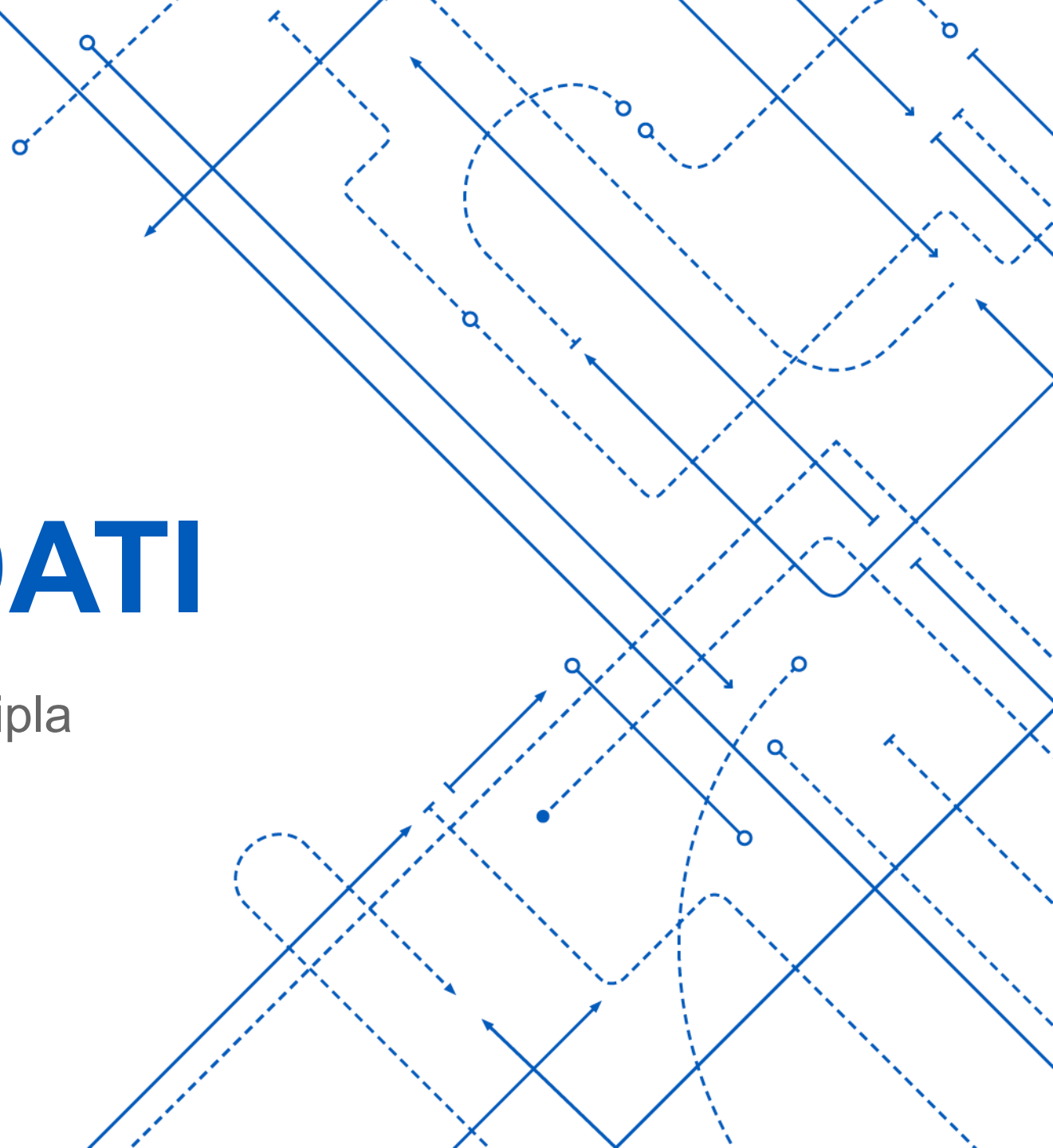
STATISTICA E ANALISI DEI DATI

Capitolo 5 – Regressione Lineare Multipla

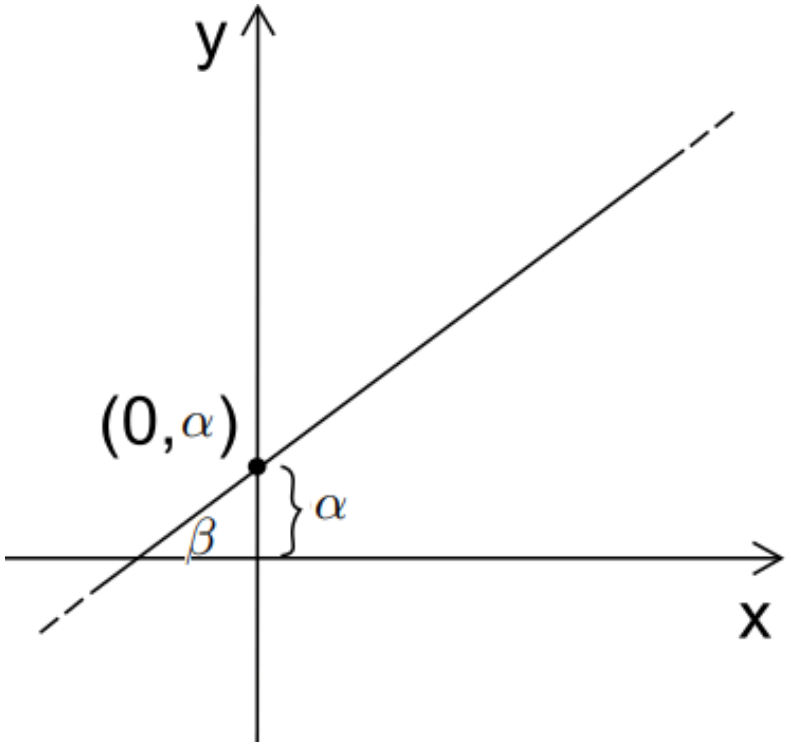
Dott. Stefano Cirillo
Dott. Luigi Di Biasi

a.a. 2025-2026



Breve Recap (Regressione semplice)

Il **modello lineare semplice** prevede **l'uso di una retta di interpolazione** per **stimare quanto più precisamente possibile** tutti i punti corrispondenti ai nostri dati (X,Y):



Breve Recap (Regressione semplice/2)

$$Q = \sum_{i=1}^{n} \left[y_i - (\alpha + \beta\, x_i) \right]^2$$

Il **metodo dei minimi quadrati** ci suggerisce un modo per calcolare intercetta e coefficiente angolare utilizzando deviazione standard e coefficiente di correlazione.

$$\beta = \frac{s_y}{s_x} r_{xy}, \qquad \alpha = \overline{y} - \beta\, \overline{x}.$$

`lm(y~x)`

In R, la funzione **lm()** fornisce i valori dell'intercetta α e del coefficiente angolare.

$$E_i = y_i - \widehat{y}_i = y_i - (\alpha + \beta\, x_i)$$

`resid(lm(y~x))`

Usando i residui possiamo osservare quanto la retta interpolante si adatta alle osservazioni. La funzione resid in R ci permette di valutare i residui.

Breve Recap (Regressione semplice/3)

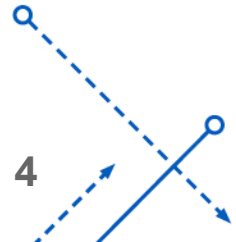
In R, le funzioni fitted() e abline() sono molto utili nell'analisi dei dati e nella rappresentazione grafica, soprattutto quando si tratta di analizzare e visualizzare modelli di regressione.

```
fitted(lm(y~x))
```

La funzione fitted() restituisce i valori predetti dal modello statistico per ogni osservazione nei dati. È usata principalmente per ottenere i valori che il modello ha stimato sulla base delle variabili indipendenti fornite.

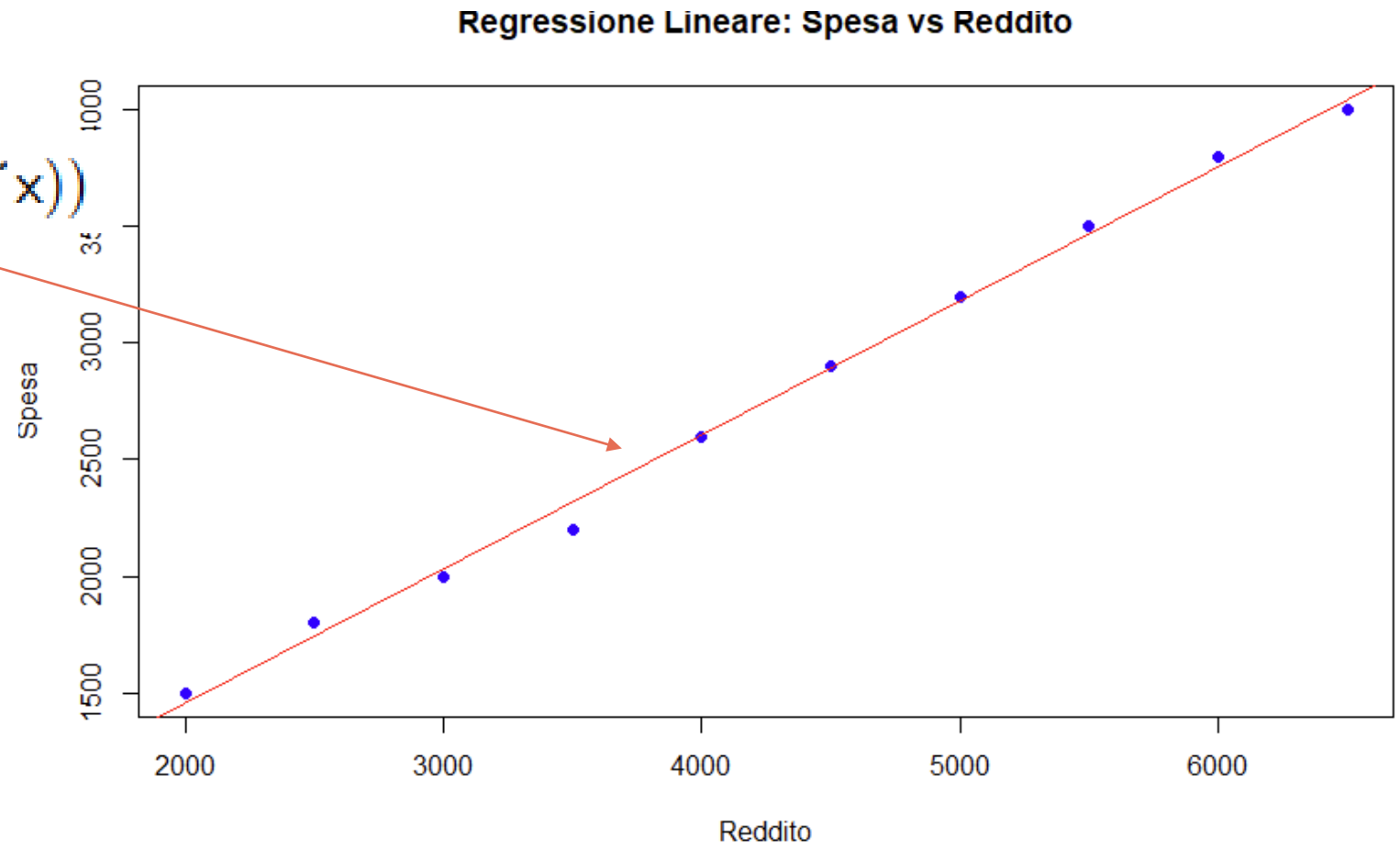
```
abline(lm(y~x))
```

La funzione abline() è utilizzata per aggiungere linee a un grafico in R. Una delle applicazioni più comuni è aggiungere la linea di regressione su un grafico a dispersione, usando un modello di regressione lineare creato con lm().



Breve Recap (Regressione semplice/4)

abline(lm(y~x))



Breve Recap (Quanto è precisa la regressione?)

Abbiamo riflettuto sul fatto che **per vedere quanto la retta si adatta ai dati (ovvero, quanto il nostro modello lineare è bravo)**, l'accento può essere posto su due valori particolari
:

- sul **quadrato del coefficiente di correlazione** (su **quanto esso si avvicini ad uno);**
- **sul coefficient di determinazione.**

$$r_{xy} = \frac{C_{xy}}{s_x\, s_y}.$$

$$D^2 = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2}{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2} = \frac{\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}.$$

Breve Recap (Quanto è precisa la regressione?)

$$r_{xy} = \frac{C_{xy}}{s_x s_y}.$$

Ci dà un'idea di **quanto e come** la variabile dipendente è collegata alla variabile indipendente:

- Se $R_{xy} > 0$ → **correlazione positiva**.
- Se $R_{xy} < 0$ → **correlazione negativa**.
- Se $R_{xy} = 0$ → NO correlazione.

Breve Recap (Quanto è precisa la regressione?)

$$D^2 = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2}{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2} = \frac{\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}.$$

Ci permette di avere un'idea **sulla proporzione della varianza della variabile dipendente che è spiegata dalla variabile indipendente.**

Come possiamo interpretarlo?

Supponiamo $D^2 = 0.8$ → **80% della varianza della variabile dipendente è spiegata dalla variabile indipendente.**

Coefficienti di correlazione campionario e di determinazione

$$D^2 = r_{xy}^2$$

Nel caso di **regressione lineare semplice**, il coefficiente di determinazione **coincide** con il quadrato del coefficiente di correlazione.

```
summary(lm(y~x))$r.square.
```

In R, la funzione summary() applicata a un modello di regressione lineare (lm) restituisce **un riepilogo dettagliato delle caratteristiche e dei risultati del modello**. Questo output è fondamentale per interpretare i risultati della regressione, comprendere l'importanza delle variabili indipendenti e valutare la bontà del modello.

Regressione Lineare Multipla

In molte applicazioni dell'analisi di regressione sono coinvolte situazioni con più di una singola variabile indipendente.

In molte applicazioni dell'analisi di regressione sono coinvolte situazioni con più di una singola variabile indipendente.

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p.$$

Intercetta

Regressori

Y quando $X_1 = X_2 = \ldots = X_p = 0$;

Regressione Lineare Multipla

In molte applicazioni dell'analisi di regressione sono coinvolte situazioni con più di una singola variabile indipendente.

In molte applicazioni dell'analisi di regressione sono coinvolte situazioni con più di una singola variabile indipendente.

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p.$$

Intercetta

Regressori

Y quando $X_1 = X_2 = \ldots = X_p = 0$;

Regressione Lineare Multipla

In molte applicazioni dell'analisi di regressione sono coinvolte situazioni con più

In particolare:

- β_1 rappresenta l'inclinazione di Y rispetto alla variabile X_1 **tenendo costanti le variabili X2, X3, . . . , Xp**

- β_p rappresenta l'inclinazione di Y rispetto alla variabile Xp tenendo costanti le variabili X1, X2, . . . , Xp−1.

Intercetta

Regressori

Y quando $X_1 = X_2 = \ldots = X_p = 0$;

Tuttavia, un singolo regressore non sempre basta!

La **spesa familiare** mensile (Y) *«potrebbe»* dipende da un insieme di variabili:

- reddito familiare mensile (X_1)
- il numero di componenti della famiglia (X_2),
- l'età del capofamiglia (X_3)

Il nostro modello di regressione lineare multipla assumerà dunque la forma seguente:

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\epsilon$$

Regressione Lineare Multipla /2

Utilizziamo R per risolvere il problema.

Ipotizziamo di aver raccolto il seguente campione di dati.

