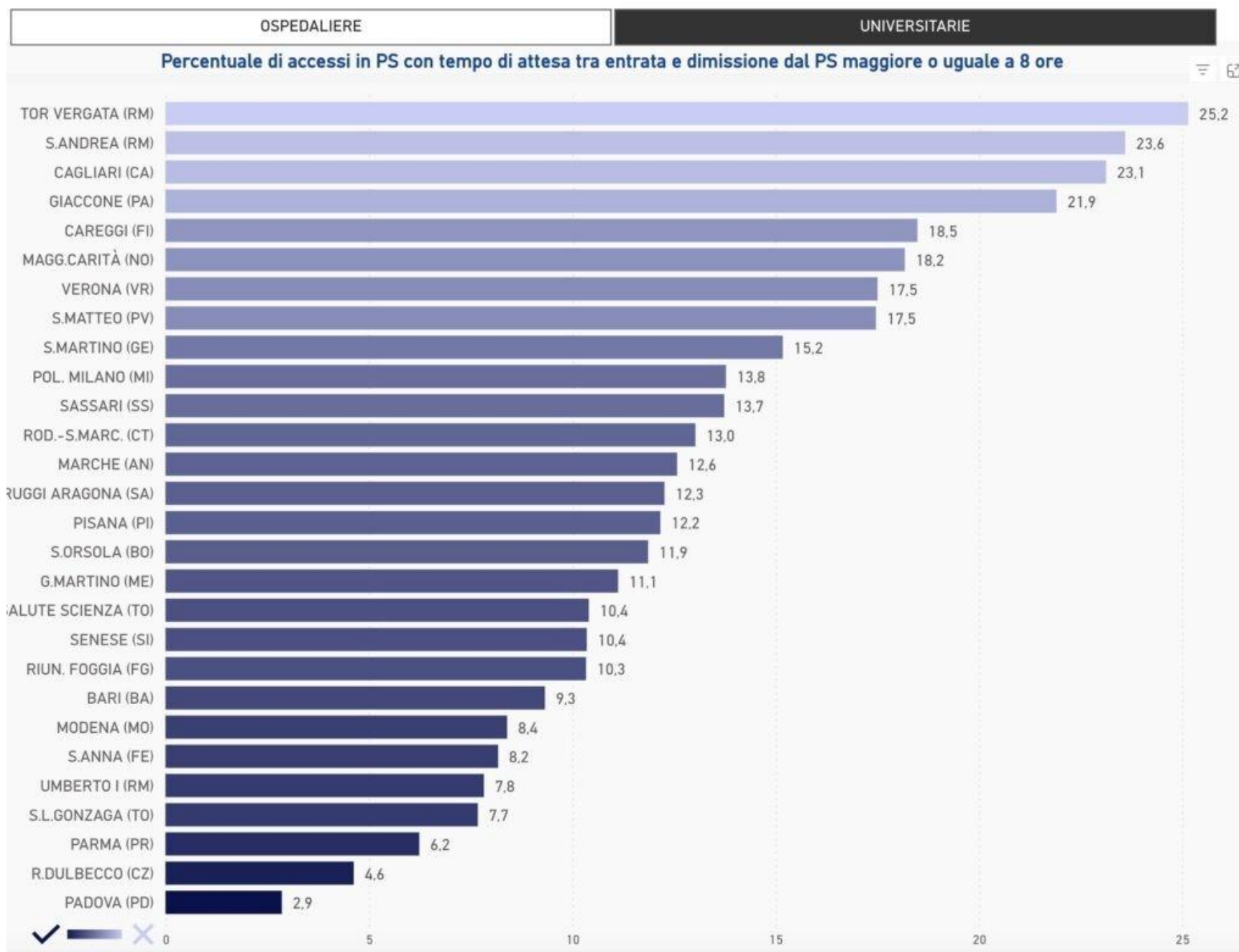


**conclusioni inaffidabili  
(tassi di mortalità, cure appropriate, ecc.)**

*Soluzione proposta:*

*utilizzo di un sistema di modelli di Markov non stazionari per gestire gli esiti mancanti, in modo da poterli includere nell'analisi*

# Dati di cui non conosciamo la mancanza



fonte: [quotidianosanità.it](http://quotidianosanità.it) del 2 dicembre 2025

# Dati di cui (non) conosciamo la mancanza

## esiti sconosciuti

### **Cartelle cliniche cartacee**

(o scansionate ma non analizzate digitalmente)

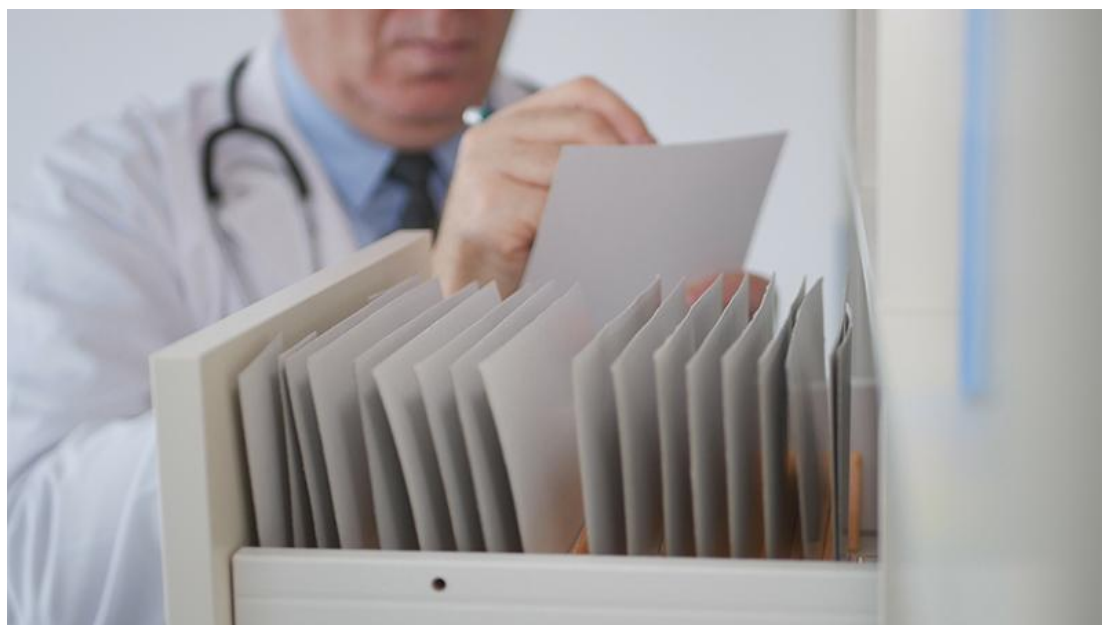
### **Note scritte a mano dai medici**

(archivate ma non strutturate)

