



Sapere utile

IFOA

Istituto Formazione Operatori Aziendali

BIG DATA e Analisi dei Dati

Lezione 5 – Rilevanza statistica


Mauro Bellone,
Robotics and AI researcher

bellonemauro@gmail.com
www.maurobellone.com

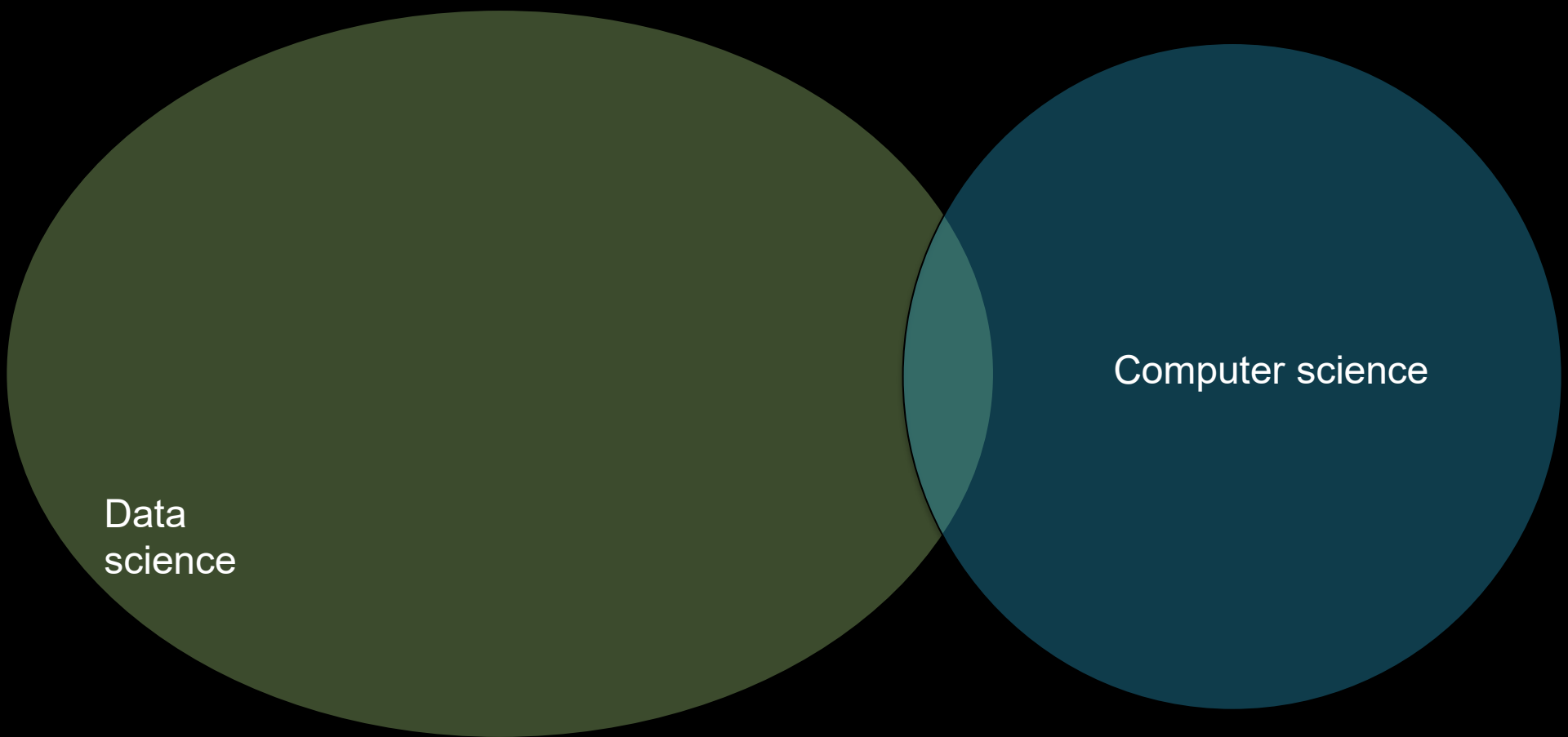
Obiettivo

- ✓ Comprensione della potenzialità statistica nei big data
- ✓ Elementi di probabilità
- ✓ Tecniche di visualizzazione e interpretazione dei risultati

*Rationality is not about knowing fact but
knowing which facts are relevant*

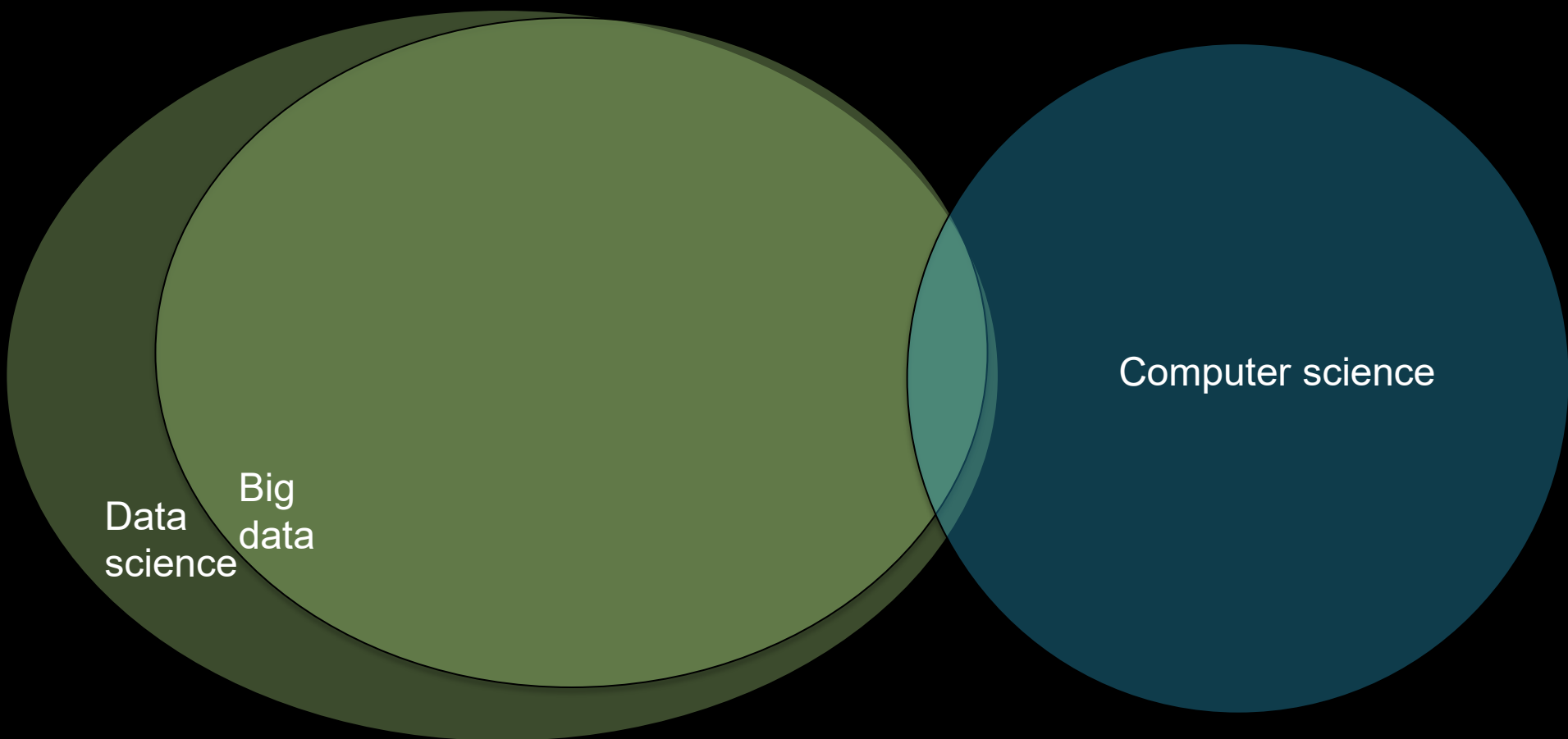


Computer science



Data
science

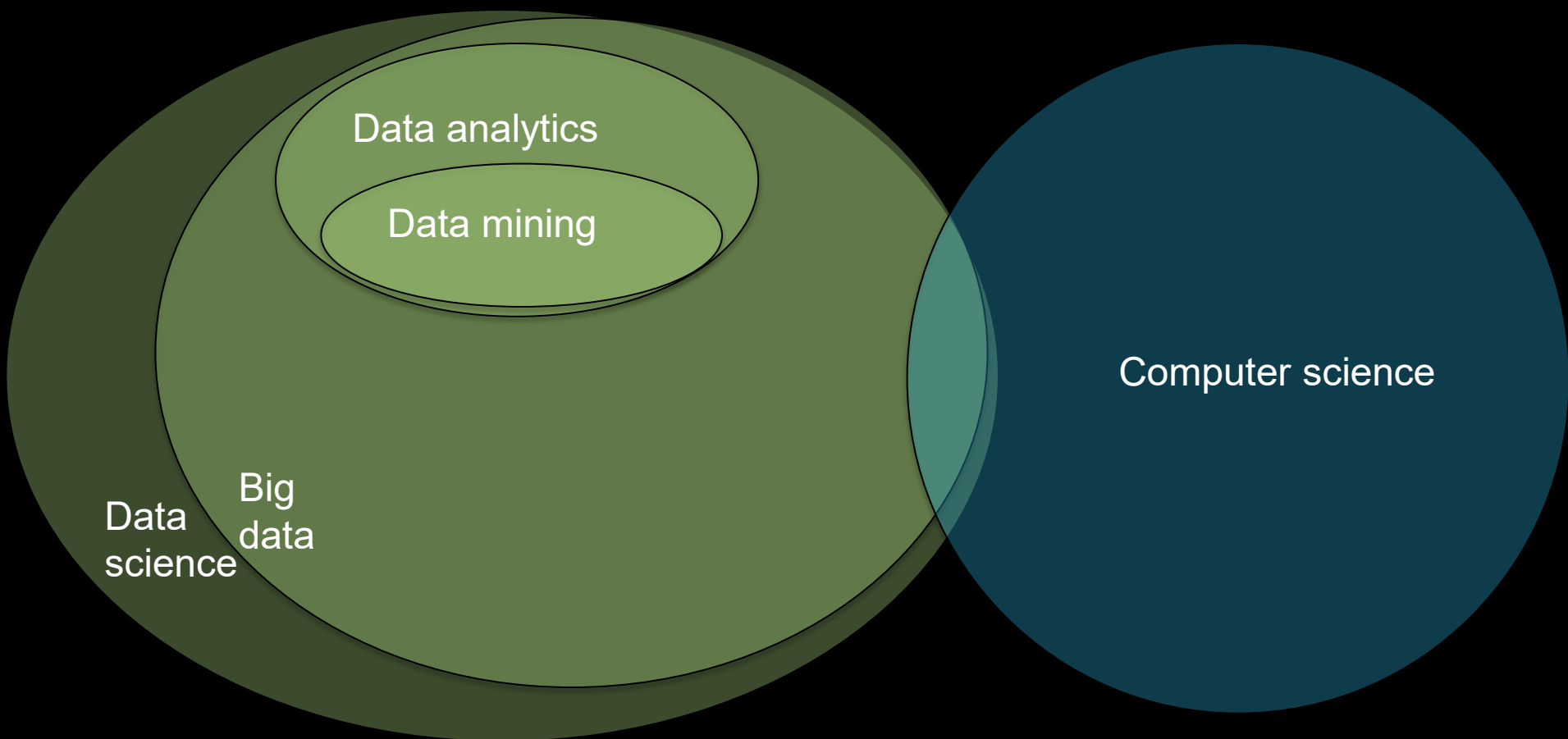
Computer science



Data
science

Big
data

Computer science



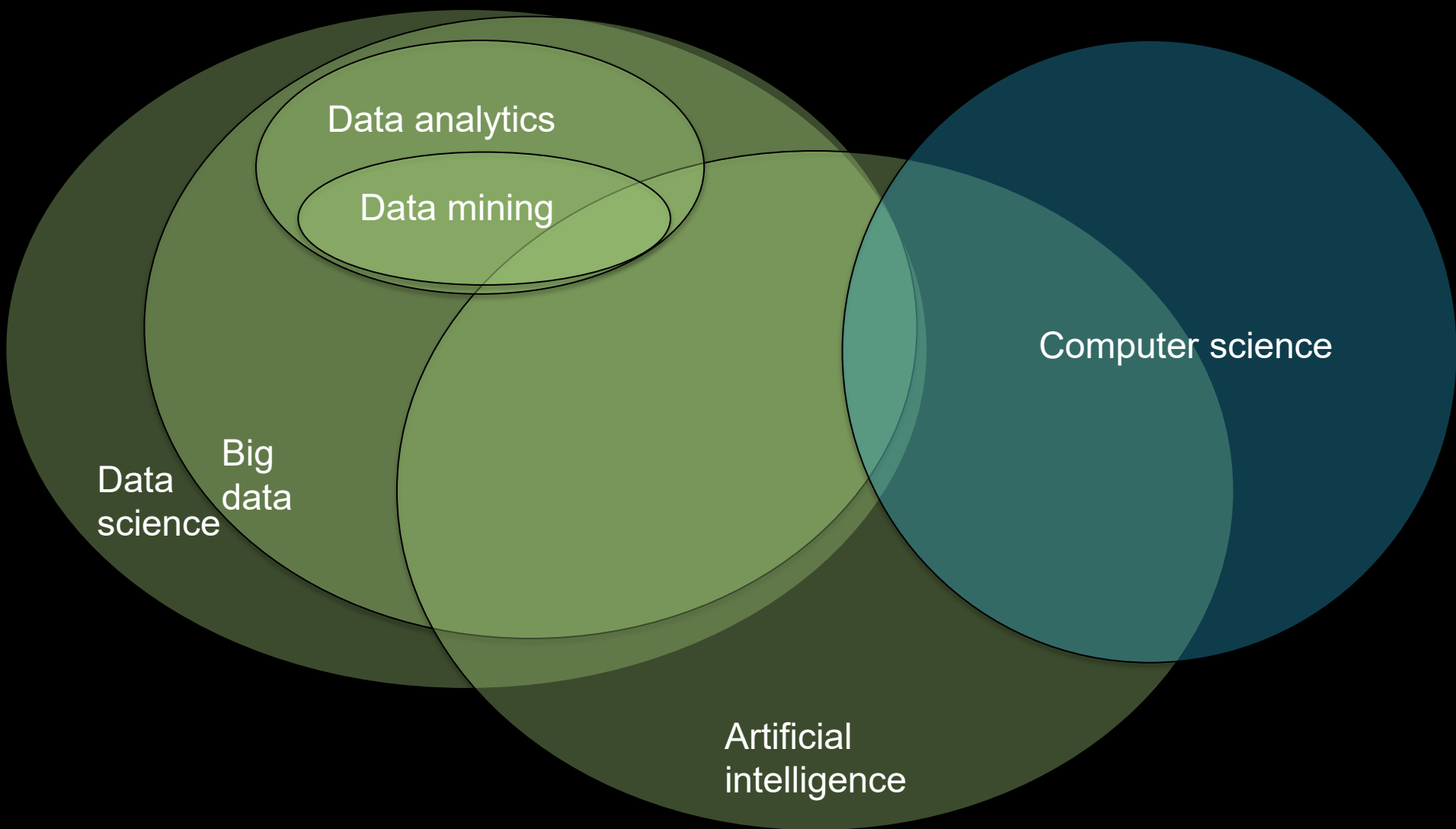
Data
science

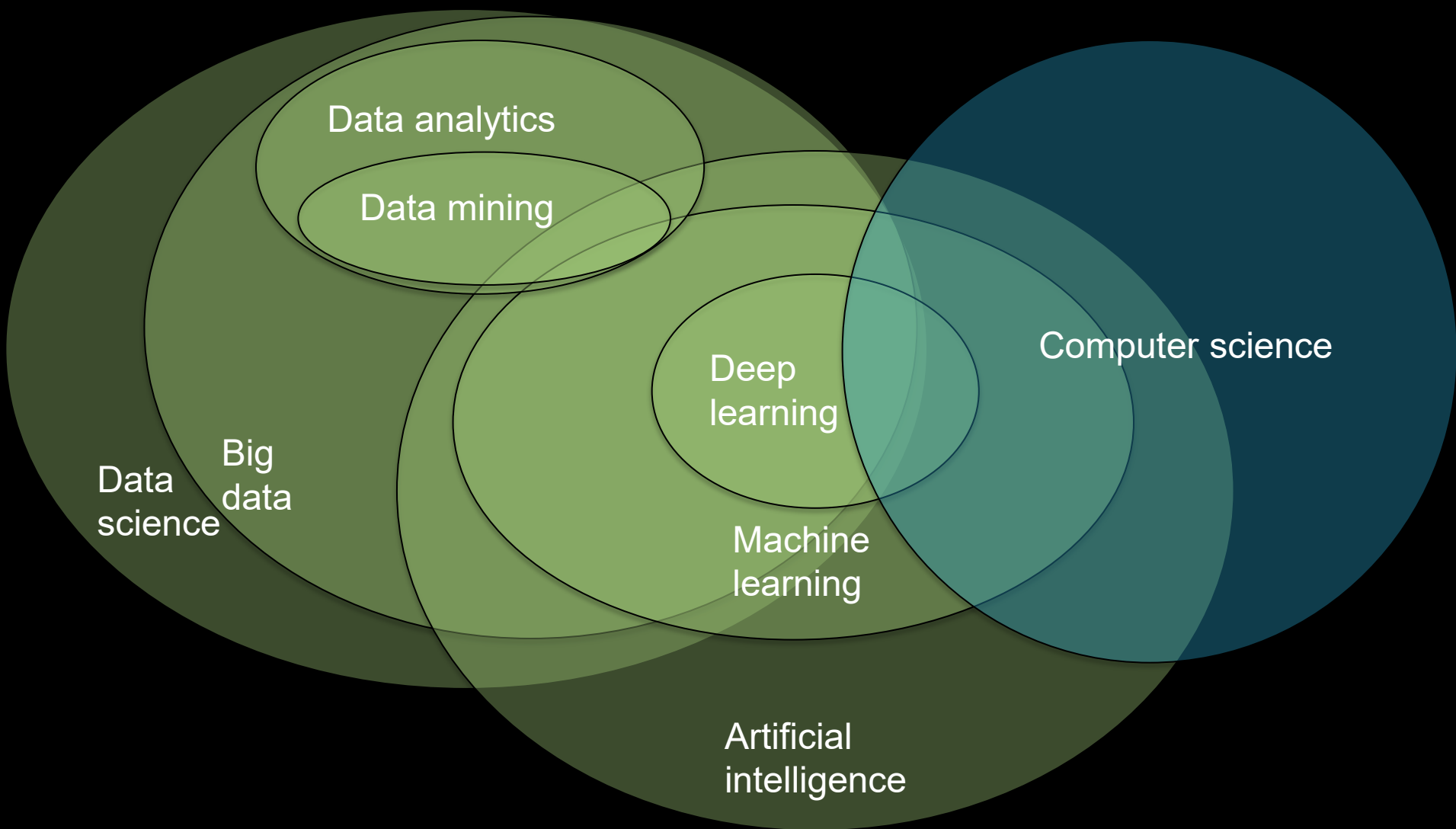
Big
data

Data analytics

Data mining

Computer science





Era dei big data

- ✓ Big data si riferisce a dataset così grandi e complessi che le moderne tecniche di processamento dati risultano inadeguate. Le sfide includono analisi, acquisizione, trattamento, ricerca, condivisione, memorizzazione, visualizzazione, interrogazione e privacy.

Era dei big data

- ✓ Big data si riferisce a dataset così grandi e complessi che le moderne tecniche di processamento dati risultano inadeguate. Le sfide includono analisi, acquisizione, trattamento, ricerca, condivisione, memorizzazione, visualizzazione, interrogazione e privacy.
- ✓ Il termine semplicemente si riferisce all'uso di analisi predittiva o metodi di estrazione del valore dai dati e solo raramente alla specifica dimensione dei dati.

Era dei big data

- ✓ Big data si riferisce a dataset così grandi e complessi che le moderne tecniche di processamento dati risultano inadeguate. Le sfide includono analisi, acquisizione, trattamento, ricerca, condivisione, memorizzazione, visualizzazione, interrogazione e privacy.
- ✓ Il termine semplicemente si riferisce all'uso di analisi predittiva o metodi di estrazione del valore dai dati e solo raramente alla specifica dimensione dei dati.
- ✓ L'accuratezza nell'ambito dei big data può portare a prendere decisioni migliori e più consapevoli, che portano minori costi e minori rischi.

Perchè la statistica è importante nell'ambito big data

- ✓ La statistica è la scienza che studia l'acquisizione, l'analisi, l'organizzazione, l'interpretazione e la presentazione di dati.