Untitled1* famiglie

Filter

	Reddito	Componenti	Eta	Spesa
1	3000	4	35	1500
2	4500	3	40	2200
3	2500	2	30	1200
4	5000	5	50	3000
5	4000	3	45	2300
6	3500	4	38	1600
7	5500	6	55	3200
8	6000	5	60	3500
9	2800	2	28	1400
10	4700	3	48	2500

```
famiglie <- data.frame(
  Reddito = c(3000, 4500, 2500, 5000, 4000, 3500, 5500, 6000, 2800, 4700),
  Componenti = c(4, 3, 2, 5, 3, 4, 6, 5, 2, 3),
  Eta = c(35, 40, 30, 50, 45, 38, 55, 60, 28, 48),
  Spesa = c(1500, 2200, 1200, 3000, 2300, 1600, 3200, 3500, 1400, 2500)
)
```

Regressione Lineare Multipla /2

Utilizziamo R per risolvere il problema.

Mappiamo il nostro modello in codice R

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\epsilon$$

```
modello <- lm(Spesa ~ Reddito + Componenti + Eta, data = famiglie)
```

Regressione Lineare Multipla /2

Utilizziamo R per risolvere il problema.

Analizziamo il nostro modello.
Usiamo summary(modello)

```
Residuals:
     Min       1Q   Median       3Q      Max
-230.261  -82.222    9.404   58.700  192.301

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -749.3755   238.9288  -3.136   0.0202 *
Reddito        0.4172     0.1726   2.417   0.0521 .
Componenti     7.0959    63.0479   0.113   0.9141
Eta           28.7120    20.9146   1.373   0.2189
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 156.2 on 6 degrees of freedom
Multiple R-squared:  0.9752,    Adjusted R-squared:  0.9628
F-statistic: 78.64 on 3 and 6 DF,  p-value: 3.307e-05
```

Proprietà del D^2 :

- assume valori tra 0 e 1. Un valore di 0 indica che il modello non spiega affatto la variabilità dei dati, mentre un valore di 1 indica che il modello spiega completamente la variabilità;
- è sempre maggiore o uguale a 0. **In modelli di regressione non lineare o in modelli mal adattati**, può anche assumere valori negativi, ma questo è raro;
- aggiungere più variabili indipendenti a un modello **non può ridurne il valore** → può solo rimanere costante o aumentare. (attenzione, un aumento non implica necessariamente che il modello sia migliore)
- Esiste una versione «aggiustata», che penalizza l'aggiunta di variabili indipendenti non significative. È utile per confrontare modelli con diverse quantità di variabili (provate ad cercarlo).
- E' sensibile agli outlier!

Correlation does not imply causation!!!

Un alto valore di D^2 non implica necessariamente una relazione causale tra le variabili.

Di conseguenza, è di estrema importanza:

1. Applicare un approccio scientifico e critico durante lo studio di un fenomeno;
2. **Evitare** chirurgicamente di inserire **deduzioni soggettive** non legate ai dati;
3. **Fare attenzione quando si analizzano i dati;**
4. Porre estrema attenzione al **contesto** analizzato;
5. Porre estrema attenzione **alle fonti utilizzate**;
6. **Per ogni risultato ottenuto, spiegare in che modo si è raggiunto tale risultato, verificare che non ci siano «flaw» o «bias» o «artefatti» che possono inficiare il risultato ottenuto**.

Perché è importante il contesto?

Vediamo con un esempio reale perché è importante tenere sempre in considerazione «il contesto» nel quale svogliamo un’analisi statistica.

Il problema del riscaldamento climatico

Il riscaldamento climatico **è un fenomeno osservabile che ci mostra l'aumento progressivo della temperatura media** della Terra, principalmente **(si assume, oggi)** a causa delle attività umane.

Questo incremento delle temperature **è associato all'aumento della concentrazione di gas serra** (GHG, greenhouse gases) nell'atmosfera come:

- l'anidride carbonica (CO_2);
- il metano (CH_4);
- il protossido di azoto (N_2O).

Questi gas trattengono il calore nella bassa atmosfera terrestre, creando un "effetto serra" naturale, ma amplificato a livelli insostenibili dall'intervento umano.

Il problema del riscaldamento climatico

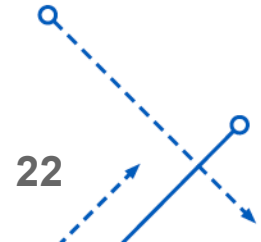
Da statistici e scienziati, la prima domanda che potremmo porci (senza sembrare negazionisti, complottisti, sciakimicari o haarp-ari) nella più pura e candida oggettività, potrebbe (e dovrebbe) essere:

Il riscaldamento climatico (ovvero, l'aumento della temperatura media) **è veramente un problema**? Se sì, perché e per chi?

Prima di rispondere, armiamoci di altri strumenti

Il mondo è un posto molto complesso.

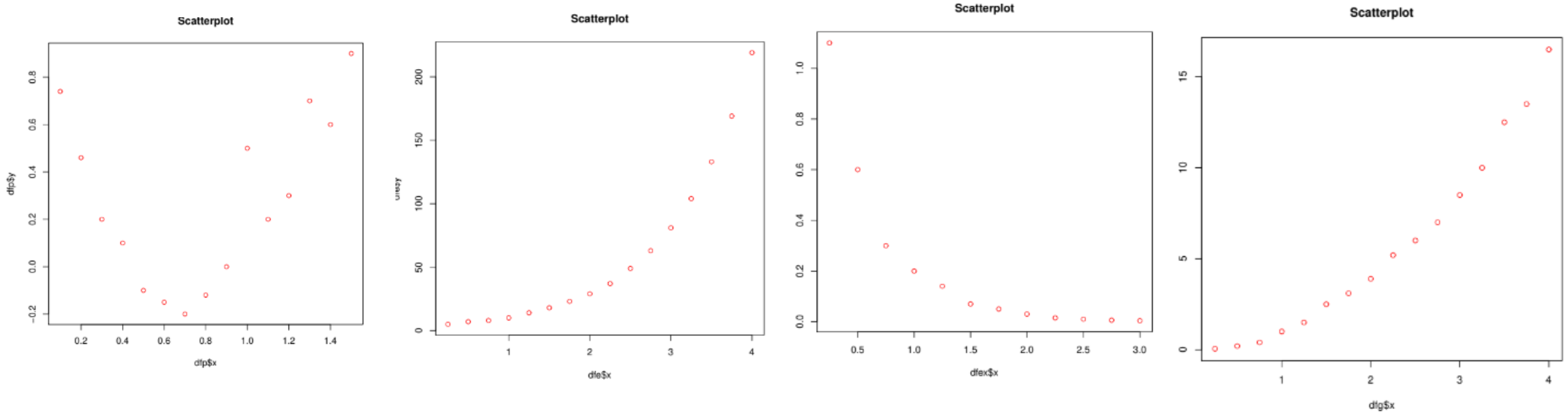
Di conseguenza, non possiamo aspettarci che tutti i fenomeni osservabili seguano un andamento lineare e/o possano essere «interpolati» tramite una retta!



Prima di rispondere, armiamoci di altri strumenti

Il mondo è un posto molto complesso.

Di conseguenza, non possiamo aspettarci che tutti i fenomeni osservabili seguano un andamento lineare e/o possano essere «interpolati» tramite una retta!



Funzioni linearizzabili

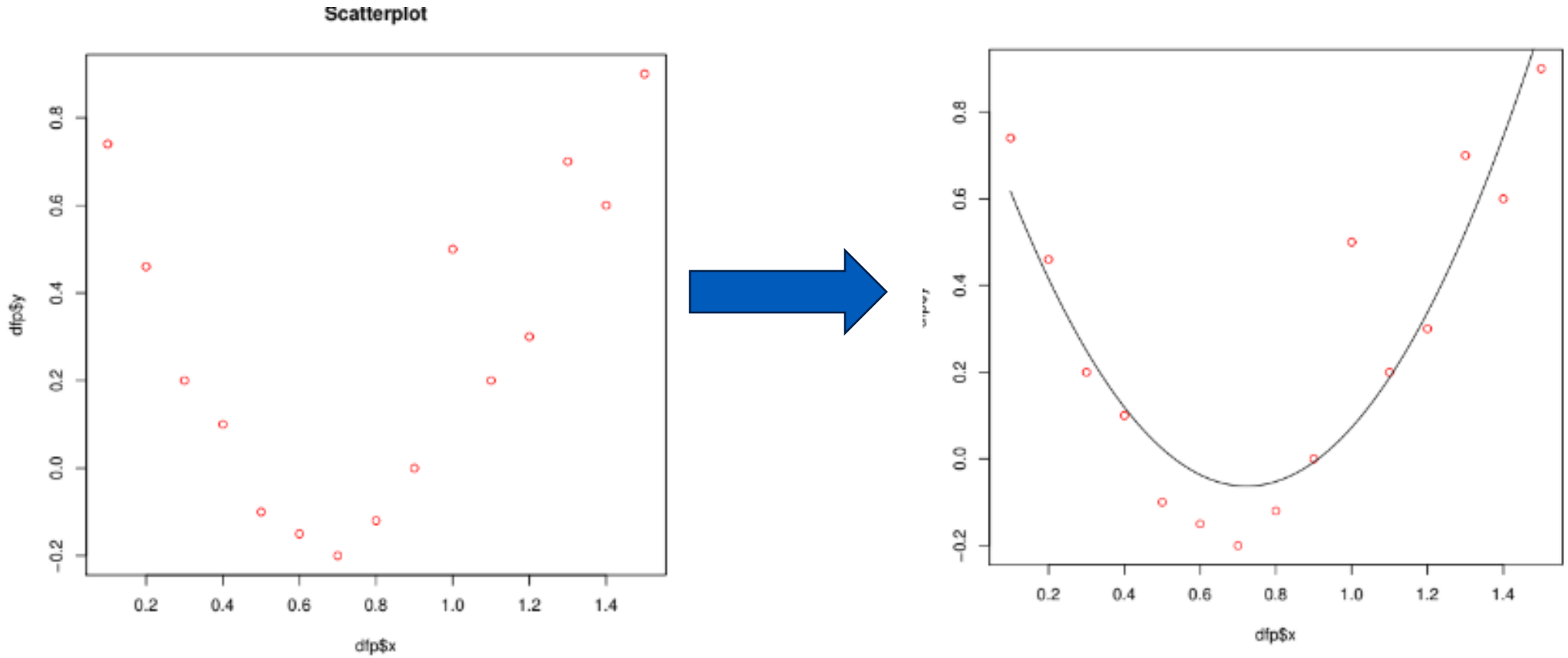
In alcuni casi, modelli che **sembrano non lineari sono linearizzabili** attraverso opportune trasformazioni.

Nei plot precedenti infatti erano rappresentate:

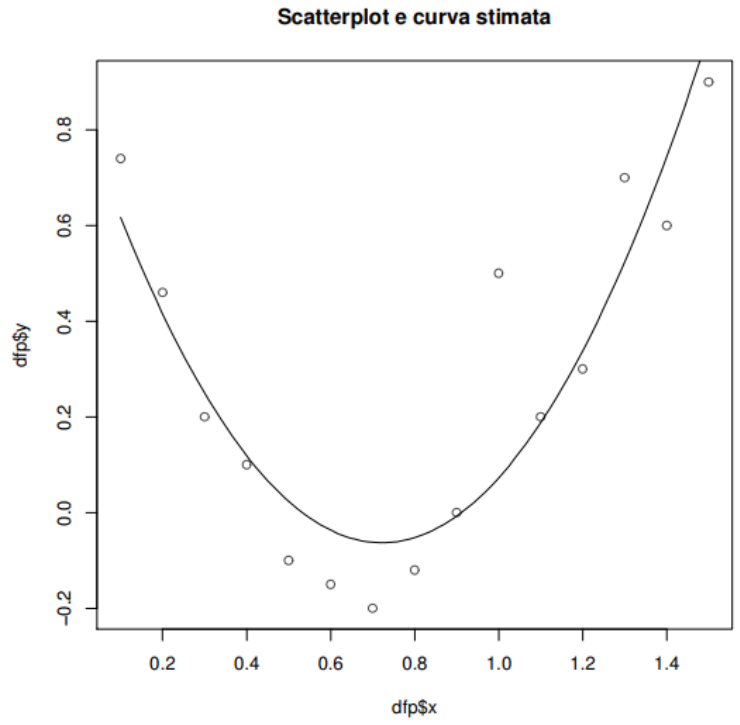
- Regressione quadratica; $Y = \alpha + \beta X + \gamma X^2$, `lm(y~x+I(x^2))`
- Regressione esponenziale; $Y = \alpha + \beta e^X$, `lm(dfe$y~I(exp(dfe$x)))`
- Regressione semi-logaritmica; $Y = e^{\alpha+\beta X}$, `lm(I(log(dfex$y))~dfex$x)`
- Regressione logaritmica. $Y = \alpha_0 X^\beta$ `lm(I(log(dfg$y))~I(log(dfg$x)))`

Statistica e analisi dei dati

Regressione quadratica $Y = \alpha + \beta X + \gamma X^2,$ lm(y~x + I(x ∧ 2))