### **Conversazioni con i pazienti**

(es. call center)

mai trascritte o analizzate



## contenuti (invisibili):

**segnali precoci di malattie**

**preferenze dei pazienti**

**inefficienze nei protocolli medici**

**ecc. ecc.**



# Il bias di sopravvivenza



**Abraham Wald**  
**1902-1950**  
*la teoria statistica delle decisioni*

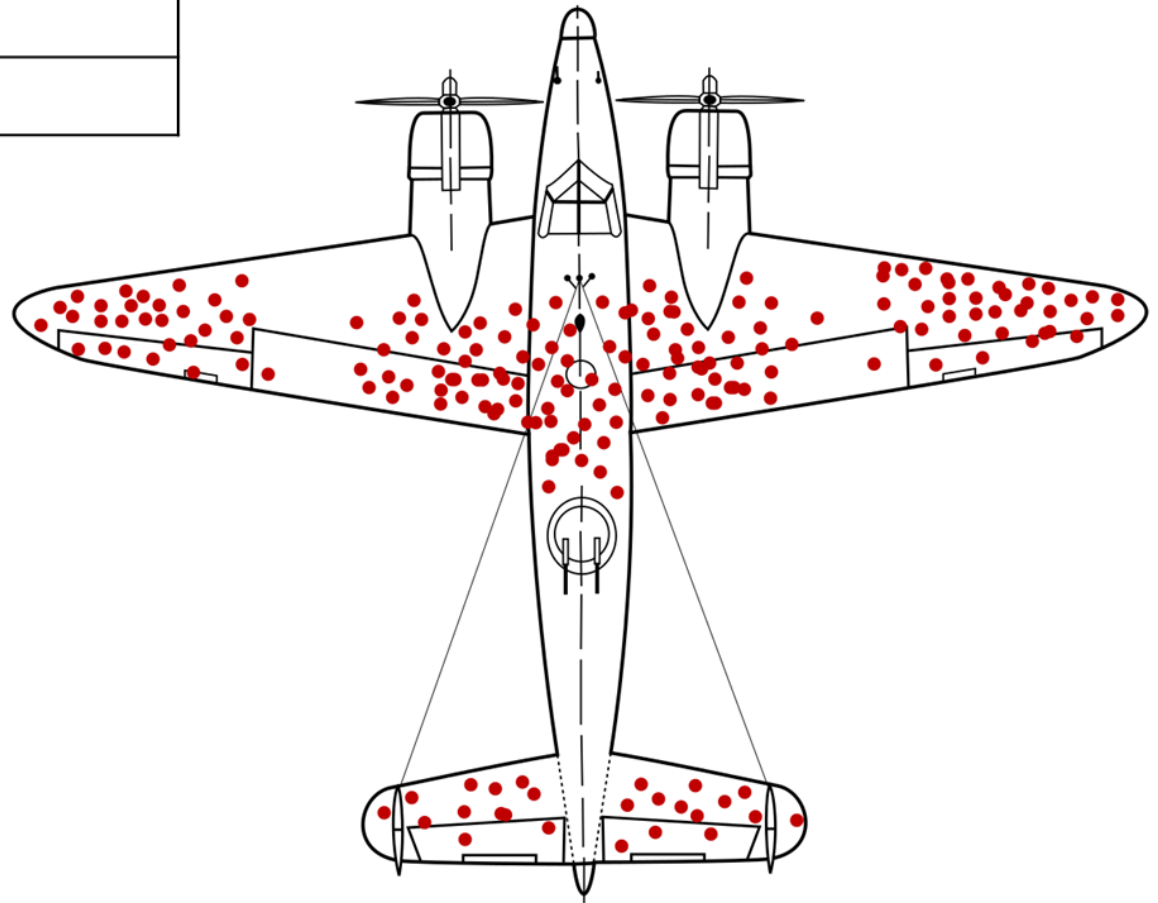
**la salvaguardia degli aerei e dei piloti  
nella seconda guerra mondiale**

**come corazzare un aereo?**



# Il bias di sopravvivenza

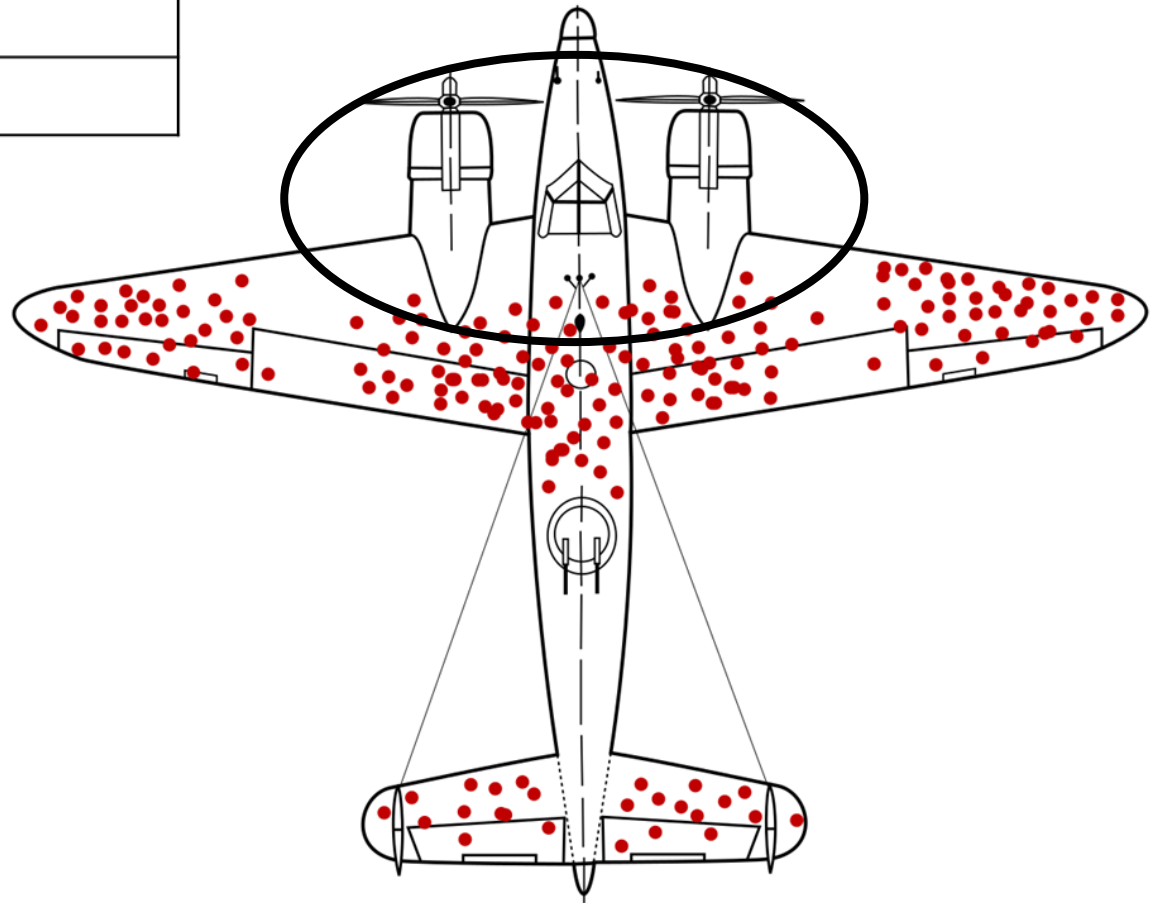
| <i>zona dell'aereo</i> | <i>numero di colpi/dm<sup>2</sup></i> |
|------------------------|---------------------------------------|
| motore                 | 0,103                                 |
| fusoliera              | 0,161                                 |
| alimentazione          | 0,143                                 |
| ali                    | 0,167                                 |



Martin Grandjean (vector), McGeddon (picture), US Air Force (hit plot concept) - opera propria, CC BY-SA 4.0, <https://commons.wikimedia.org/w/index.php?curid=102017718>