Perchè la statistica è importante nell'ambito big data

- ✓ La statistica è la scienza che studia l'acquisizione, l'analisi, l'organizzazione, l'interpretazione e la presentazione di dati.
- ✓ I big data si occupano di gestire grandi volume di dati con alta varietà e velocità di generazione

Perchè la statistica è importante nell'ambito big data

- ✓ La statistica è la scienza che studia l'acquisizione, l'analisi, l'organizzazione, l'interpretazione e la presentazione di dati.
- ✓ I big data si occupano di gestire grandi volume di dati con alta varietà e velocità di generazione
- ✓ Il maggiore valore del collezionare dati sta proprio nella possibilità di effettuare inferenza statistica altamente accurata

Perchè la statistica è importante nell'ambito big data

- ✓ La statistica è la scienza che studia l'acquisizione, l'analisi, l'organizzazione, l'interpretazione e la presentazione di dati.
- ✓ I big data si occupano di gestire grandi volume di dati con alta varietà e velocità di generazione
- ✓ Il maggiore valore del collezionare dati sta proprio nella possibilità di effettuare inferenza statistica altamente accurata
- ✓ Spesso si suppone di avere della conoscenza, o creare delle associazioni che in realtà non hanno un nesso di causalità

Problemi statistici nel big data

- ✓ Dati non indipendenti identicamente distribuiti (iid)
- ✓ I dati non soddisfano le nostre aspettative
- ✓ Non siamo riusciti ad osservare tutti i casi possibili

Legge dei grandi numeri

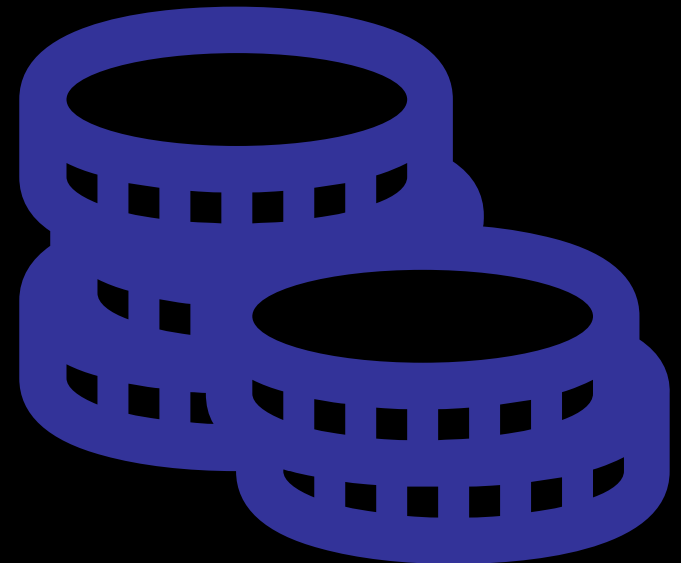
Il risultato di un esperimento eseguito n volte tende al suo valore atteso

Legge dei grandi numeri

Il risultato di un esperimento eseguito n volte tende al suo valore atteso

es.

lancio una moneta n volte, per n che tende a infinito, $1/2$ degli esperimenti avrà come esito “testa” e $1/2$ “croce”



Esempio - misura

Supponiamo di voler misurare una certa quantità, es. peso di una persona

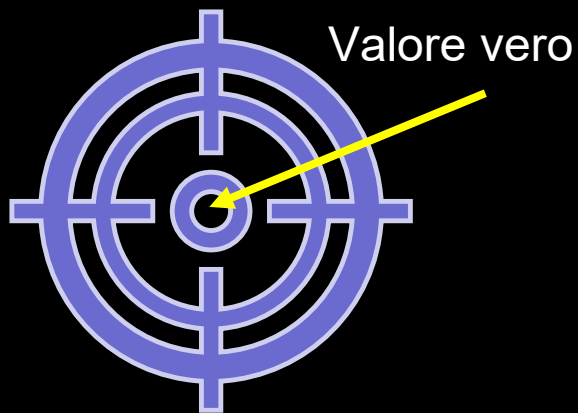


Tuttavia la bilancia è affetta da diversi tipi di errore:

- ✓ errore di sensibilità (quanto lo strumento è sensibile)
- ✓ errore sistematico (taratura dello strumento)
- ✓ errore causale o accidentale

Esempio - misura

Supponiamo di voler misurare una certa quantità, es. peso di una persona

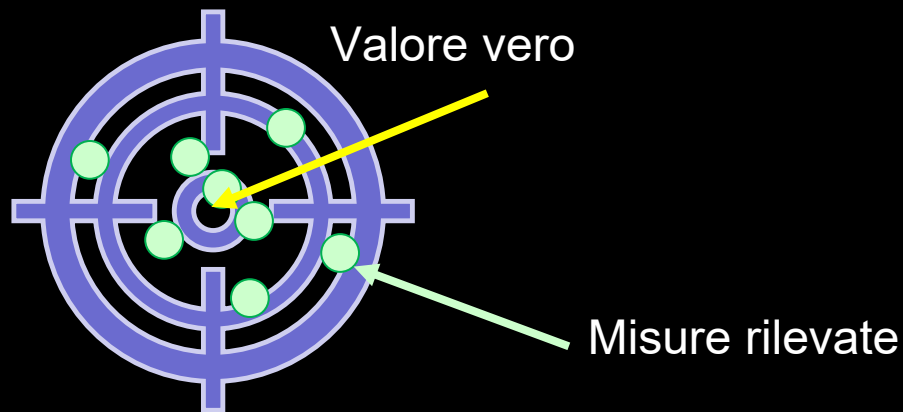


Tuttavia la bilancia è affetta da diversi tipi di errore:

- ✓ errore di sensibilità (quanto lo strumento è sensibile)
- ✓ errore sistematico (taratura dello strumento)
- ✓ errore causale o accidentale

Esempio - misura

Supponiamo di voler misurare una certa quantità, es. peso di una persona

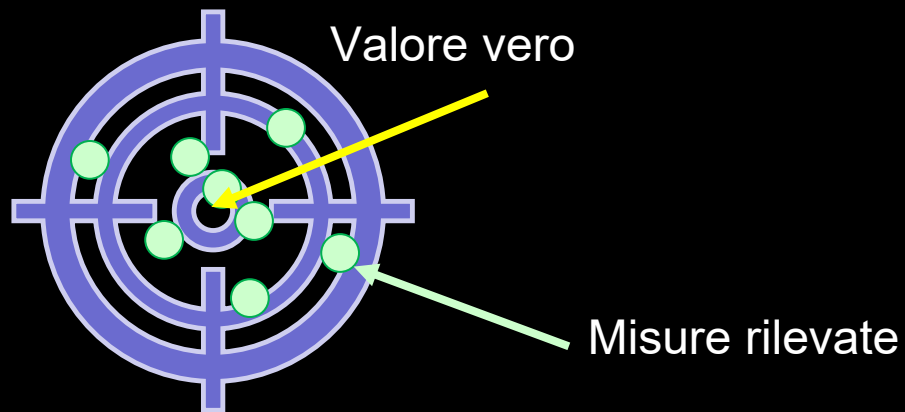


Tuttavia la bilancia è affetta da diversi tipi di errore:

- ✓ errore di sensibilità (quanto lo strumento è sensibile)
- ✓ errore sistematico (taratura dello strumento)
- ✓ errore causale o accidentale

Esempio - misura

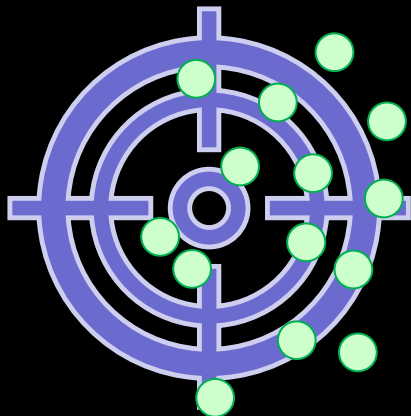
Supponiamo di voler misurare una certa quantità, es. peso di una persona



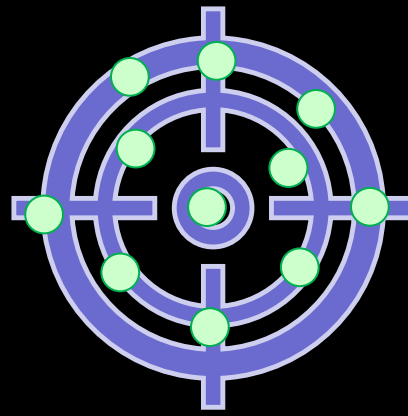
Al crescere del numero di misure, la media tra tutte le misure tende al valore vero

Esempio - misura

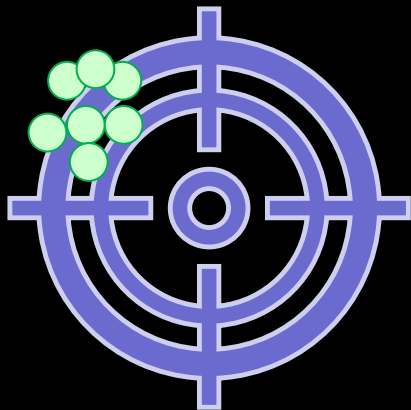
Non accurato
e non preciso



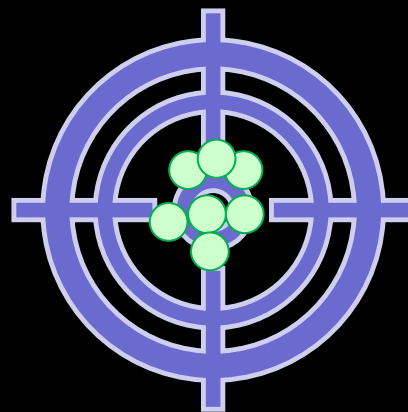
accurato ma
non preciso



Non accurato
ma preciso



Accurato e
preciso



Probabilità ed eventi

Dato un esperimento casuale (es. lancio dadi) la probabilità di verifica dell'evento corrisponde al rapporto tra il numero di volte che l'evento si verifica e la popolazione (numero di volte che lancio il dado)

$$p_evento_1 = num_eventi / popolazione$$

Probabilità ed eventi

Dato un esperimento casuale (es. lancio dadi) la probabilità di verifica dell'evento corrisponde al rapporto tra il numero di volte che l'evento si verifica e la popolazione (numero di volte che lancio il dado)

$$p_evento_1 = num_eventi / popolazione$$

popolazione = numero di volte che il dado è lanciato

evento_1 = il dado si ferma sul lato 1

evento_2 = il dado si ferma sul lato 2

evento_n = il dado si ferma sul lato n

Regole

- ✓ La probabilità che non succeda nulla è 0
- ✓ La probabilità che qualcosa succeda è 1
- ✓ La probabilità di qualcosa è 1 meno la probabilità che l'opposto accada $p = 1 - q$

Regole

- ✓ La probabilità che almeno una delle due (o più cose) che non possono accadere contemporaneamente (mutuamente esclusive) è la somma delle loro rispettive probabilità

$$P(A \cup B) = P(A) + P(B) \quad \text{se } P(A \cap B) = \emptyset$$

- ✓ Se un evento A implica l'occorezza dell'evento B, allora la probabilità di A è minore o uguale alla probabilità di B

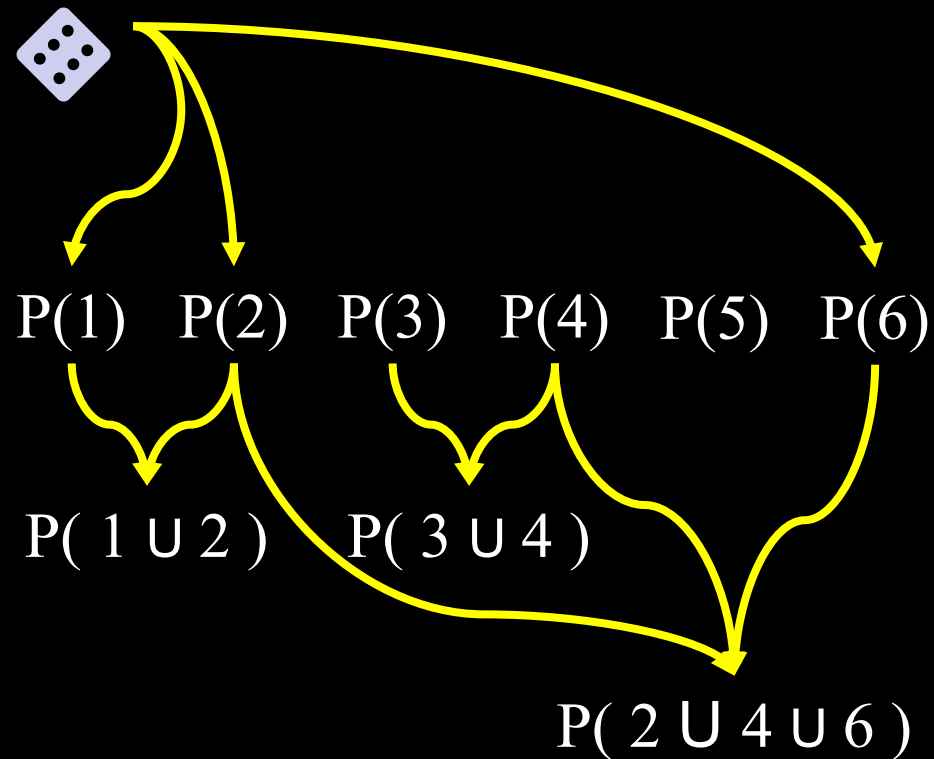
$$\text{Se } P(A) \in P(B) \text{ allora } P(A) \leq P(B)$$

- ✓ Per qualunque di due eventi, la probabilità che almeno uno accada è la somma delle loro probabilità meno la loro intersezione

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Probabilità ed eventi

Le probabilità possono essere di fatto indicate come degli insiemi mutuamente esclusivi



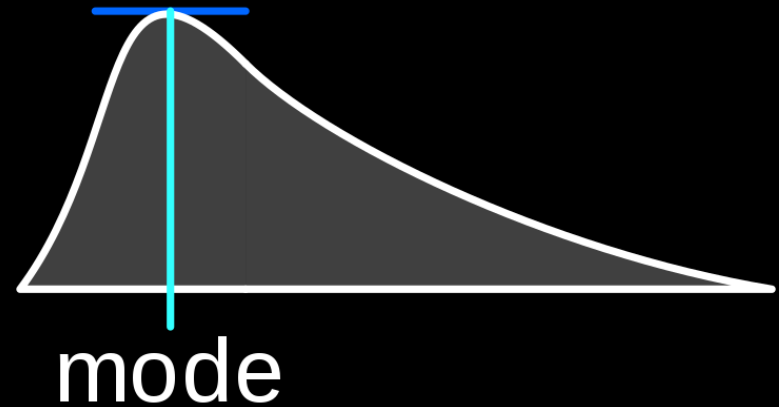
Probabilità ed eventi

Le probabilità possono essere di fatto indicate come degli insiemi mutuamente esclusivi

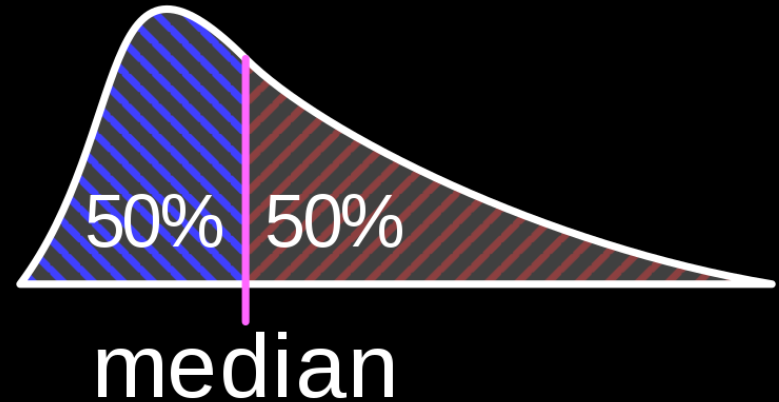
In ambito di probabilità, gli eventi si indicano con il termine “variabili casuali”, ed effettuando numerose prove sul mio campione ottengo una distribuzione della variabile casuale x

Metriche

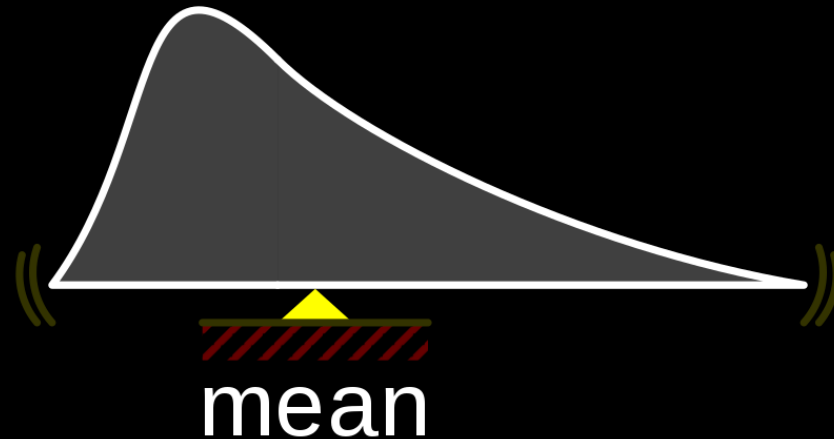
✓ Moda $\max[x]$



✓ Mediana $(x) = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}$



✓ Media $E[x] = \mu = \frac{\sum x}{N}$



Metriche

✓ Varianza $\sigma^2(x)$ e deviazione standard $\sigma(x)$

$$\sigma^2(x) = \sqrt{E[x - E[x]]^2}$$

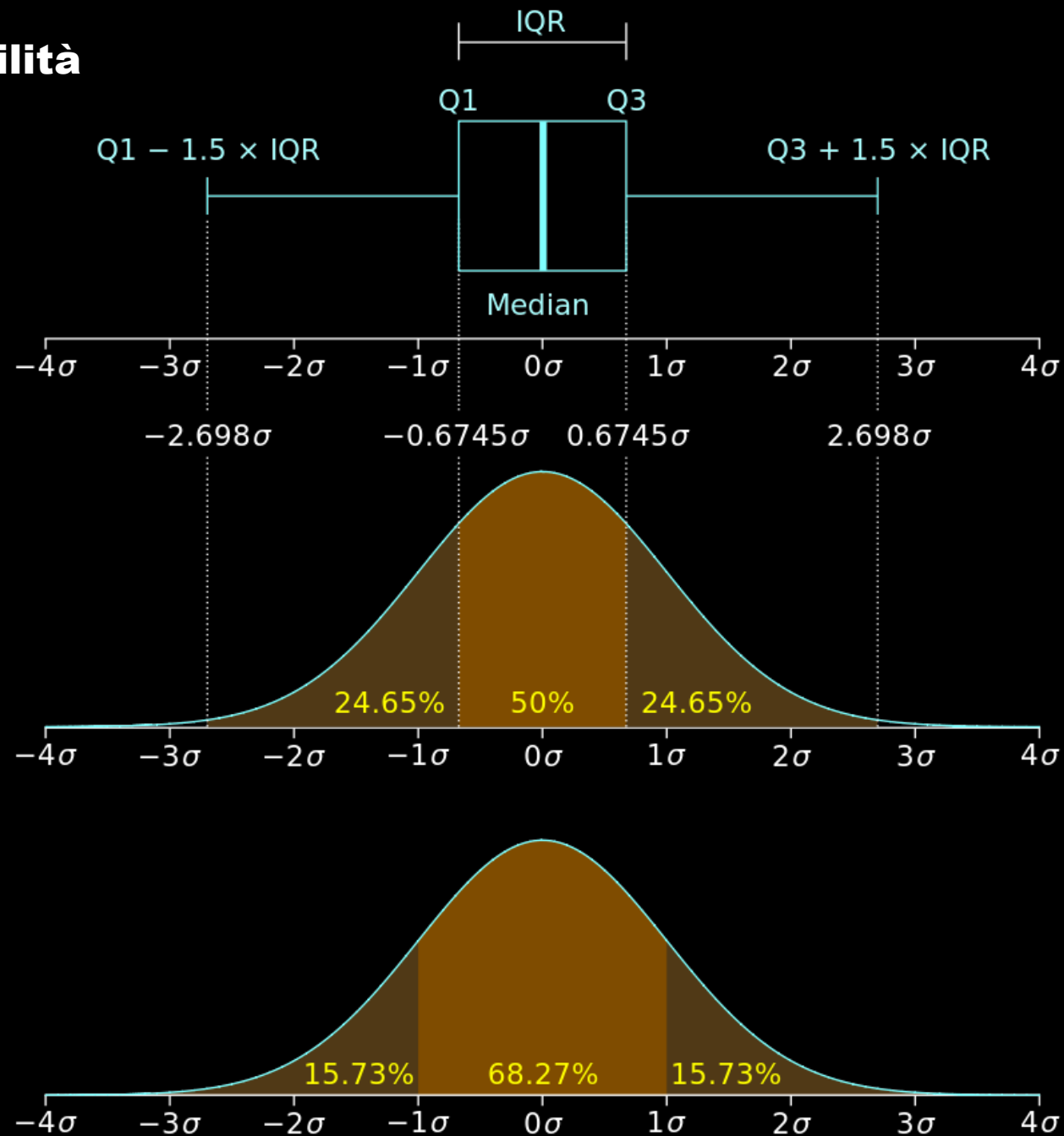
$$\sigma(x) = \sqrt{E[x - E[x]]^2}$$

Funzione di probabilità

La funzione di probabilità valutata su un valore corrisponde alla probabilità che una variabile causale prenda quel valore.

