Selezione di caratteristiche in dati biomedici con metodi di ensemble

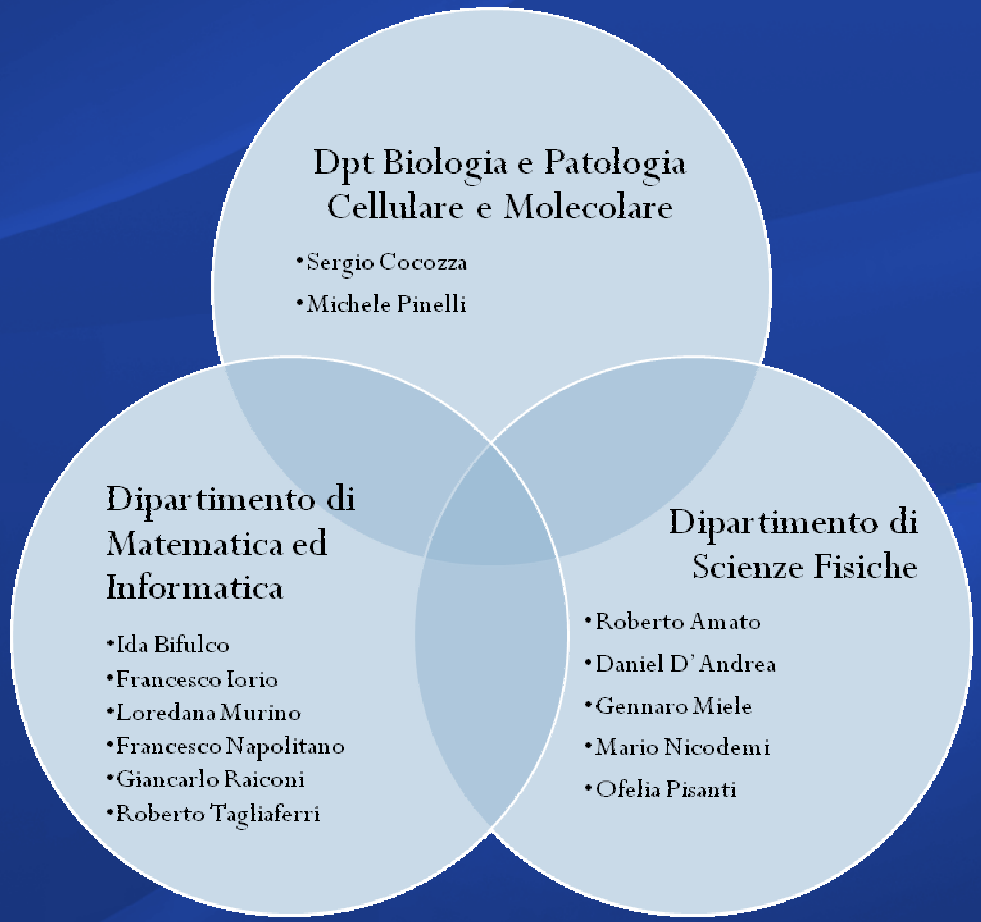
Gennaro Miele

Dipartimento di Scienze Fisiche

Università degli Studi di Napoli "Federico II"

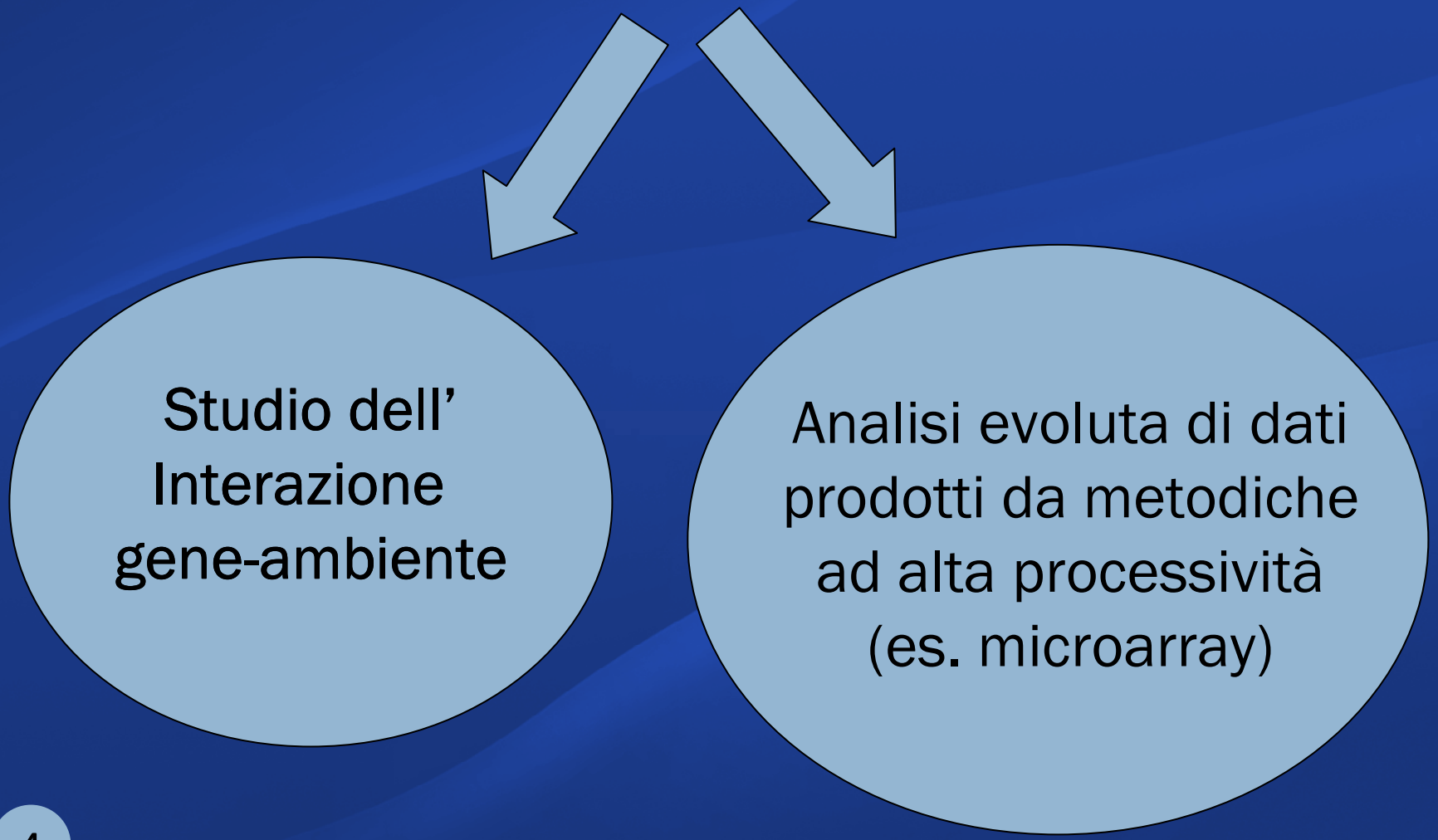
Napoli, 20/II/2009

Gruppo Interdipartimentale di Bioinformatica e Biologia computazionale



Research Doctorate (PhD) in Computational Biology and Bioinformatics at the University of Naples ”Federico II”

Caratteri Complessi (prodotto finale di una serie di interazioni) in biologia e patologia:



Sommario

- **Definizione del problema**
- **Materiali e metodi**
- **Generazione dei dati sintetici**
- **Risultati**
- **Un caso di studio reale**
- **Conclusioni**