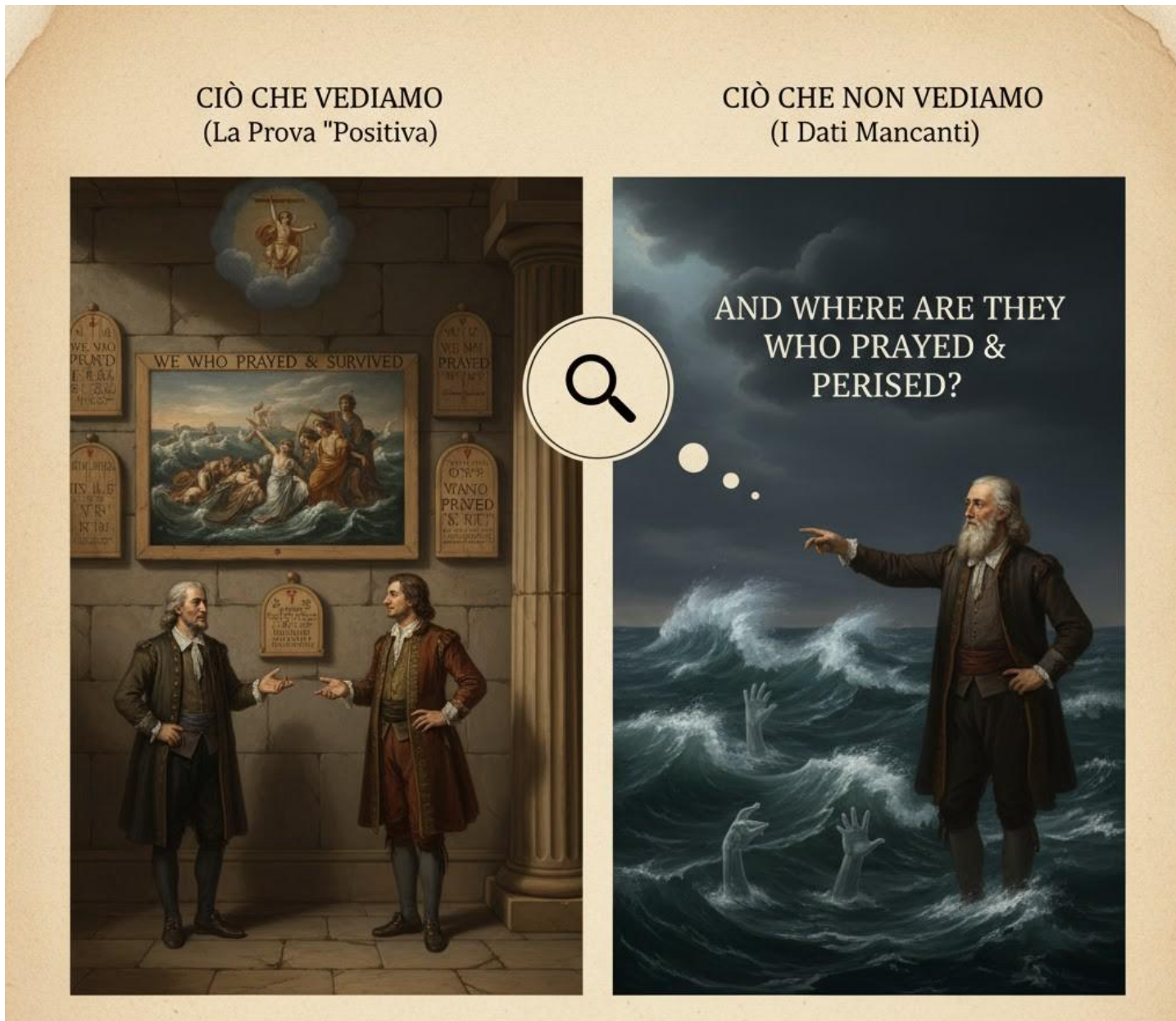
# Il bias di sopravvivenza

| <i>zona dell'aereo</i> | <i>numero di colpi/dm<sup>2</sup></i> |
|------------------------|---------------------------------------|
| motore                 | 0,103                                 |
| fusoliera              | 0,161                                 |
| alimentazione          | 0,143                                 |
| ali                    | 0,167                                 |



Martin Grandjean (vector), McGeddon (picture), US Air Force (hit plot concept) - opera propria, CC BY-SA 4.0, <https://commons.wikimedia.org/w/index.php?curid=102017718>

# Il bias di sopravvivenza

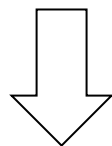


**L'aneddoto di Francis Bacon sulla pericolosità di ignorare i dati**

# Il bias di sopravvivenza

tendenza naturale della mente umana

prendiamo decisioni basandoci solo  
sugli elementi che sono "sopravvissuti" a un processo,  
ignorando quelli che sono stati eliminati o falliti



ci concentriamo sugli esempi che confermano le nostre credenze  
dimenticando gli esempi che le contraddirebbero