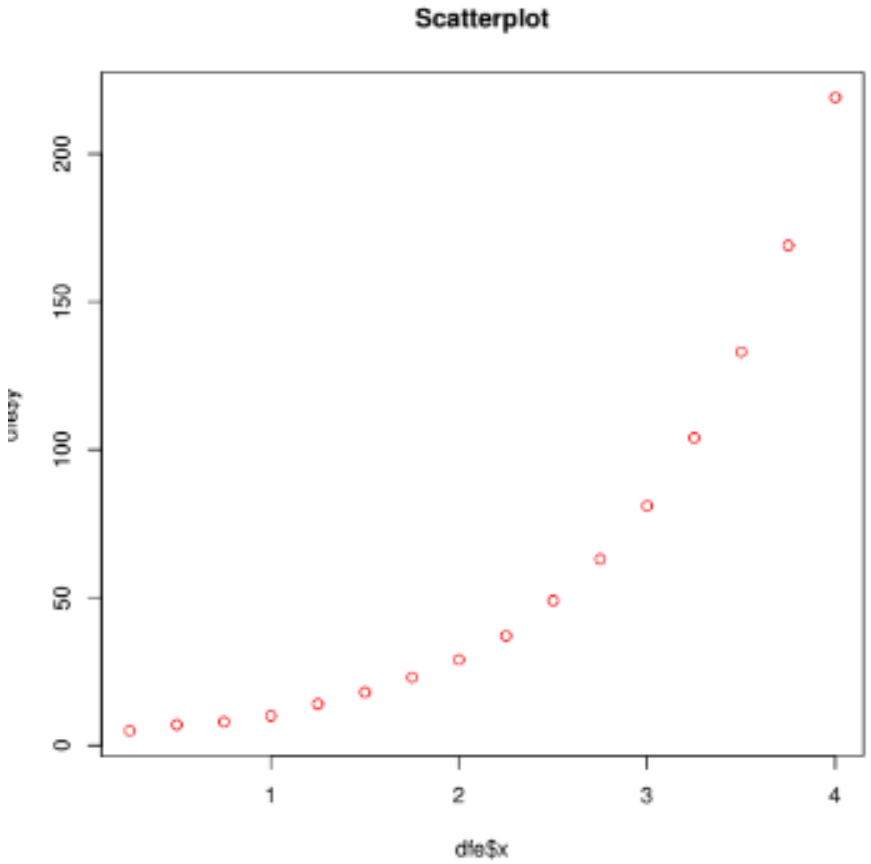
Regressione quadratica

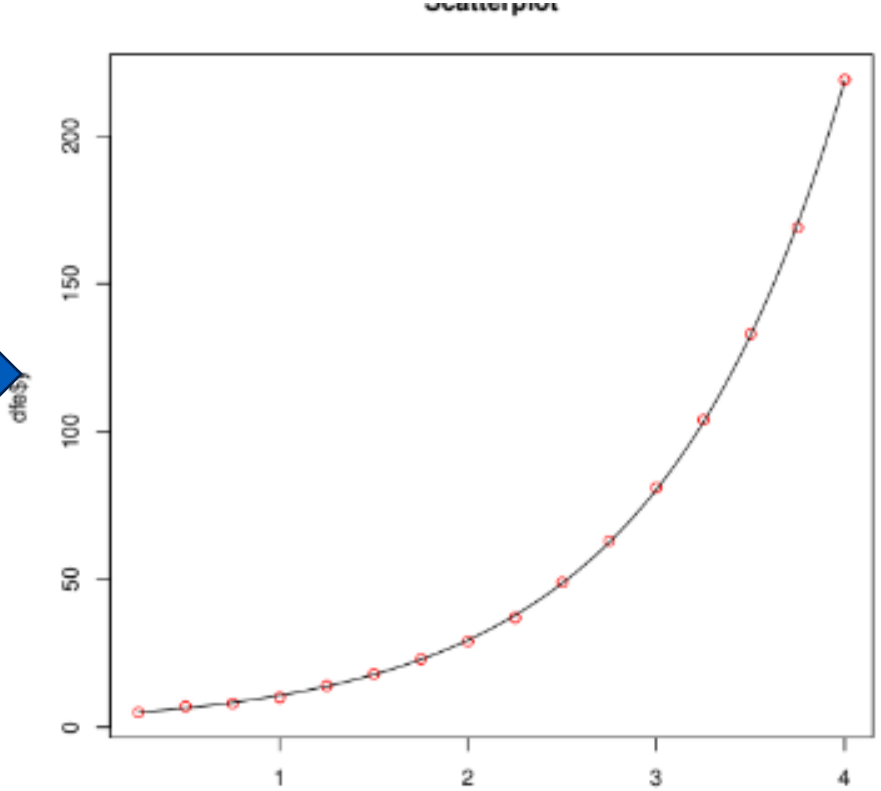


Statistica e analisi dei dati

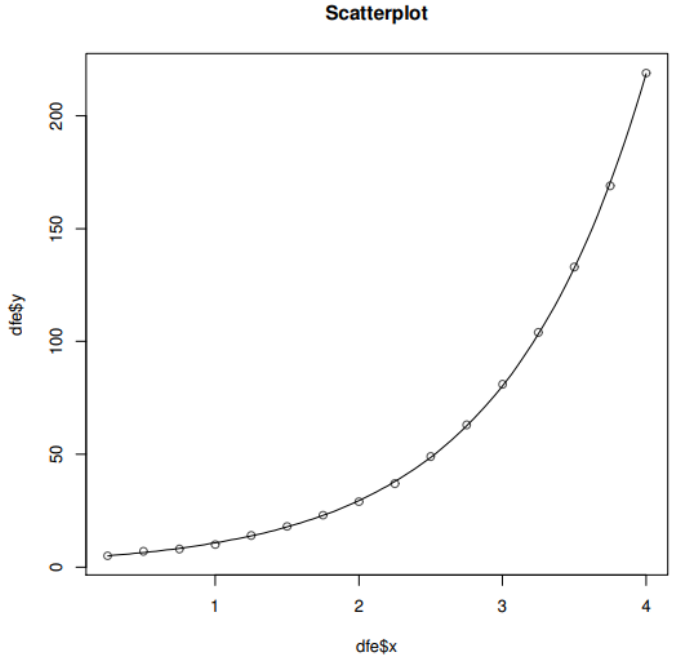
Regressione esponenziale $Y = \alpha + \beta\, e^{X},$

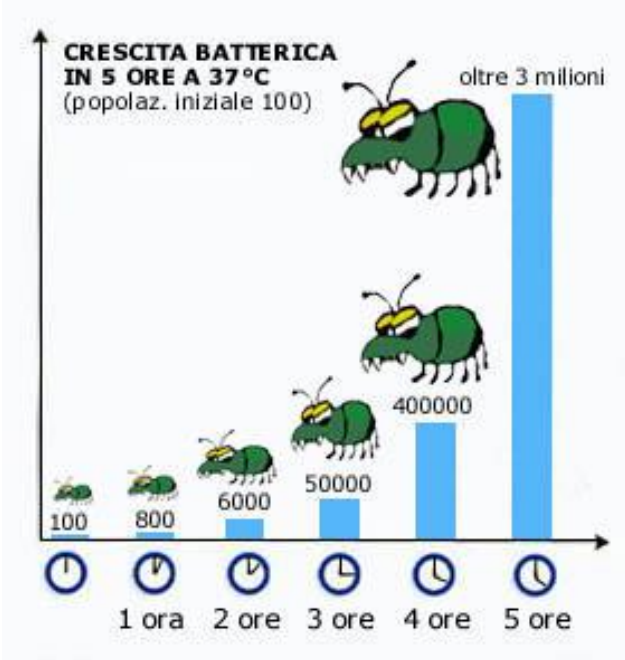
```
lm(dfe$y~I(exp(dfe$x)))
```

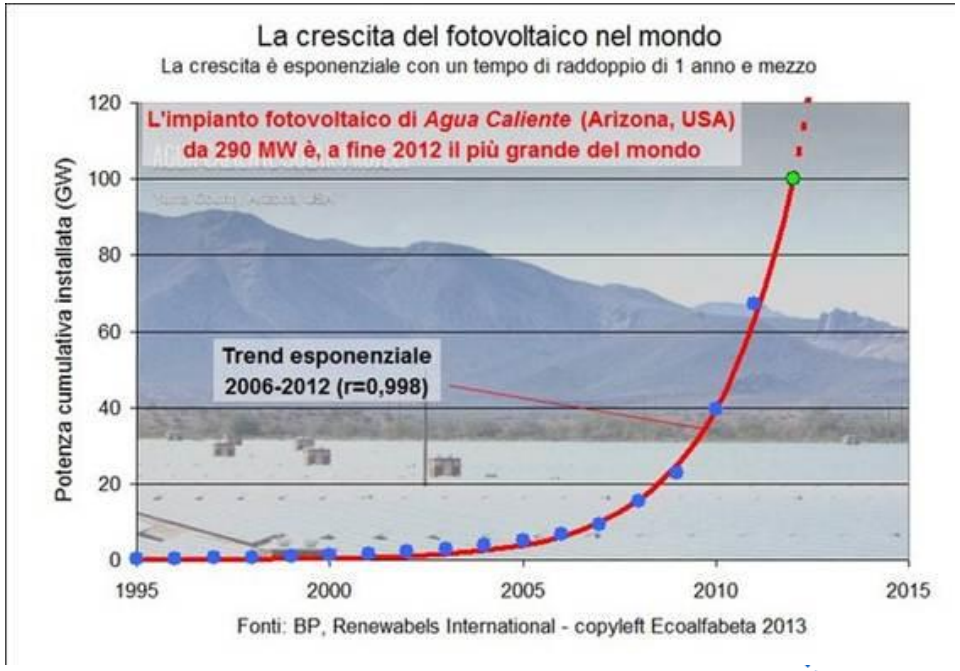




Regressione esponenziale



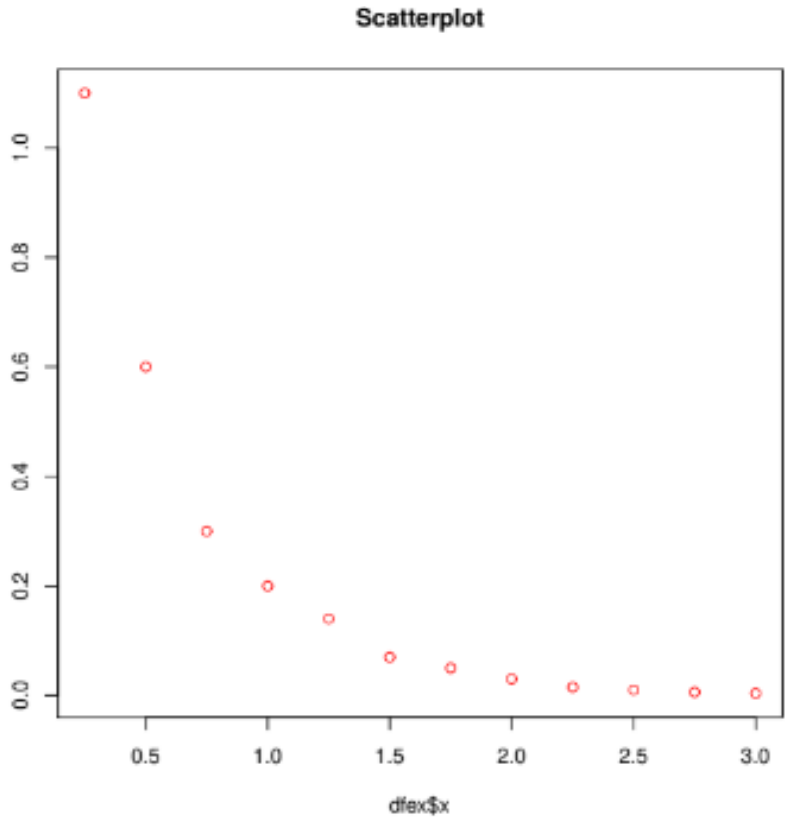


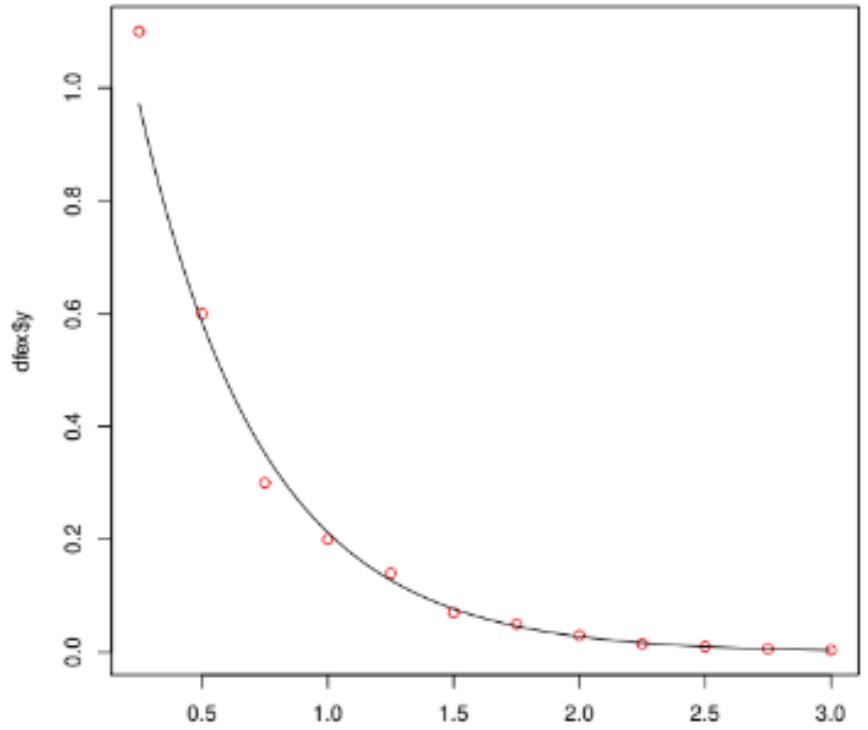


Statistica e analisi dei dati

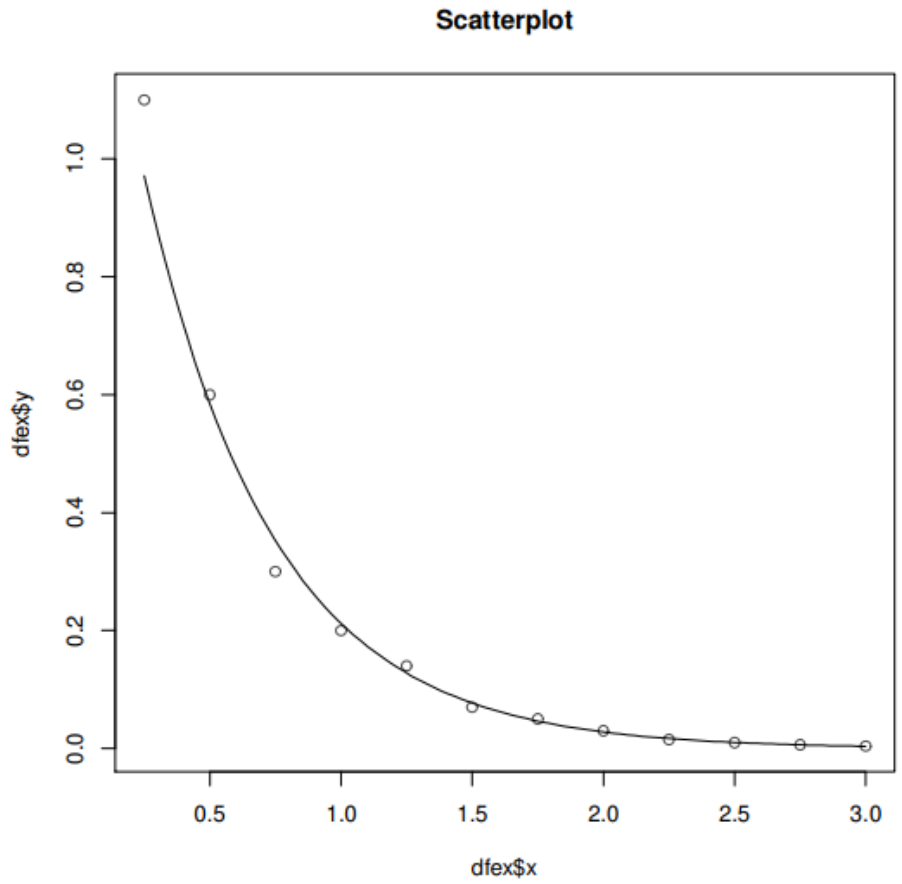
Regressione semi-log $Y = e^{\alpha+\beta X}$,