Definizione del Problema

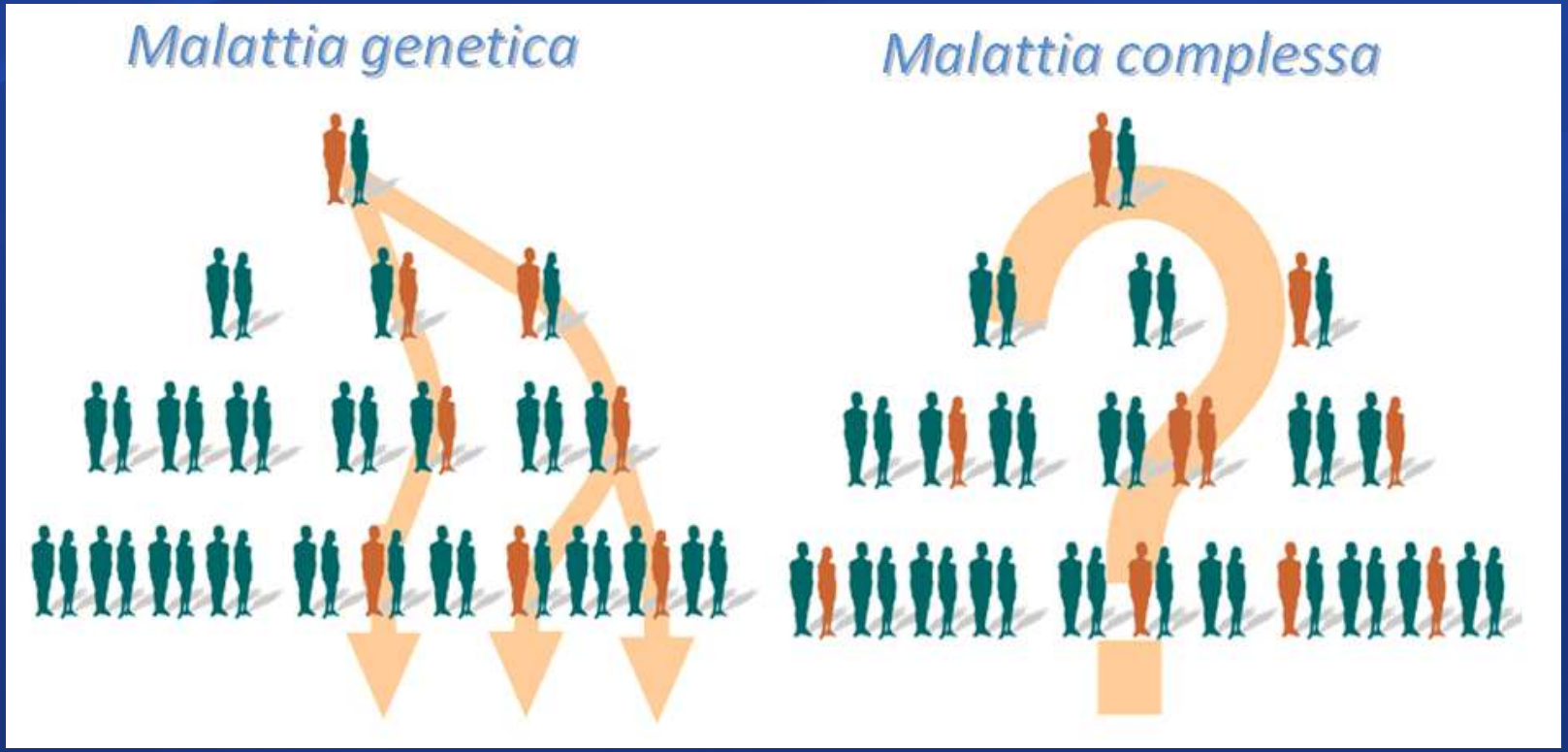
L'insieme delle patologie umane è stato storicamente suddiviso in:

- **Patologie puramente genetiche o Mendeliane (es. Fibrosi Cistica)**
- **Patologie non genetiche - puramente ambientali (es. Traumi)**

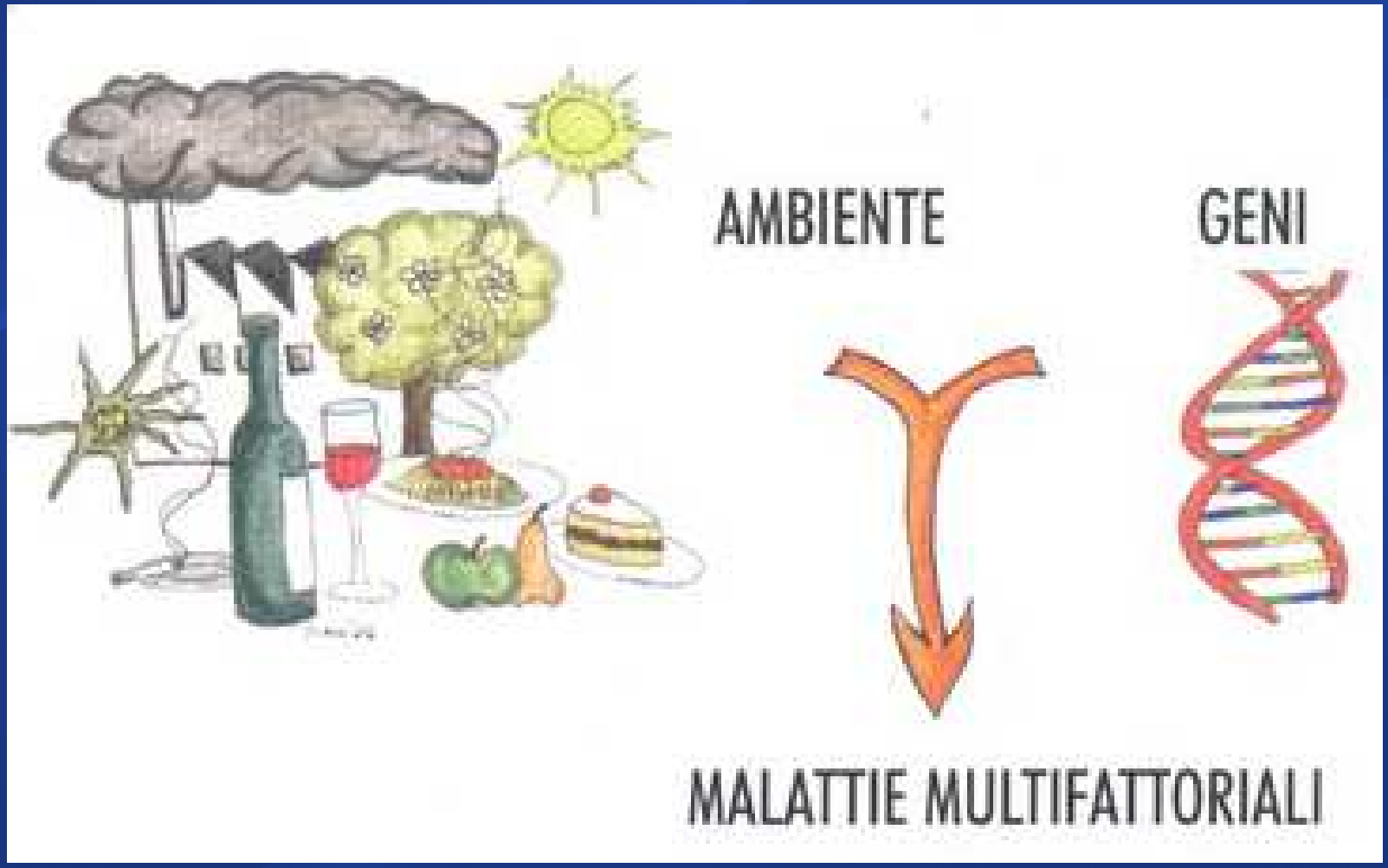
ma oggi è evidente che la maggioranza ricade in una terza categoria ….

Le malattie complesse

- **Costituiscono la maggioranza delle patologie che colpiscono l'uomo (infarto, diabete, ipertensione, obesità,…)**



Interazione gene/ambiente



Dato che l'esposizione ambientale aumenta con l'età ciò implica:

Frequenze delle Malattie Multifattoriali

Autosomiche dominanti (su 1000 nati)	
Iperlipidemia familiare combinata	5.0
Ipercolesterolemia familiare	2.0
Otosclerosi dominante	1.0
Rene policistico dell'adulto	0.8
Esostosi multipla	0.5
Morbo di Huntington	0.5
Neurofibromatosi	0.4
Distrofia miotonica	0.2
Sferocitosi congenita	0.2
Poliposi al colon	0.1

Recessive X-Linked	
Sindrome da Fragile X	0.5
Distrofia muscolare Duchenne	0.3
Ittiosi X-linked	0.2
Emofilia A	0.1
Distrofia muscolare Becker	0.05
Emofila B	0.03