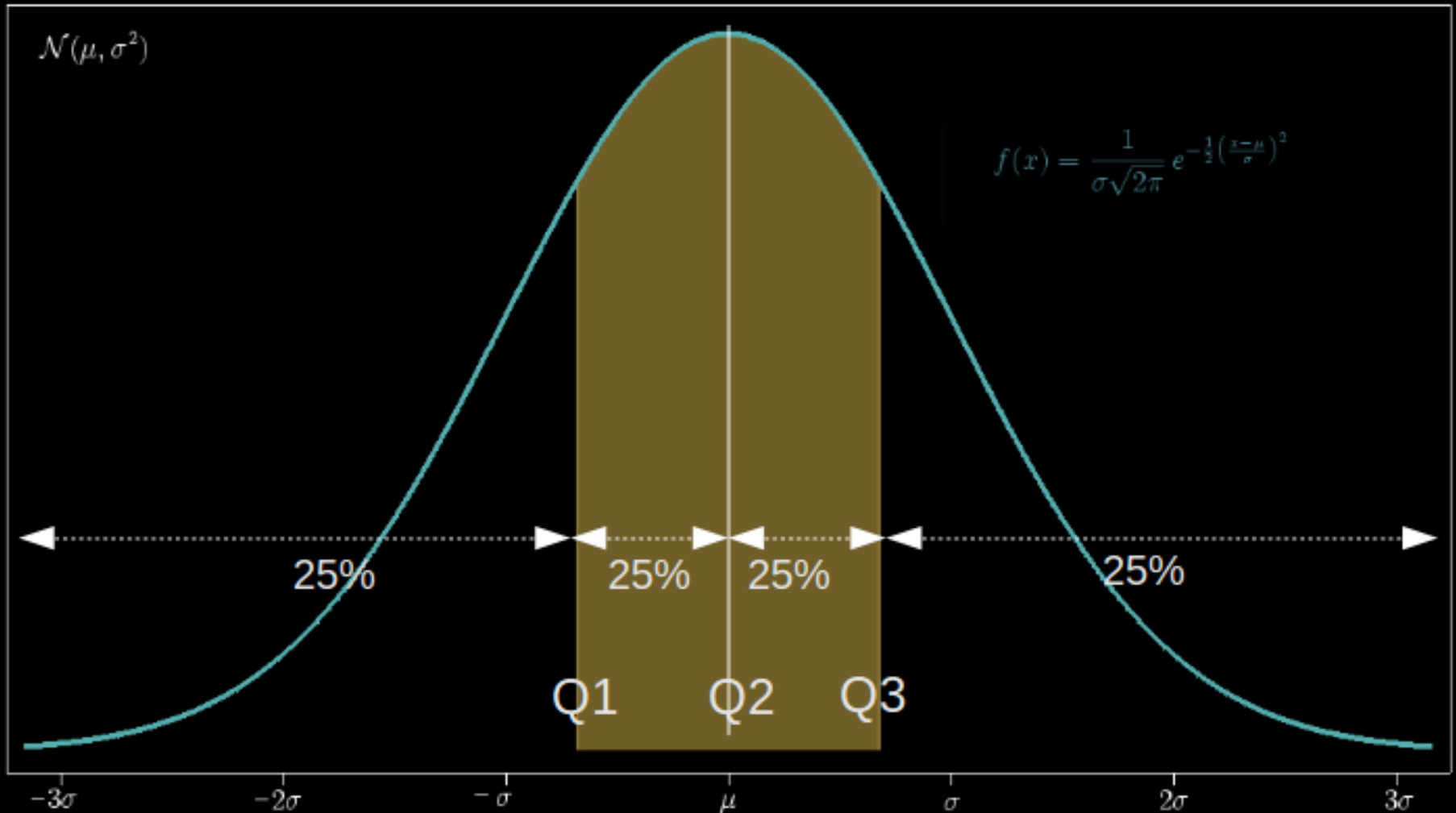
deve:

- ✓ essere maggiore o uguale a zero
- ✓ la somma di tutti i possibili valori deve fare uno



Quantili

I quantili sono punti di taglio della distribuzione tali da divider la distribuzione in parti uguali



Probabilità condizionata

Sia B un evento tale che $P(B) > 0$

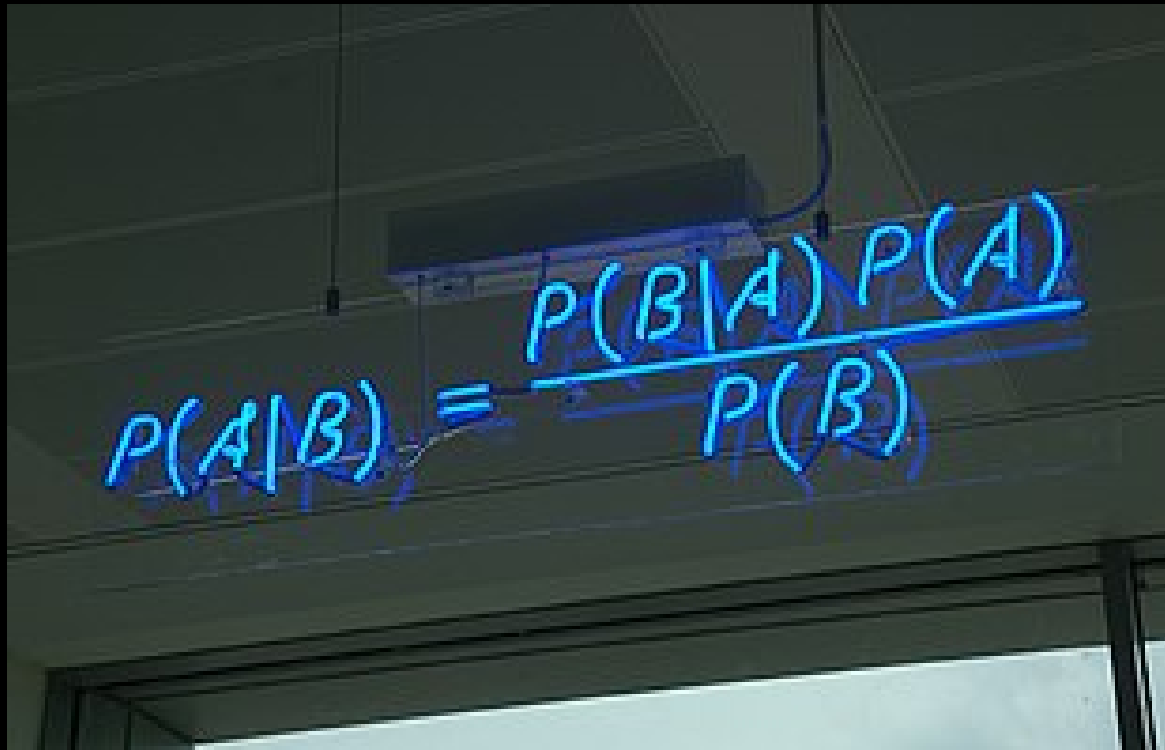
$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Nel caso in cui gli eventi A e B siano indipendenti:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A) * P(B)}{P(B)} = P(A)$$

Teorema di Bayes

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$



Esempio: Test diagnostico (alcol test)

Siano + e – gli eventi risultati da un alcol test che può essere positivo o negativo

Quindi D e D^c sono gli eventi che risultano dal soggetto di aver bevuto o no

La **sensibilità** è la probabilità che il test sia positivo dato che il soggetto ha bevuto

$$P(+|D)$$

La **specificità** è la probabilità che il test sia negativo dato che il soggetto **non** ha bevuto

$$P(-|D^c)$$

Esempio: Test diagnostico (alcol test)

Siano + e – gli eventi risultati da un alcol test che può essere positivo o negativo

Quindi U e D^c sono gli eventi che risultano dal soggetto di aver bevuto o no

La **sensibilità** è la probabilità che il test sia positivo dato che il soggetto ha bevuto

$$P(+|D)$$

La **specificità** è la probabilità che il test sia negativo dato che il soggetto **non** ha bevuto

$$P(-|D^c)$$

Se hai avuto il test positivo sei interessato alla probabilità che $P(D|+)$, cioè la probabilità di avere un test positivo e di aver effettivamente bevuto

Se hai avuto un test negativo sei interessato alla probabilità che $P(D^c|-)$, cioè la probabilità di aver avuto un test negativo e di non aver bevuto

Esempio: Test diagnostico (alcol test)

	$P(D)$	$P(D^c)$
$P(+)$	95	5
$P(-)$	8	92

sensibilità dell'alcol test = 95% $P(+|D)$

specificità dell'alcol test = 92% $P(-|D^c)$

prevalenza nella popolazione = 10% $P(+) \Rightarrow 90\% P(-)$

$$P(D|+) = \frac{P(+|D)P(+)}{P(D)} = \frac{P(+|D)P(+)}{P(+|D)P(+) + P(D|-)P(-)}$$

Esempio: Test diagnostico (alcol test)

		$P(D)$	$P(D^c)$
$CM =$	$P(+)$	95	5
	$P(-)$	8	92

sensibilità dell'alcol test = 95% $P(+|D)$

specificità dell'alcol test = 92% $P(-|D^c)$

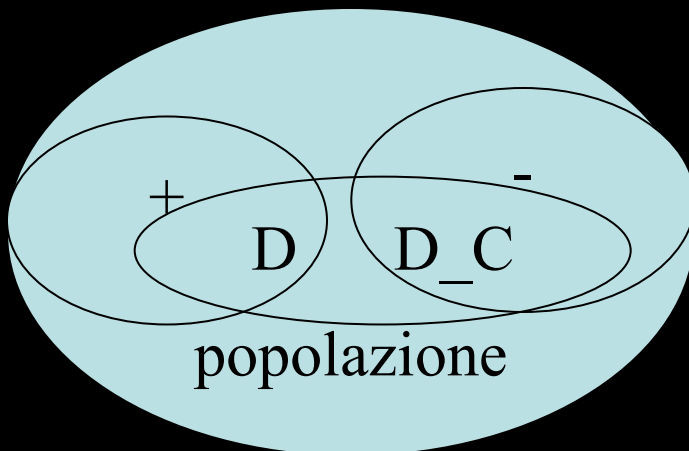
prevalenza nella popolazione = 10% $P(+)$ => 90% $P(-)$

$$P(D|+) = \frac{P(+|D)P(+)}{P(+|D)P(+) + P(D|-)P(-)}$$

Legge della probabilità totale

- ✓ Popolazione si divide in $\{+\}$ e $\{-\}$, non esiste popolazione che non appartenga a queste categorie
- ✓ D è l'insieme delle persone che hanno fatto il test quindi di fatto

$$P(D) = P(+ \cup D) + P(- \cup D)$$



Esempio: Test diagnostico (alcol test)

	$P(D)$	$P(D^c)$
$CM = P(+)$	95	5
$P(-)$	8	92

sensibilità dell'alcol test = 95% $P(+|D)$

specificità dell'alcol test = 92% $P(-|D^c)$

prevalenza nella popolazione = 10% $P(+) \Rightarrow 90\% P(-)$

$$\begin{aligned}
 P(D|+) &= \frac{P(+|D)P(+)}{P(D)} = \frac{P(+|D)P(+)}{P(+|D)P(+) + P(D|-)P(-)} \\
 &= \frac{0.95 * 0.1}{0.95 * 0.1 + 0.08 * 0.90} = \frac{0.095}{0.095 + 0.072} = 56\%
 \end{aligned}$$

Esempio: Test diagnostico (alcol test)

	$P(D)$	$P(D^c)$
$P(+)$	95	5
$P(-)$	8	92

sensibilità dell'alcol test = 95% $P(+|D)$

specificità dell'alcol test = 92% $P(-|D^c)$

prevalenza nella popolazione = 10% $P(+)$ => 90% $P(-)$

$$\begin{aligned} P(D|+) &= \frac{P(+|D)P(+)}{P(D)} = \frac{P(+|D)P(+)}{P(+|D)P(+)+P(D|-)P(-)} \\ &= \frac{0.95 * 0.1}{0.95 * 0.1 + 0.08 * 0.90} = \frac{0.095}{0.095 + 0.072} = 56\% \end{aligned}$$

Questo significa che la probabilità di essere positivo, avendo ricevuto un test con esito positivo è appena del 56%

Esempi di probabilità

Soggetti che hanno bevuto e
hanno avuto un test positivo:

95% pari a 95 persone

Soggetti che non hanno bevuto e hanno
avuto un test positivo:

8% pari a 72 soggetti

Supponiamo di avere una popolazione
di mille persone selezionate in
maniera causale:

il 10% ha bevuto (ciano) 100 soggetti

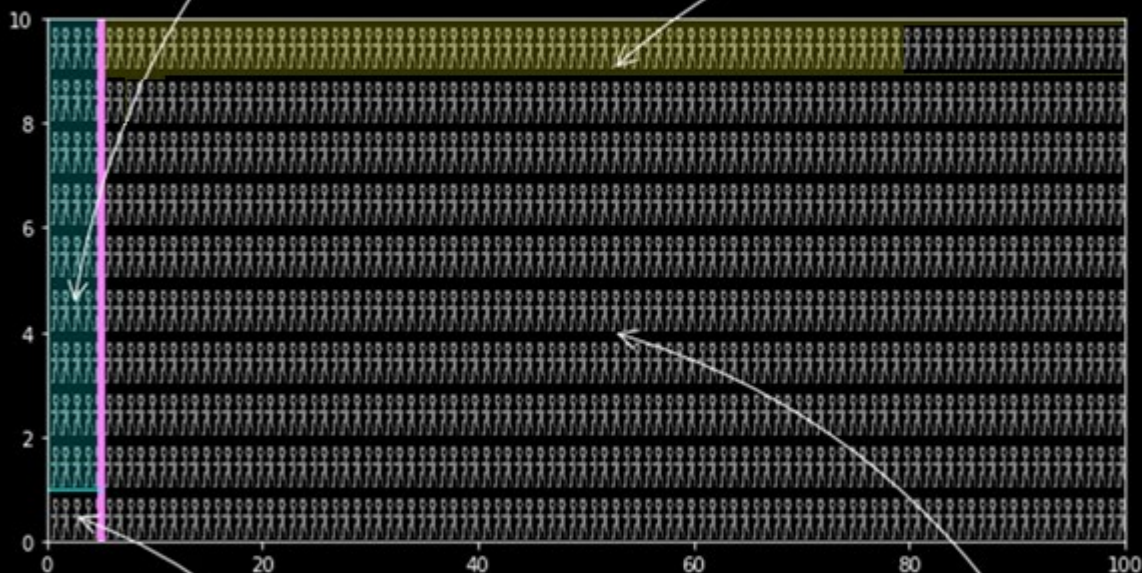
il 90% restante è sobrio 900 soggetti

Soggetti che non hanno bevuto e hanno
avuto un test negativo:

92% pari a 828 soggetti

Soggetti che hanno bevuto e
hanno avuto un test negativo:

5% pari a 5 soggetti



Esempio: Test diagnostico (alcol test)

		$P(D)$	$P(D^c)$
$CM =$	$P(+)$	95	5
	$P(-)$	8	92

sensibilità dell'alcol test = 99% $P(+|D)$

specificità dell'alcol test = 99% $P(-|D^c)$

prevalenza nella popolazione = 10% $P(+) \Rightarrow 90\% P(-)$

$$P(D|+) = \frac{P(+|D)P(+)}{P(D)} = \frac{P(+|D)P(+)}{P(+|D)P(+) + P(D|-)P(-)}$$

$$= \frac{0.95 * 0.1}{0.95 * 0.1 + 0.08 * 0.90} = \frac{0.095}{0.095 + 0.072} = 56\%$$

$$= \frac{0.99 * 0.1}{0.99 * 0.1 + 0.01 * 0.9} = \frac{0.099}{0.099 + 0.009} = 91\%$$

Anche supponendo di aumentare sensibilità e specificità del mio test al 99%, la probabilità di essere positivo, avendo ricevuto un test con esito positivo è del 91%

Esempio: Test diagnostico (alcol test)

	$P(D)$	$P(D^c)$
$CM = P(+)$	95	5
$P(-)$	8	92

sensibilità dell'alcol test = 99% $P(+|D)$

specificità dell'alcol test = 99% $P(-|D^c)$

prevalenza nella popolazione = 2% $P(+)$ => 98% $P(-)$

$$P(D|+) = \frac{P(+|D)P(+)}{P(D)} = \frac{P(+|D)P(+)}{P(+|D)P(+) + P(D|-)P(-)}$$

$$= \frac{0.95 * 0.1}{0.95 * 0.1 + 0.08 * 0.90} = \frac{0.095}{0.095 + 0.072} = 56\%$$

$$= \frac{0.99 * 0.02}{0.99 * 0.02 + 0.01 * 0.98} = \frac{0.0198}{0.0198 + 0.0098} = 16\%$$

Assumendo che la prevalenza nella popolazione sia pari al 2%,
la probabilità di essere positivo, avendo ricevuto un test con esito positivo è del 16%

Tutorial

- ✓ Visualizzazione su dashboard della legge di Bayes

Correlazione

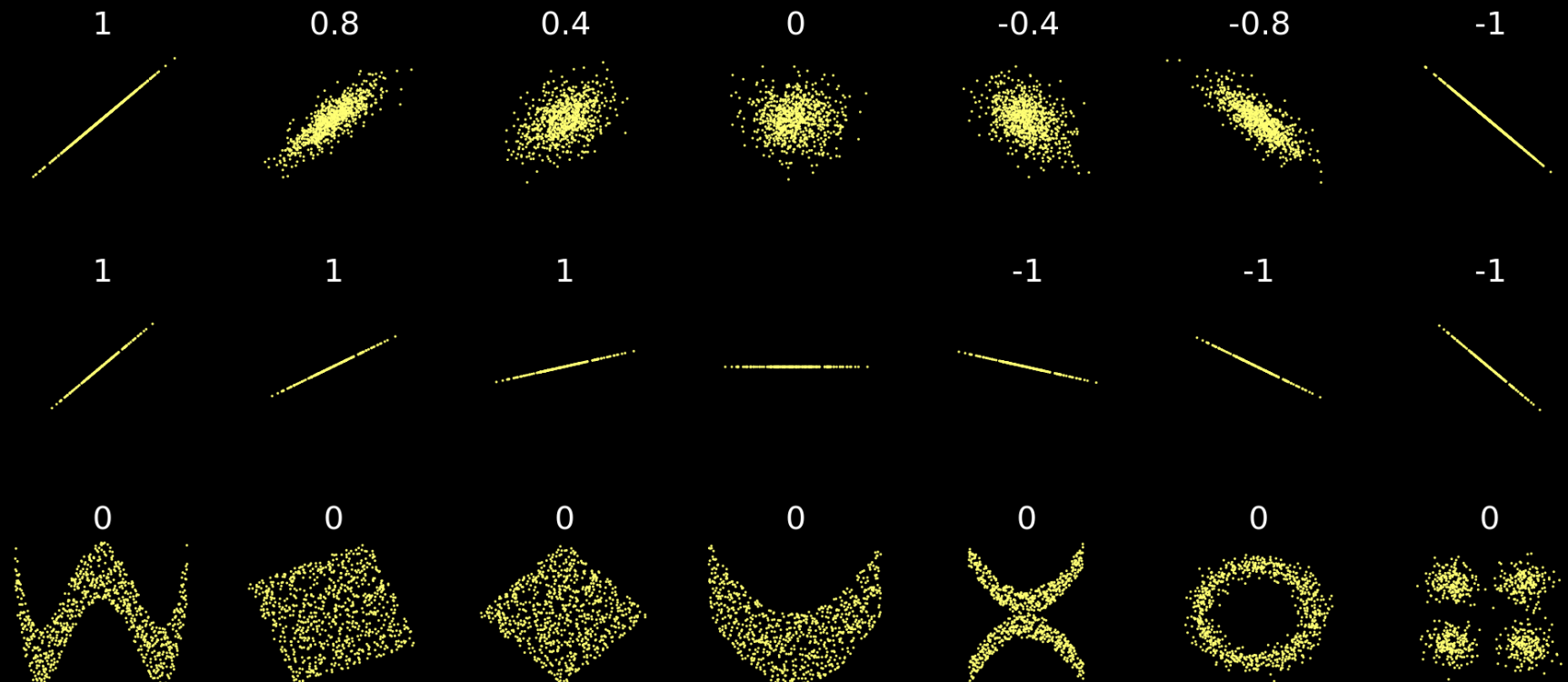
In statistica la correlazione ci dice, date due variabili casuali X e Y , quanto esse sono correlate, quindi al crescere di X cresce Y o viceversa.

$$\text{corr}(X, Y) = \frac{\text{COV}(X, Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