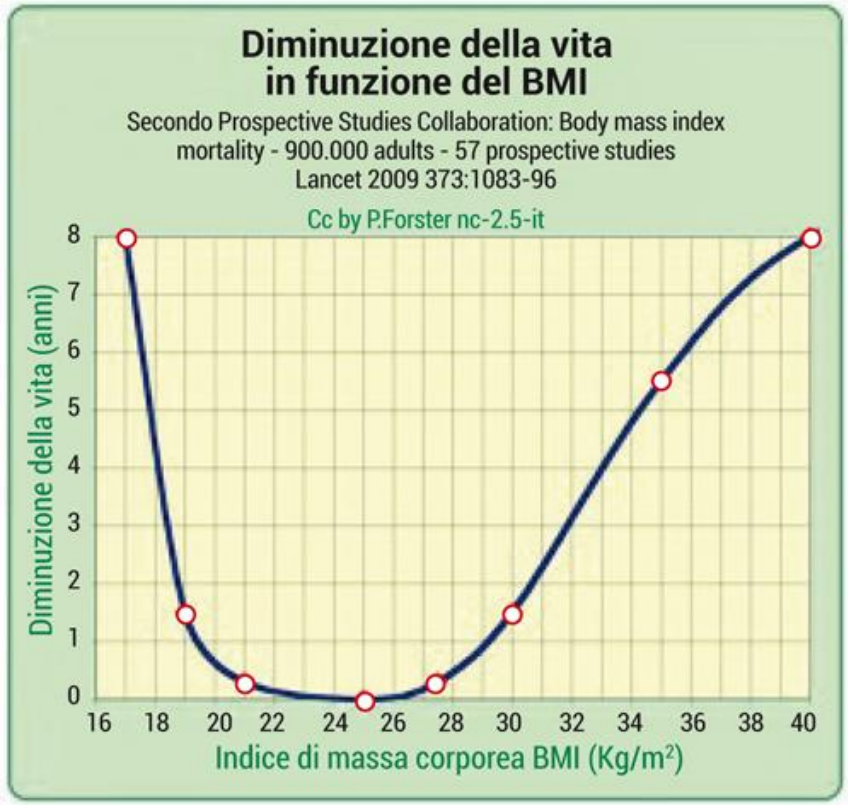
```
lm(I(log(dfex$y))~dfex$x)
```





Regressione logaritmica



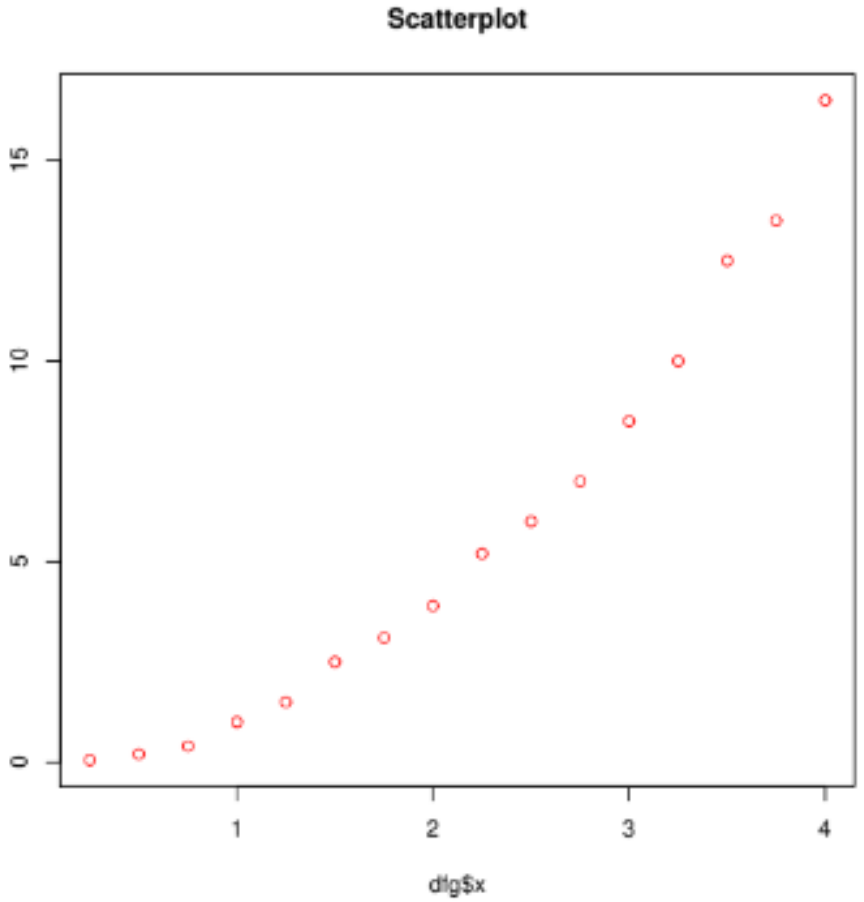


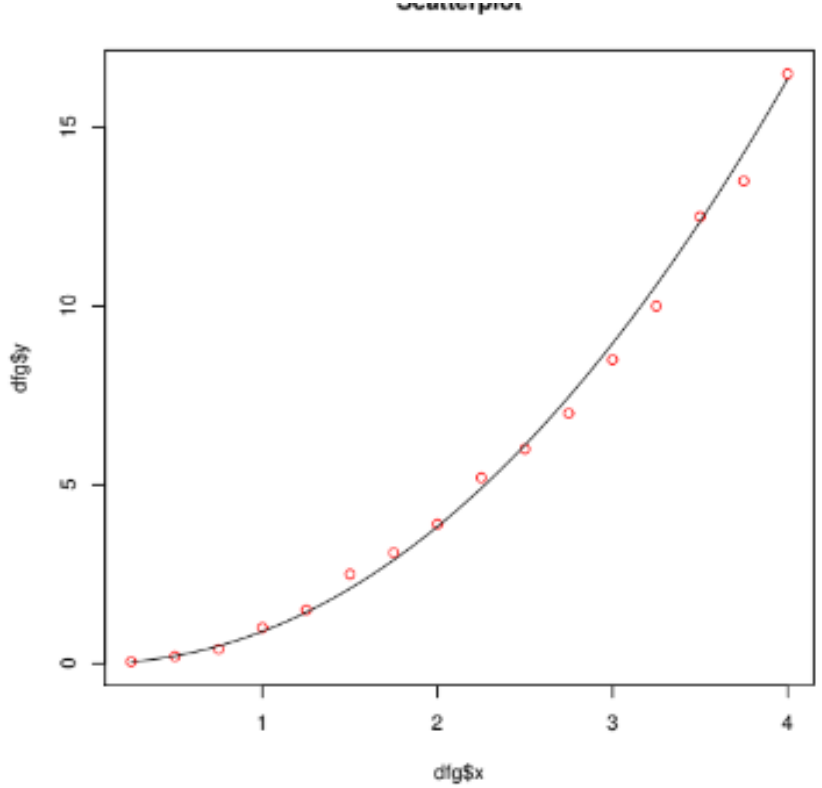
Statistica e analisi dei dati

Regressione logaritmica

$$Y = \alpha_0 X^{\beta}$$

```
lm(I(log(dfg$y))~I(log(dfg$x)))
```

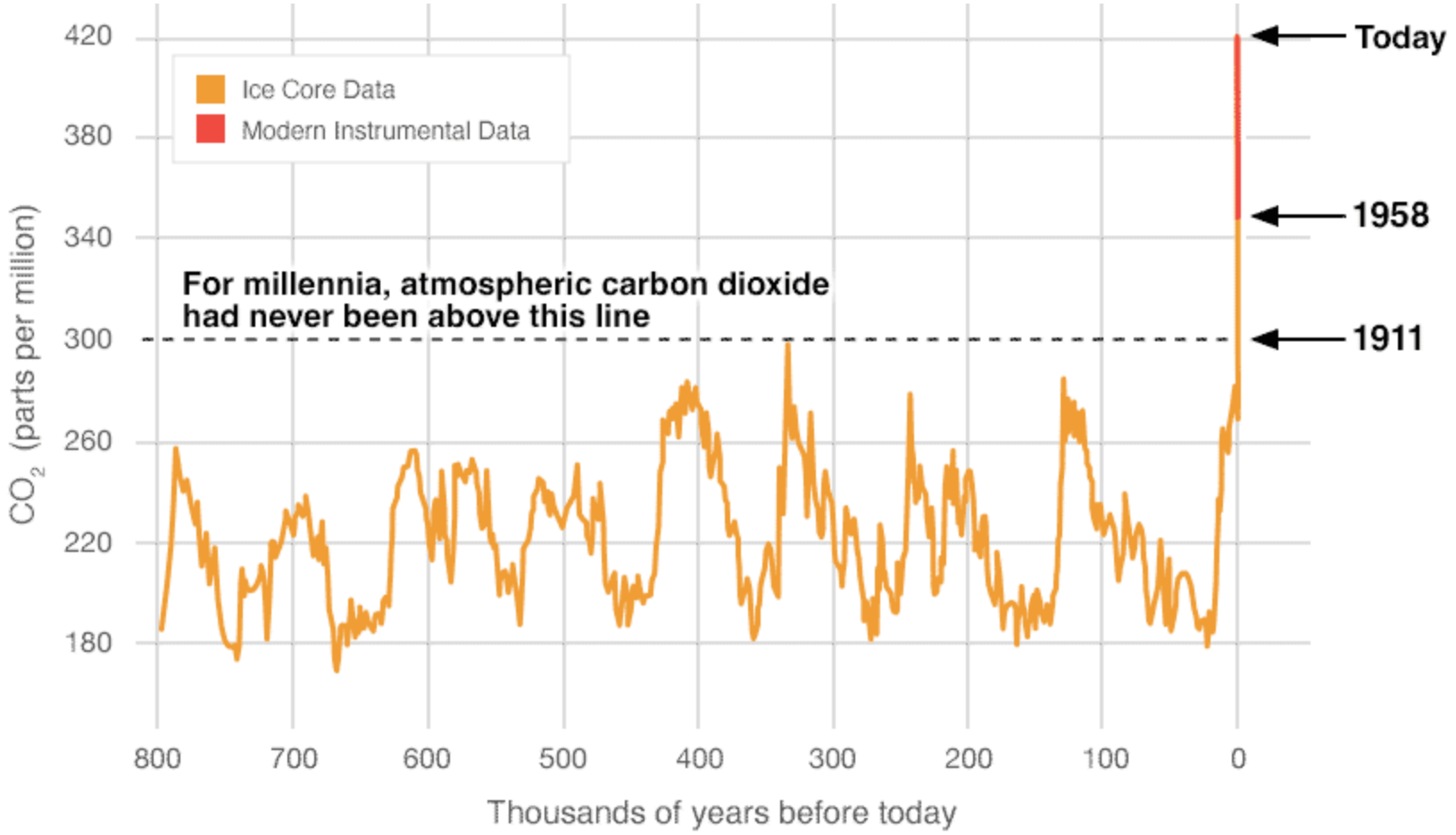




Altre funzioni linearizzabili

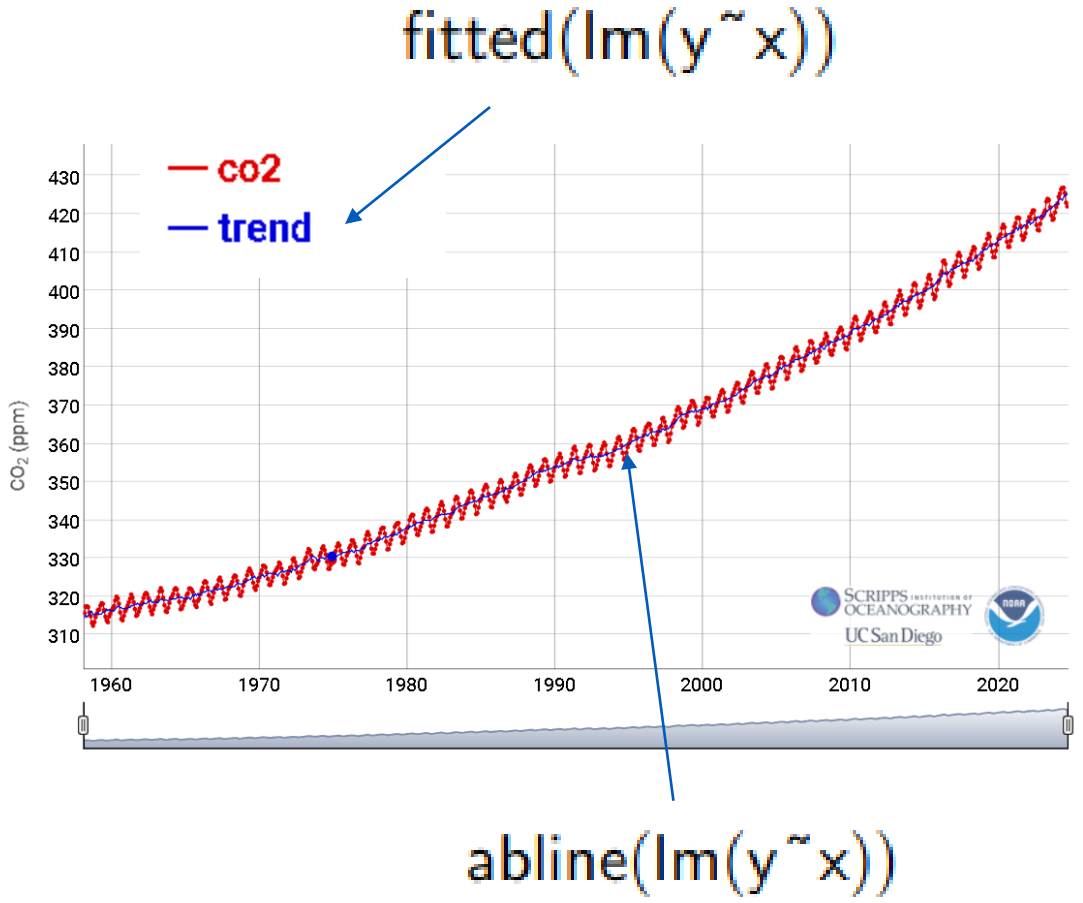
Funzione $y = f(x)$	Forma linearizzata $Y = \alpha + \beta X$	Cambiamenti di variabili e costanti
$y = C \cdot x^{\beta}$	$\log y = \log C + \beta \log x$	$X = \log x, \quad Y = \log y$ $\alpha = \log C$
$y = C \cdot e^{\beta x}$	$\log y = \log C + \beta \cdot x$	$X = x, \quad Y = \log y$ $\alpha = \log C$
$y = \alpha + \beta \cdot \log x$	$y = \alpha + \beta \cdot \log x$	$X = \log x, \quad Y = y$
$y = \frac{\beta}{x} + \alpha$	$y = \alpha + \beta \cdot \frac{1}{x}$	$X = \frac{1}{x}, \quad Y = y$
$y = \frac{H}{C\,x + D}$	$\frac{1}{y} = \frac{D}{H} + \frac{C}{H}\,x$	$X = x, \quad Y = \frac{1}{y}$ $\alpha = \frac{D}{H}, \quad \beta = \frac{C}{H}$
$y = \frac{x}{\beta + \alpha\,x}$	$\frac{1}{y} = \alpha + \beta \frac{1}{x}$	$X = \frac{1}{x}, \quad Y = \frac{1}{y}$
$y = \frac{1}{\alpha + \beta\,e^{-x}}$	$\frac{1}{y} = \alpha + \beta\,e^{-x}$	$X = e^{-x}, \quad Y = \frac{1}{y}$