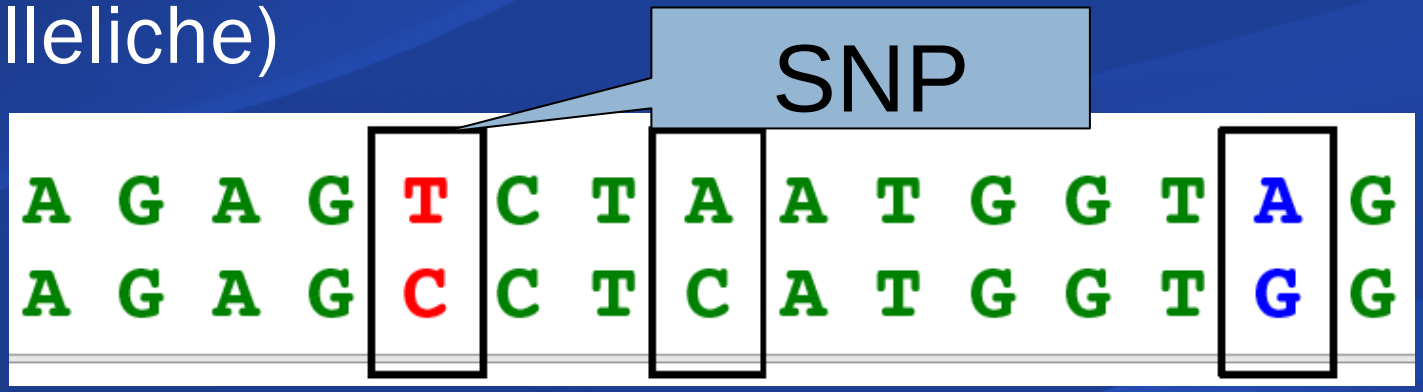
Malattie Multifattoriali	
Labbro Leporino	1.5
Spina Bifida	0.5
Difetti del Tubo Neurale	1
Diabete *(20 – 79 anni)*	52
Diabete T1 (*tutte le età*)	0.9
Asma Br. (UK - Nuova Guinea)	50 - 3
Asma Br. (Italia, adulti)	20 - 50
Asma Br. (Italia, età pediatrica)	70 - 90
Ictus	13
Ictus (Italia, pop anziana)	68
Parkinsonismi	5.3

Autosomiche recessive	
Fibrosi Cistica	0.4
Deficienza di alpha-1-antitripsina	0.2
Fenilchetonuria	0.1
Iperplasia congenita del surrene	0.1
Atrofia Spino-muscolare	0.1
Anemia Falciforme	0.1
beta-Talassemia	0.05

Le malattie complesse

Il fenotipo dipende da:

1) effetto ereditabile (genotipo – variazioni alleliche)



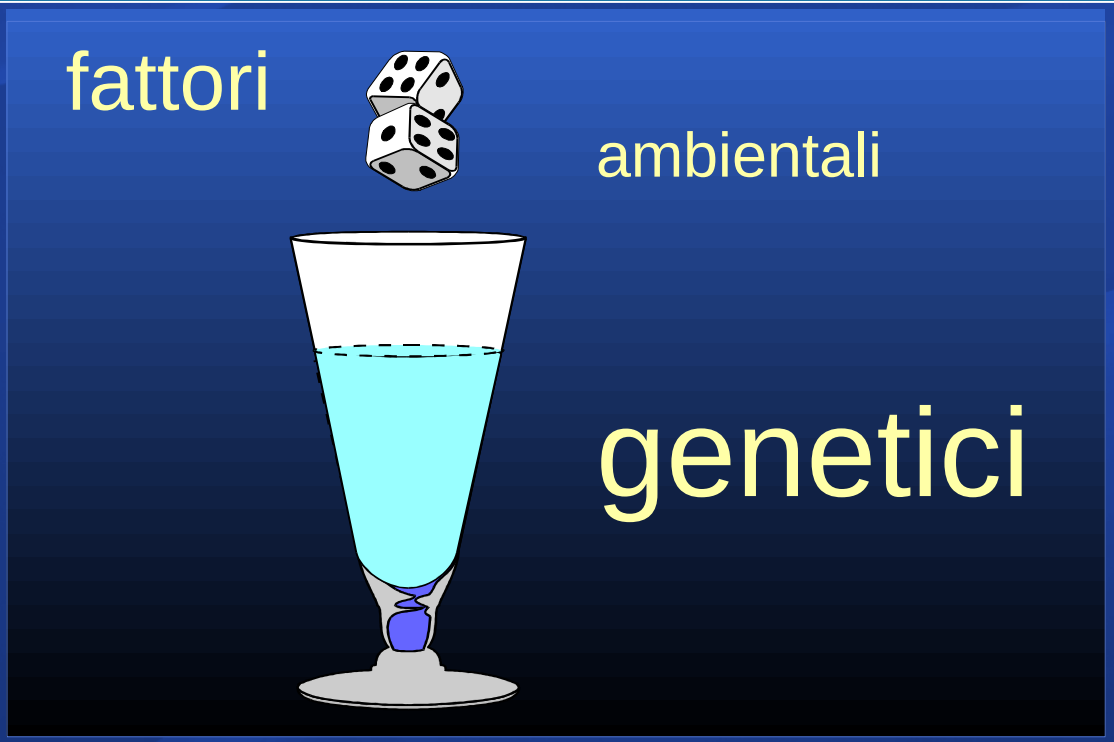
2) effetto ambientale (fattori ambientali)

Si noti che in questo contesto per “fattore ambientale" è inteso tutto ciò che non sia direttamente ed esplicitamente riconducibile alla genetica della persona: esposizione ad inquinanti, farmaci, età, ecc.

3) interazione genotipo-ambiente

Un Modello semplice

La suscettibilità viene normalmente trattata con un modello additivo a soglia

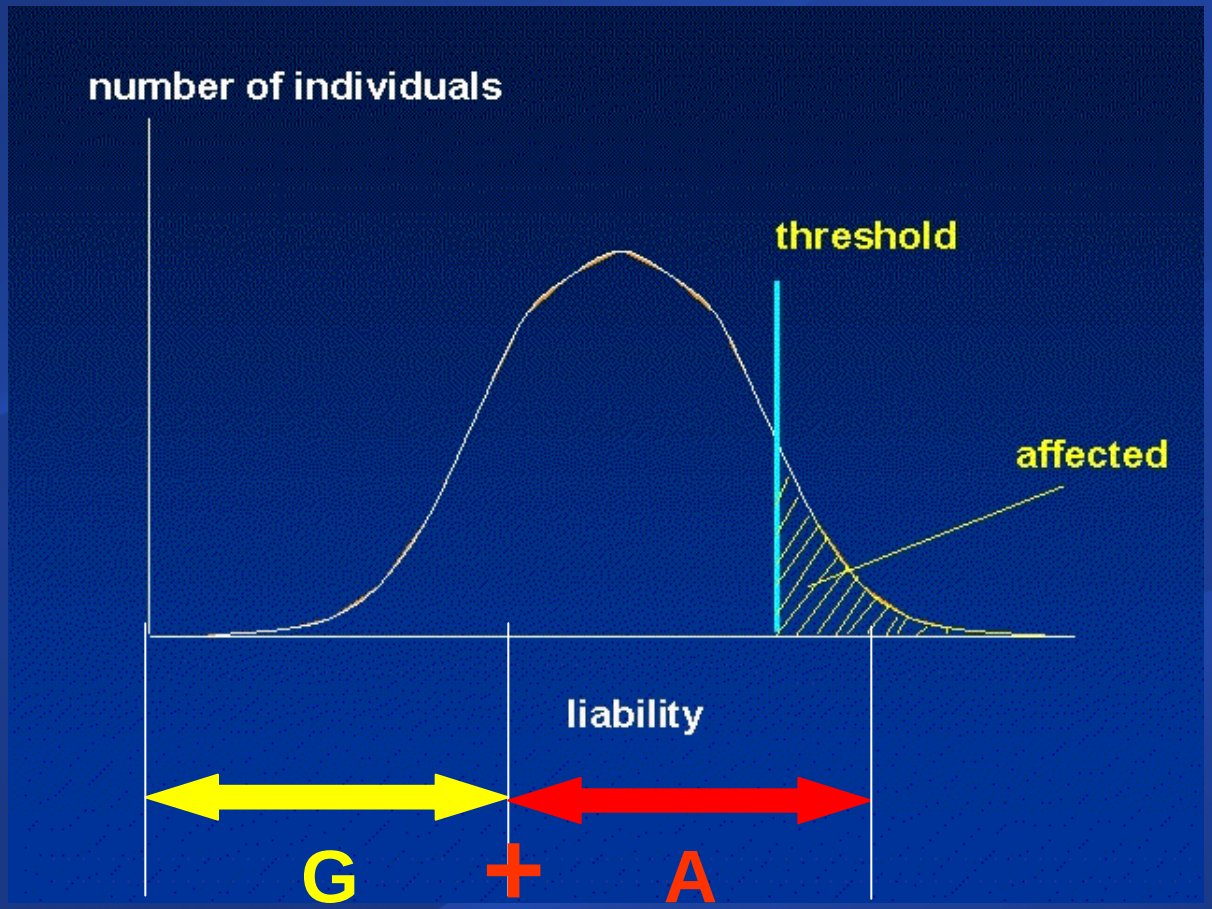


La soglia non viene superata.....: assenza della malattia

La soglia della suscettibilità viene superata… : presenza della malattia

La soglia della suscettibilità viene superata... i fattori ambientali pesano maggiormente: presenza della malattia

Il modello additivo a soglia ben riproduce il fattore di rischio



L'effetto in generale non è puramente additivo!