**vediamo:**  
i successi, i dati positivi,  
i casi che sono "tornati indietro"  
(i sopravvissuti)



**ignoriamo:**  
i fallimenti, i dati mancanti,  
i casi che sono stati eliminati  
(i non-sopravvissuti)

# Dati di cui non conosciamo la mancanza

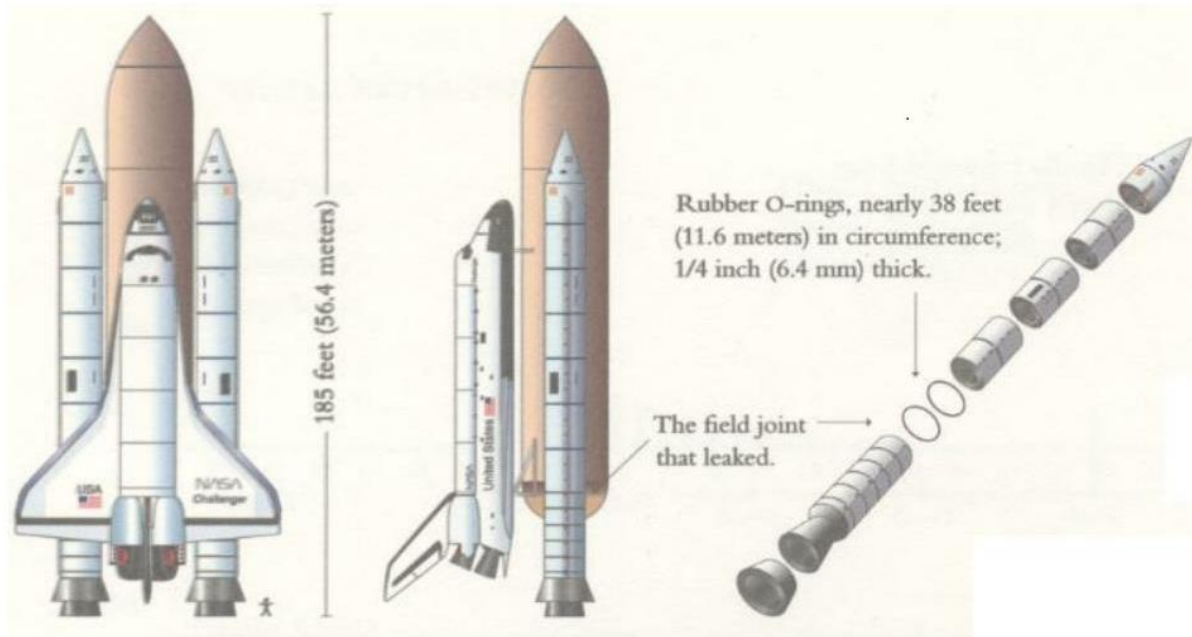


**CHALLENGER**

**28 gennaio del 1986**

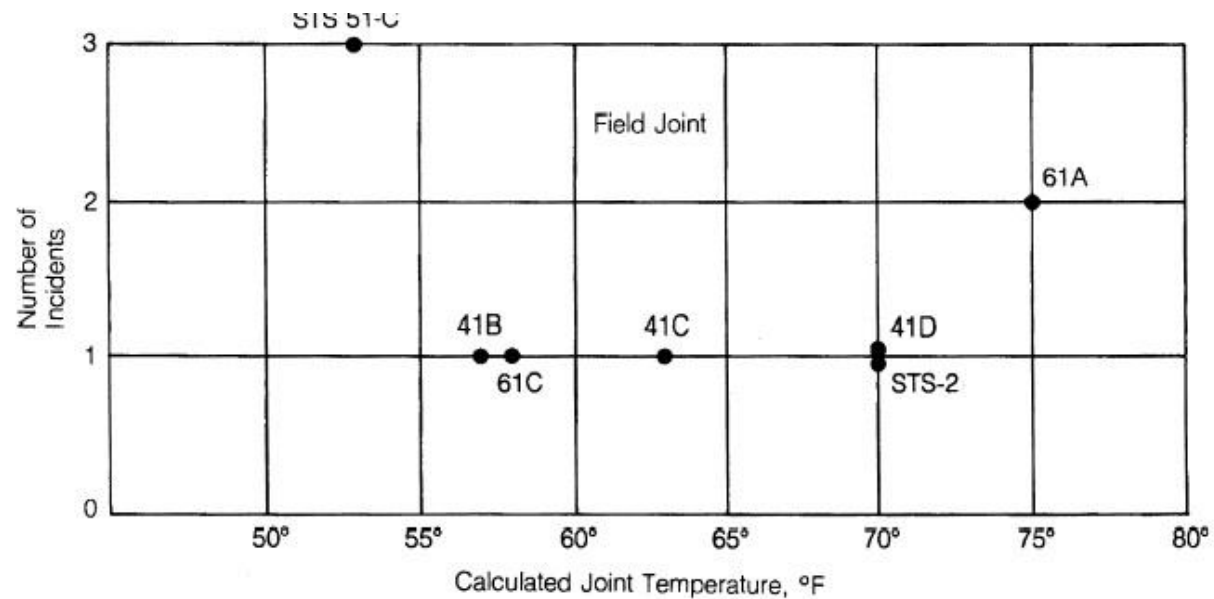


# Dati di cui non conosciamo la mancanza

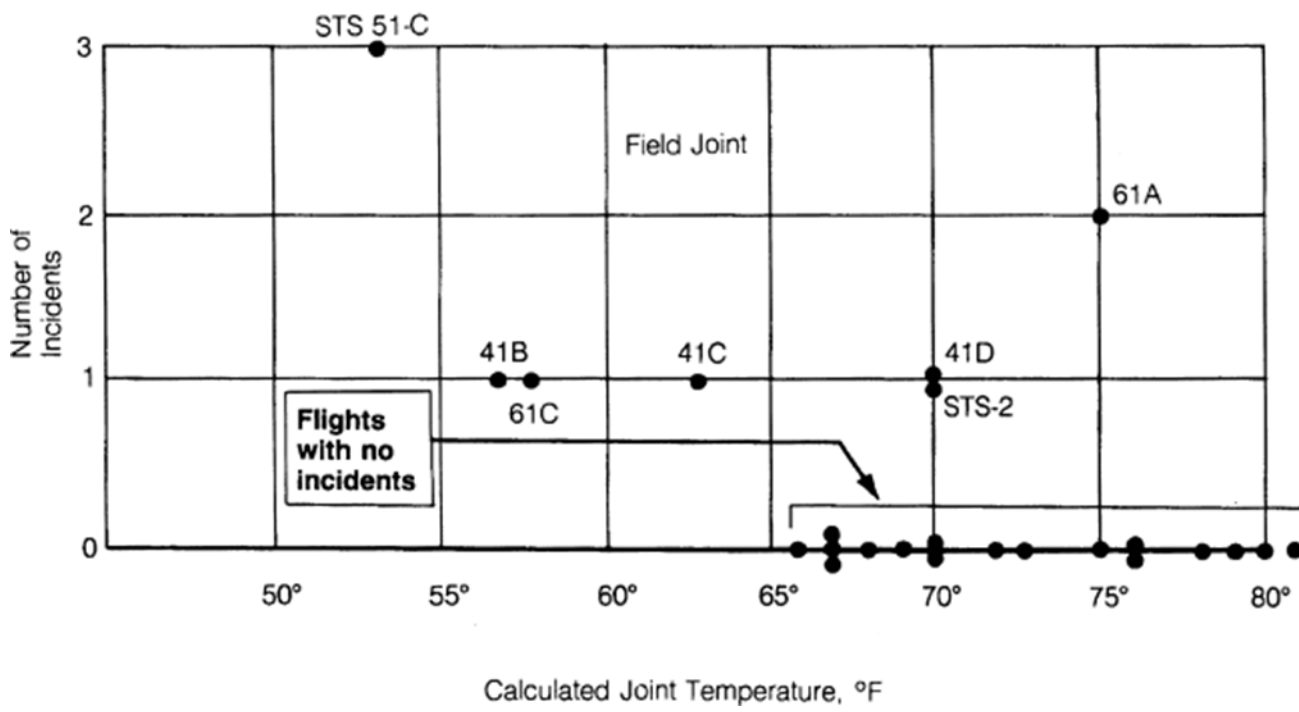
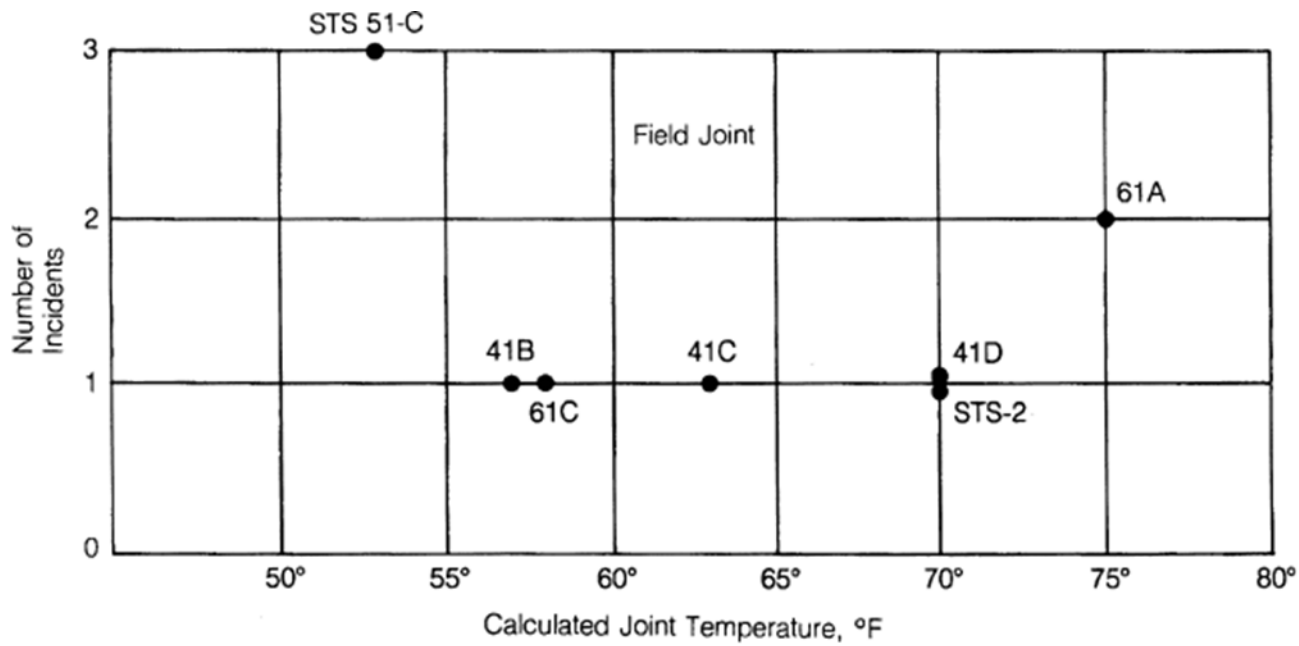