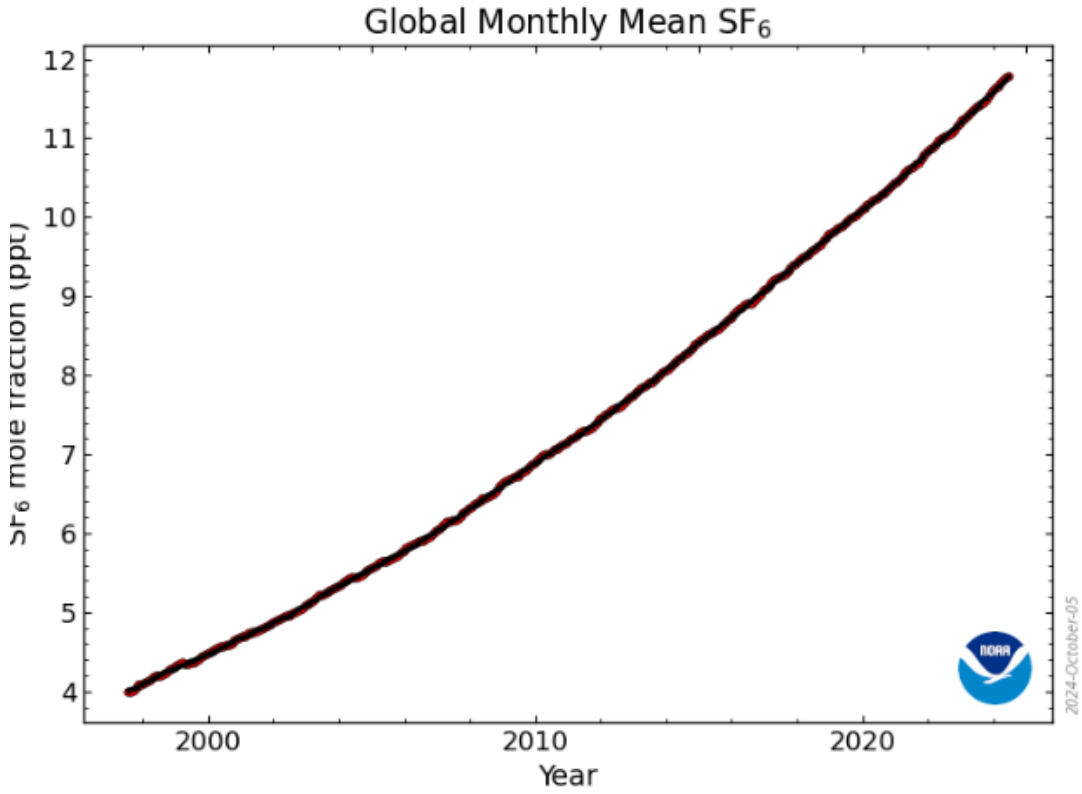
Il problema del riscaldamento climatico



Statistica e analisi dei dati

Il problema del riscaldamento climatico

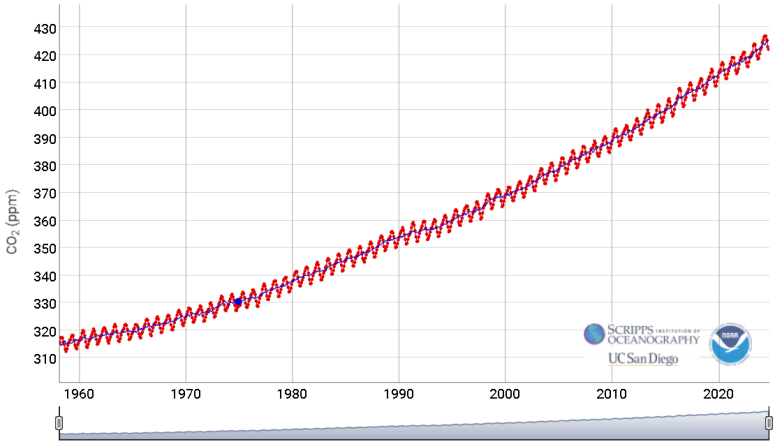


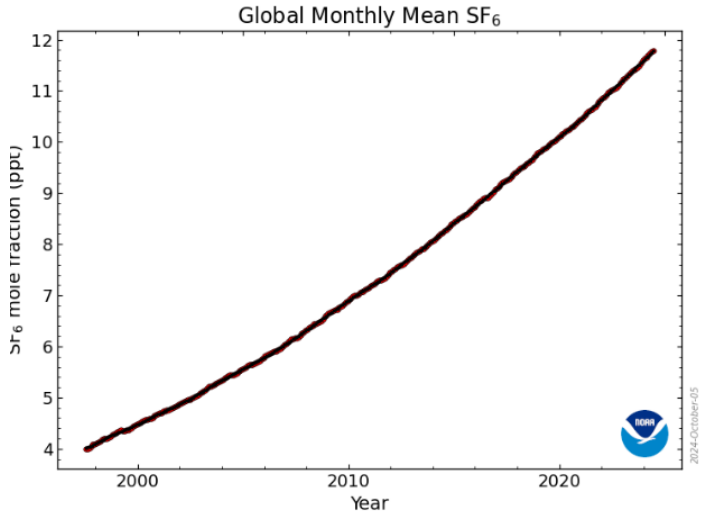


Il problema del riscaldamento climatico

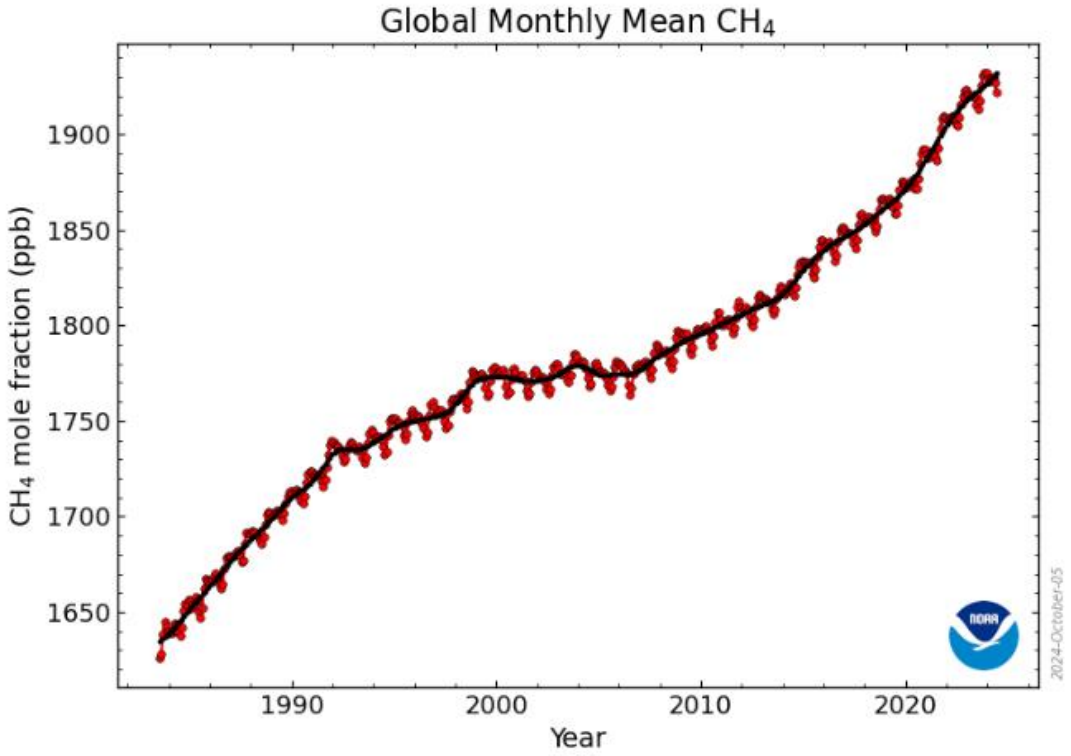
$$r_{xy} = \frac{C_{xy}}{s_x\, s_y}.$$

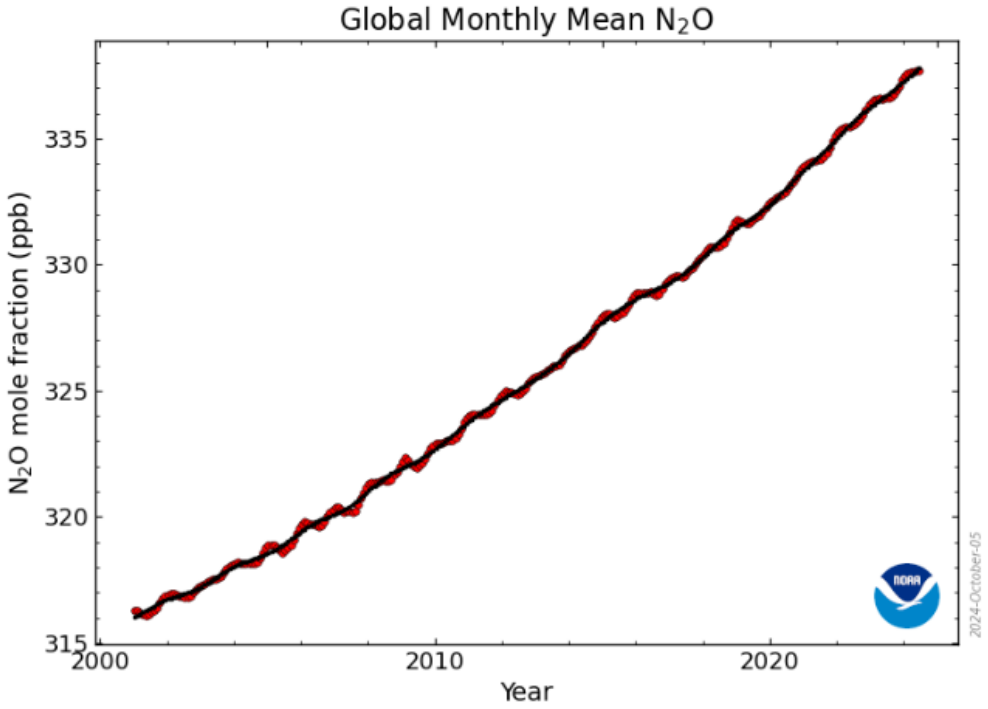
L'andamento della variabile e della variabile SF_6 CO_2 sembrano assumere un trend simile. Che coefficiente di correlazione posso aspettarmi? (Attenzione: considerate solo dal 2000 in poi poiché la registrazione della SF6 è iniziata molto dopo).





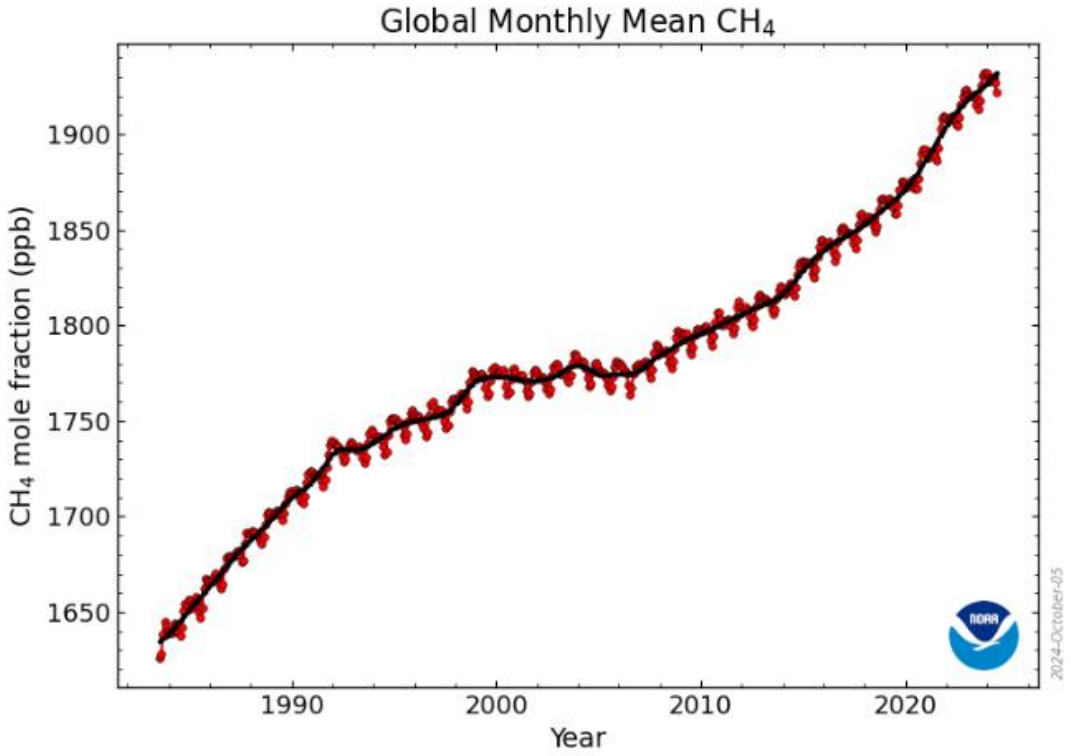
Il problema del riscaldamento climatico

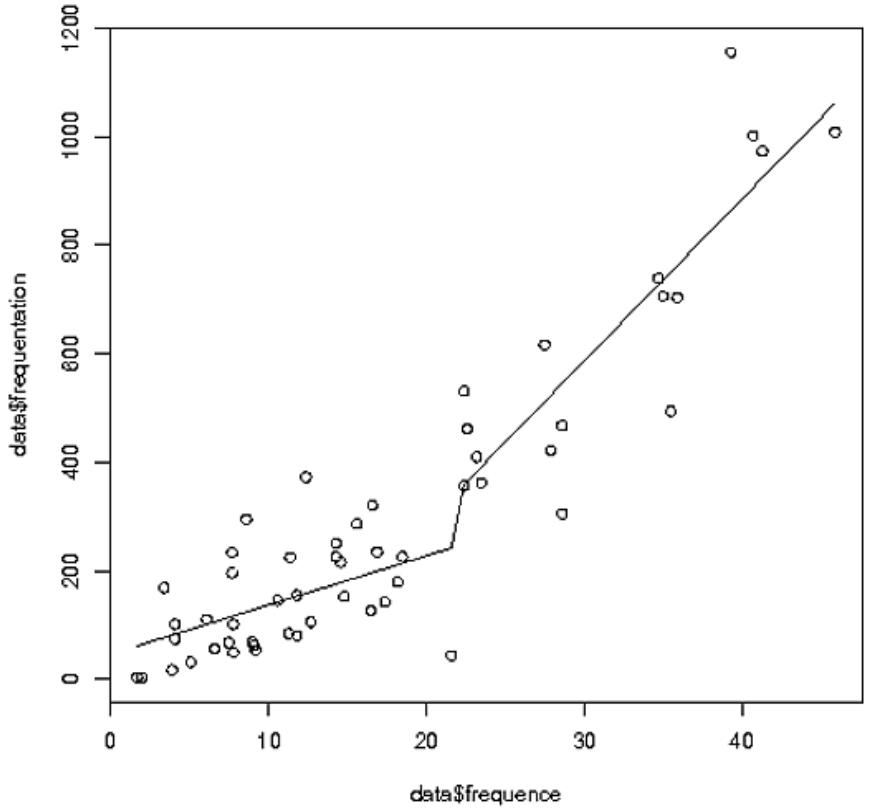


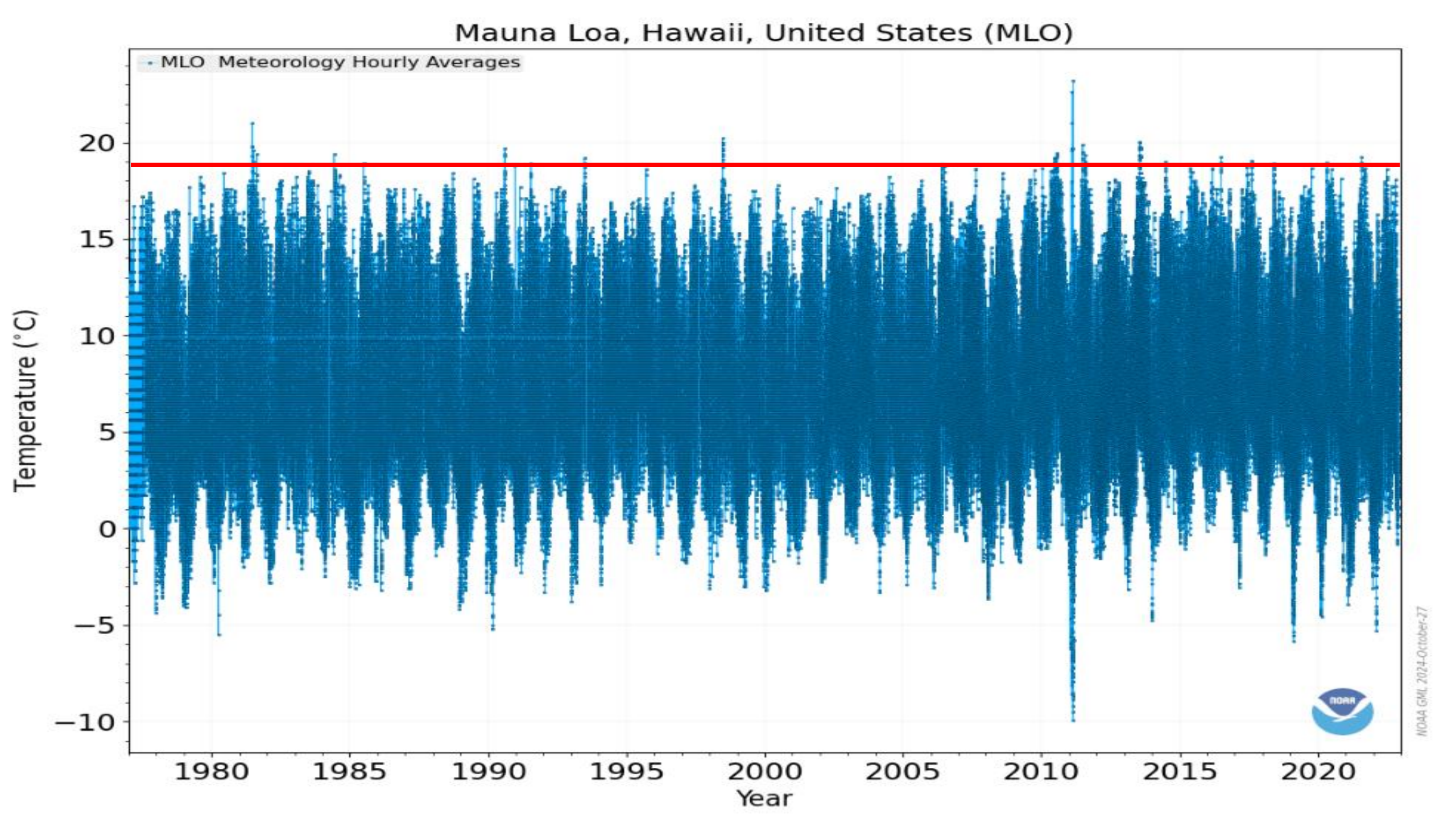


Statistica e analisi dei dati

Il problema del riscaldamento climatico






Mauna Loa, Hawaii, United States (MLO)
MLO Meteorology Hourly Averages
Temperature (°C)
20
15
10
5
0
−5
−10
1980
1985
1990
1995
2000
2005
2010
2015
2020
Year
NOAA GML 2024-October-27

Il problema del riscaldamento climatico

Dopo aver visto i dati che risposta dareste alla domanda: ma è un problema il riscaldamento climatico?