Correlazione

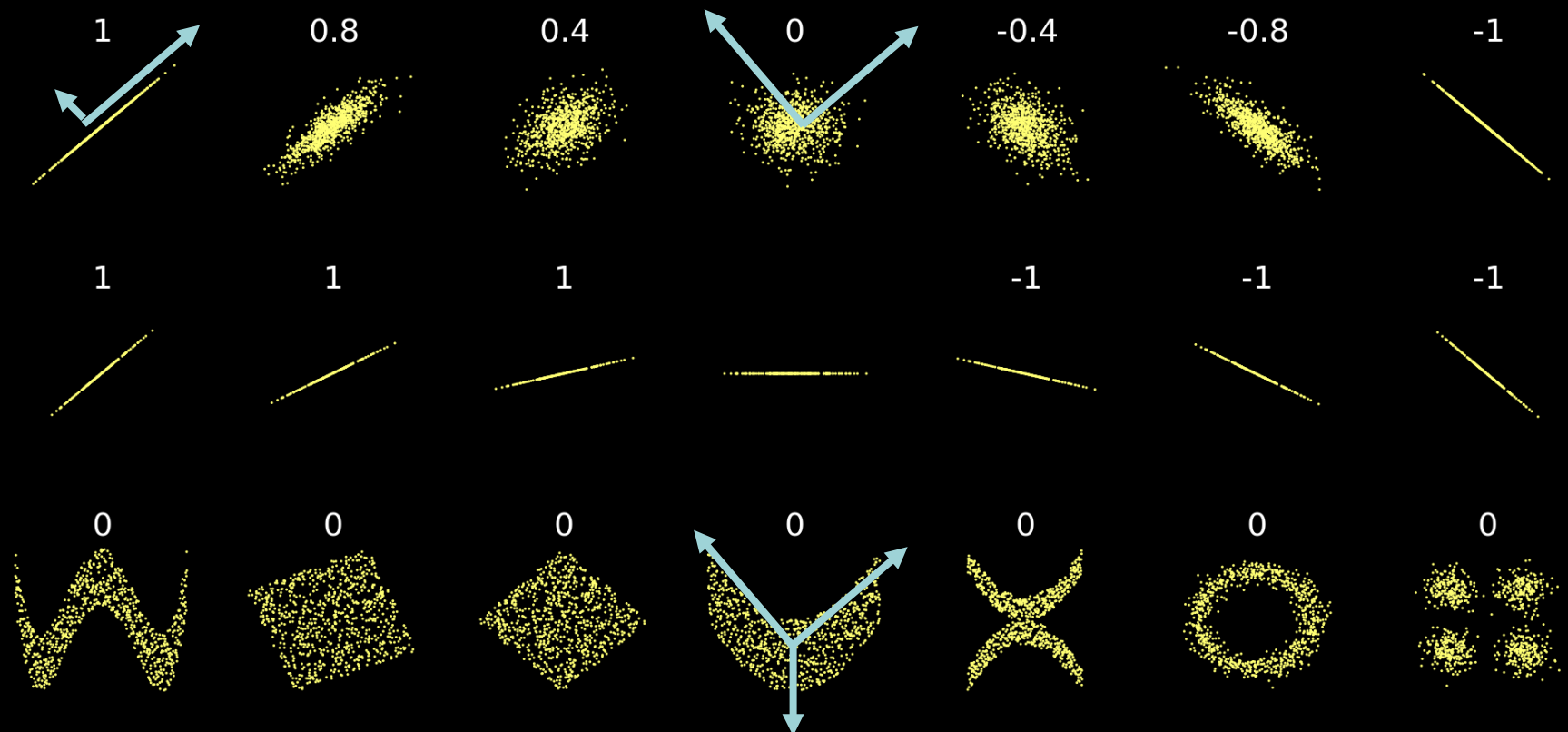
In statistica la correlazione ci dice, date due variabili casuali X e Y, quanto esse sono correlate, quindi al crescere di X cresce Y o viceversa.

$$\text{corr}(X, Y) = \frac{\text{COV}(X, Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$



Analisi delle componenti principali (PCA)

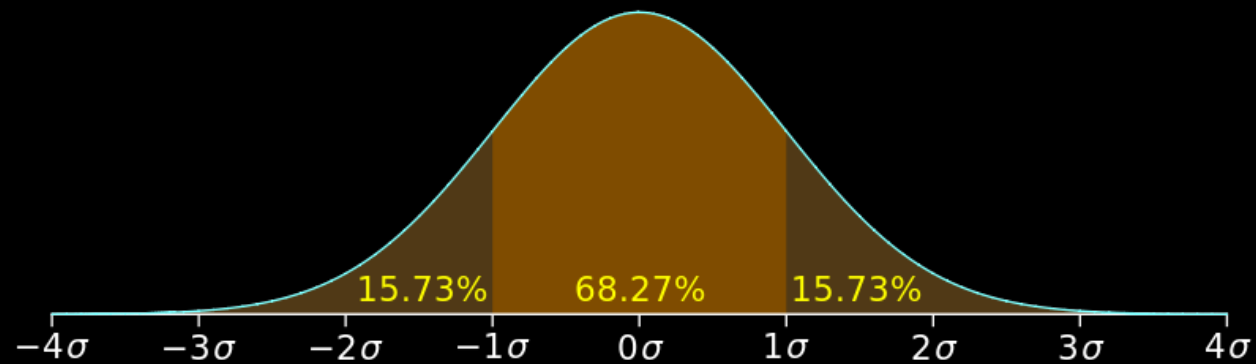
L'analisi delle componenti principali valuta la lunghezza di vettori in alcune regioni dello spazio



Z-score

é il numero di di deviazioni standard di un determinato punto della distribuzione rispetto alla media

$$z = \frac{x - E[x]}{\sigma[x]}$$

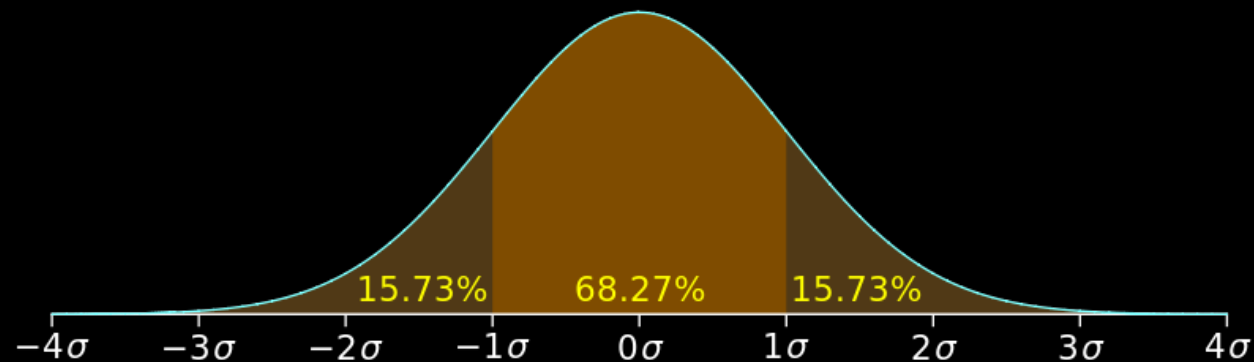


Z-score

é il numero di di deviazioni standard di un determinato punto della distribuzione rispetto alla media

$$z = \frac{x - E[x]}{\sigma[x]}$$

essenzialmente mi dice dove la mia misura è allocate rispetto alla distribuzione, se sono vicino alla media con buona probabilità sto osservando il fenomeno, se sono nella coda della distribuzione sto probabilmente osservando una eccezione

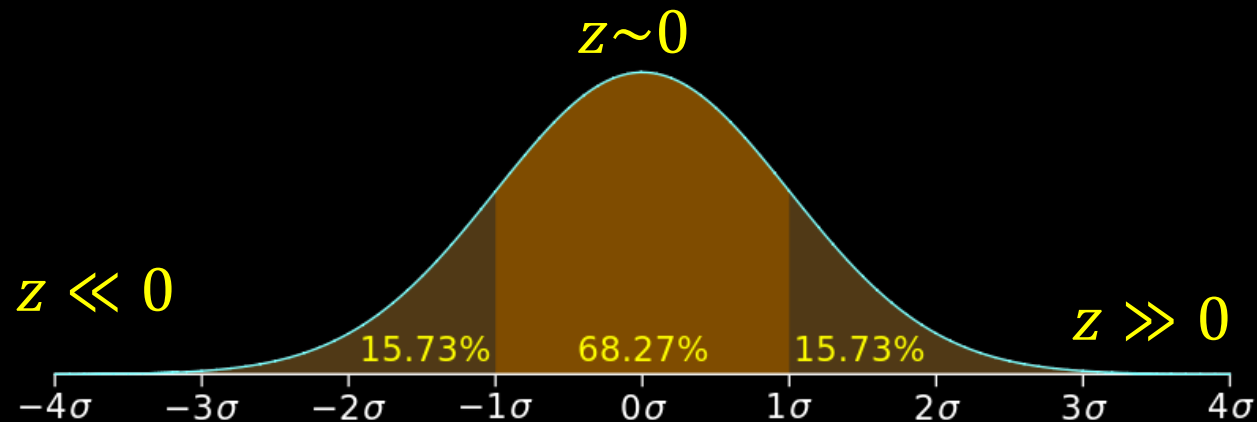


Z-score

é il numero di di deviazioni standard di un determinato punto della distribuzione rispetto alla media

$$z = \frac{x - E[x]}{\sigma[x]}$$

essenzialmente mi dice dove la mia misura è allocate rispetto alla distribuzione, se sono vicino alla media con buona probabilità sto osservando il fenomeno, se sono nella coda della distribuzione sto probabilmente osservando una eccezione

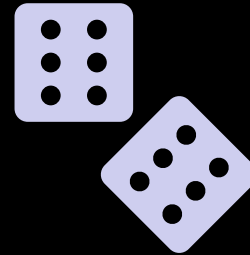


Test di ipotesi statistica

Fornisce una indicazione che un determinato risultato è chiaramente significativo oppure è generato semplicemente dal caso

Test di ipotesi statistica

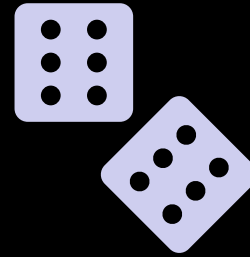
Fornisce una indicazione che un determinato risultato è chiaramente significativo oppure è generato semplicemente dal caso



Supponiamo di lanciare i dadi 1000 volte:

Test di ipotesi statistica

Fornisce una indicazione che un determinato risultato è chiaramente significativo oppure è generato semplicemente dal caso



Supponiamo di lanciare i dadi 1000 volte:

$$P(2)=P(12) = 1/36$$

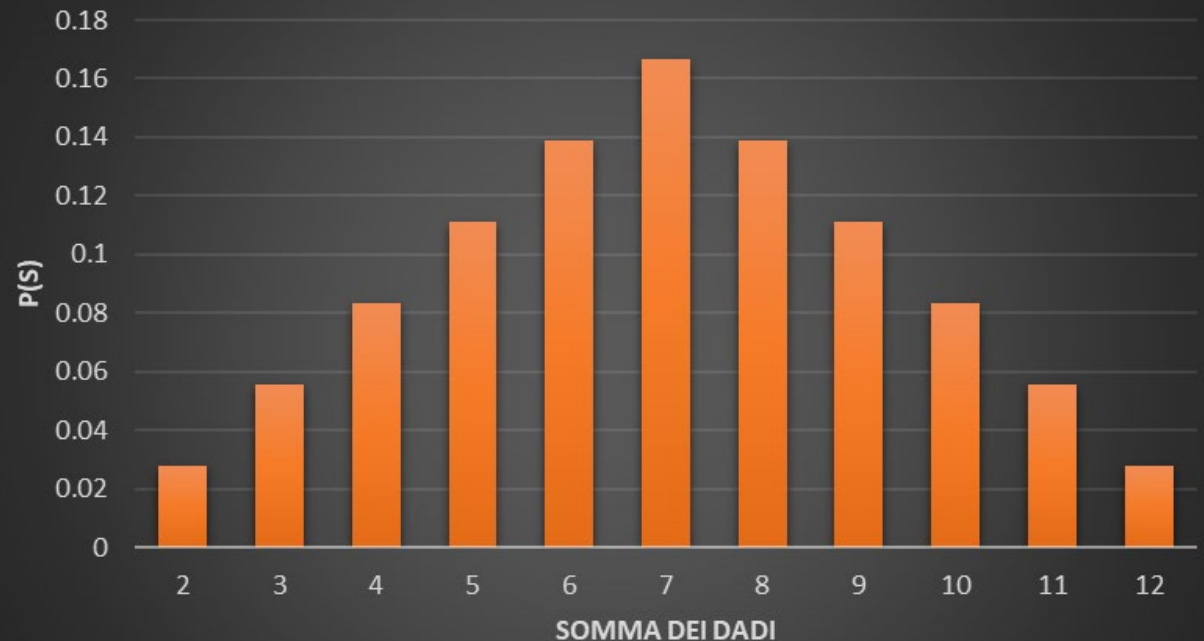
$$P(3)=P(11) = 2/36$$

$$P(4)=P(10) = 3/36$$

$$P(5)=P(9) = 4/36$$

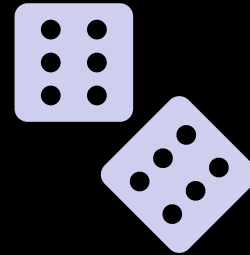
$$P(6)=P(8) = 5/36$$

$$p(7) = 6/36$$



Test di ipotesi statistica

Fornisce una indicazione che un determinato risultato è chiaramente significativo oppure è generato semplicemente dal caso



Supponiamo di lanciare i dadi 1000 volte:

$$P(2)=P(12) = 1/36$$

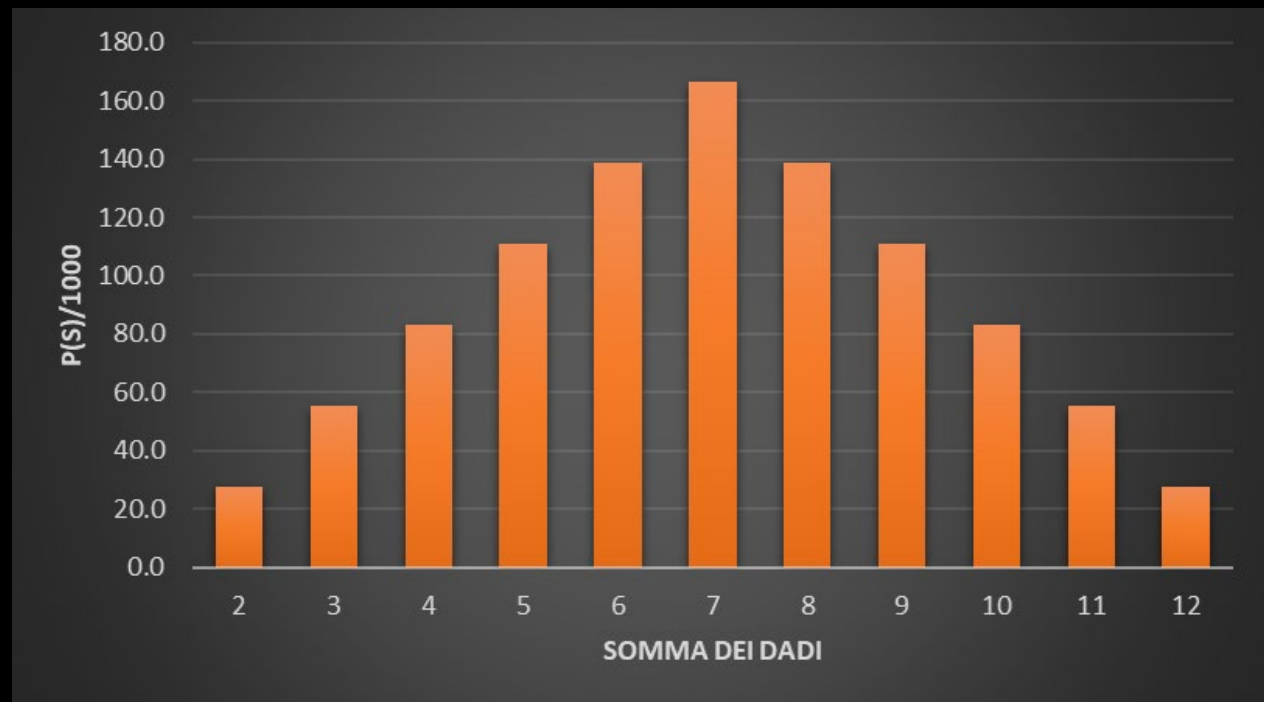
$$P(3)=P(11) = 2/36$$

$$P(4)=P(10) = 3/36$$

$$P(5)=P(9) = 4/36$$

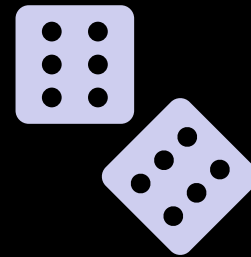
$$P(6)=P(8) = 5/36$$

$$p(7) = 6/36$$



Test di ipotesi statistica

Fornisce una indicazione che un determinato risultato è chiaramente significativo oppure è generato semplicemente dal caso



Posso affermare che i dadi sono truccati in base a questa statistica o quello che vedo è solo frutto del caso?

Supponiamo di lanciare i dadi 1000 volte:

$$P(2)=P(12) = 1/36$$

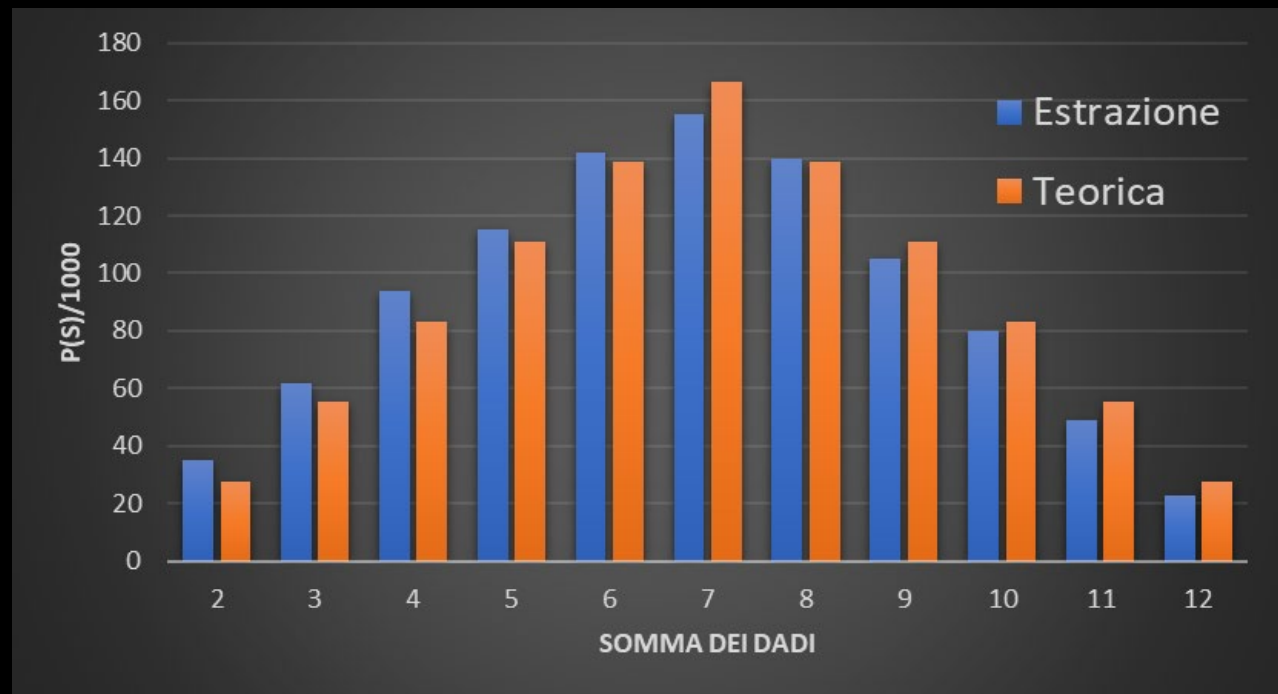
$$P(3)=P(11) = 2/36$$

$$P(4)=P(10) = 3/36$$

$$P(5)=P(9) = 4/36$$

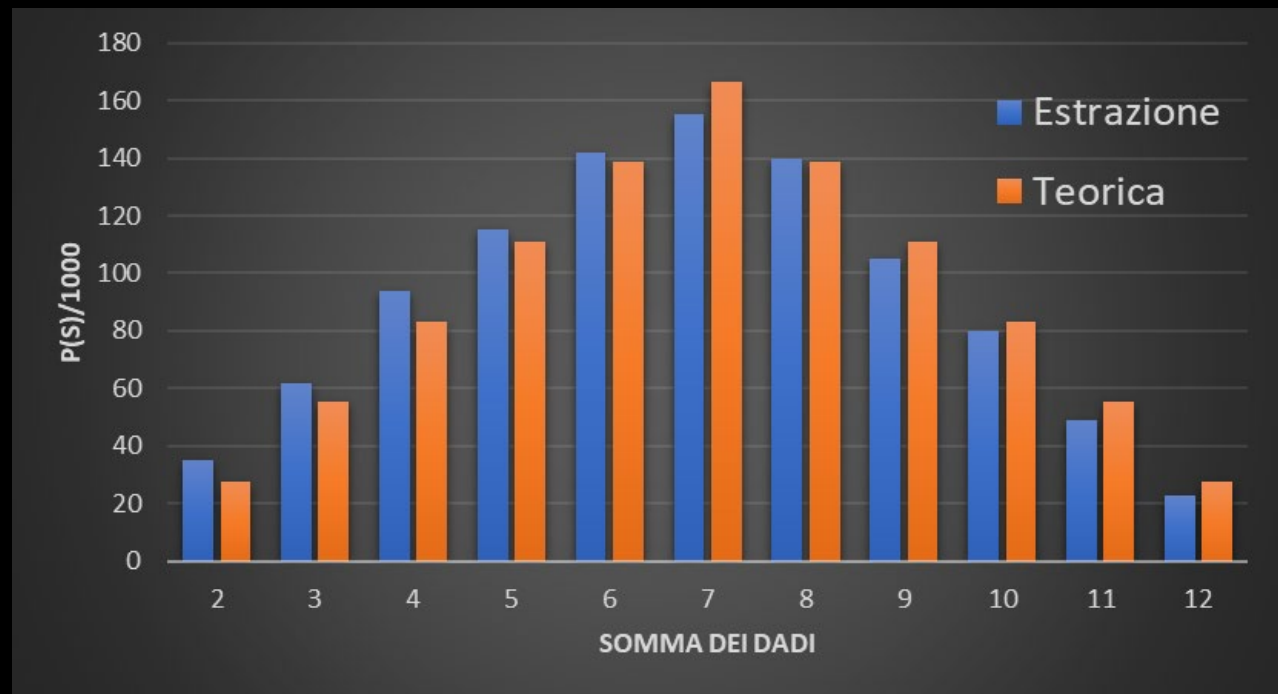
$$P(6)=P(8) = 5/36$$

$$p(7) = 6/36$$



Test di ipotesi statistica

è chiaro che per il numero di prove che tende a infinito (BIG DATA) la distribuzione blu (estrazioni) tende alla distribuzione arancio (teorica)



Test di ipotesi statistica

Null-hypothesis (ipotesi nulla): date due variabili causali, si assume che esse siano completamente indipendenti ($\text{corr}=0$) e si cerca di confutare questa ipotesi.