# Dati di cui non conosciamo la mancanza

## Voli che hanno presentato danni agli o-ring

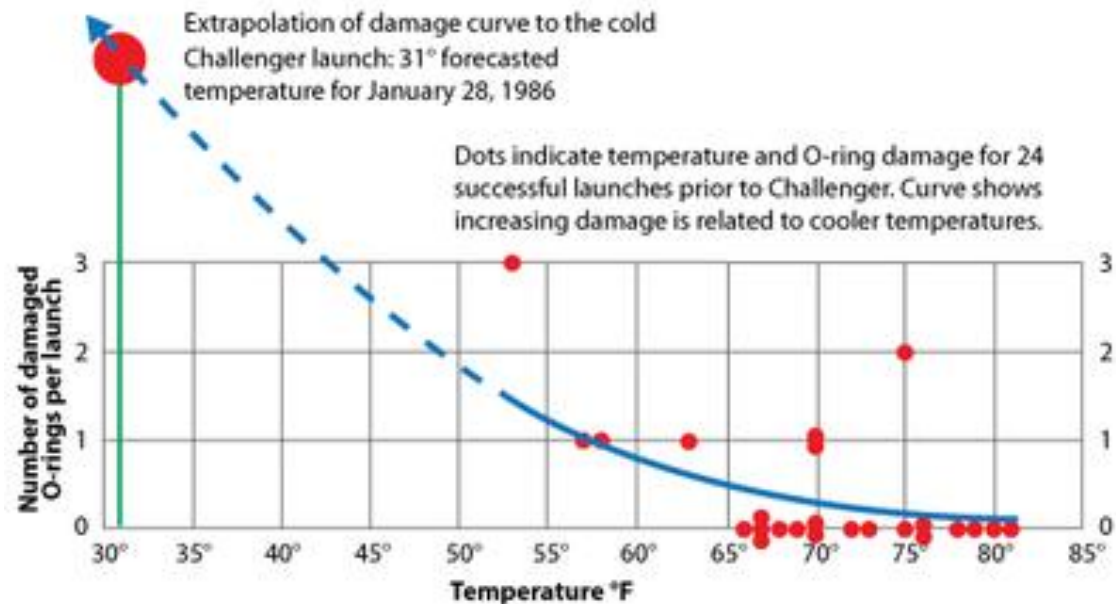
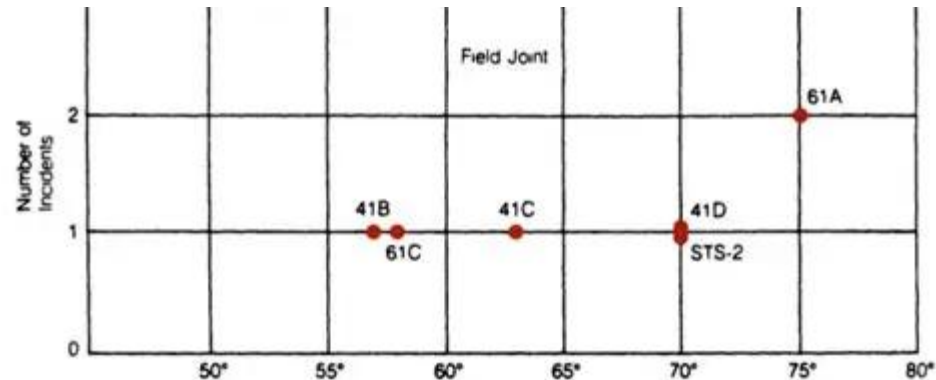


# Dati di cui non conosciamo la mancanza



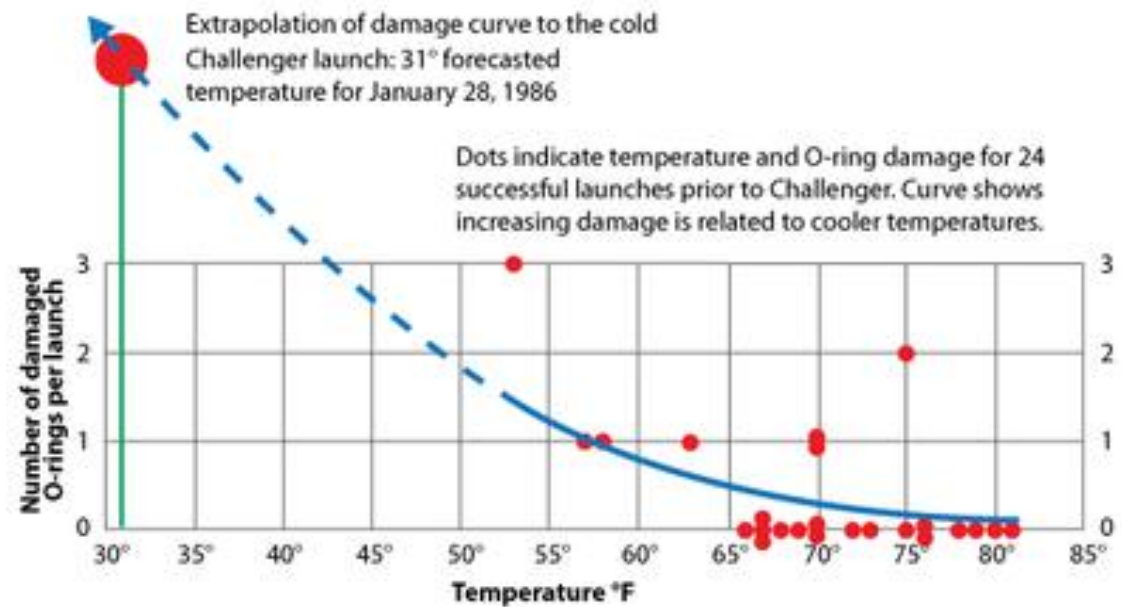
# Dati di cui non conosciamo la mancanza