Non tiene conto di fenomeni di epistasi

L'epistasi

L'epistasi è l'effetto mascherante di un gene su un altro. In generale si è portati a pensare che più geni influiscano additivamente nel determinare la suscettibilità ad una data patologia.

Ma ciò non avviene sempre! Assumiamo che gli alleli abbiano probabilità identiche del 50%

	BB	Bb	bb	*Penetranza*
AA	0.1	0.9	0.1	0.5
Aa	0.9	0.1	0.9	0.5
aa	0.1	0.9	0.1	0.5
Penetranza	0.5	0.5	0.5	

$$P(aff.|g_\alpha) = \sum_\beta P(\tilde{g}_\beta) P(aff.|g_\alpha, \tilde{g}_\beta)$$

- Se si considerano i singoli loci (marginali), non si trova alcuna associazione tra il genotipo e la malattia.
- Nel caso del primo locus, la penetranza nei tre genotipi AA, Aa e aa e sempre uguale a 0.5. La stessa situazione si ha per l'altro locus.
- L'analisi dei genotipi derivati dalle combinazioni rivela: alcuni genotipi come AA/BB sono a basso rischio (penetranza pari a 0.1) mentre altri come AA/Bb ad alto rischio (penetranza 0.9). Ciò rende necessaria la caratterizzazione contemporanea dei due loci per ogni individuo.
- Il ruolo di un fattore ambientale si manifesterebbe solamente in alcuni genotipi sensibili.

Sono comunque stati sviluppati modelli di suscettibliità molto più complessi:

Framingham Hearth Study

(circa 2000 papers pubblicati)

Uno studio del rischio di insorgenza di malattie Cardiovascolari (CVD), iniziato negli anni ’50, su tre generazioni di circa 5000 persone tra i 30 ed i 60 anni di età di Framingham (Massachusetts) seguendone stile di vita e storia clinica.

Framingham equations for 10 year risk of event

Equation 1: $\mu = \sum_{10}^{i=0} \beta_i x_i$ where β_i and x_i are defined in table 1.

Different values of β_i are used for CHD, stroke, and CVD

Equation 2: $\sigma = \theta_0 + \mu\theta_1$ where θ_0 and θ_1 are defined in table 2 and μ is defined from equation 1

Equation 3: $u = \frac{ln(10) - \mu}{\sigma}$ μ is defined in equation 1 and σ is defined in equation 2; for other time periods ln(10) can be replaced with the number of years

Equation 4: 10 year risk of event = $1 - e^{-e^u}$ where u is defined in equation 3

i	x_i
1	Female=1, male=0
2	Ln age (years)
3	(Ln age (years))^2
4	Ln (age) if female, 0 if male
5	(Ln age)^2 if female, 0 if male
6	Ln systolic blood pressure
7	Current smoker=1, 0 otherwise
8	Ln (total cholesterol)/(HDL)
9	Diabetic=1, non-diabetic=0
10	Female and diabetic=1, 0 otherwise

Cappuccio et al. (2002) BMJ, 325, 1-6

Riassumendo: nelle malattie complesse..

- la relazione tra genotipo e patologia non è semplicemente causale ma di tipo "probabilistico"
- i geni rappresentano dei "fattori di rischio". Aumentata familiarità ma trasmissione non mendeliana

	M. mendeliane	M.Complesse
Ereditarietà	Monogenica	Multigenica
Frequenza	rare	comuni
Variazione	bimodale(si/no)	continua
Fattore genetico	causale	predisponente
Influenza dell'ambiente	no	sì

Che vantaggi possiamo trarre da una comprensione migliore dell'interazione gene-ambiente?

- Fornire stime di rischio personalizzate e di conseguenza più precise
- Offrire cure e trattamenti personalizzati
  - Riduzione degli effetti collaterali
- Migliore comprensione dei processi biologici sottostanti

Un Esempio

BMC Medical Genetics

Research article

β2-adrenergic receptor and UCP3 variants modulate the relationship between age and type 2 diabetes mellitus

Michele Pinelli*†[1], Manuela Giacchetti†[1], Fabio Acquaviva[1], Sergio Cocozza[1,3], Giovanna Donnarumma[2], Emanuela Lapice[2], Gabriele Riccardi[2], Geremia Romano[2], Olga Vaccaro[2] and Antonella Monticelli[1,3]

Le difficoltà che si incontrano nell'identificare i fattori di rischio

- Variano nella gravita dei sintomi e nei tempi in cui si manifestano;
- Possono variare nei loro meccanismi eziologici (cause che portano alla malattia)
- Sono solitamente causate da più, ed a volte numerosi, geni, ognuno dei quali contribuisce in piccola misura al manifestarsi della malattia
- Difficoltà tecniche legate alla natura dei dati

Fattori di rischio genetici:
Studi di associazione a gene-candidato

- Si cerca di identicare una correlazione statistica tra specifiche varianti genetiche e la malattia:

  si candida un gene la cui alterazione può essere responsabile della malattia in base alla sua funzione e si procede nel ricercare variazioni del gene candidato che siano presenti nella popolazione degli affetti (casi) e assenti nella popolazione dei non affetti (controlli).
- Se fosse presente una variante (allelica) e questa variante fosse maggiormente presente tra gli affetti in maniera statisticamente signicativa, si potrebbe suggerire un suo ruolo nella suscettibilita alla malattia.

La “maledizione della dimensionalita"

- La necessità di considerare ogni possibile combinazione di fattori porta ad un numero di combinazioni che aumenta esponenzialmente con il numero di fattori considerati.
- Il gran numero di variabili da esaminare combinato al numero generalmente molto piccolo di osservazioni, porta al problema conosciuto in statistica come curse of dimensionality.
- Tale situazione porta, nelle metodiche classiche come la regressione, a valori elevati dell'errore standard ed aumenta la probabilita di commettere errori di tipo I (rigetto dell'ipotesi nulla anche se vera)

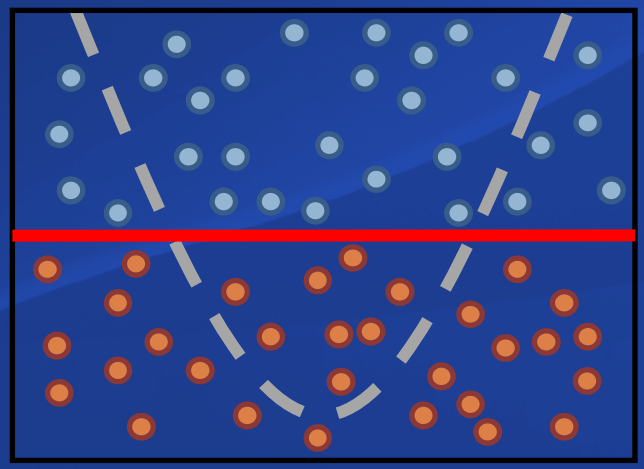
Il problema dal punto di vista informatico e statistico