Test di ipotesi statistica

Null-hypothesis (ipotesi nulla): date due variabili causali, si assume che esse siano completamente indipendenti ($\text{corr}=0$) e si cerca di confutare questa ipotesi.

Ipotesi nulla

$$H_0$$

Affermazione su un parametro
in una popolazione

Si testa la probabilità che questa
affermazione sia corretta in maniera
tale da confutare o accettare la
nostra ipotesi alternativa

Ipotesi alternativa

$$H_a$$

Affermazione in completa
contraddizione con l'ipotesi nulla

Si determina se accettare o no
l'ipotesi alternativa basandosi sulla
probabilità che l'ipotesi nulla sia
vera

p-value

- ✓ Il p-value è la probabilità che l'ipotesi nulla sia vera
- ✓ $p < 0.05$ è tipicamente preso come soglia di significatività statistica, ma questo non è sempre scontato
- ✓ piccoli valori di p ci indicano una probabilità alta che l'ipotesi nulla sia falsa, con $p\text{-value} < 0.0001$ si indica un risultato altamente significativo
- ✓ $p > 0.05$ significa che il risultato è “non significativo”, ma questo non prova che la metodologia è errata, potrebbe semplicemente richiedere più dati per essere significativo

p-value

- ✓ Il p-value è la probabilità che l'ipotesi nulla sia vera
- ✓ $p < 0.05$ è tipicamente preso come soglia di significatività statistica, ma questo non è sempre scontato
- ✓ piccoli valori di p ci indicano una probabilità alta che l'ipotesi nulla sia falsa, con $p\text{-value} < 0.0001$ si indica un risultato altamente significativo
- ✓ $p > 0.05$ significa che il risultato è “non significativo”, ma questo non prova che la metodologia è errata, potrebbe semplicemente richiedere più dati per essere significativo

ATTENZIONE: il p-value è spesso interpretato a supporto o rigetto dell'ipotesi alternativa. Tuttavia, questo non è il caso. Il p-value ci dice se l'ipotesi nulla è vera, ma non ci dice niente sulla validità dell'ipotesi alternativa.

p-value: esempio


Pursuing the art of technical creativity

[HOME](#) [PUBLICATIONS](#) [RESUME](#)

Mauro Bellone: Roboticist, entrepreneur and innovation technologies specialist

About

I received the M.S. degree in Automation Engineering from the [University of Salento](#), Lecce, Italy, where I received the Ph.D. in Mechanical and Industrial Engineering in 2014. In 2009, I had the pleasure to visit the [Space Robotics Lab](#) of Tohoku University, Sendai, Japan. In 2013-14, as visiting researcher at University of Almeria, Spain, I have worked with the Automatic, Electronic and Robotics Research Group [TEP-197](#) studying new advanced navigation techniques for mobile robotics and autonomous driving. From 2015 to 2020, I have worked with the [applied artificial intelligence](#) research group of Chalmers University of Technology in Sweden, in which I have actively contributed with several research projects in the field of autonomous robots. Currently, I am studying artificial intelligence methods applied to healthcare with [Echolight Spa](#) in Italy. As a senior researcher, since 2021, I am supporting [Tallinn University](#) of technology in the area of smart transportation systems.



Contact Info

E-mail: bellonemauro@gmail.com
Skype: bellonemauro

" we are part of a 10 billion people brain "

Supponiamo che il tempo medio trascorso dagli utenti sul sito sia 20 minuti e di voler fare un cambiamento che porti più utenti sul sito,

p-value: esempio

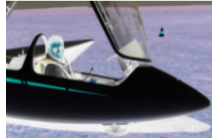
Pursuing the art of technical creativity

[HOME](#) [PUBLICATIONS](#) [RESUME](#)

Mauro Bellone: Roboticist, entrepreneur and innovation technologies specialist








About

I received the M.S. degree in Automation Engineering from the [University of Salento](#), Lecce, Italy, where I received the Ph.D. in Mechanical and Industrial Engineering in 2014. In 2009, I had the pleasure to visit the [Space Robotics Lab](#) of Tohoku University, Sendai, Japan. In 2013-14, as visiting researcher at University of Almeria, Spain, I have worked with the Automatic, Electronic and Robotics Research Group [TEP-197](#) studying new advanced navigation techniques for mobile robotics and autonomous driving. From 2015 to 2020, I have worked with the [applied artificial intelligence](#) research group of Chalmers University of Technology in Sweden, in which I have actively contributed with several research projects in the field of autonomous robots. Currently, I am studying artificial intelligence methods applied to healthcare with [Echolight Spa](#) in Italy. As a senior researcher, since 2021, I am supporting [Tallinn University](#) of technology in the area of smart transportation systems.



Contact Info

E-mail: bellonemauro@gmail.com
Skype: bellonemauro



" we are part of a 10 billion people brain "

Supponiamo che il tempo medio trascorso dagli utenti sul sito sia 20 minuti e di voler fare un cambiamento che porti più utenti sul sito, quindi lo cambio e lo faccio bianco.

Come faccio a capire che il mio cambiamento ha effettivamente avuto la conseguenza cercata?

p-value: esempio

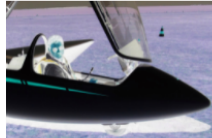
Pursuing the art of technical creativity

[HOME](#) [PUBLICATIONS](#) [RESUME](#)

Mauro Bellone: Roboticist, entrepreneur and innovation technologies specialist


About

I received the M.S. degree in Automation Engineering from the [University of Salento](#), Lecce, Italy, where I received the Ph.D. in Mechanical and Industrial Engineering in 2014. In 2009, I had the pleasure to visit the [Space Robotics Lab](#) of Tohoku University, Sendai, Japan. In 2013-14, as visiting researcher at University of Almeria, Spain, I have worked with the Automatic, Electronic and Robotics Research Group [TEP-197](#) studying new advanced navigation techniques for mobile robotics and autonomous driving. From 2015 to 2020, I have worked with the [applied artificial intelligence](#) research group of Chalmers University of Technology in Sweden, in which I have actively contributed with several research projects in the field of autonomous robots. Currently, I am studying artificial intelligence methods applied to healthcare with [Echolight Spa](#) in Italy. As a senior researcher, since 2021, I am supporting [Tallinn University](#) of technology in the area of smart transportation systems.



Contact Info

E-mail: bellonemauro@gmail.com
Skype: bellonemauro



" we are part of a 10 billion people brain "

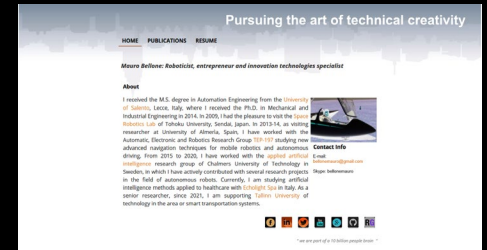
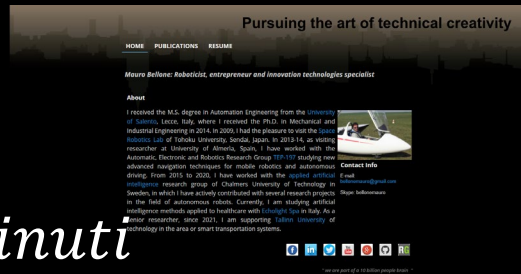
Supponiamo che il tempo medio trascorso dagli utenti sul sito sia 20 minuti e di voler fare un cambiamento che porti più utenti sul sito, quindi lo cambio e lo faccio bianco.

Come faccio a capire che il mio cambiamento ha effettivamente avuto la conseguenza cercata?

Questo è un classico test di significatività statistica

p-value: esempio

$\mu = 20\text{minuti}$



p-value: esempio

$$\mu = 20 \text{ minuti}$$

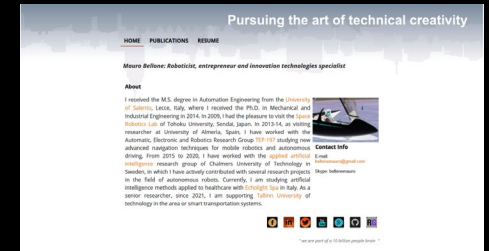
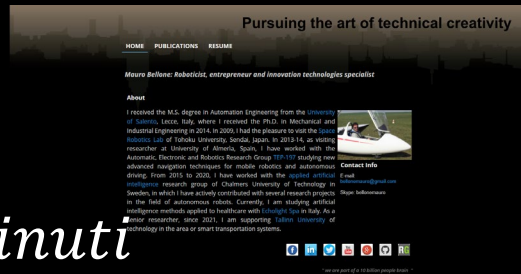
Impostiamo il test di significatività statistica partendo dall'ipotesi nulla e l'ipotesi alternativa

1. $H_0 : \mu = 20 \text{ minuti}$ ipotesi nulla

Significa che il cambiamento di fatto non ha portato effetti sul traffico quindi la mia media sarebbe sempre $E[x] = 20 \text{ minuti}$

p-value: esempio

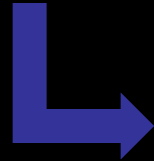
$$\mu = 20 \text{ minuti}$$



Impostiamo il test di significatività statistica partendo dall'ipotesi nulla e l'ipotesi alternativa

1. $H_0 : \mu = 20 \text{ minuti}$ ipotesi nulla

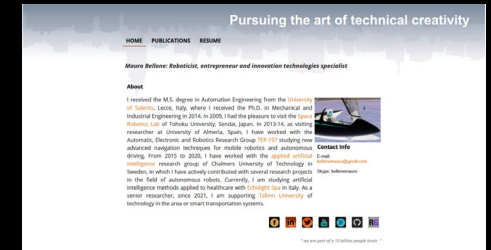
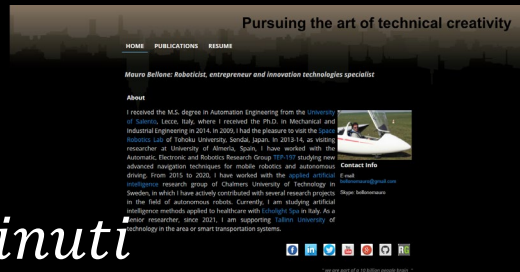
$H_a: \mu > 20 \text{ minuti}$ ipotesi alternativa



significa che il cambiamento di fatto ha portato effetti sul traffico e quindi la media è maggiore della precedente

p-value: esempio

$$\mu = 20 \text{ minuti}$$

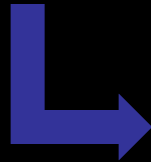


Impostiamo il test di significatività statistica partendo dall'ipotesi nulla e l'ipotesi alternativa

1. $H_0 : \mu = 20 \text{ minuti}$ ipotesi nulla

$H_a: \mu > 20 \text{ minuti}$ ipotesi alternativa

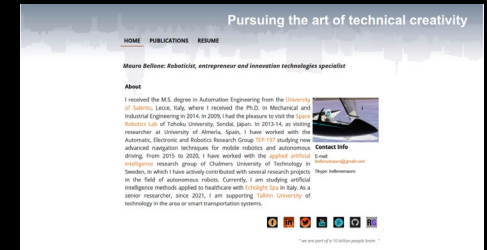
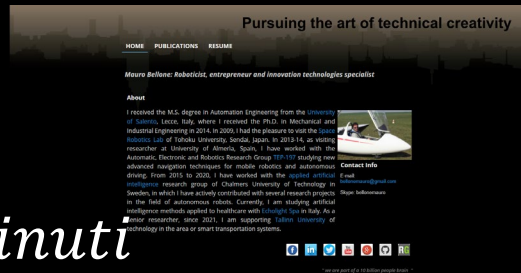
2. $\alpha = 0.05$ significatività statistica



Scelgo una soglia di significatività statistica, in questo caso il 5%

p-value: esempio

$$\mu = 20 \text{ minuti}$$



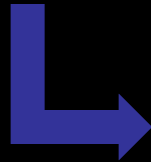
Impostiamo il test di significatività statistica partendo dall'ipotesi nulla e l'ipotesi alternativa

1. $H_0 : \mu = 20 \text{ minuti}$ ipotesi nulla

$H_a: \mu > 20 \text{ minuti}$ ipotesi alternativa

2. $\alpha = 0.05$ significatività statistica

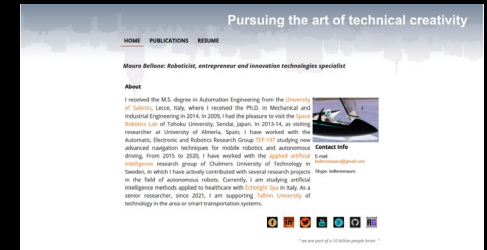
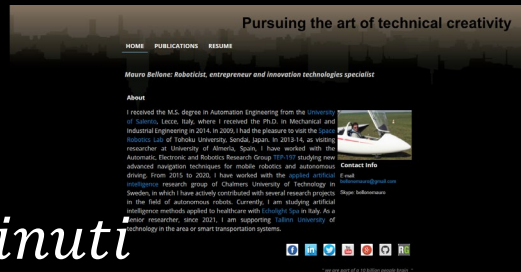
3. $n = 100$ campioni con $\mu = 25$ minuti



Eseguo un test a campione (es. 100 campioni) e calcolo le statistiche, media, varianza ect.

p-value: esempio

$$\mu = 20 \text{ minuti}$$



Impostiamo il test di significatività statistica partendo dall'ipotesi nulla e l'ipotesi alternativa

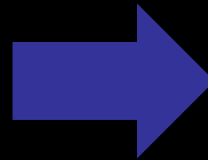
1. $H_0 : \mu = 20 \text{ minuti}$ ipotesi nulla

$H_a: \mu > 20 \text{ minuti}$ ipotesi alternativa

2. $\alpha = 0.05$ significatività statistica

3. $n = 100$ campioni con $\mu = 25$ minuti

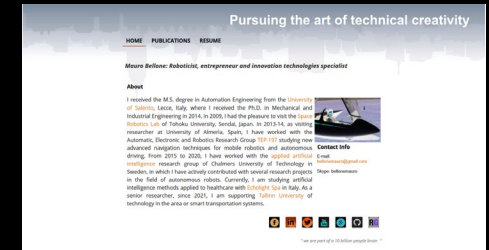
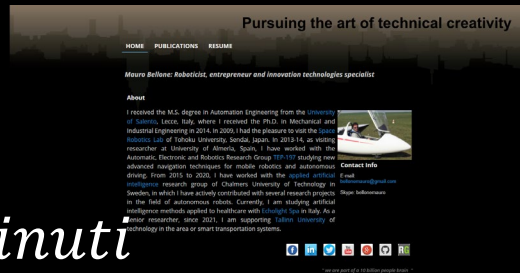
4. $p_{value} = P(\mu \geq 25 \text{ min} | H_0 = \text{true})$



il p-value è la probabilità condizionata che la media sia maggiore del valore iniziale dato che l'ipotesi nulla è vera

p-value: esempio

$$\mu = 20 \text{ minuti}$$



Impostiamo il test di significatività statistica partendo dall'ipotesi nulla e l'ipotesi alternativa

1. $H_0 : \mu = 20 \text{ minuti}$ ipotesi nulla

$H_a: \mu > 20 \text{ minuti}$ ipotesi alternativa

2. $\alpha = 0.05$ significatività statistica

3. $n = 100$ campioni con $\mu = 25$ minuti

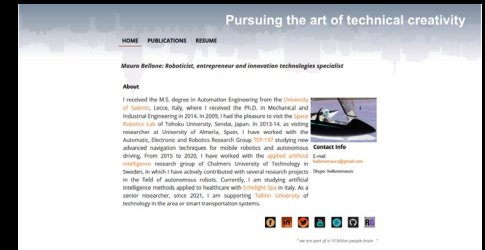
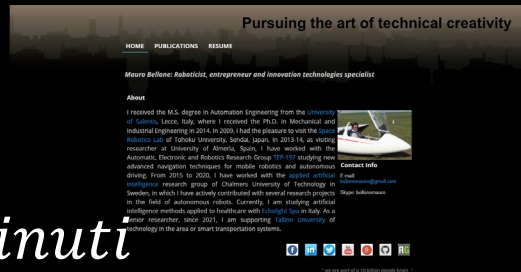
4. $p_{value} = P(\mu \geq 25 \text{ min} | H_0 = \text{true})$

5. $\begin{cases} \text{if } p_{value} < \alpha \Rightarrow \text{rigettare } H_0 \\ \text{if } p_{value} \geq \alpha \Rightarrow \text{NON rigettare } H_0 \end{cases}$

posso scegliere se accettare o no l'ipotesi nulla ma non posso affermare che l'ipotesi alternativa sia vera!

p-value: esempio

$$\mu = 20 \text{ minuti}$$



Impostiamo il test di significatività statistica partendo dall'ipotesi nulla e l'ipotesi alternativa

1. $H_0 : \mu = 20 \text{ minuti}$ ipotesi nulla

$H_a: \mu > 20 \text{ minuti}$ ipotesi alternativa

2. $\alpha = 0.05$ significatività statistica

3. $n = 100$ campioni con $\mu = 25 \text{ minuti}$

4. $p_{value} = P(\mu \geq 25 \text{ min} | H_0 = \text{true})$

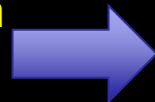
5.
$$\begin{cases} \text{if } p_{value} < \alpha \Rightarrow \text{rigettare } H_0 \\ \text{if } p_{value} \geq \alpha \Rightarrow \text{NON rigettare } H_0 \end{cases}$$

Aspetti rilevanti in una analisi statistica

- ✓ Campioni analizzati
- ✓ Metriche utilizzate per la misura



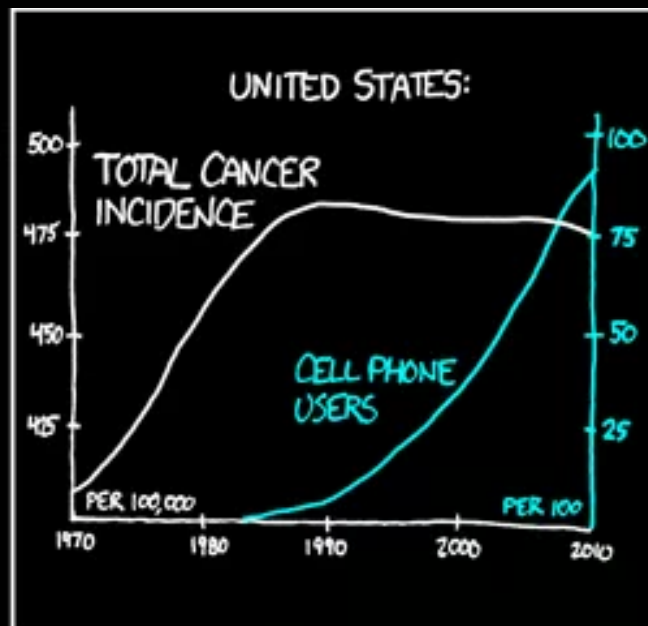
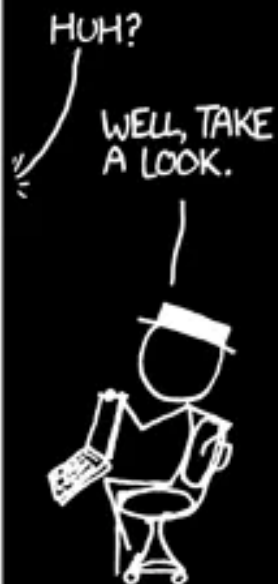
Una correlazione data senza una misura di affidabilità (errore probabilistico, errore standard) non è da prendere sul serio.



You won't always be told how many cases. The absence of such a figure, particularly when the source is an interested one, is enough to throw suspicion on the whole thing. Similarly a correlation given without a measure of reliability (probable error, standard error) is not to be taken very seriously.

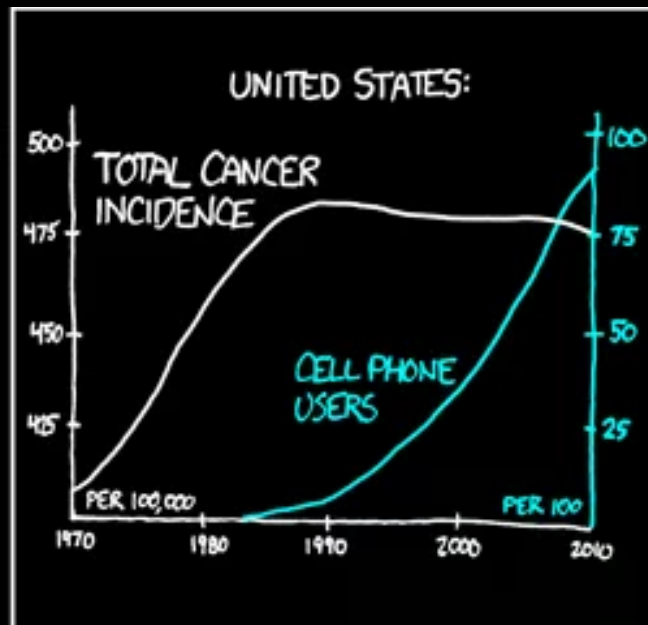
Watch out for an average, variety unspecified, in any matter where mean and median might be expected to differ substantially.