Non era successo niente quindi non l'abbiamo preso in considerazione



# Dati di cui non conosciamo la mancanza

Non era successo niente quindi non l'abbiamo preso in considerazione



# Dati di cui non conosciamo la mancanza



**Boston**

**la manutenzione  
stradale**

**e il progetto  
Street Bump**



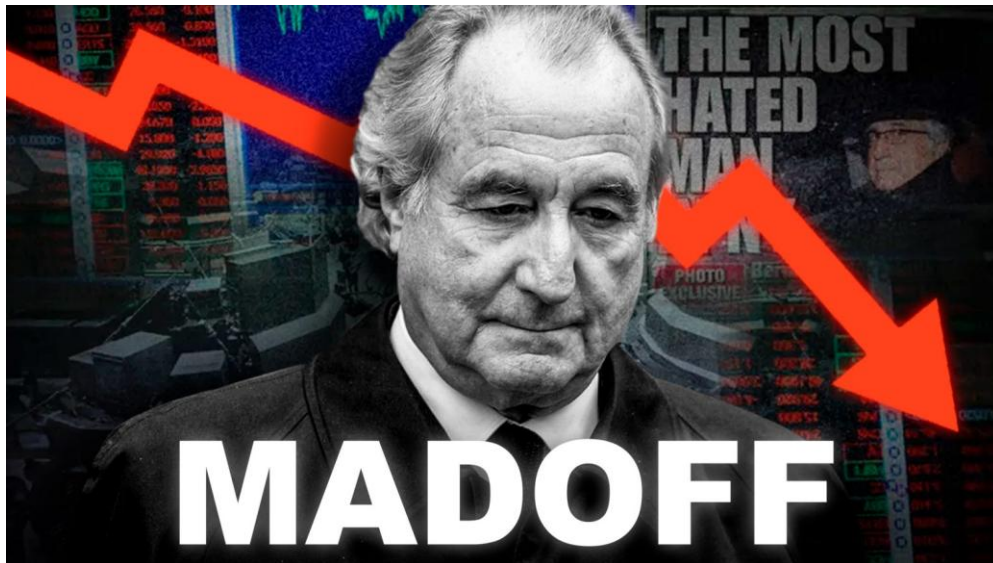
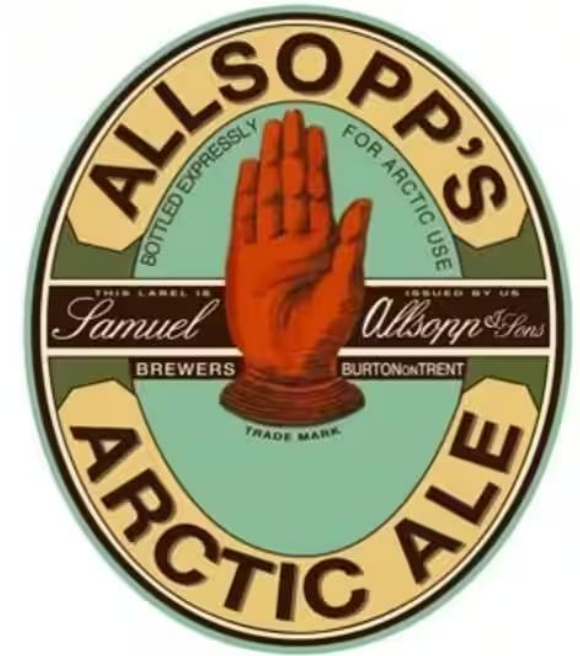
# Dati di cui non conosciamo la mancanza

**Uragano  
Sandy  
Ottobre 2012**



# Dati di cui non conosciamo la mancanza

la birra  
**ALLSOPP**



il caso  
**BERNIE MADOFF**

# Come raccogliere i dati



**Raccogliere i dati solo su alcuni dei membri (es. campionamenti)**

**CENSIMENTO  
POPOLAZIONE  
e ABITAZIONI**



**censimenti  
inventari**



**Cambiare le condizioni**

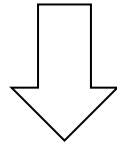


# Il caso dell'oscuramento

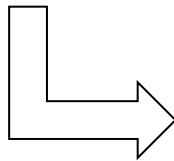


# Dati di cui non conosciamo la mancanza

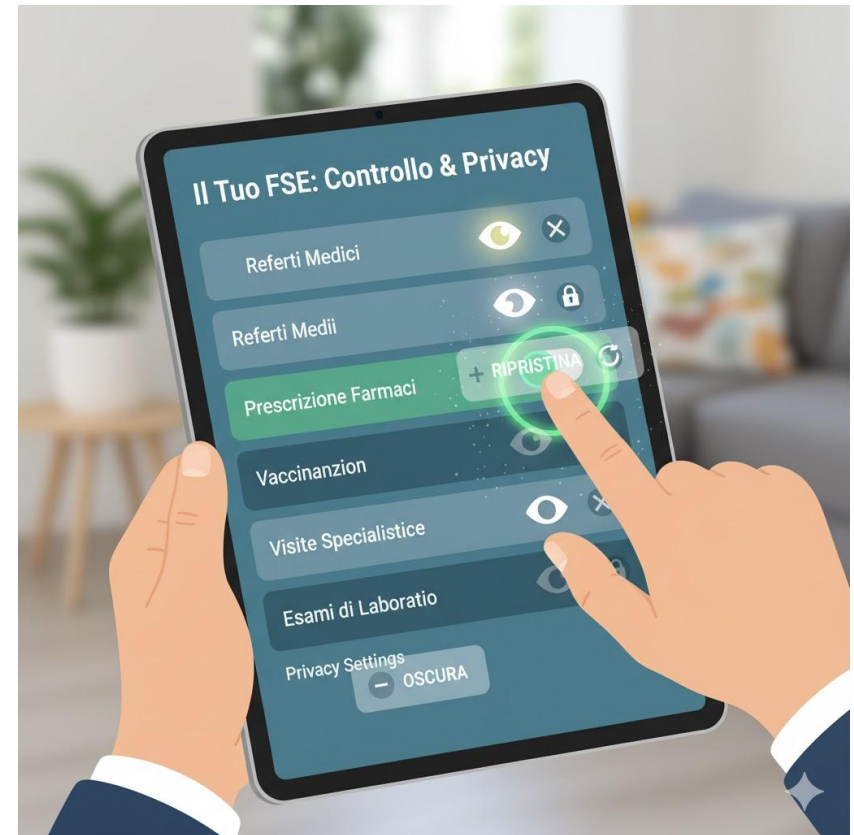
**possibilità per il cittadino  
di oscurare alcuni propri dati  
(e oscuramento a catena)**



**rischi clinici**  
**qualità dell'assistenza**  
**situazioni di emergenza**  
**possibilità di ripristino**



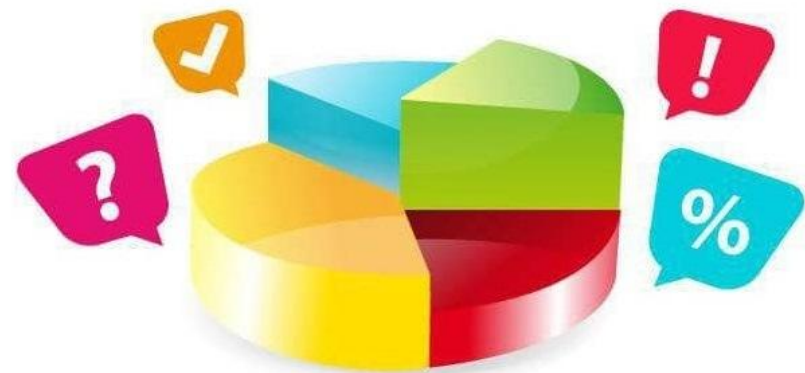
*delicato bilanciamento tra*  
**qualità dell'assistenza sanitaria**  
**e**  
**riservatezza dell'individuo**