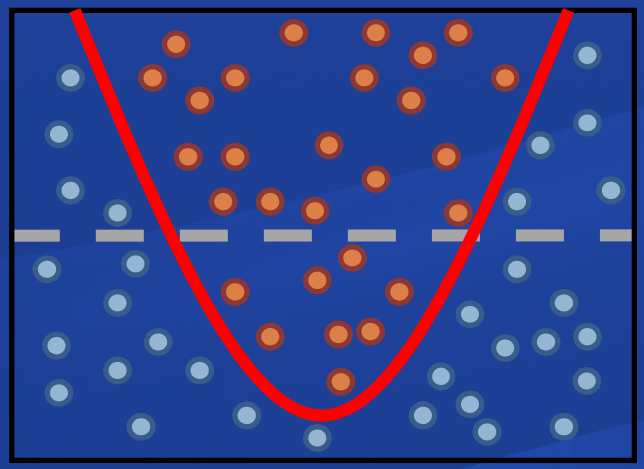
- Mancanza di metodi specifici
- Può essere visto come un problema di selezione delle caratteristiche (*feature selection*)

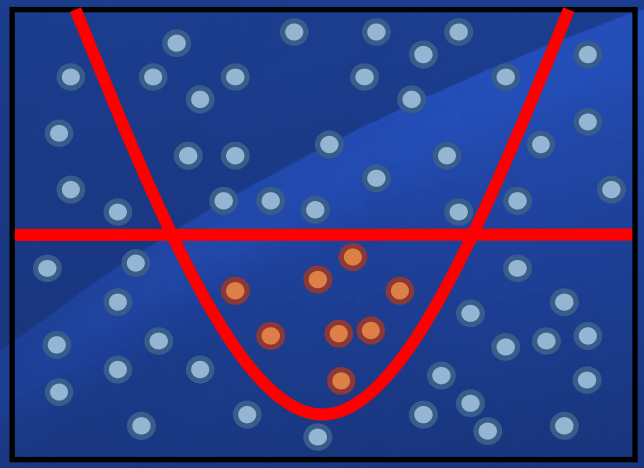
Selezione di Caratteristiche (Feature selection)

- E' la tecnica, comunemente usata in machine learning, di selezionare un sottoinsieme di caratteristiche rilevanti (correlate allo status dei singoli dati) tra tutte quelle disponibili nel data set
- Rimuovendo dai dati le caratteristiche più irrilevanti o ridondanti essa migliora le prestazioni del modello di apprendimento in quanto:
  - Migliora l'interpretabilità modello
  - Allevia l'effetto del *curse of dimensionality*
  - Migliora la capacità di *generalizzazione* del modello
- La selezione di caratteristiche è un *key issue* per le analisi biomediche
  - Migliora le prestazioni di classificazione
  - Può contribuire a chiarire il background biologico, per esempio identificando i fattori maggiormente correlati ad un fenotipo

Un esempio pittorico







- Da questo punto di vista, esistono molti metodi di feature selection, ma spesso manca una loro convalida ed un confronto sistematico in ambito biomedico
  - Solo in pochissimi studi sono state misurate contemporaneamente sia le variabili ambientali che quelle genetiche degli individui
  - Anche quando le informazioni sono raccolte, non si conosce il vero fenomeno sottostante

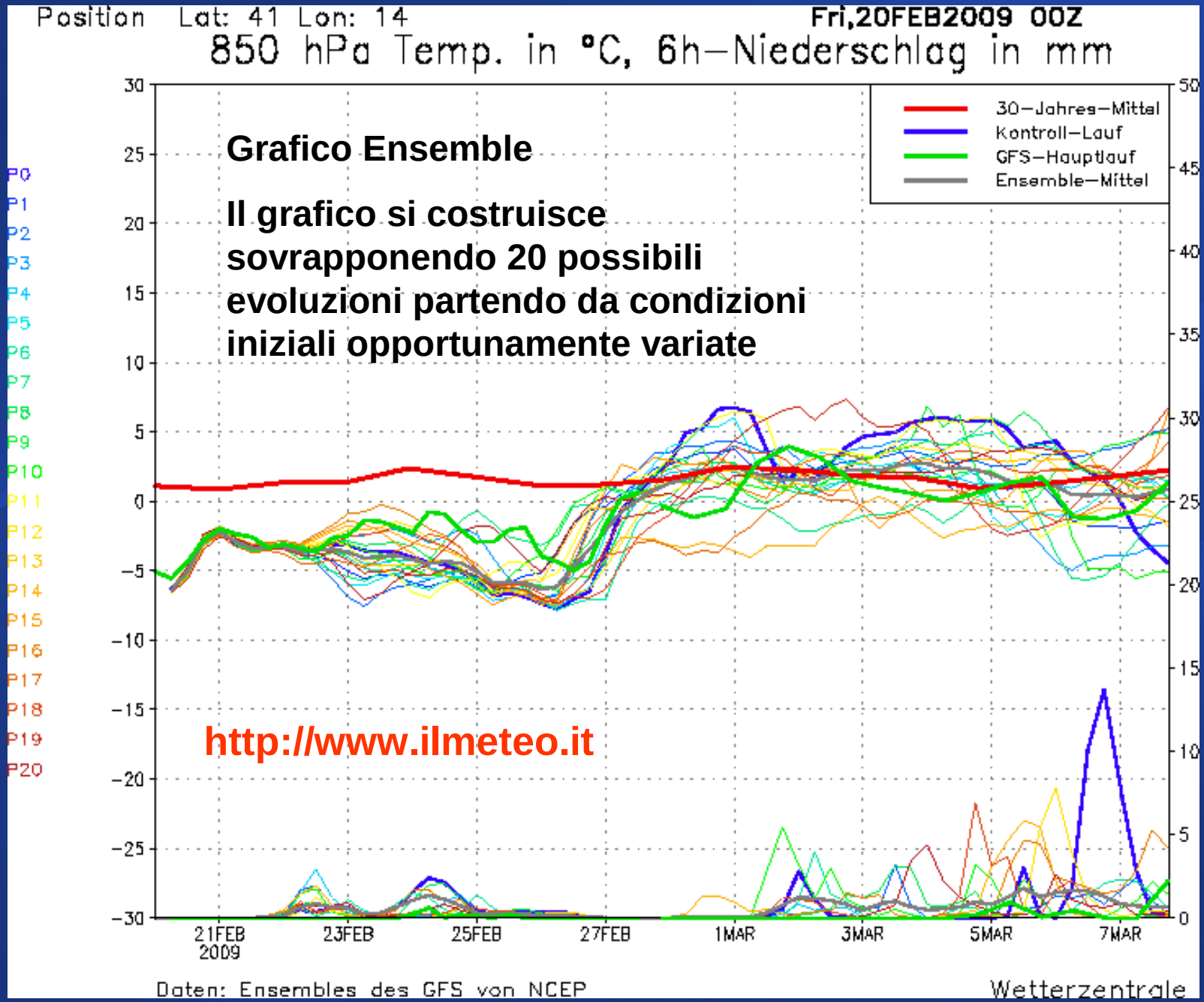
Obiettivi

- Ogni metodo di analisi ha i propri pro e contro e, quindi, funziona meglio in determinate situazioni piuttosto che in altre
  - **Una soluzione può essere utilizzare contemporaneamente diversi metodi allo scopo di rafforzare i pregi di ognuno e mitigarne i difetti**
- Per fare ciò, però, è necessario confrontare e valutare i vari metodi su dati per i quali sia noto il fenomeno sottostante
  - **Mancando in letteratura dati adatti allo scopo, si può ricorrere a dati sintetici ma che riproducano con fedeltà il problema studiato e, soprattutto, siano biologicamente plausibili ed interpretabili**

Materiali e Metodi

Ensemble

- Tradizionalmente usati per problemi di classificazione, gli "ensemble" mettono assieme diversi metodi per produrre una risposta globale più accurata
  - Sotto alcune ipotesi, un ensemble è migliore di ciascun metodo che lo compone
- E' possibile combinare le risposte di
  - differenti metodi di analisi
  - lo stesso metodo di analisi ma con diversi insiemi di addestramento
  - lo stesso metodo di analisi ma variandone i parametri di addestramento
- utilizzando un
  - voto di maggioranza
  - voto pesato (es: rispetto alla significatività della risposta)


Position Lat: 41 Lon: 14
Fri,20FEB2009 00Z
850 hPa Temp. in °C, 6h-Niederschlag in mm
30-Jahres-Mittel
Kontroll-Lauf
GFS-Hauptlauf
Ensemble-Mittel
P0
P1
P2
P3
P4
P5
P6
P7
P8
P9
P10
P11
P12
P13
P14
P15
P16
P17
P18
P19
P20
Grafico Ensemble
Il grafico si costruisce sovrapponendo 20 possibili evoluzioni partendo da condizioni iniziali opportunamente variate
http://www.ilmeteo.it
21FEB 2009
23FEB
25FEB
27FEB
1MAR
3MAR
5MAR
7MAR
Daten: Ensembles des GFS von NCEP
Wetterzentrale