Causalità Vs associazione



Causalità Vs associazione

- ✓ Il fumo causa il cancro?
- ✓ é una specifica campagna di vendita ad aumentare le vendite di un prodotto?
- ✓ è uno specifico trattamento a far ridurre la progressione di una malattia?



Per diventare dei bravi data scientists

Qualità:

- ✓ Scientifiche: Bisogna essere in grado di imparare velocemente i fatti sulle aree di applicazione ed applicarle con rigore
- ✓ Statistiche: Bisogna avere conoscenza e esperienza per manipolare in maniera appropriate i dati per progettare dei sistemi efficaci
- ✓ Computazionali: bisogna essere in grado di programmare e usare efficacemente il software
- ✓ Comunicazionali: bisogna saper comunicare i dati in maniera efficace e precisa

Tutorial: Correlazione Vs causalità

- ✓ Dato che l'azione X e l'azione Y risultano strettamente correlate, è l'innalzamento della azione X che genera l'innalzamento dell'azione Y?

Visualizzazione

La visualizzazione dati è fondamentale per individuare patterns tra i dati e interpretare risultati anche in formati diversi

- ✓ Semplifica informazioni quantitative tendenzialmente complesse
- ✓ Permette di analizzare ed esplorare big data facilmente
- ✓ Permette di identificare le aree di miglioramento
- ✓ Fa emergere informazioni non evidenti
- ✓ Permette di individuare relazioni tra data points e variabili

Visualizzazione

Caratteristiche importanti per una buona visualizzazione dati

- ✓ Chiarezza
- ✓ Efficacia/efficienza
- ✓ Accuratezza

Visualizzazione

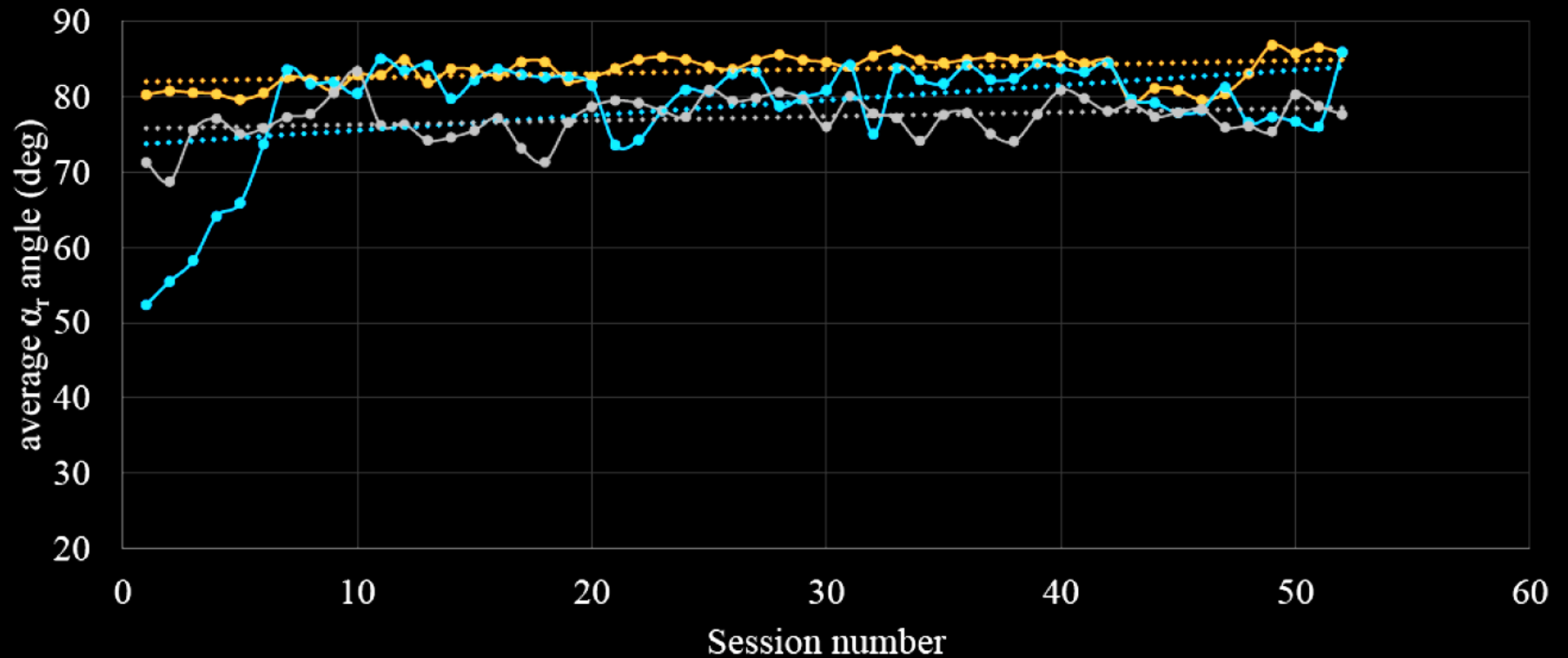
Caratteristiche importanti per una buona visualizzazione dati

- ✓ Effetto visivo (formato grafico appropriato, colori, dimensione etc.)
- ✓ Unità di misura (incluso tipo di dati, numerico, categorie etc.)
- ✓ Sistema di riferimento (descrizione degli assi)
- ✓ Interpretazioni (annotazioni, titoli, legende etc.)

Visualizzazione

- ✓ Grafici a line
- ✓ Scatterplot
- ✓ Istogrammi
- ✓ Boxplot

Visualizzazione - linee



—●— Mean α - Patient A

—●— Mean α - Patient B

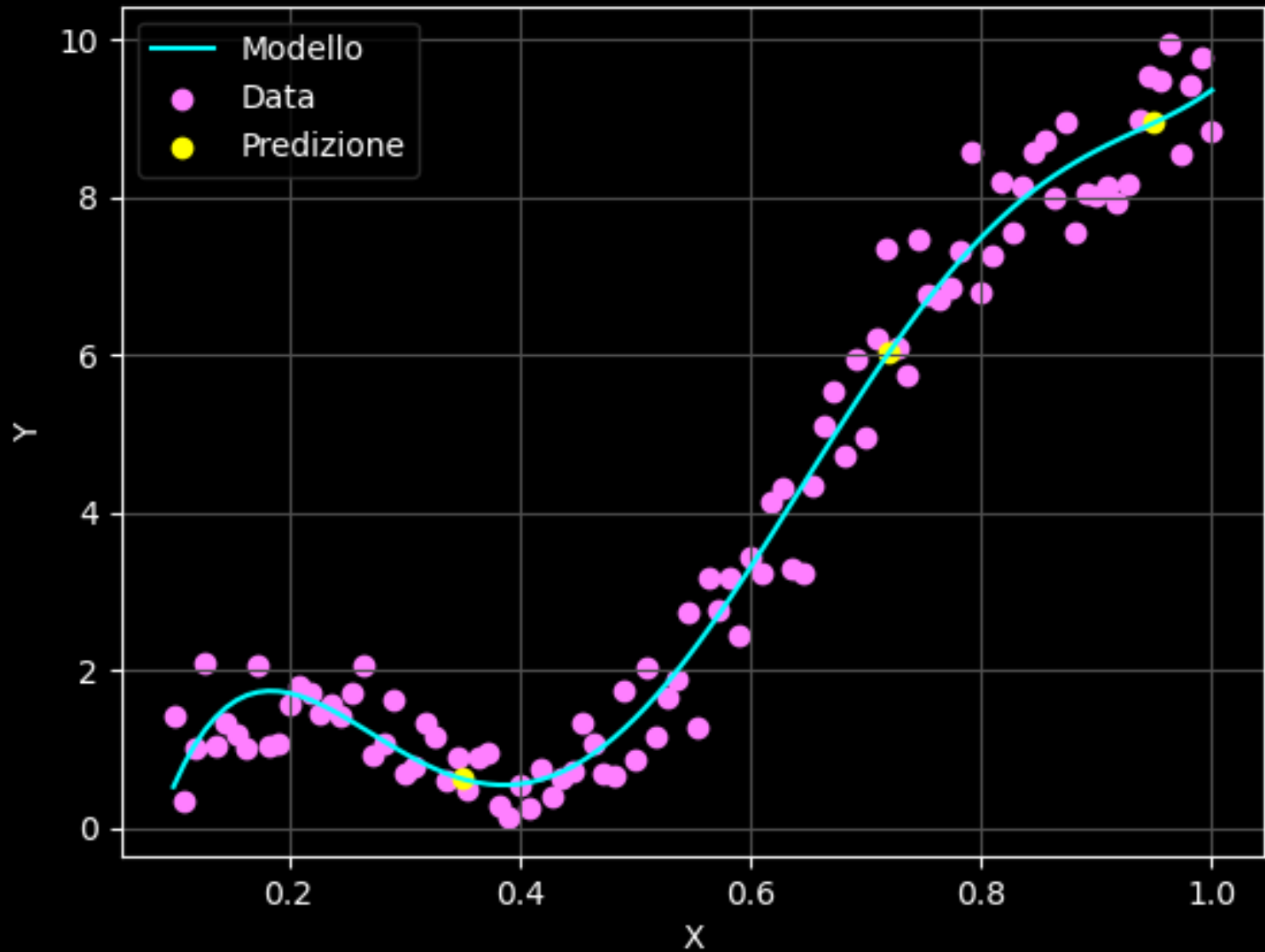
—●— Mean α - Patient C

..... Linear (Mean α - Patient A)

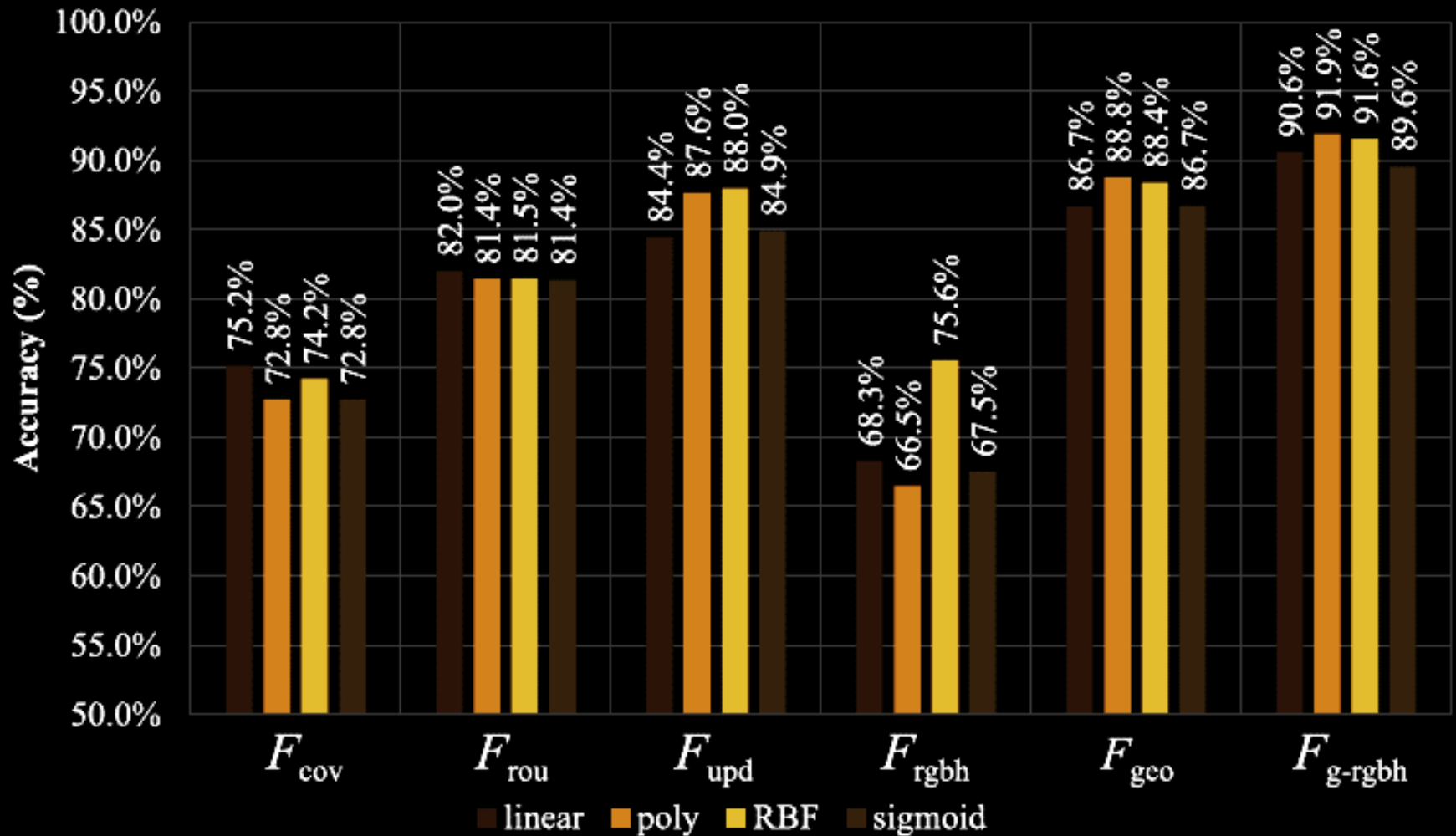
..... Linear (Mean α - Patient B)

..... Linear (Mean α - Patient C)

Visualizzazione - scatterplot



Visualizzazione - istogrammi



source:

M. Bellone, G. Reina, L. Caltagirone, and M. Wahde - "Learning Traversability from Point Clouds in Challenging Scenarios" IEEE Transactions on Intelligent Transportation Systems, Volume: PP, Issue: 99

Visualizzazione - torte

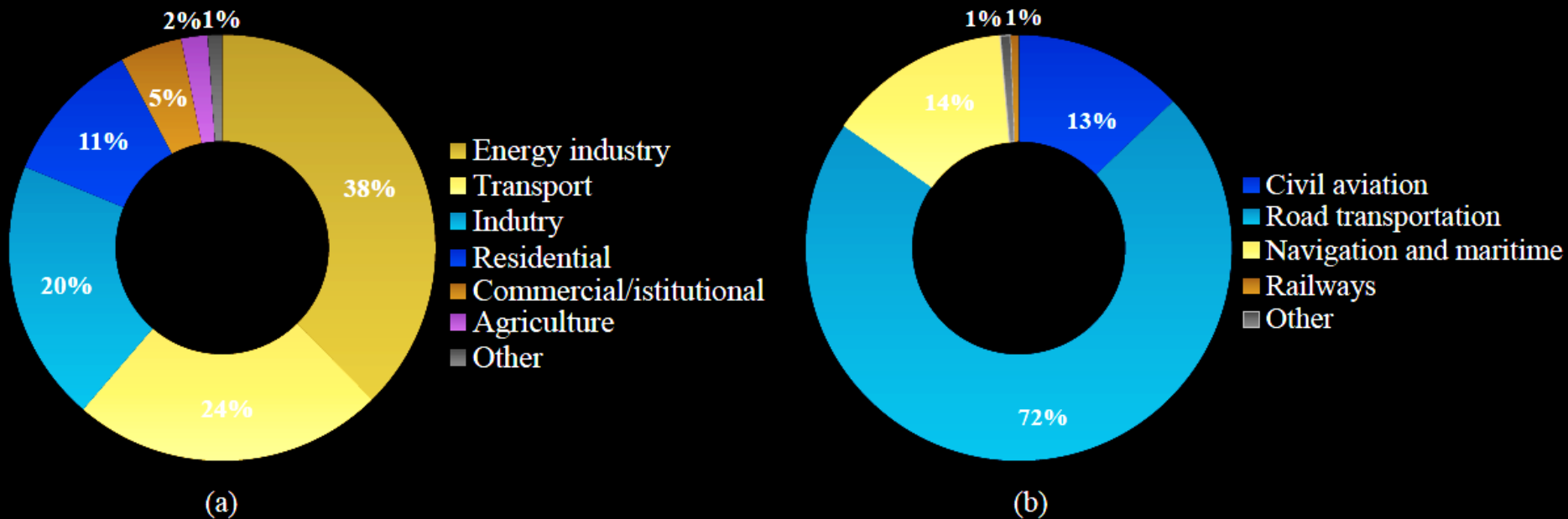
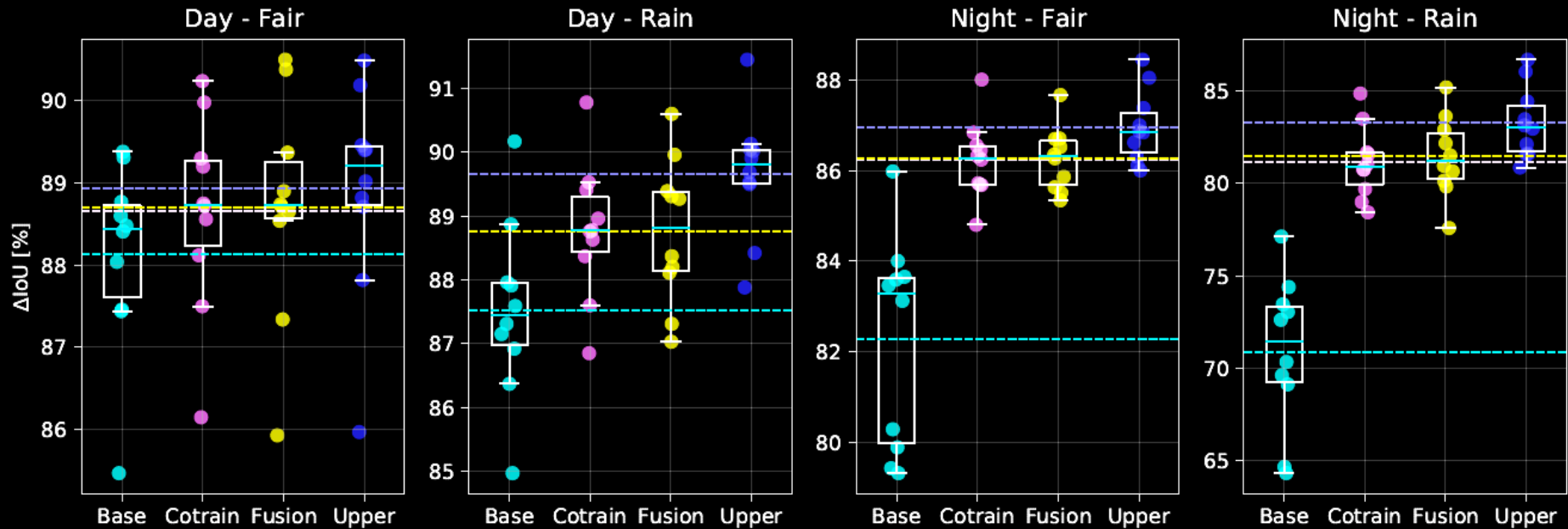


Figure 4. Impact of CO₂ emissions in 2012 by sector in EU-28. The global CO₂ emissions, in percentage, are shown in (a); whereas the specific contribution of the transportation sector is reported in (b). Transport causes 24% of total emissions, 72% of which comes from road transportation. Data source: EEA [7].