# L'affidabilità dei sondaggi



**SONDAGGIO**



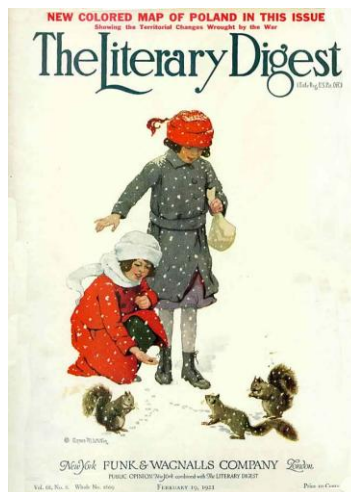
# L'affidabilità dei sondaggi



**Il problema  
delle non risposte**



# L'affidabilità dei sondaggi



elezioni presidenziali USA 1936

il sondaggio Literary Digest

# L'affidabilità dei sondaggi

elezioni presidenziali USA 1936

il sondaggio Gallup

Today  
Election Forecast

AMERICA SPEAKS  
THE NATIONAL WEEKLY POLL  
PUBLIC OPINION

Next Sunday  
The Election Results

INSTITUTE FORECASTS THE RE-ELECTION OF FRANKLIN D. ROOSEVELT,  
GIVES HIM 54% OF POPULAR VOTE, MINIMUM OF 315 ELECTORS

**Major Party Percent  
Is 55.7; New York in  
F.D.R. 'Sure' Column**

**AMERICAN INSTITUTE OF PUBLIC OPINION**  
NEW YORK, N.Y., Sept. 16. (AP)—The Institute of Public Opinion, which has been conducting a series of nationwide polls for the American Institute of Public Opinion, today announced that it has forecast the re-election of Franklin D. Roosevelt to the White House. The forecast is based on a poll of 10,000 people taken in 12 states and the District of Columbia, and is the first of a series of weekly polls to be conducted by the institute.

The poll shows that Roosevelt has the support of 55.7% of the people, while his opponent, Alf Landon, has the support of 44.3%. The poll also shows that Roosevelt has the support of 54% of the people in New York, which is considered a "sure" column for him.

The poll was conducted by the American Institute of Public Opinion, which was founded by George Gallup. The institute has been conducting polls for the American Institute of Public Opinion since 1935.

**Election Forecast**

1.—The American Institute of Public Opinion predicts the re-election of Franklin D. Roosevelt and John N. Garner.

2.—The Institute's latest presidential poll indicates that Roosevelt will receive approximately 55% of the major party vote (about 54% of the total vote) in 1936. Landon will receive 44% of the major party vote.

3.—With minor parties included, President Roosevelt's percentage of the total popular vote will be approximately 54%, to 46% for Landon.

4.—The President will receive a minimum of 315 electoral votes. The majority necessary to win is 270. Should last-minute shifts in the front of states where the vote is close and dark, this would give Roosevelt a margin of 315 to 285 electoral votes over Landon, who he polled 47%.

5.—William Larkin, candidate of the Union Party, will poll fewer than 1,000,000 popular votes, and carry no state.

6.—Harold Ickes, Socialist candidate, will poll about half as many votes as in 1932, when he polled 2,000,000.

**Election Will Test Clashing Poll Methods**

A clash of two types of poll taking is likely to occur in 1936. The American Institute of Public Opinion, which has been conducting a series of nationwide polls for the American Institute of Public Opinion, today announced that it has forecast the re-election of Franklin D. Roosevelt to the White House. The forecast is based on a poll of 10,000 people taken in 12 states and the District of Columbia, and is the first of a series of weekly polls to be conducted by the institute.

The poll shows that Roosevelt has the support of 55.7% of the people, while his opponent, Alf Landon, has the support of 44.3%. The poll also shows that Roosevelt has the support of 54% of the people in New York, which is considered a "sure" column for him.

The poll was conducted by the American Institute of Public Opinion, which was founded by George Gallup. The institute has been conducting polls for the American Institute of Public Opinion since 1935.

# L'affidabilità dei sondaggi

**che affidabilità  
possono avere  
i sondaggi?**



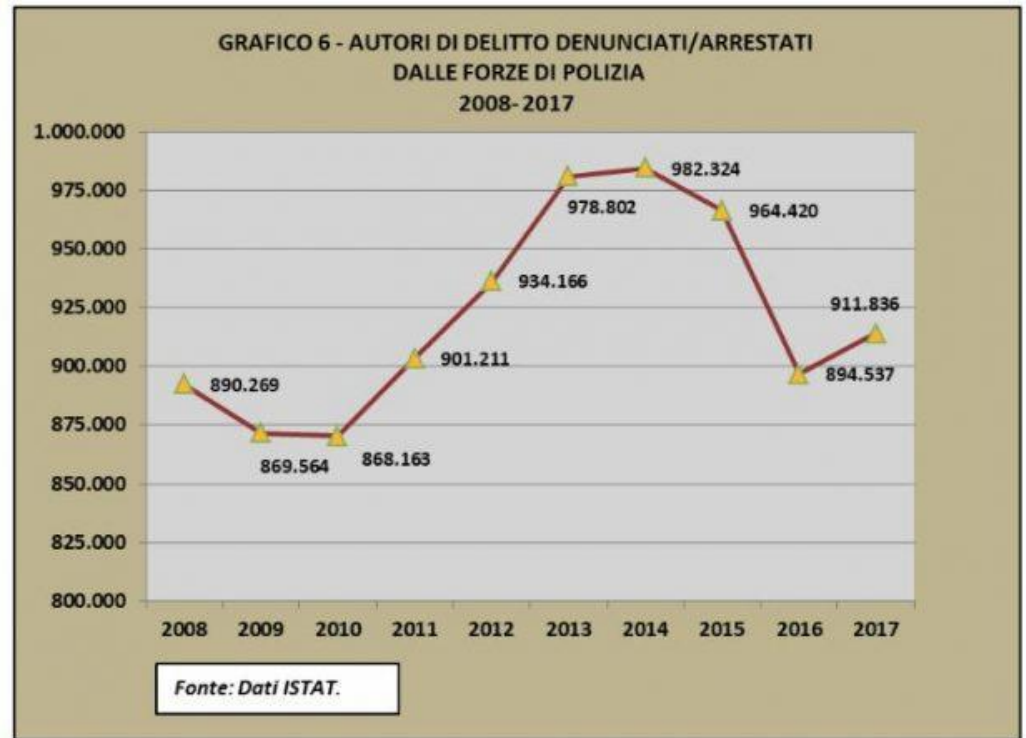
# Dati sperimentali e dark data



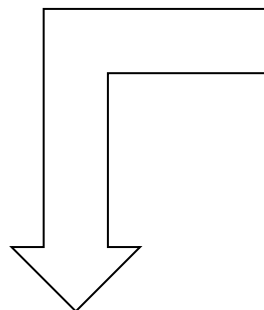
**sperimentazioni e abbandoni:  
la distorsione  
del sopravvissuto**

# Distorsione del sopravvissuto

riduzione del crimine?