Il problema del riscaldamento climatico

Dopo aver visto i dati che risposta dareste alla domanda: ma è un problema il riscaldamento climatico?

Potrebbe sembrare strano ma… **non esiste una risposta univoca**!

La risposta dipende «dal contesto», dalla «definizione di problema» e da chi «definisce un problema» il riscaldamento climatico.

Il problema del riscaldamento climatico

Potrebbe sembrare strano ma… non esiste una risposta univoca!

La risposta dipende «dal contesto», dalla «definizione di problema» e da chi «definisce un problema» il riscaldamento climatico.

Dal punto di vista «**antropocentrico**» **è un problema reale**, perché oltre una certa temperatura il nostro habitat (la nostra confort-zone, il mondo al quale siamo abituati) ne risentirebbe gravemente (avete visto Interstellar?)

Il problema del riscaldamento climatico

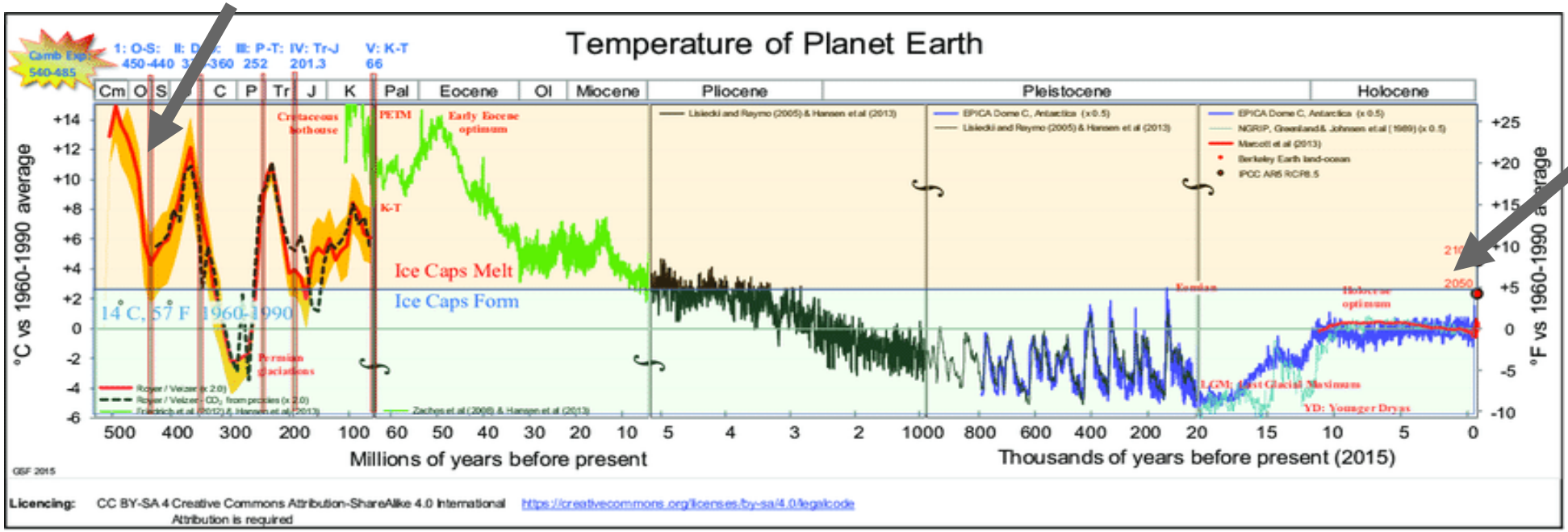
Tuttavia, dal punto di vista «globale» e della «natura» ciò non è vero… anzi…noi esistiamo proprio **«GRAZIE AL RISCALDAMENTO CLIMATICO»**

All'inizio del periodo paleozoico (nel cambriano), circa 538 milioni di anni fa, «temperature estremamente alte» e «alti livelli di gas serra» comportarono **un'improvvisa esplosione di vita complessa** e praticamente tutti i principali phylum animali cominciarono ad apparire nei registri fossili (Cambrian explosion).

Il problema (umano) del riscaldamento climatico

Siamo «nati» più o meno qui

Nel 2015 eravamo qui



https://www.coursera.org/learn/emergence-of-life

https://www.researchgate.net/figure/Temperature-on-Earth-over-the-last-500-Ma-26-27-adapted-and-licensed-under-CC-BY-SA-40_fig2_363429184

Dunque è o non è un problema?

Dal punto di vista prettamente «antropocentrico» è un problema perché potremmo dover fronteggiare:

Scioglimento dei ghiacciai e innalzamento del livello del mare: Con l'aumento delle temperature, i ghiacciai e le calotte polari si sciolgono più velocemente, causando l'innalzamento del livello del mare e **mettendo a rischio le aree costiere e le isole;**

Eventi meteorologici estremi: Il cambiamento climatico rende più frequenti e intensi eventi meteorologici estremi come uragani, siccità, ondate di calore e piogge torrenziali;

Impatti su salute e sicurezza alimentare: Le temperature elevate, unite all'aumento delle malattie trasmesse da vettori come zanzare e altri insetti, influenzano negativamente la salute umana;

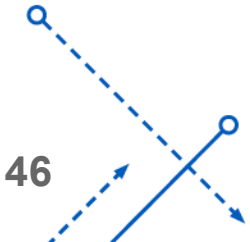
Ricordando Interstellar: La produzione agricola è inoltre messa a rischio da eventi estremi e cambiamenti nei modelli di pioggia, influendo sulla sicurezza alimentare globale.

Dunque è o non è un problema?

Da un punto di vista globale, poiché stiamo sperimentando una graduale perdita di biodiversità, una nuova esplosione in stile «cambriano» non potrebbe far altro che bene alla «vita» in generale.

Perché è importante fare attenzione ai dati?

Vediamo con un esempio reale perché è importante valutare attentamente «i dati» che utilizziamo per le nostre analisi.



Des Moines Sunday Register

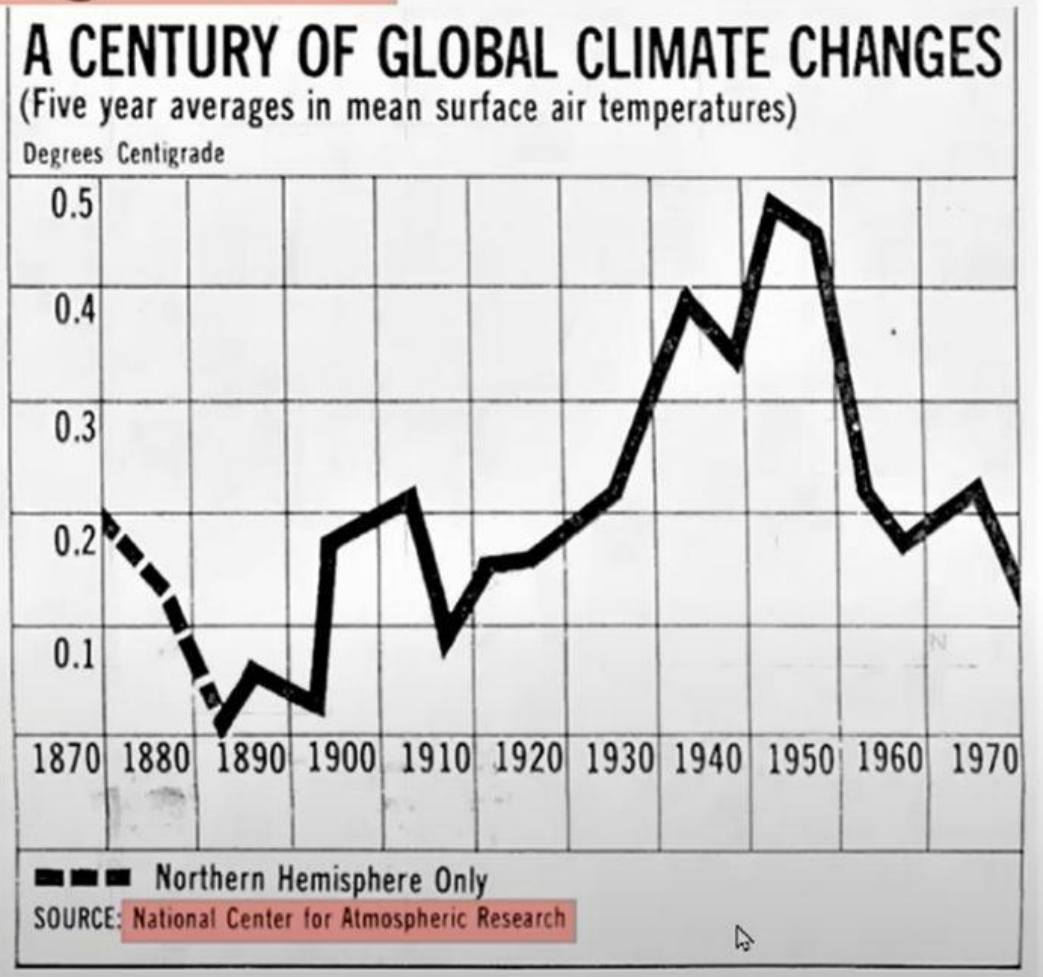
July 21, 1974

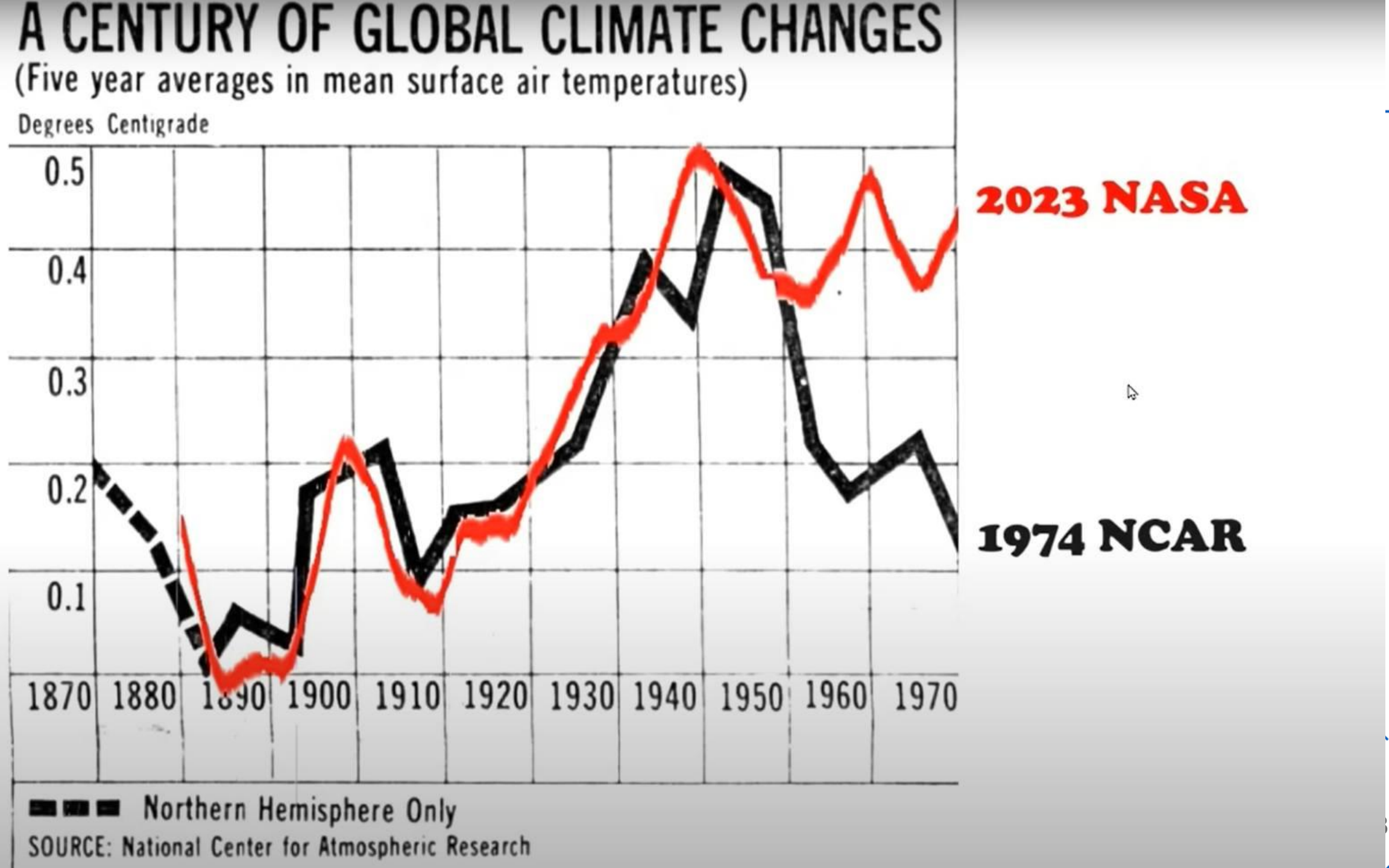
World is getting cooler

Droughts, floods, blizzards, tornadoes, typhoons and hurricanes have plagued much of the nation and the world in recent years. Most people considered these weather conditions to be abnormal and temporary, but instead, climatologists now believe that the first half of the Twentieth Century was blessed with unusually mild weather and that the global climate has begun returning to a harsher — but more normal — state.

For the long run, there is mounting evidence of a worldwide cooling trend. The average temperature of the world as a whole has dropped by one-third to one-half a degree Centigrade in the last 30 years. "The decline of prevailing temperatures since about 1945 appears to be the longest-continued downward trend since temperature records began," says Professor Hubert H. Lamb of the University of East Anglia in Britain.

Global cooling may be a cause of the devastating African drought, now in its sixth year. Some scientists believe that expansion of the cold polar air caps pushed the monsoon rain belt southward, causing many of the life-giving rains to fall on already fertile lands or into the sea. Dry weather conditions also prevail in parts of India, China, Kenya, Bolivia and other countries on both sides of the equator, raising the specter of even more serious drought and famine. Drought has hit the United States regularly about every 20 years, and is due again in the mid-1970s.