Visualizzazione - boxplot



Visualizzazione

Caratteristiche importanti per una buona visualizzazione dati

- ✓ Chiarezza
- ✓ Efficacia/efficienza
- ✓ Accuratezza

Inferenza statistica

Inferenza statistica

- ✓ Generazione di una conclusione su una popolazione da un insieme di dati rumorosi
- ✓ Vediamo i nostri esempi come esperienza, stiamo quindi cercando di creare conoscenza dalla nostra esperienza
- ✓ L'inferenza statistica è, di fatto, l'unico modo che abbiamo per creare conoscenza

Usi dell'inferenza statistica

- ✓ Regressione
- ✓ Classificazione
- ✓ Clusterizzazione

Usi dell'inferenza statistica

- ✓ Regressione – stima di un valore continuo
- ✓ Classificazione
- ✓ Clusterizzazione

Usi dell'inferenza statistica

- ✓ Regressione – stima di un valore continuo
- ✓ Classificazione – stima di un valore discreto (appartenenza a una classe)
- ✓ Clusterizzazione

Usi dell'inferenza statistica

- ✓ Regressione – stima di un valore continuo
- ✓ Classificazione – stima di un valore discreto (appartenenza a una classe)
- ✓ Clusterizzazione – Individuazione e divisione in classi

Valutazione degli algoritmi di inferenza statistica

- ✓ Consideriamo un insieme di dati D

Valutazione degli algoritmi di inferenza statistica

- ✓ Consideriamo un insieme di dati D
- ✓ Dividiamo il nostro insieme di dati in 3 parti
 - ✓ Dati di addestramento D_a
 - ✓ Dati di validazione D_v
 - ✓ Dati di test D_t

Valutazione degli algoritmi di inferenza statistica

- ✓ Consideriamo un insieme di dati D
- ✓ Dividiamo il nostro insieme di dati in 3 parti
 - ✓ Dati di addestramento D_a
 - ✓ Dati di validazione D_v
 - ✓ Dati di test D_t
- ✓ Progettiamo il sistema di inferenza statistica

Valutazione degli algoritmi di inferenza statistica

- ✓ Consideriamo un insieme di dati D
- ✓ Dividiamo il nostro insieme di dati in 3 parti
 - ✓ Dati di addestramento D_a
 - ✓ Dati di validazione D_v
 - ✓ Dati di test D_t
- ✓ Progettiamo il sistema di inferenza statistica
- ✓ Addestramento

Valutazione degli algoritmi di inferenza statistica

- ✓ Consideriamo un insieme di dati D
- ✓ Dividiamo il nostro insieme di dati in 3 parti
 - ✓ Dati di addestramento D_a
 - ✓ Dati di validazione D_v
 - ✓ Dati di test D_t
- ✓ Progettiamo il sistema di inferenza statistica
- ✓ Addestramento
- ✓ Validazione

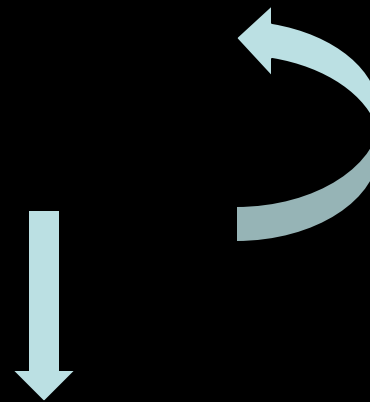
Valutazione degli algoritmi di inferenza statistica

- ✓ Consideriamo un insieme di dati D
- ✓ Dividiamo il nostro insieme di dati in 3 parti
 - ✓ Dati di addestramento D_a
 - ✓ Dati di validazione D_v
 - ✓ Dati di test D_t
- ✓ Progettiamo il sistema di inferenza statistica
- ✓ Addestramento
- ✓ Validazione



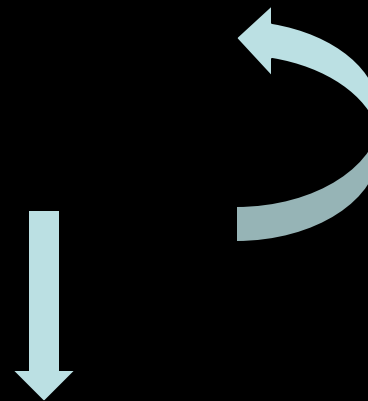
Valutazione degli algoritmi di inferenza statistica

- ✓ Consideriamo un insieme di dati D
- ✓ Dividiamo il nostro insieme di dati in 3 parti
 - ✓ Dati di addestramento D_a
 - ✓ Dati di validazione D_v
 - ✓ Dati di test D_t
- ✓ Progettiamo il sistema di inferenza statistica
- ✓ Addestramento
- ✓ Validazione
- ✓ Test



Valutazione degli algoritmi di inferenza statistica

- ✓ Consideriamo un insieme di dati D
- ✓ Dividiamo il nostro insieme di dati in 3 parti
 - ✓ Dati di addestramento D_a -- **Training set**
 - ✓ Dati di validazione D_v -- **Validation set**
 - ✓ Dati di test D_t -- **Test set**
- ✓ Progettiamo il sistema di inferenza statistica
- ✓ Addestramento
- ✓ Validazione
- ✓ Test



R² coefficiente di determinazione statistica

il coefficiente di determinazione statistica R^2 indica il legame tra i dati e la correttezza di un modello generato

$$R^2 = 1 - \frac{RSS}{TSS}$$

Dove:

TSS = devianza totale

$$RSS = \sum (x_i - E[x])^2$$

RSS = devianza residua

$$RSS = \sum (x_i - \hat{x}_i)^2$$

con x_i , \hat{x}_i rispettivamente le osservazioni e la stima del modello

R^2 coefficiente di determinazione statistica

il coefficiente di determinazione statistica R^2 indica il legame tra i dati e la correttezza di un modello generato

$$R^2 = 1 - \frac{RSS}{TSS}$$

- ❖ $R^2 \approx 1$ Significa che le previsioni del modello sono attendibili
- ❖ $R^2 \approx 0$ Significa che le previsioni del modello NON sono attendibili

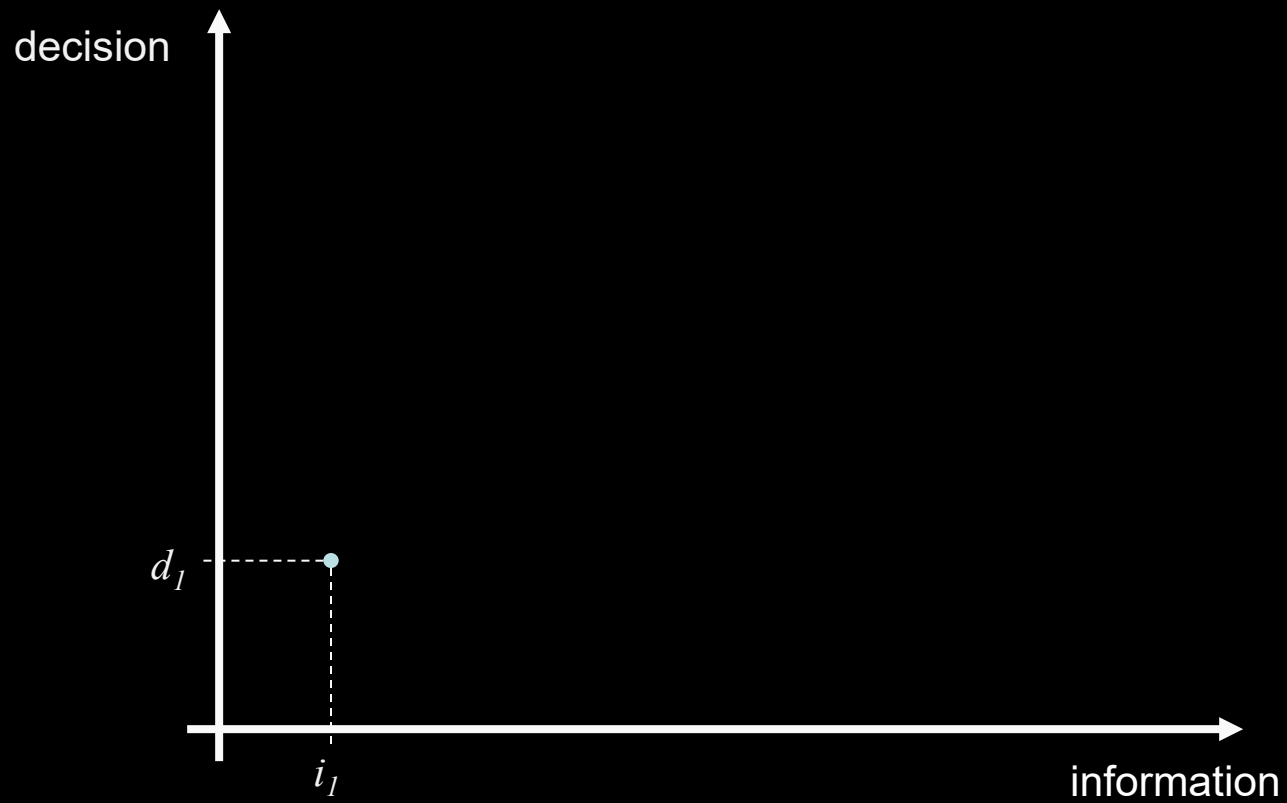
Astrazione di conoscenza

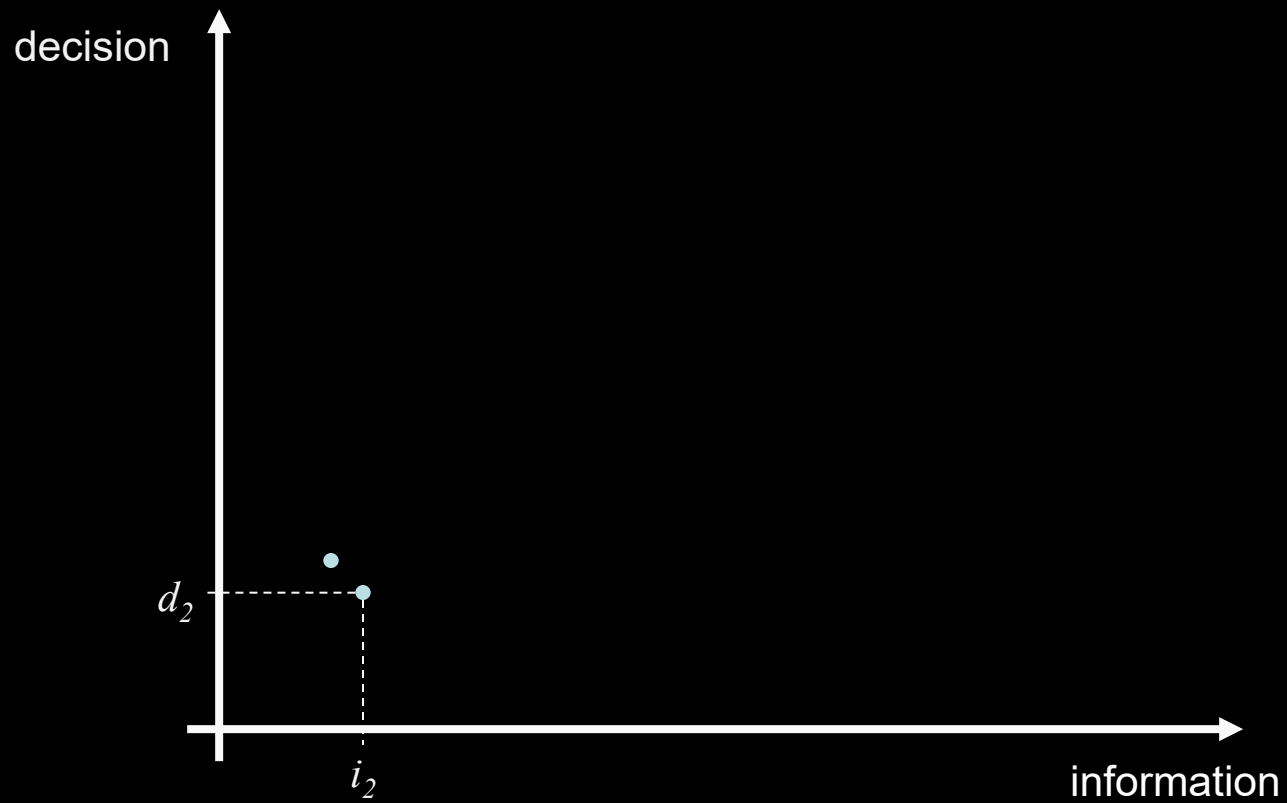
Per astrazione si intende la capacità di un sistema di inferenza statistica di effettuare previsioni con alta accuratezza su nuovi dati

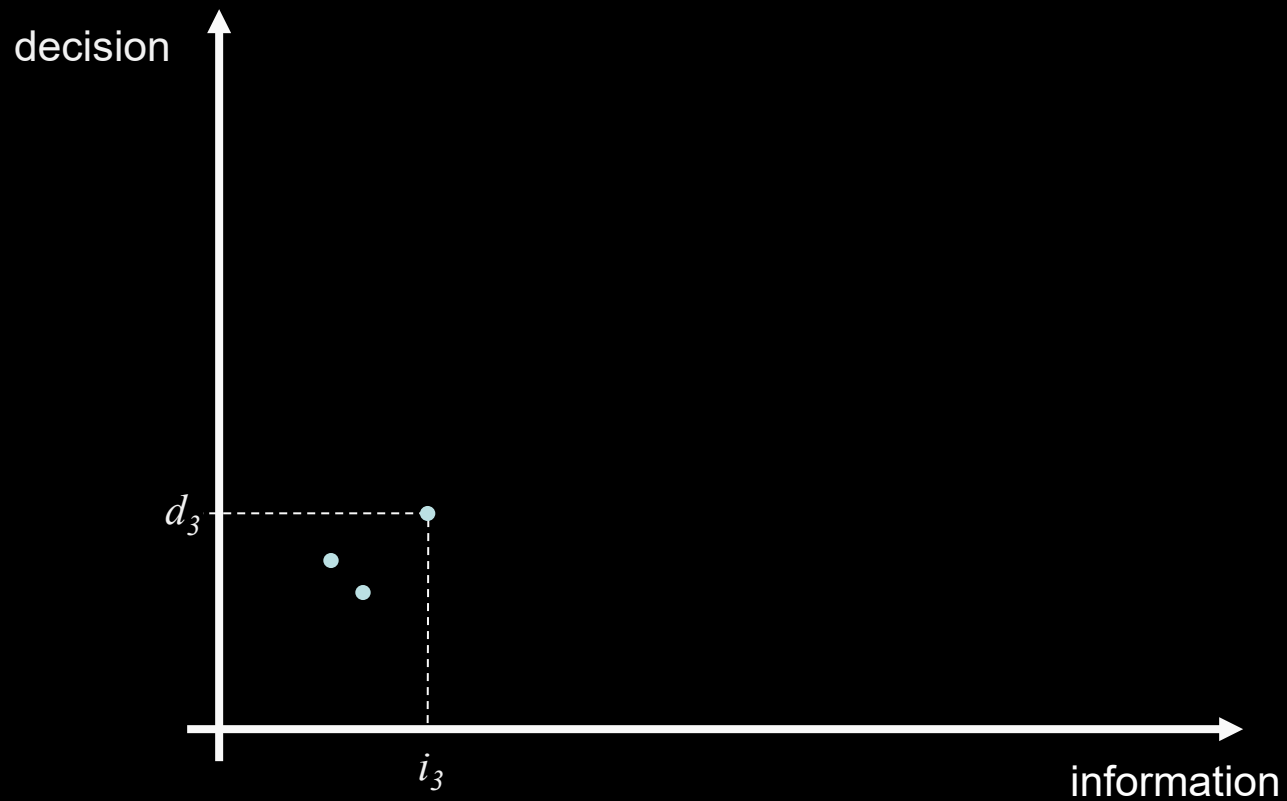
Tutti i sistemi di analisi dati hanno una capacità di memorizzazione del dataset, pochi hanno capacità di astrazione

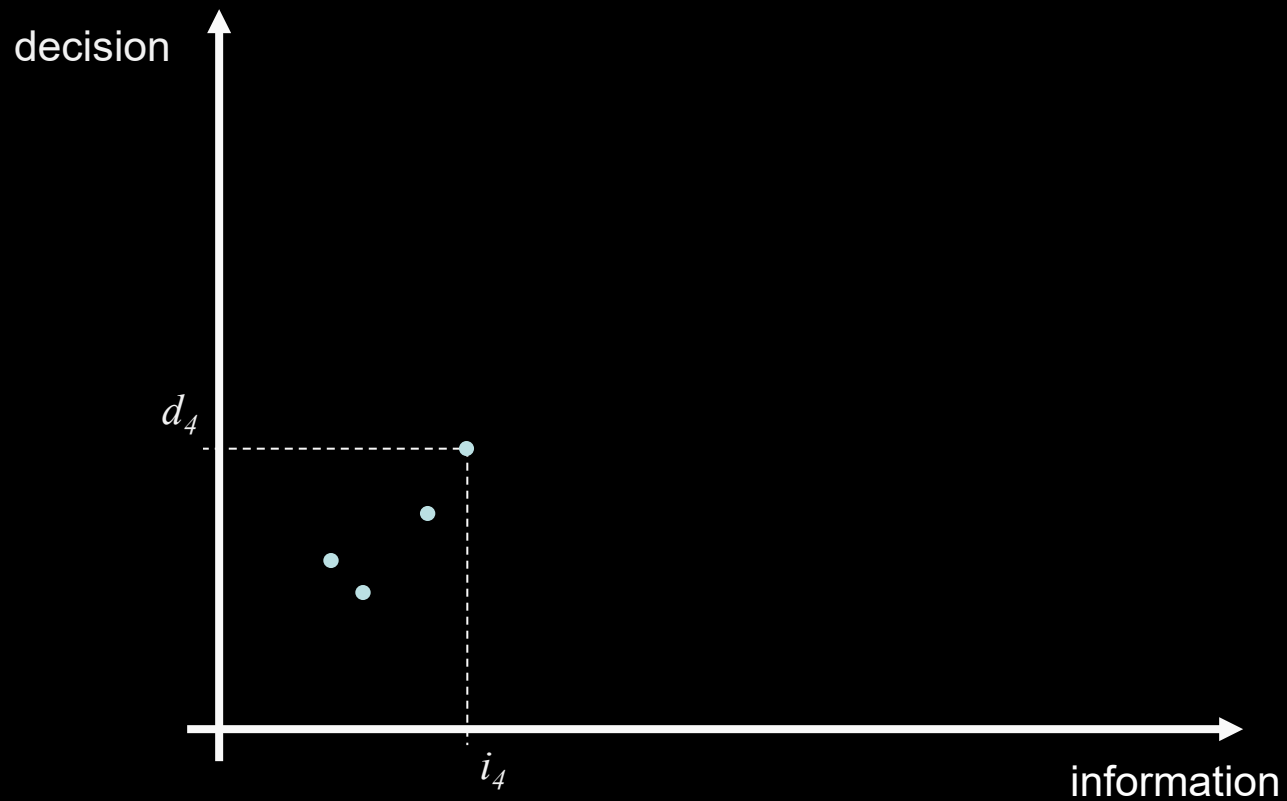


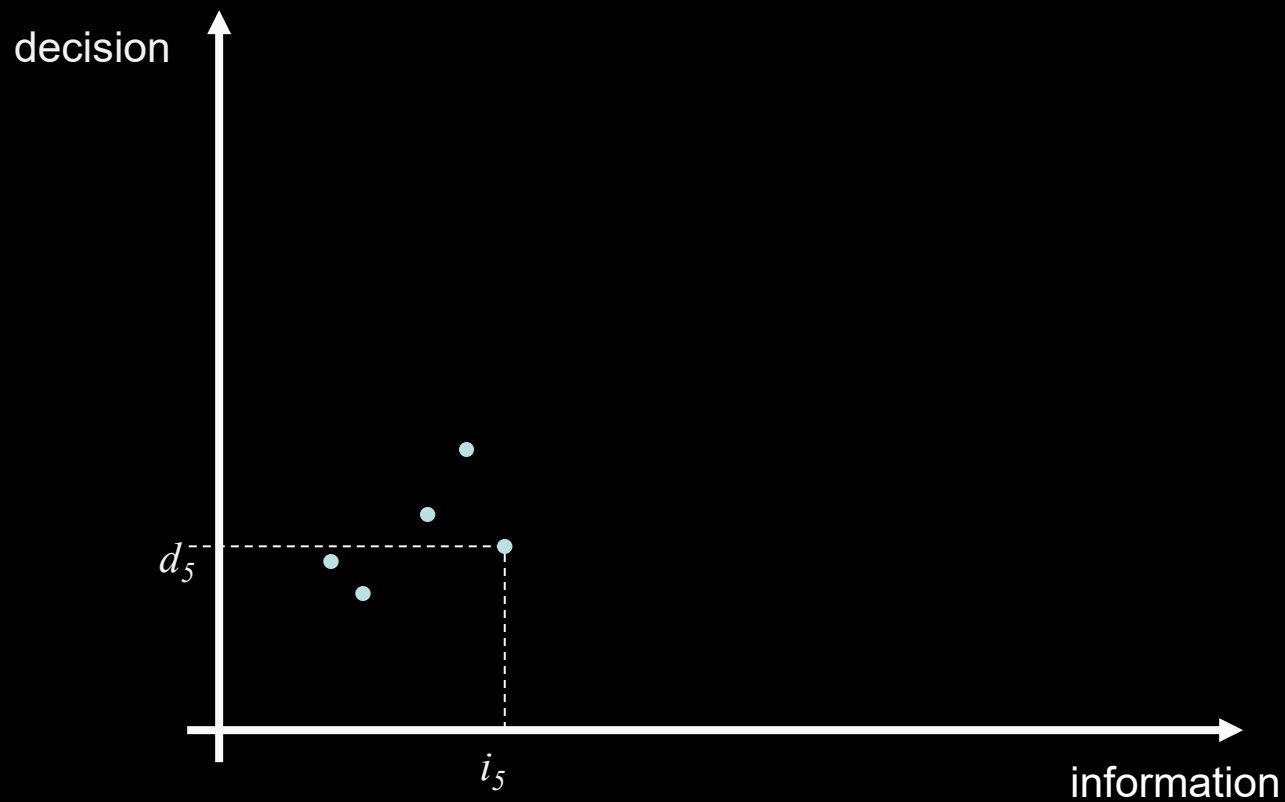
$$\text{decision} = f(\text{information})$$