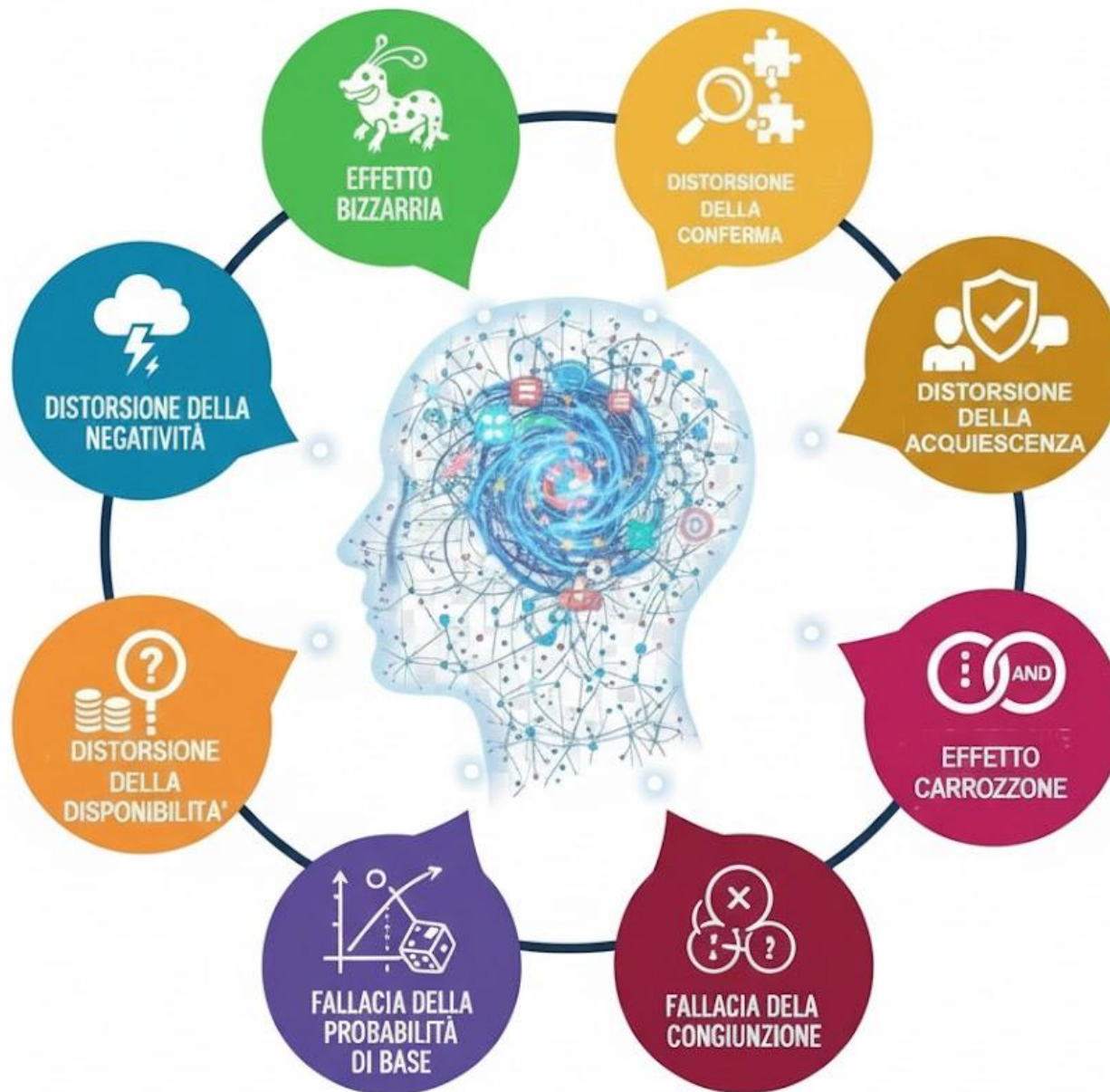
# Validità delle statistiche



| <i><b>tipo<br/>di reato</b></i>        | <i><b>tasso<br/>di non denuncia</b></i> |
|----------------------------------------|-----------------------------------------|
| <b>Borseggi .....</b>                  | <b>56,3%</b>                            |
| <b>Furti di oggetti personali.....</b> | <b>31,8%</b>                            |
| <b>Scippi.....</b>                     | <b>47,6%</b>                            |
| <b>Rapine.....</b>                     | <b>45,8%</b>                            |
| <b>Truffe su internet.....</b>         | <b>89,9%</b>                            |
| <i>(acquisti on-line)</i>              |                                         |
| <b>Clonazione di carte/conto.....</b>  | <b>40,6%</b>                            |

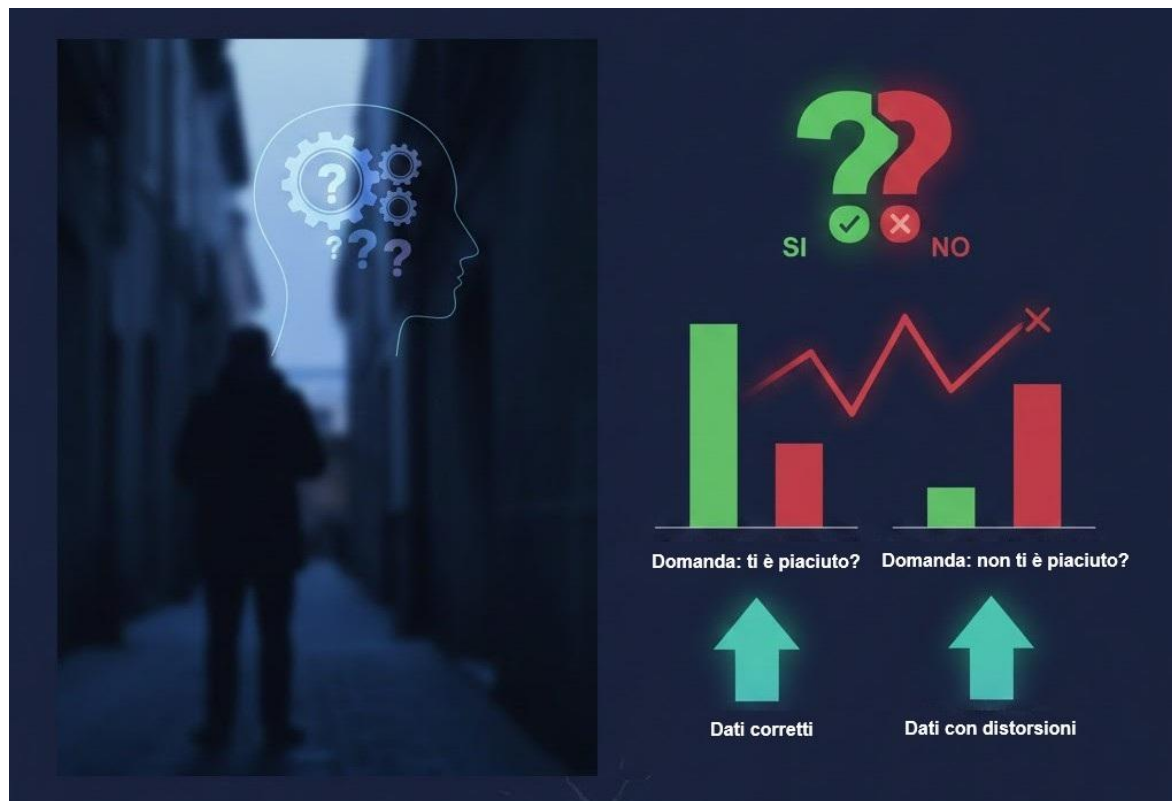
**Fonte:** Istat, Indagine sulla sicurezza dei cittadini, Report "Reati contro la persona e contro la proprietà: vittime ed eventi" (2022-2023).

# Fattore umano e distorsioni



# Altri tipi di dark data?

presentazione,  
riduzione,  
correlazione,  
definizione





# DARK DATA

RISCHI NASCOSTI. DECISIONI DISTORTE.

**STATE ALL'ERTA  
DAVANTI AI DATI!**