A CENTURY OF GLOBAL CLIMATE CHANGES
(Five year averages in mean surface air temperatures)
Degrees Centigrade
0.5
0.4
0.3
0.2
0.1
1870
1880
1890
1900
1910
1920
1930
1940
1950
1960
1970
2023 NASA
1974 NCAR
Northern Hemisphere Only
SOURCE: National Center for Atmospheric Research

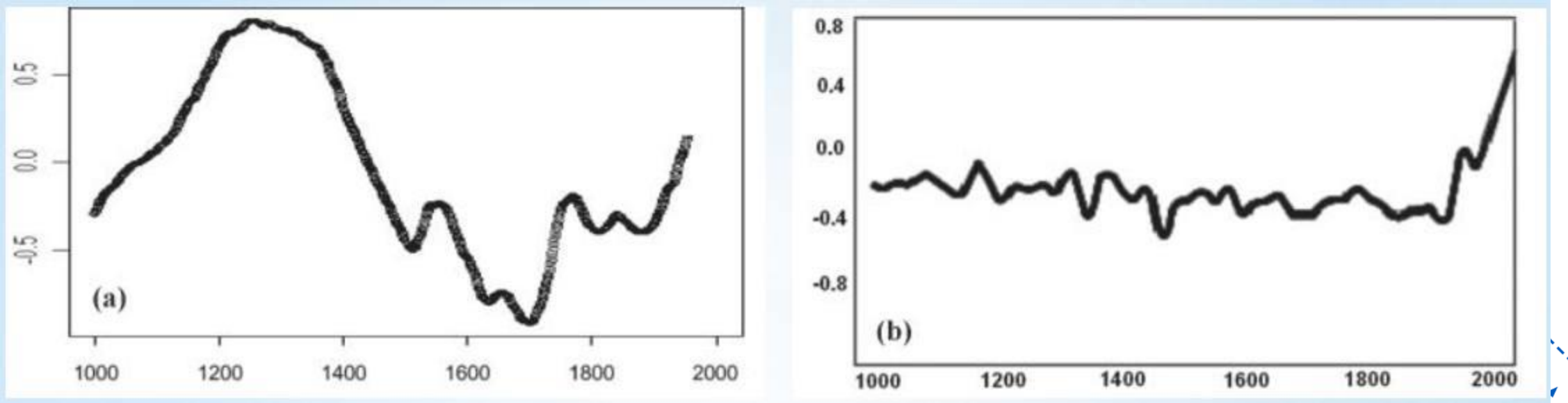
Perché è importante essere critici?

Esiste un gruppo intergovernativo, l'IPCC (**https://www.ipcc.ch/**) che ha proprio il compito di raccogliere i dati e offrirne una lettura.

Tuttavia, guardate a volte cosa accade quando inizia a prendere piede una certa deriva (o moda).

Perché è importante essere critici?

Guardate i due grafici qui sotto. **Vengono entrambi dall'IPCC** e rappresentano l'andamento delle temperature globali secondo il report IPCC del 1990 (sinistra) e del 2001 (destra).



Altre metodologie di curve fitting

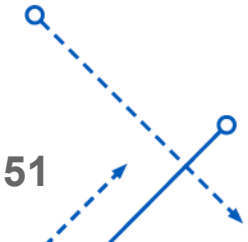
Potrebbe capitare di non riuscire immediatamente a trovare una funzione che interpoli correttamente i punti del campione di osservazioni.

In quel caso si entra nel campo del curve fitting generico (la regressione lineare fa parte del curve fitting, ne è una delle applicazioni così come la regressione polinomiale e la non-lineare).

Un metodo avanzato di curve fitting è detto «spline»:

Una spline è una funzione definita a tratti, composta da segmenti polinomiali (spesso polinomi di basso grado come le parabole o le cubiche) che si uniscono in modo fluido ai punti dati. Il punto principale è che ogni segmento polinomiale copre un piccolo intervallo, e tutti i segmenti sono "incollati" insieme in modo che la curva complessiva sia liscia.

Esistono spline lineari, cubiche e B-spline.



Altre metodologie di curve fitting

Potrebbe capitare di non riuscire immediatamente a trovare una funzione che interpoli correttamente i punti del campione di osservazioni.

In quel caso si entra nel campo del curve fitting generico (la regressione lineare fa parte del curve fitting, ne è una delle applicazioni così come la regressione polinomiale e la non-lineare).

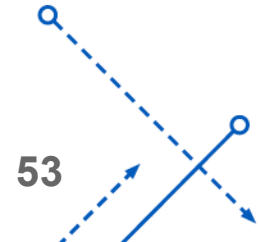
Regression Model Type	Interpretability
Linear Regression Models	Easy
Regression Trees	Easy
Support Vector Machines	Easy for linear SVMs. Hard for other kernels.
Efficiently Trained Linear Regression Models	Easy
Gaussian Process Regression Models	Hard
Kernel Approximation Models	Hard
Ensembles of Trees	Hard
Neural Networks	Hard

Le spline

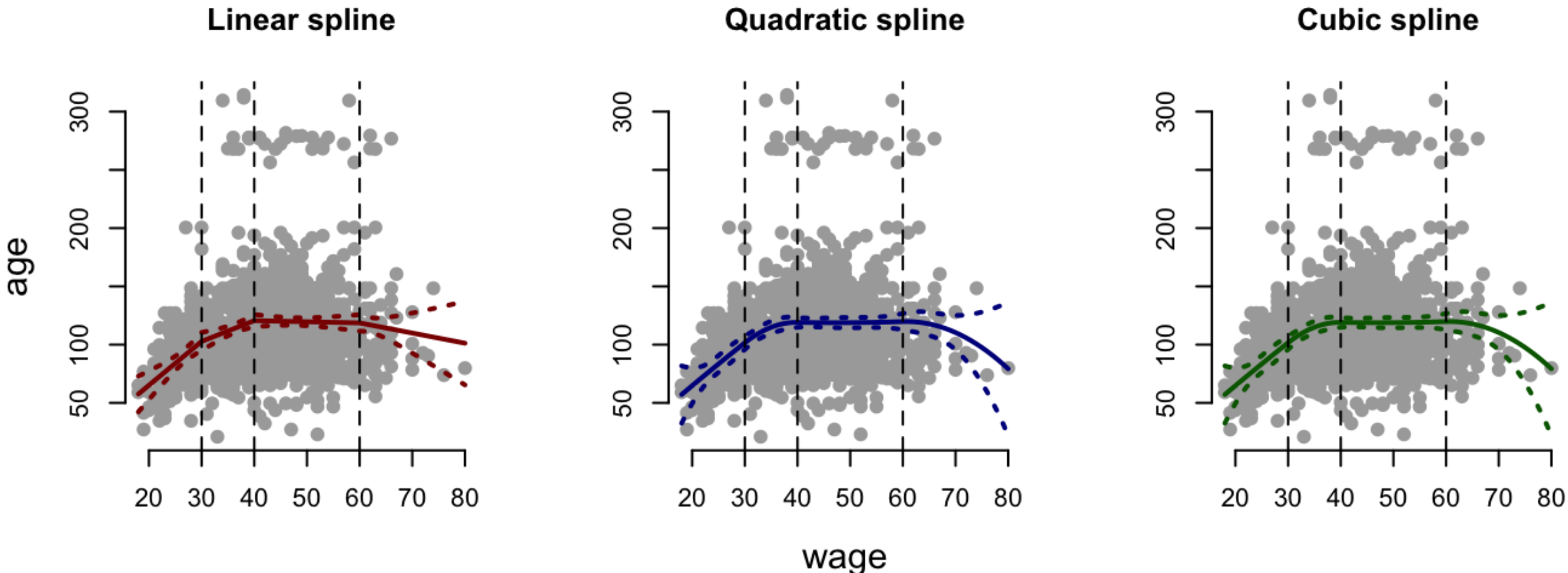
Un metodo avanzato di curve fitting è detto «spline»:

Una spline è una funzione definita a tratti, composta da segmenti polinomiali (spesso polinomi di basso grado come le parabole o le cubiche) che si uniscono in modo fluido ai punti dati. Il punto principale è che ogni segmento polinomiale copre un piccolo intervallo, e tutti i segmenti sono "incollati" insieme in modo che la curva complessiva sia liscia.

Esistono spline lineari, cubiche e B-spline.



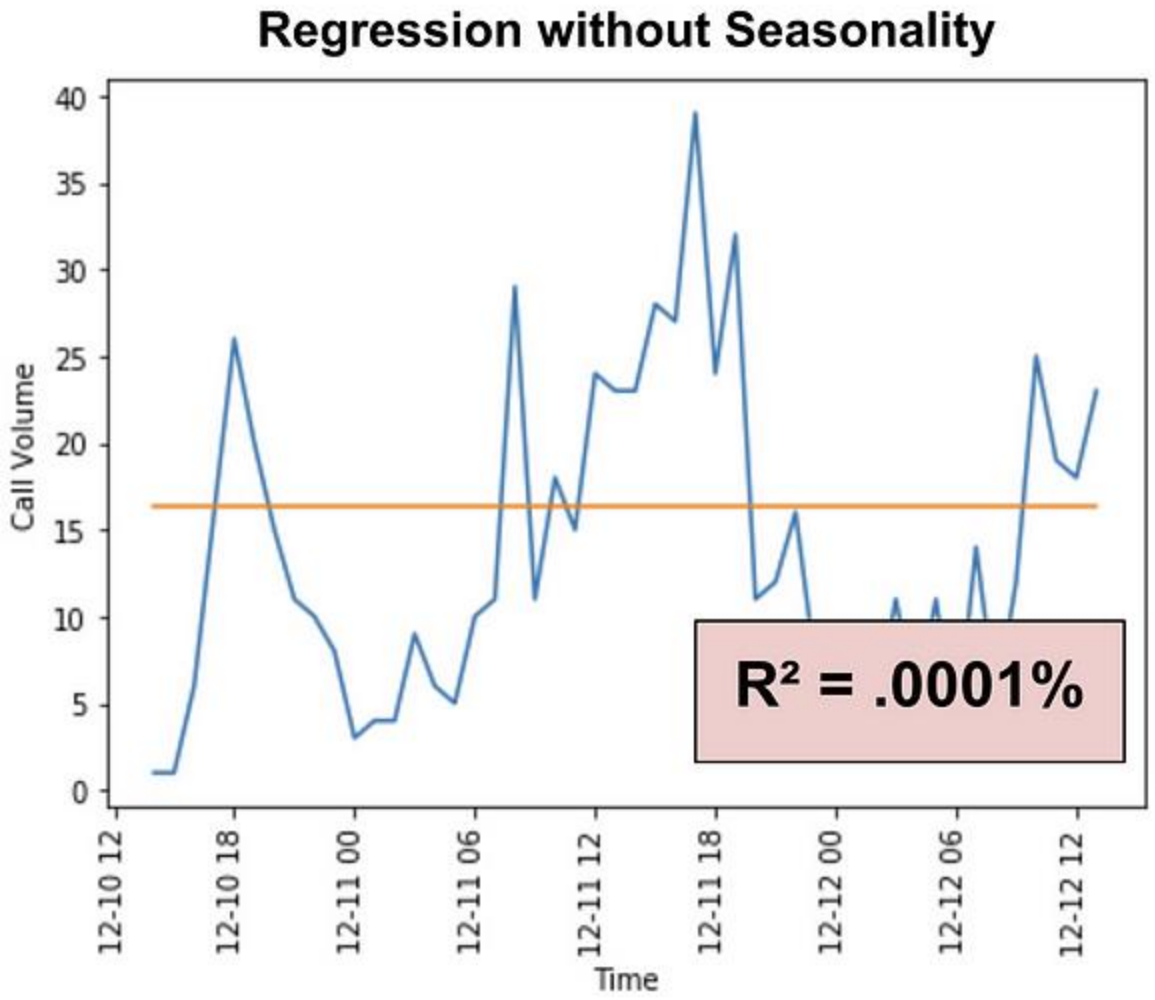
Le spline

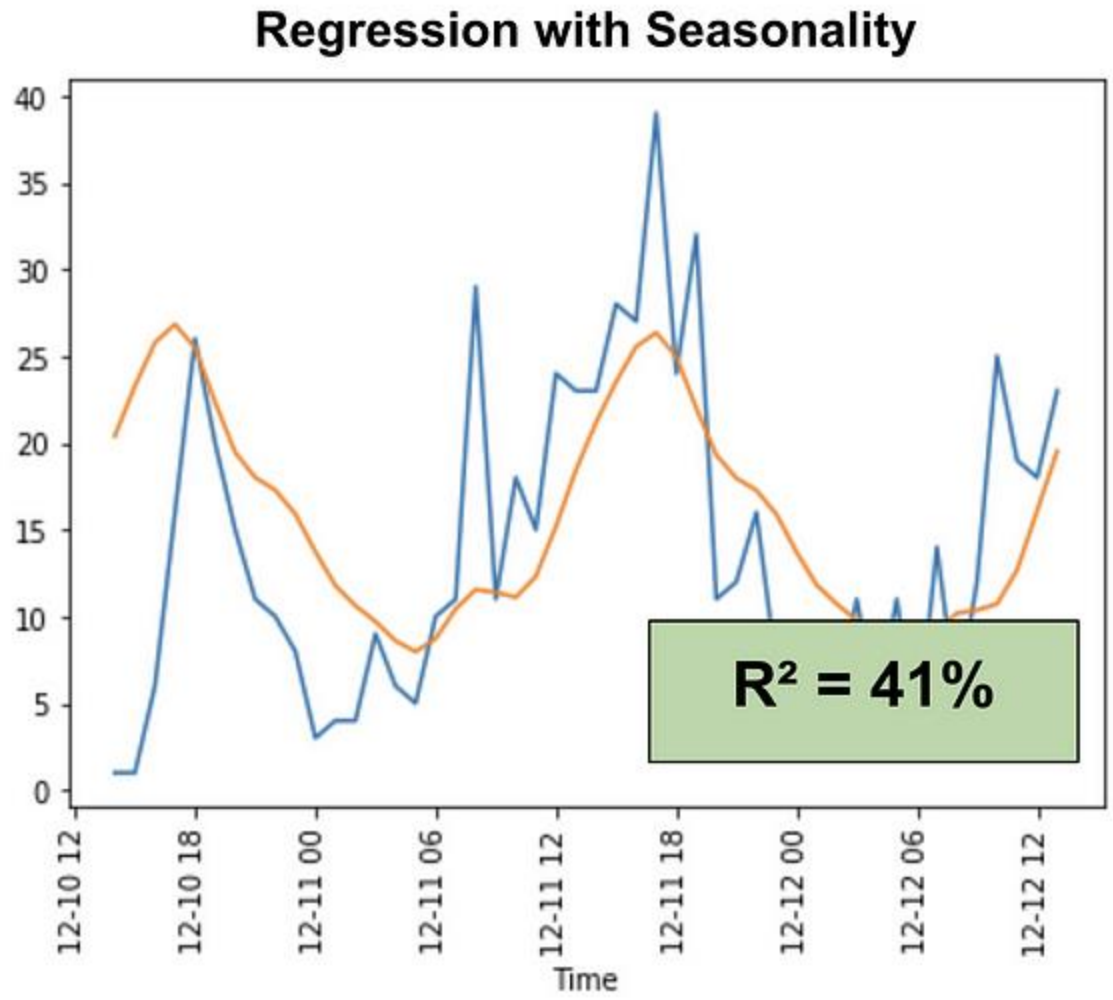

Linear spline
Quadratic spline
Cubic spline
age
wage
50
100
200
300
20
30
40
50
60
70
80

Regressione con basi ortogonali (Fourier e ondelette)

Usa funzioni base come le sinusoidi (serie di Fourier) o le ondelette per adattare i dati, particolarmente efficace quando si lavora con dati periodici o che contengono frequenze diverse.