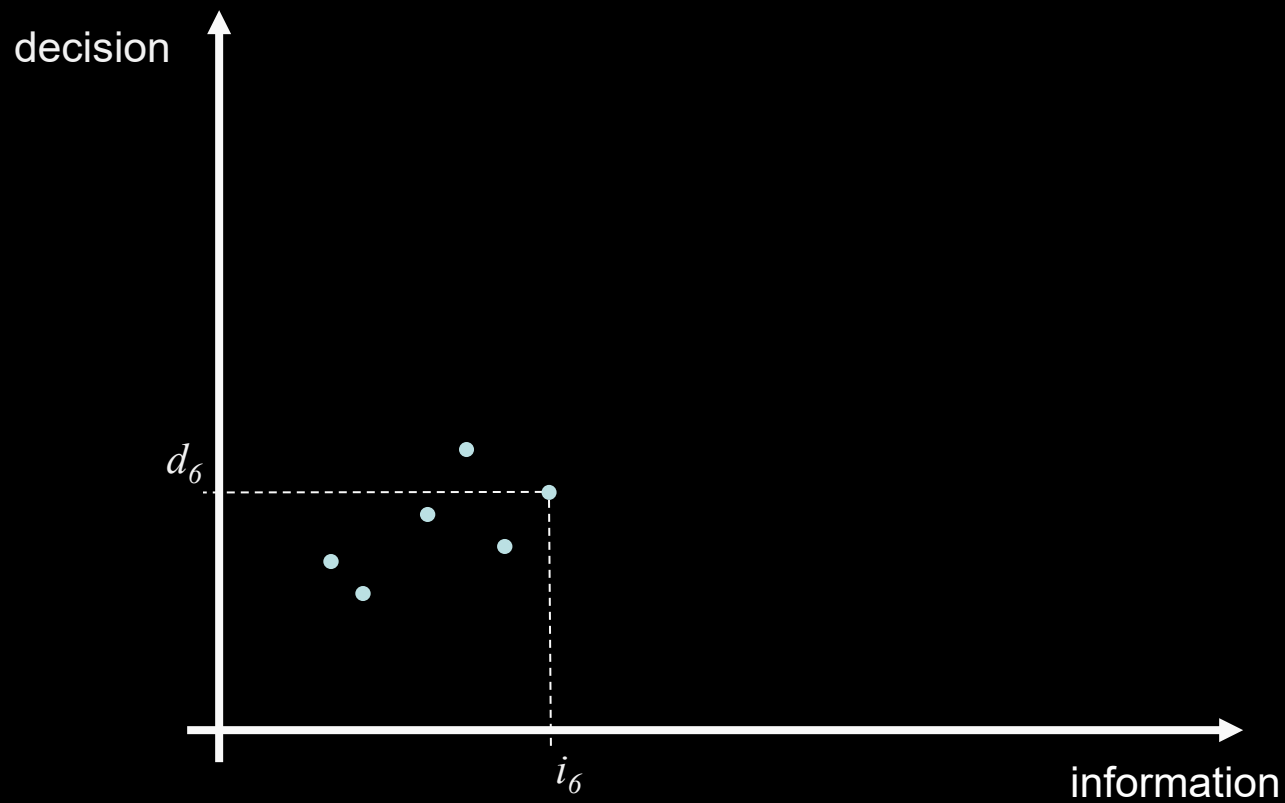


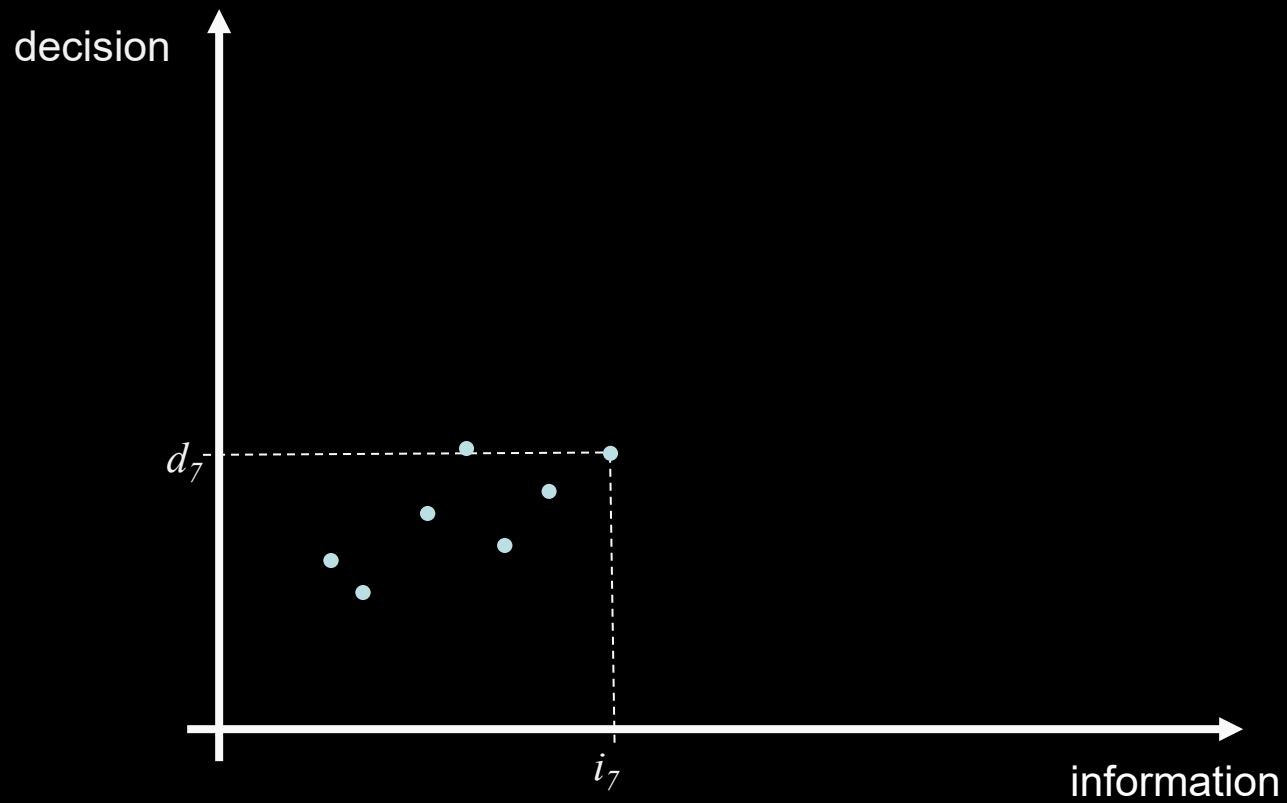


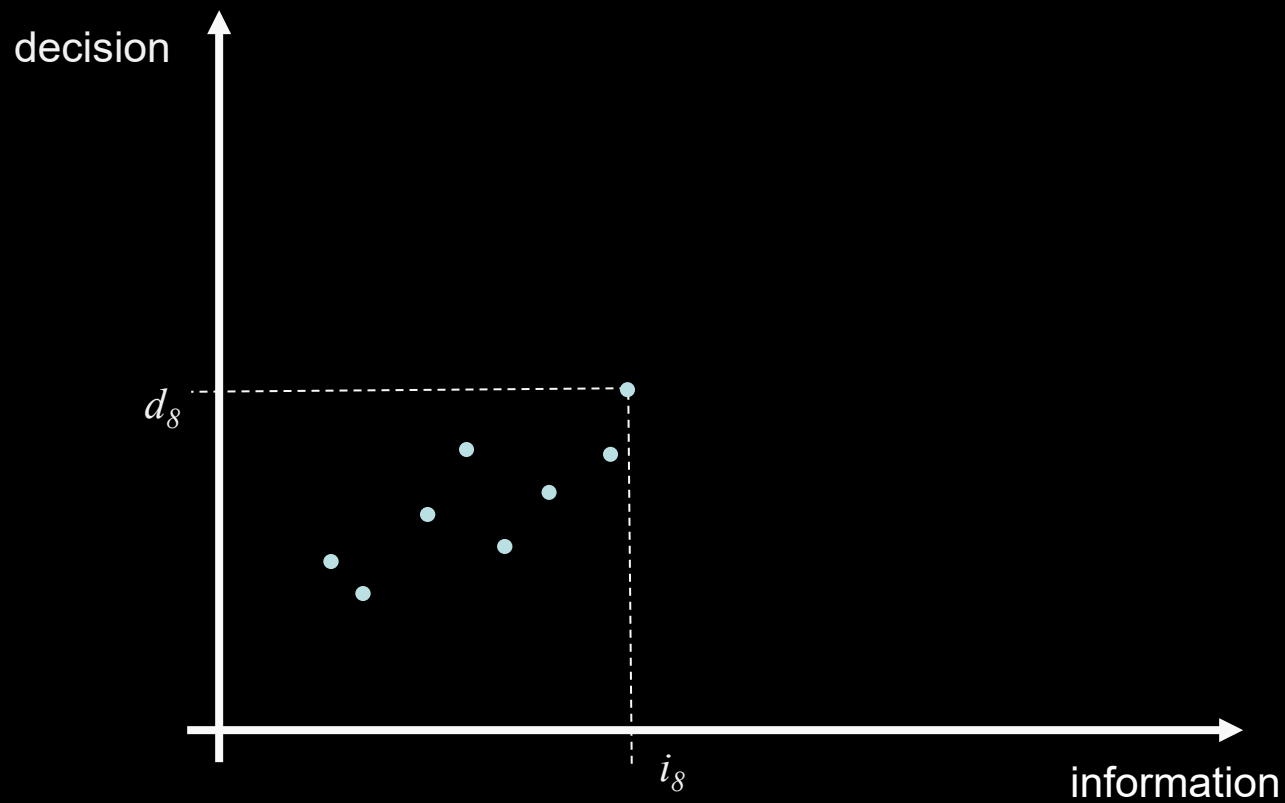


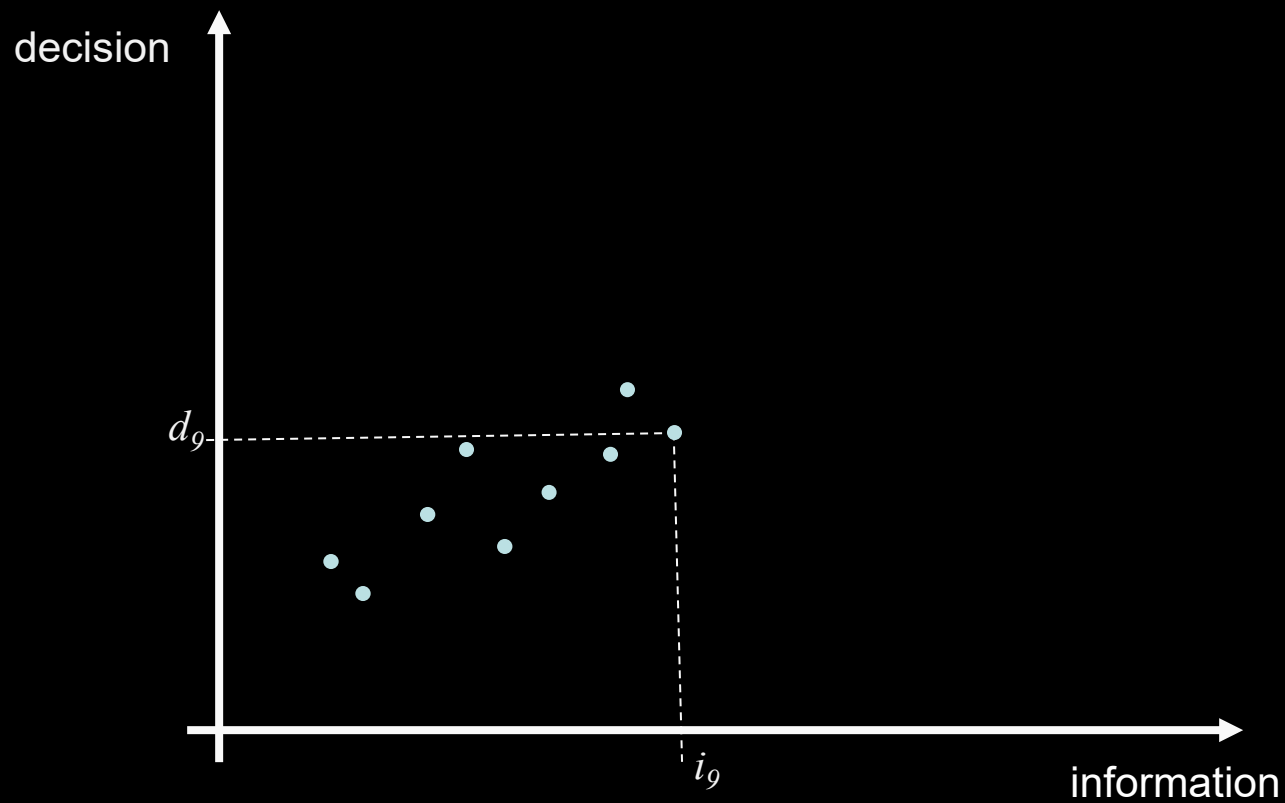


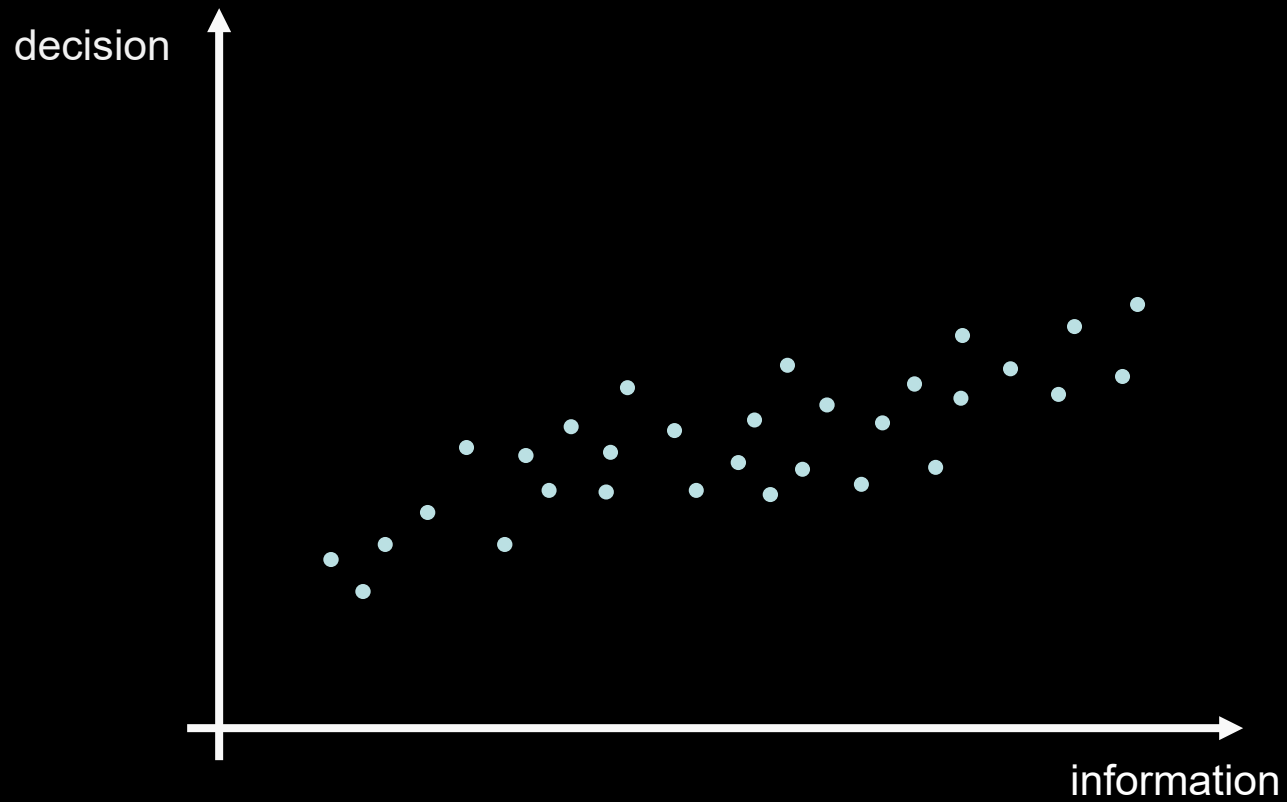




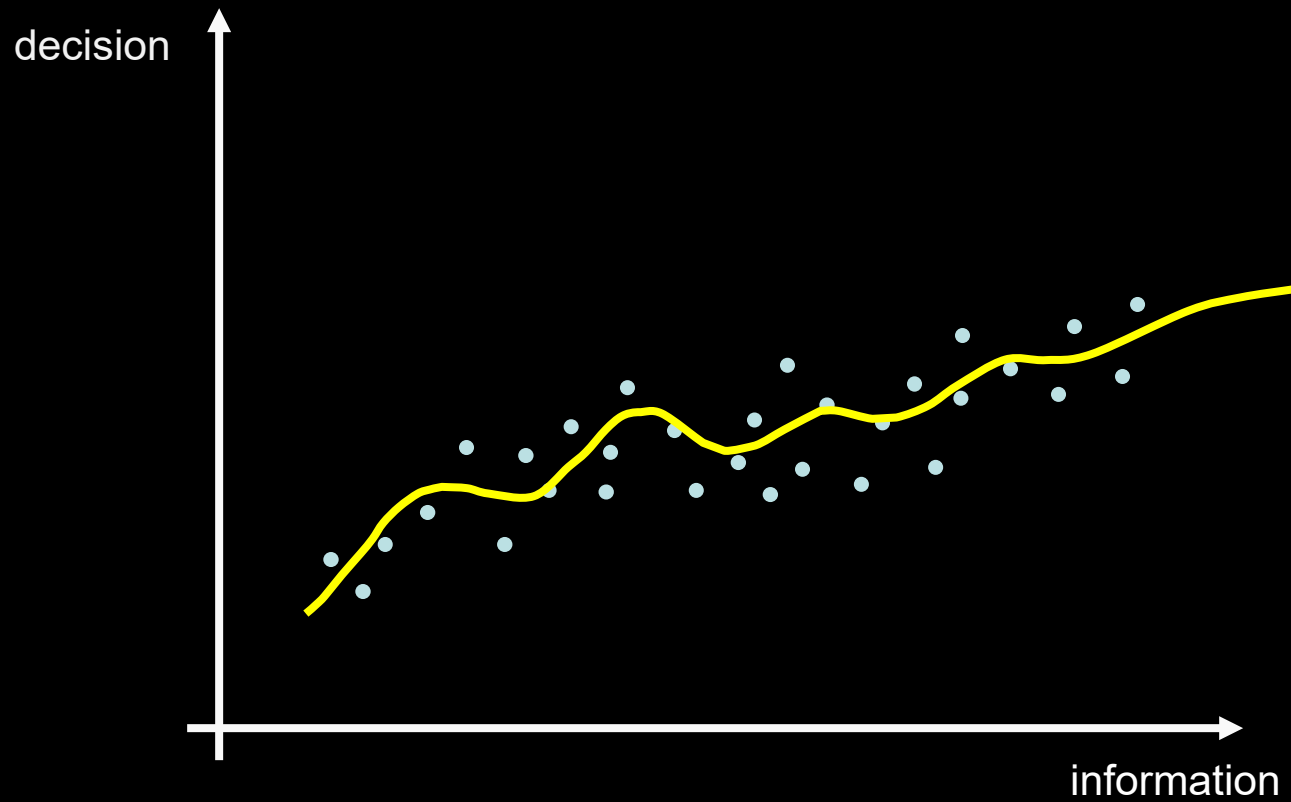






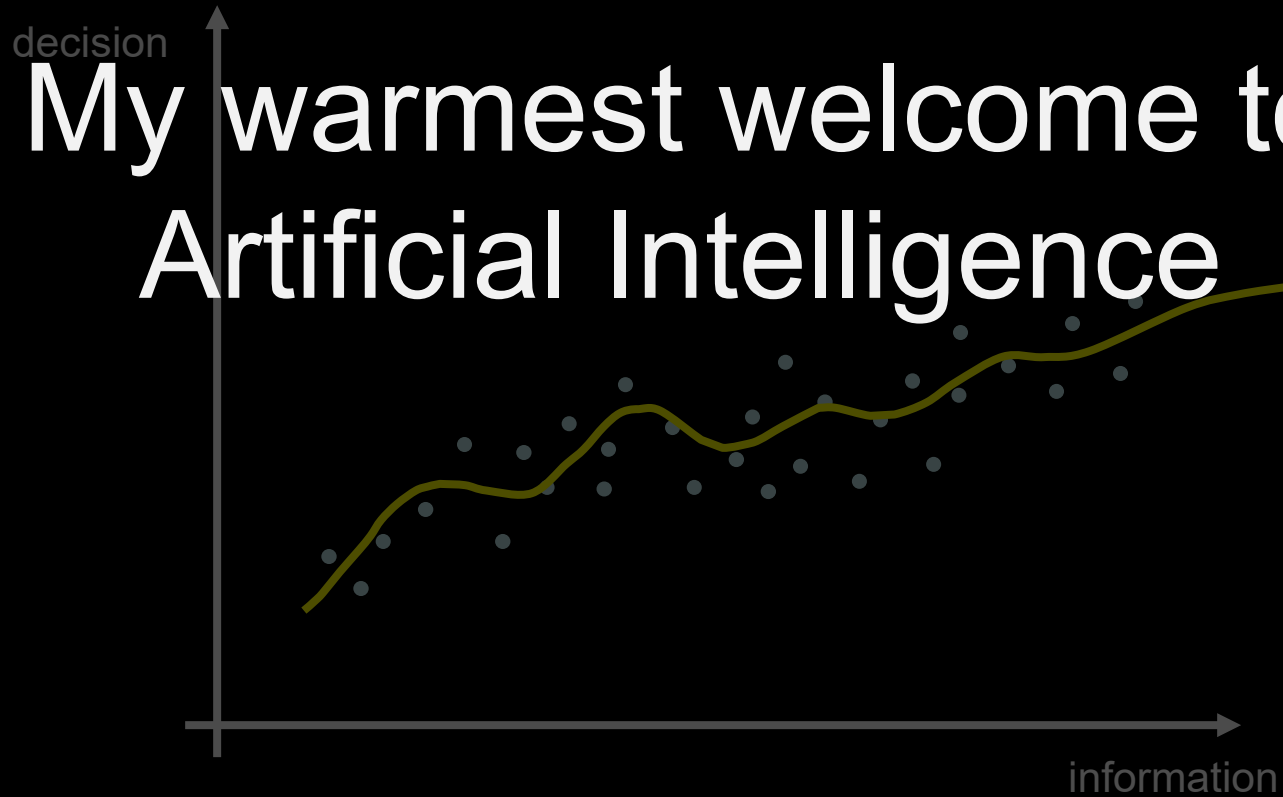


$$\text{decision} = f(\text{information})$$



$$\text{decision} = f(\text{information})$$

My warmest welcome to Artificial Intelligence



$$\text{decision} = f(\text{information})$$