Estensione a problemi di feature selection

- *Feature-by-feature* è possibile calcolare, a partire dalle risposte dei vari metodi di base, la probabilità che la variabile sia correlata allo status
- In un contesto bayesiano, essa è pari a

$$P(F_i \mid T_1,\ldots,T_k) = \prod_{j=1}^{k} \frac{P(T_j \mid F_i)P(F_i)}{P(T_j)}$$

dove con F è inteso l'evento "la variabile è correlata" e con T_i l'evento "la variabile è presente nella risposta del test i-esimo"

$$P(T \mid F) \equiv \textbf{sensibilità=TP/(Positivi)}$$

$$P(\neg T \mid \neg F) \equiv \textbf{specificità=TN/(Negativi)}$$

ovvero

la risposta di ciascun componente dell'ensemble è pesata rispetto al comportamento tipico atteso dal metodo

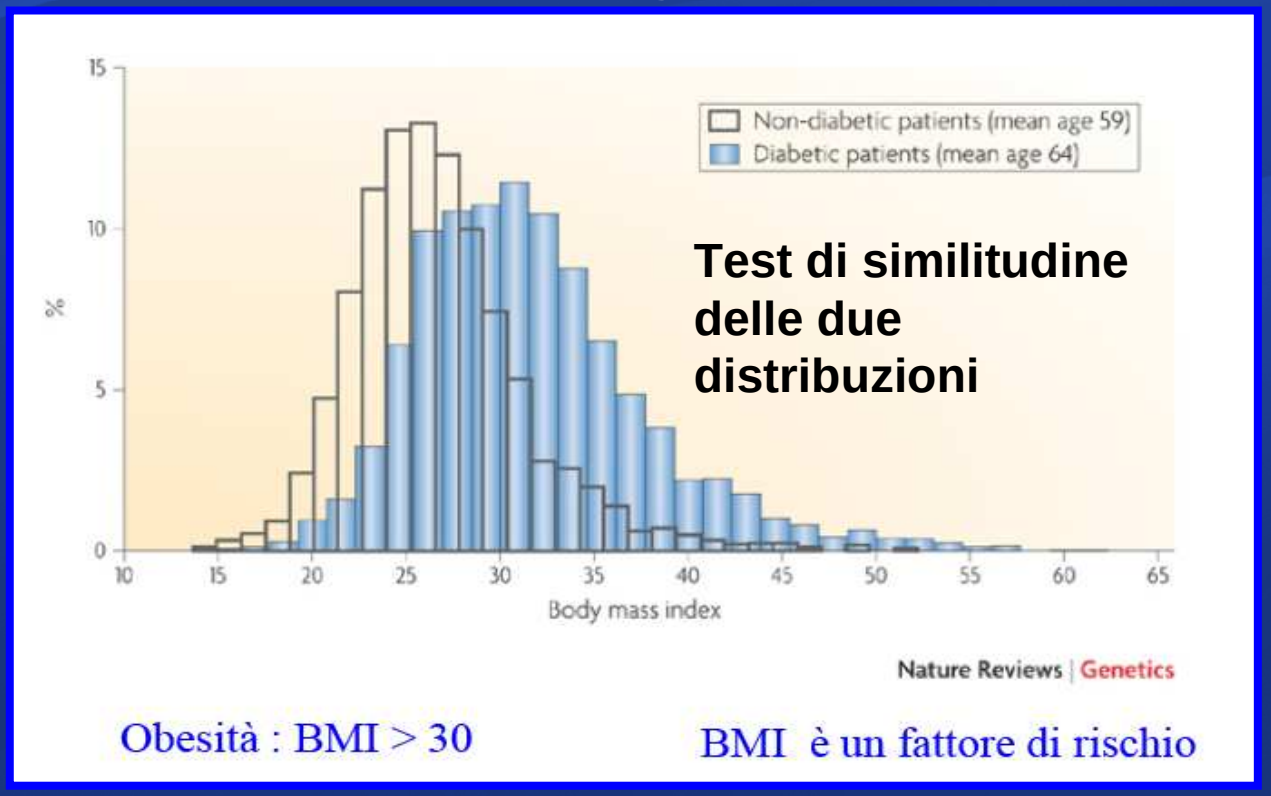
Metodi di base

Principale suddivisione: metodi “univariati”, che indagano l'influenza di singole variabili sullo status, e metodi “multivariati" che ricercano interazioni tra piu variabili.

- Criteri di scelta:
  - Diffusione in ambito biomedico
  - Semplicità
  - “Completezza”
- Scelta dei parametri:
  - Ogni metodo calcola la significatività della risposta utilizzando indici differenti (p-value, cross-validation,…)
  - Per rendere comparabile il livello di confidenza delle risposte, è stato utilizzato un filtro differente per ogni metodo che permettesse di ottenere una specificità non inferiore al 90%

Metodi univariati (filtri)

- Si esegue un test (χ^2, t-test) separatamente su ciascuna variabile per misurarne la correlazione con lo status: l'ipotesi di associazione contro la cosiddetta l' ipotesi nulla" ovvero di assenza di associazioni reali tra gli eventi.



- La significatività del risultato è stimata secondo la correzione di Bonferroni per test multipli (si modifica la soglia di significatività in relazione al numero di ipotesi valutate affinché l'intera batteria di test, considerata nel complesso, abbia la significatività prescelta).

- Non è computazionalmente immaginabile applicare il test monovariato a tutte le 2^N combinazioni di N variabili in principo correlate allo status.

Metodi multivariati

Regressione Logistica Binaria

- Il metodo piu comunemente utilizzato dagli epidemiologi per modellare le relazioni tra un gruppo di “predittori" ed una variabile discreta, spesso binaria come negli studi caso-controllo, é la regressione logistica (la Fermi/Dirac per i fisici)
- Si costruisce la funzione logistica che meglio approssima i rapporti sani/malati al variare delle covariate

$$P(Y=1|\mathbf{x}) = \frac{1}{1+\exp\{\alpha+\sum_i \beta_i x_i\}}$$

$$P(Y=0|\mathbf{x}) = 1-P(Y=1|\mathbf{x}) = \frac{\exp\{\alpha+\sum_i \beta_i x_i\}}{1+\exp\{\alpha+\sum_i \beta_i x_i\}}$$

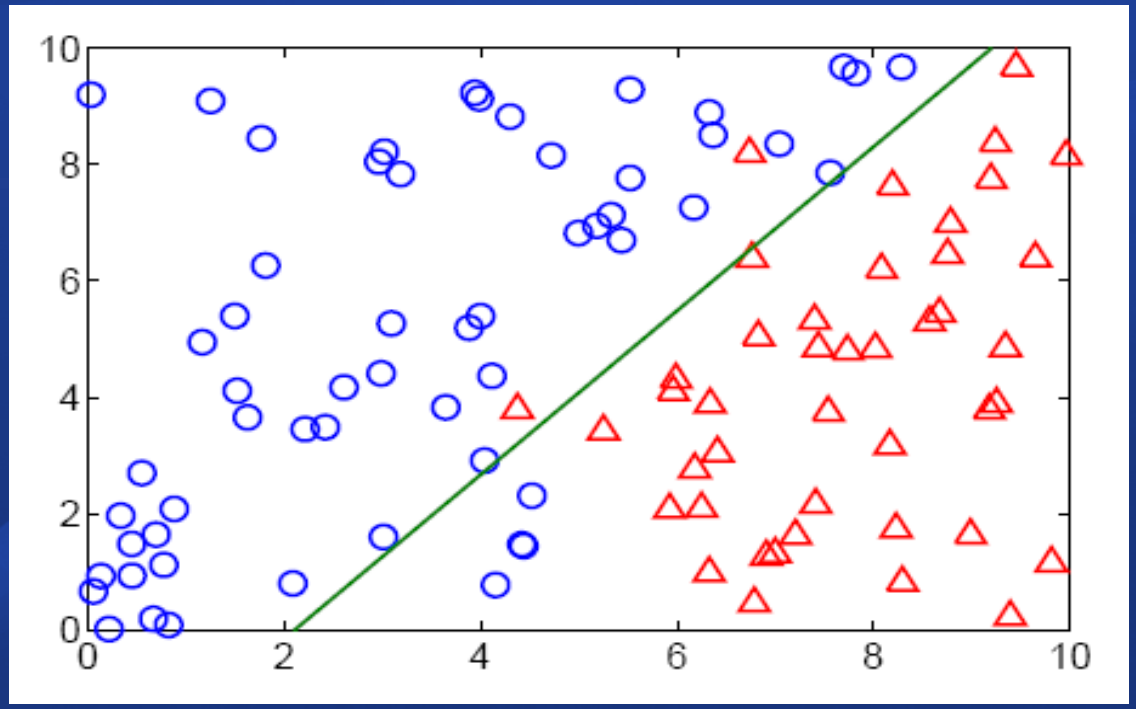
- Partendo dal modello che include tutte le variabili, ad ogni passo si elimina la variabile meno significativa (coefficiente più piccolo) fino a che la significatività resta sopra una soglia fissata "backward regression". In questo modo si ottengono insiemi di caratteristiche che possono essere anche molto più grandi di quello *ottimale*.
- Nell'approccio con "forward regression" le interazioni sono testate solo per quelle variabili che hanno un effetto statisticamente significativo indipendente dalle altre: le variabili che hanno un effetto di interazione ma non un effetto importante (significativo) da sole saranno scartate.