Statistica e analisi dei dati





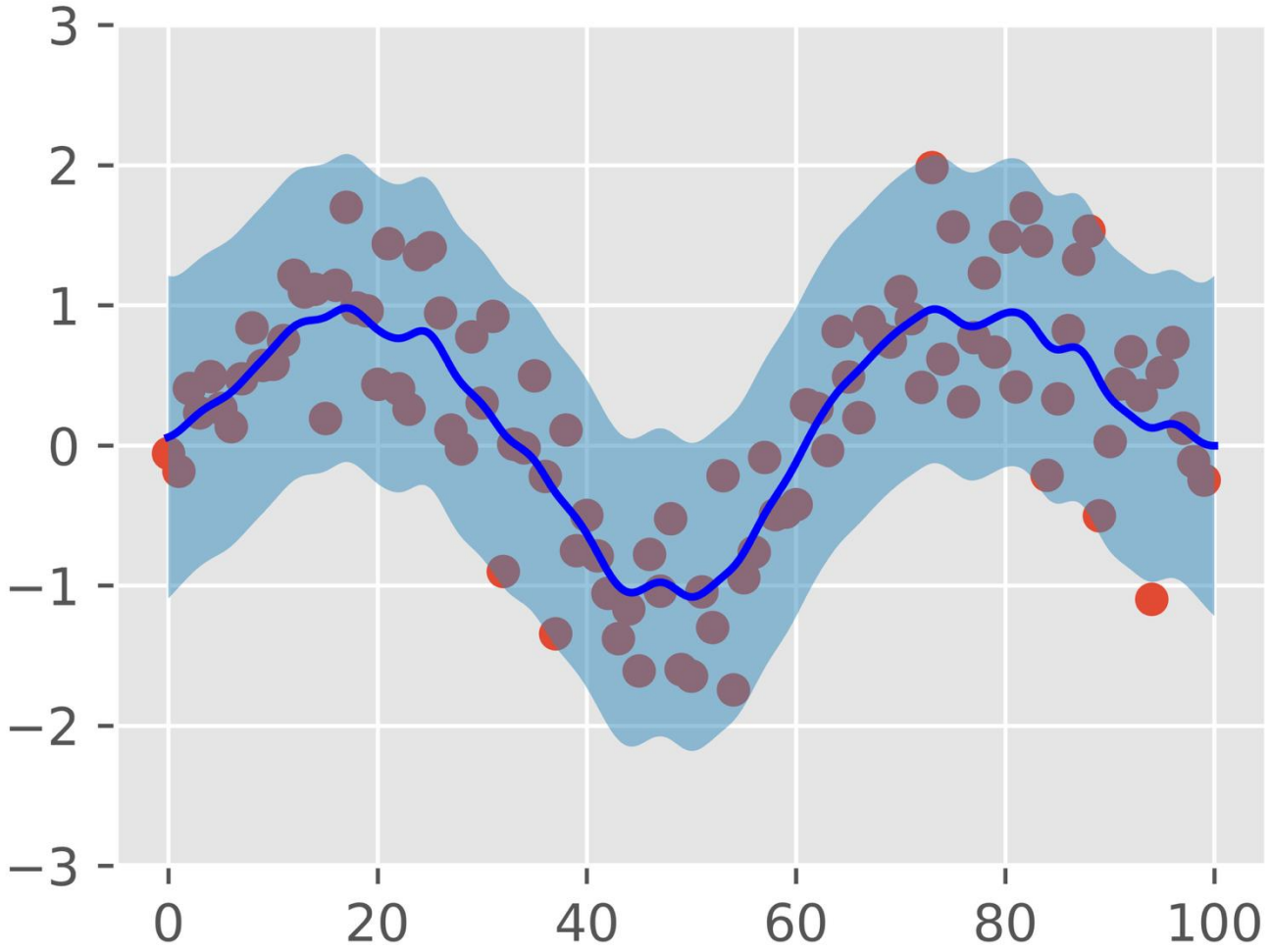
Gaussian Process Regression

La Gaussian Process Regression (GPR), comunemente nota come Kriging, è una tecnica avanzata di regressione e interpolazione utilizzata per modellare e predire dati in situazioni in cui c'è una certa quantità di incertezza.

È molto utile per dati spaziali o temporali, poiché tiene conto delle correlazioni tra punti vicini.

Può adattarsi a diverse forme di dati grazie alla varietà di kernel disponibili e fornisce non solo una previsione ma anche un intervallo di incertezza, rappresentato dalla varianza delle stime.

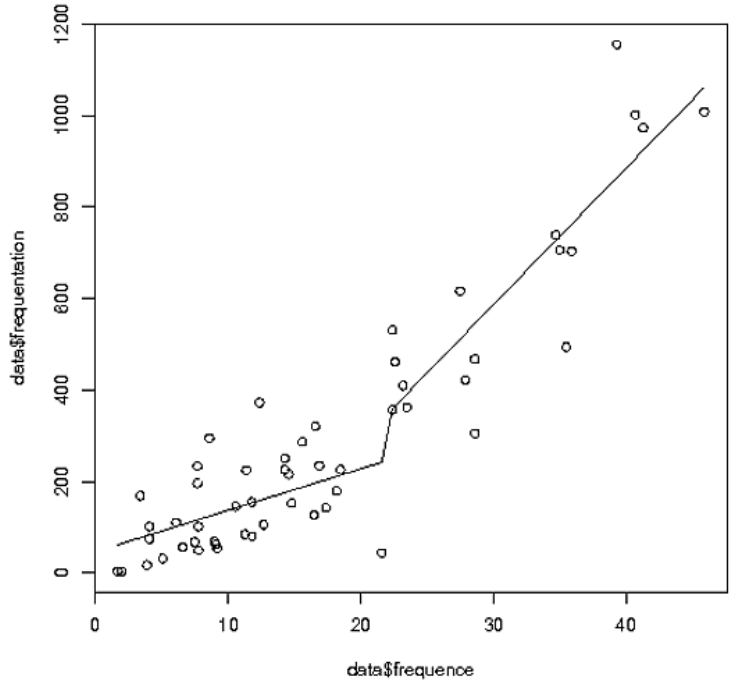
Statistica e analisi dei dati



Metodi a pezzi (piecewise)

Divide i dati in sezioni, applicando una diversa funzione di adattamento a ciascuna. Può utilizzare segmenti lineari, polinomiali o spline.

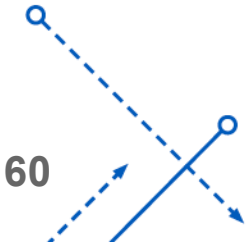
Tenete nella vostra cache mentale questa immagine (per dopo)



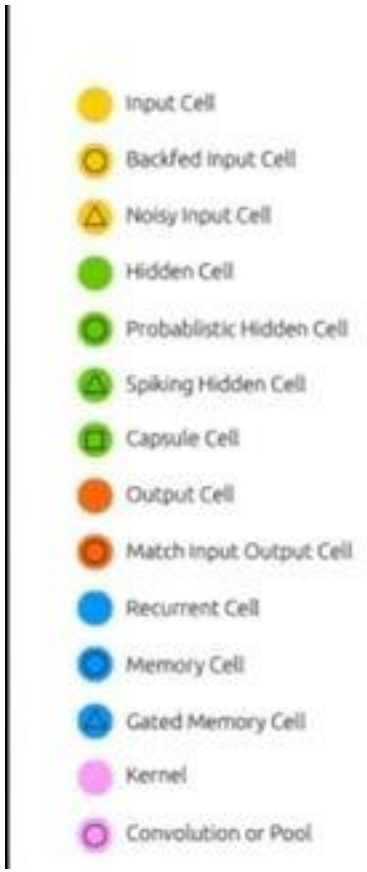
Reti neurali

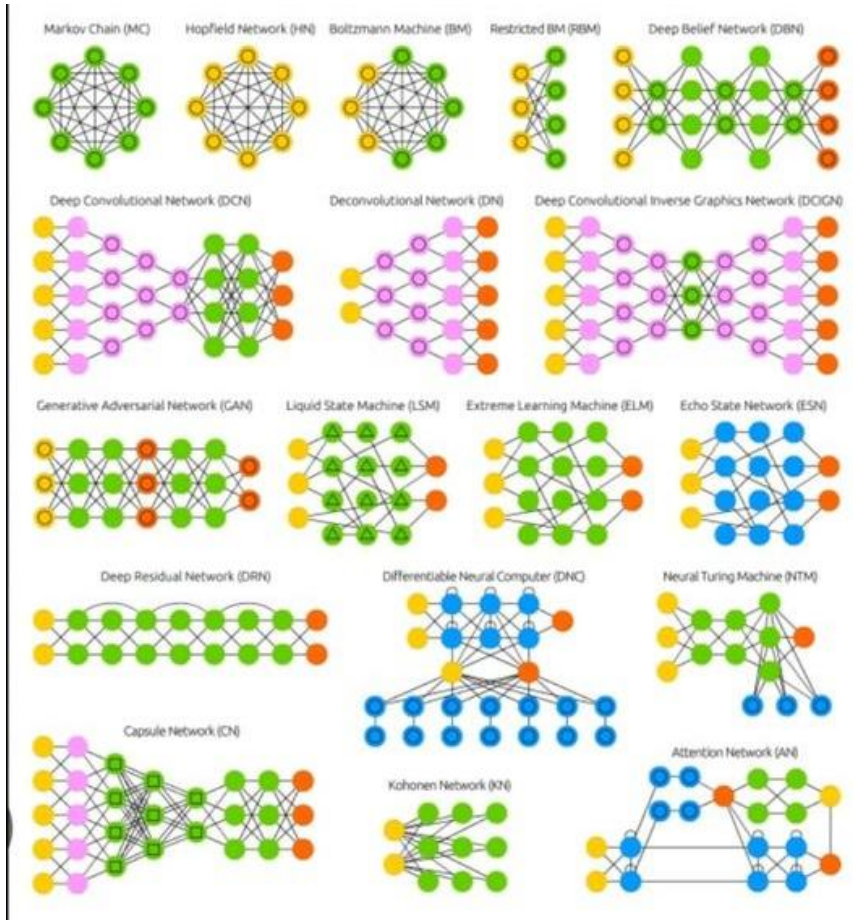
Quando capire le relazioni tra i dati è veramente difficile… lasciamo farlo alle macchine che sicuramente sono più rapide e «oggettive» di noi!

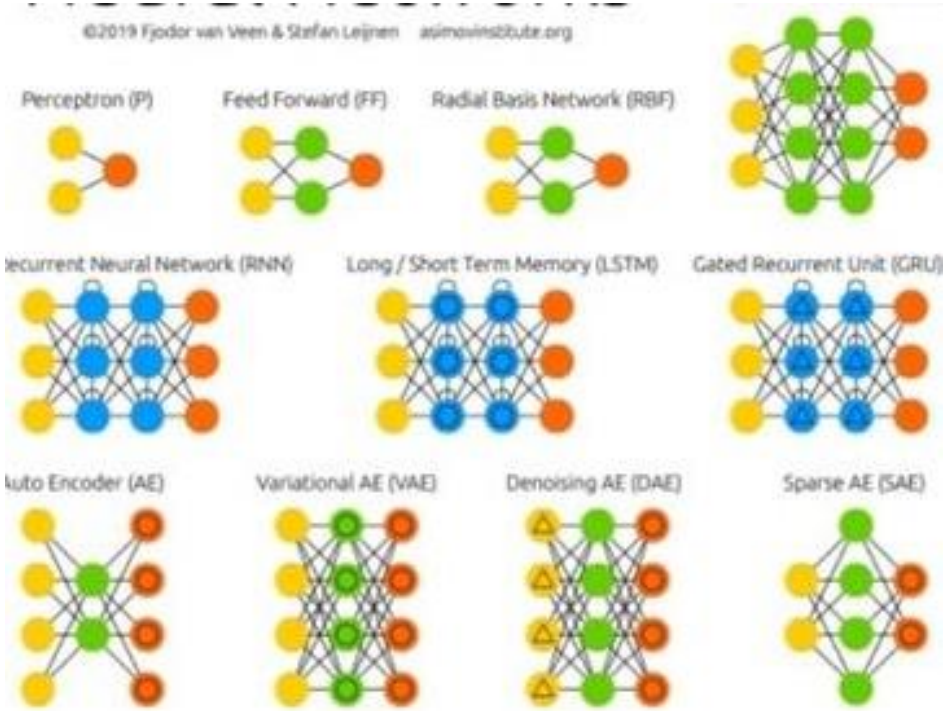
Reti neurali come le CNN (convoluzionali), RNN (residuali) e LSTM (long-short term memory) ci vengono in aiuto quando la complessità della relazione tra variabili è elevata, come riconoscimento di immagini, serie temporali e big data.



Statistica e analisi dei dati



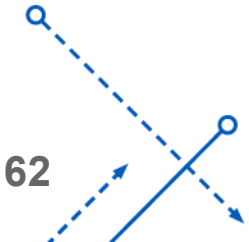




Spiegabilità dei modelli

Al crescere della complessità del modello di regressione usato, inizia a farsi più complesso (non sempre) il problema della «spiegabilità» del modello:

- Il modello come è giunto al risultato?
- Il modello che scelte ha fatto?)
- Con che grado di confidenza posso fidarmi del modello?



Spiegabilità dei modelli

Al momento possediamo solo due strumenti che possono fornirci informazioni per rispondere a queste domande (solo alla terza in realtà).

$$r_{xy} = \frac{C_{xy}}{s_x s_y}.$$

Q**uanto e come** la variabile dipendente è collegata alla variabile indipendente?

$$D^2 = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2}{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2} = \frac{\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}.$$

Se $D^2 = 0.8$ → **80% della varianza della variabile dipendente è spiegata dalla variabile indipendente.**

Spiegabilità dei modelli

Il coefficiente di determinazione per regressioni non lineari `e stato definito in letteratura in tre modi differenti.

$$\text{Definizione 1}: \quad D_1^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2},$$

$$\text{Definizione 2}: \quad D_2^2 = \frac{\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2},$$

$$\text{Definizione 3}: \quad D_3^2 = \frac{\sum_{i=1}^{n}(\widehat{y}_i - \overline{\widehat{y}})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2},$$

Nel caso di regressione lineare semplice e multipla le tre definizioni coincidono.

Nel caso di regressioni non lineari le tre definizioni di coefficiente di determinazione possono condurre a risultati differenti (la 2 e la 3 possono assumere anche valori maggiori di 1).

Spiegabilità dei modelli

Il coefficiente di determinazione per regressioni non lineari `e stato definito in letteratura in tre modi differenti.

LIME (Local Interpretable Model-agnostic Explanations)

Nel caso della regressione, LIME stima quanto ogni variabile di input contribuisca a influenzare il valore numerico predetto. L'interprete può visualizzare queste contribuzioni per ottenere un quadro chiaro su come i cambiamenti nei dati d'ingresso influenzino i risultati.

LIME si concentra su come il modello risponde a piccole variazioni dei dati di input. Il suo scopo è creare spiegazioni locali, cioè spiegare il comportamento del modello attorno a una particolare previsione.

Spiegabilità dei modelli

Il coefficiente di determinazione per regressioni non lineari `e stato definito in letteratura in tre modi differenti.

SHAP (SHapley Additive exPlanations)

SHAP si basa sui concetti della teoria dei giochi, in particolare sul valore di Shapley, che permette di determinare l'importanza di ogni variabile d'ingresso calcolando il "valore marginale" che aggiunge alla previsione complessiva del modello. A differenza di LIME, SHAP fornisce spiegazioni globali e locali con una solida base teorica.

Nel caso della regressione, SHAP fornisce valori che indicano esattamente quanto ciascuna variabile d'ingresso influisce sul valore predetto. Ad esempio, in un modello che stima il prezzo di una casa, SHAP potrebbe mostrare quanto le dimensioni, la posizione e l'anno di costruzione contribuiscano alla previsione specifica del prezzo di una casa.