Analisi Discriminanti Lineari (LDA)

Esempio di metodo embedded: il modello predittivo è parte del processo di selezione

- LDA determina funzioni lineari che dividono lo spazio dominio in regioni (E' naturalmente un classificatore)
- Realizza un mapping lineare $f(\mathbf{x})=\mathbf{w}^T\mathbf{x} + b$ tra le variabili in input e lo status tale che sia **minimizzata**

la varianza all'interno della classe e massimizzata quella tra le classi



LDA diventa un Selezionatore di Caratteristiche attraverso un modello di *selezione backward*

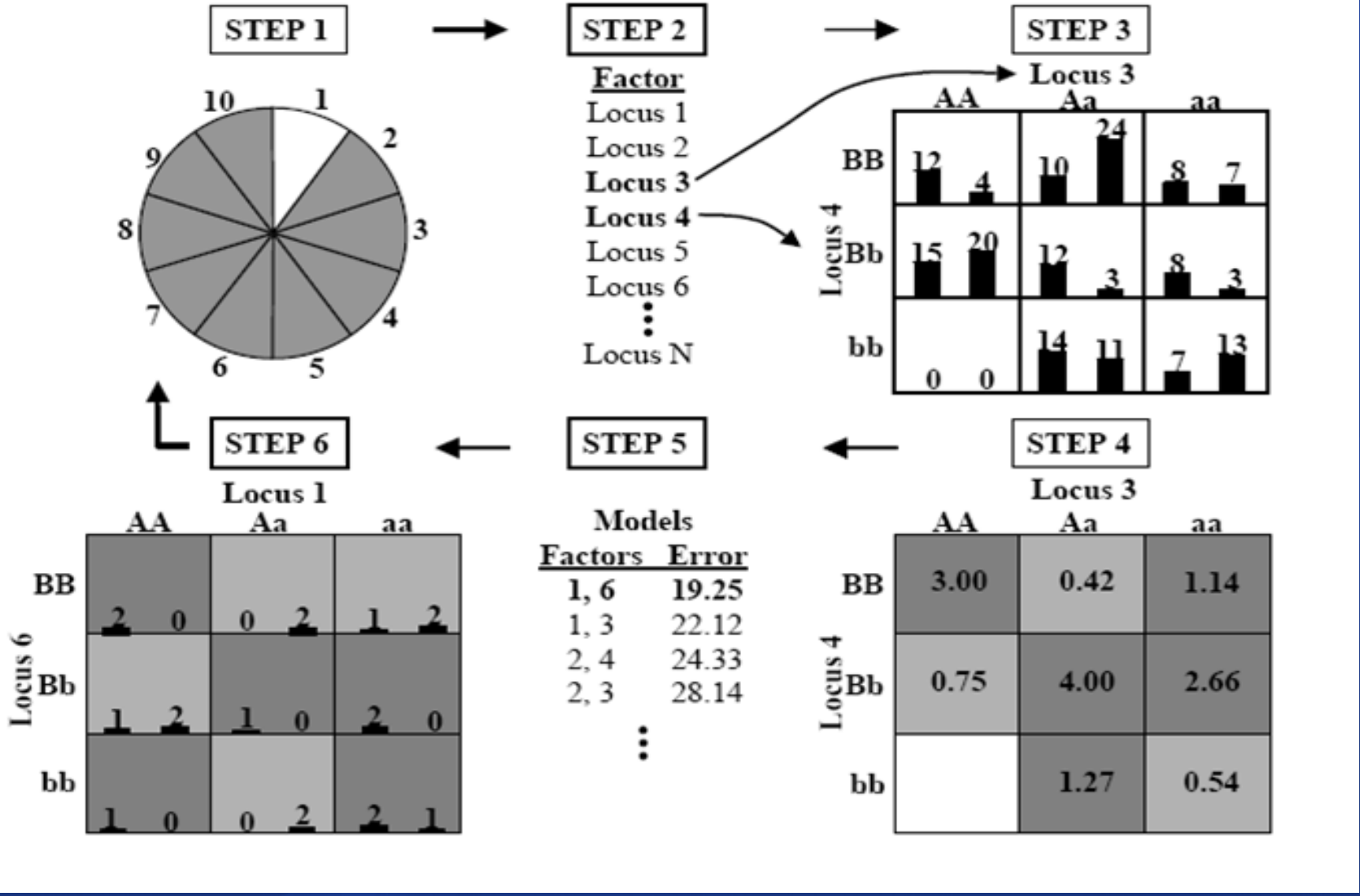
- Partendo dal modello che include tutte le variabili, ad ogni passo si elimina la variabile meno significativa (coefficiente più piccolo) fino a che la capacità di classificazione resta sopra una soglia fissata
- Il data set viene suddiviso in 10 parti (9 training + 1 test) e per ogni modello si calcola il rischio empirico medio valutato sul test set.
- Si seleziona l'insieme di features (modello) con il minor rischio empirico

42

Multifactor Dimensionality Reduction (MDR)

- Realizzato in ambito biomedico per l'analisi di malattie complesse
- Ad ogni passo si costruisce una nuova *metavariabile* unendo due o più variabili e se ne misura la capacità di classificare nuovi esempi mediante cross-validation (training and test)
- Non è parametrico ed è model-free
- E' pensato per funzionare con data set di taglia ridotta

MDR


STEP 1
STEP 2
STEP 3
Factor
Locus 1
Locus 2
Locus 3
Locus 4
Locus 5
Locus 6
Locus N
Locus 3
AA
Aa
aa
BB
Bb
bb
Locus 4
12 4
10 24
8 7
15 20
12 3
8 3
0 0
14 11
7 13
STEP 4
STEP 5
STEP 6
Locus 3
Locus 4
3.00
0.42
1.14
0.75
4.00
2.66
1.27
0.54
Models
Factors Error
1, 6 19.25
1, 3 22.12
2, 4 24.33
2, 3 28.14
Locus 1
Locus 6
2 0
0 2
1 2
1 2
1 0
2 0
1 0
0 2
2 1

L'intera procedura, dal passo uno al passo sei, è ripetuta N volte dividendo i dati, ad ogni iterazione, in diversi insiemi di training e di testing. Si otterrà, così, un insieme di N modelli tra cui sarà scelto quello che massimizza la consistenza della **cross-validation** e minimizza **l'errore di predizione**.

La consistenza della cross-validation è il numero di volte che un modello è identificato. Questa è calcolata ricordando il numero di volte che lo stesso insieme di geni o fattori è stato identicato come miglior modello tra gli N sottoinsiemi di dati.

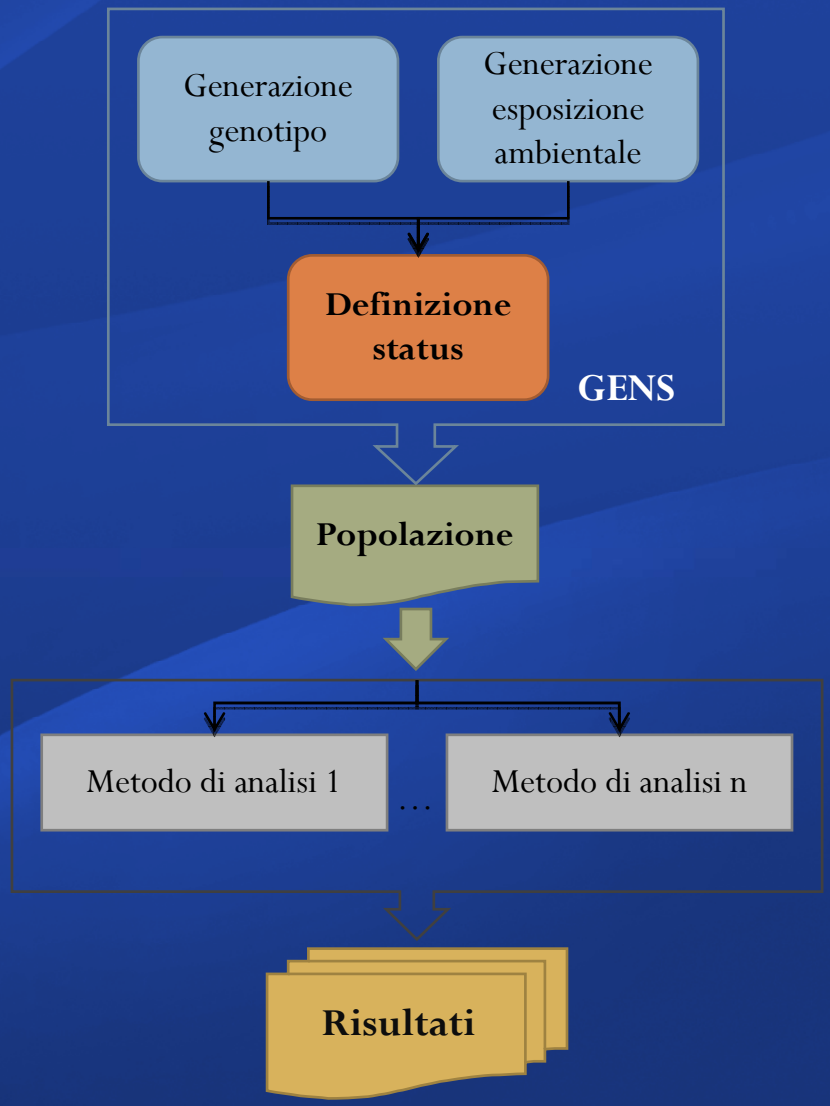
L'errore di predizione, invece, misura la capacita dell'MDR di predire lo stato della malattia nell'insieme indipendente di testing. L'errore di predizione è calcolato come la media degli errori di predizione per ognuno degli N sottoinsiemi della cross-validation.

Generazione dei dati sintetici

- Nei datasets reali non è nota a priori la forma dell'interazione gene-ambiente
- I datasets reali non hanno numerosità tali da permettere sempre uno studio comparativo delle prestazioni dei diversi metodi di feature selection
- Un approccio alternativo è fornito da tecniche Monte Carlo di generazione di popolazioni sintetiche caso/controllo

Gene-*E*nvironment i*N*teraction Simulator

- Simulazione di studi caso/controllo
- Fenomeno sottostante facilmente governabile e intellegibile
- Possibilità di introdurre "rumore" (variabilità tra individui, fenomeno non completamente noto,…)
- Plausibilità biologica
- Possibilità di variare facilmente i parametri di controllo



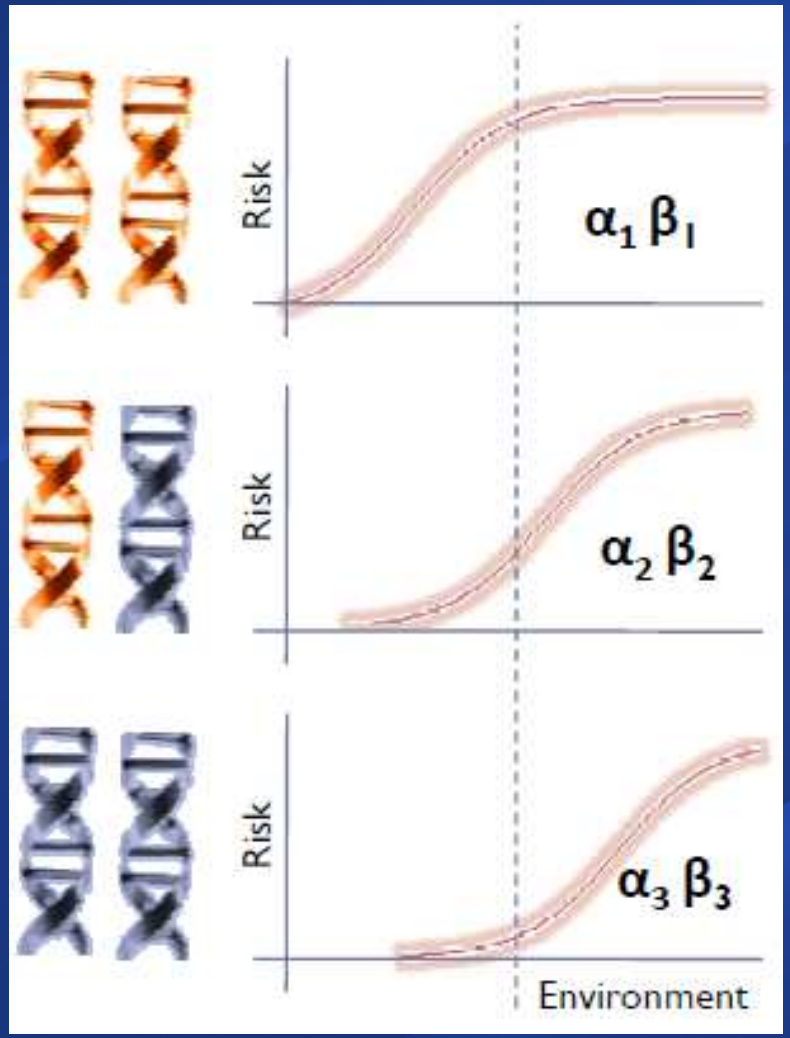
Generazione degli individui

ID	Gene 1	Gene 2	Gene 3	Amb 1	Amb 2	Status
AB23	Aa	AA	AA	3.2	8	0
XY34	BB	Bb	bb	1.1	2	1

Nota la frequenza allelica di ciascun gene è possibile generare il genotipo di ogni paziente nel rispetto di queste frequenze

La distribuzione dell'esposizione ai fattori ambientali, empirica o teorica che sia, consente di generarne i valori

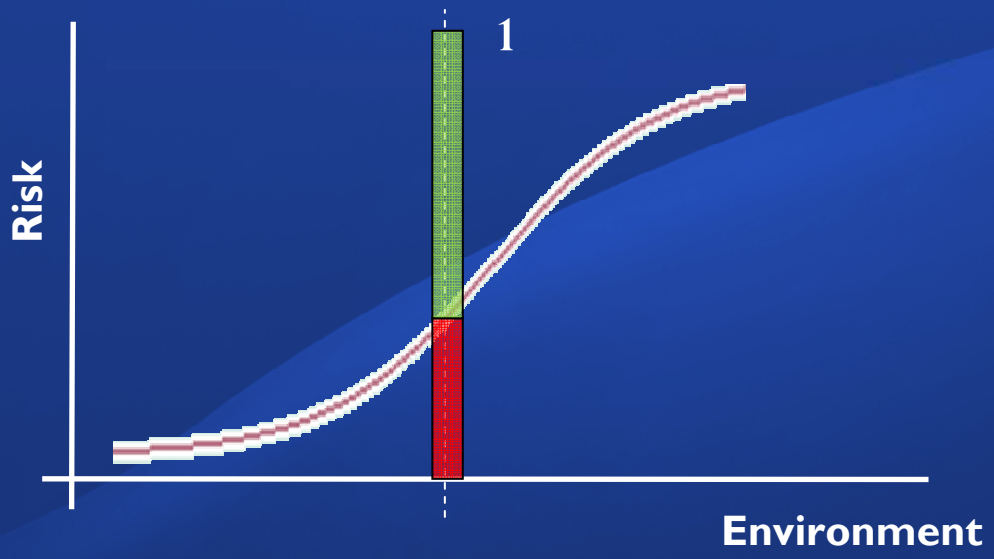
Modello matematico dell'interazione



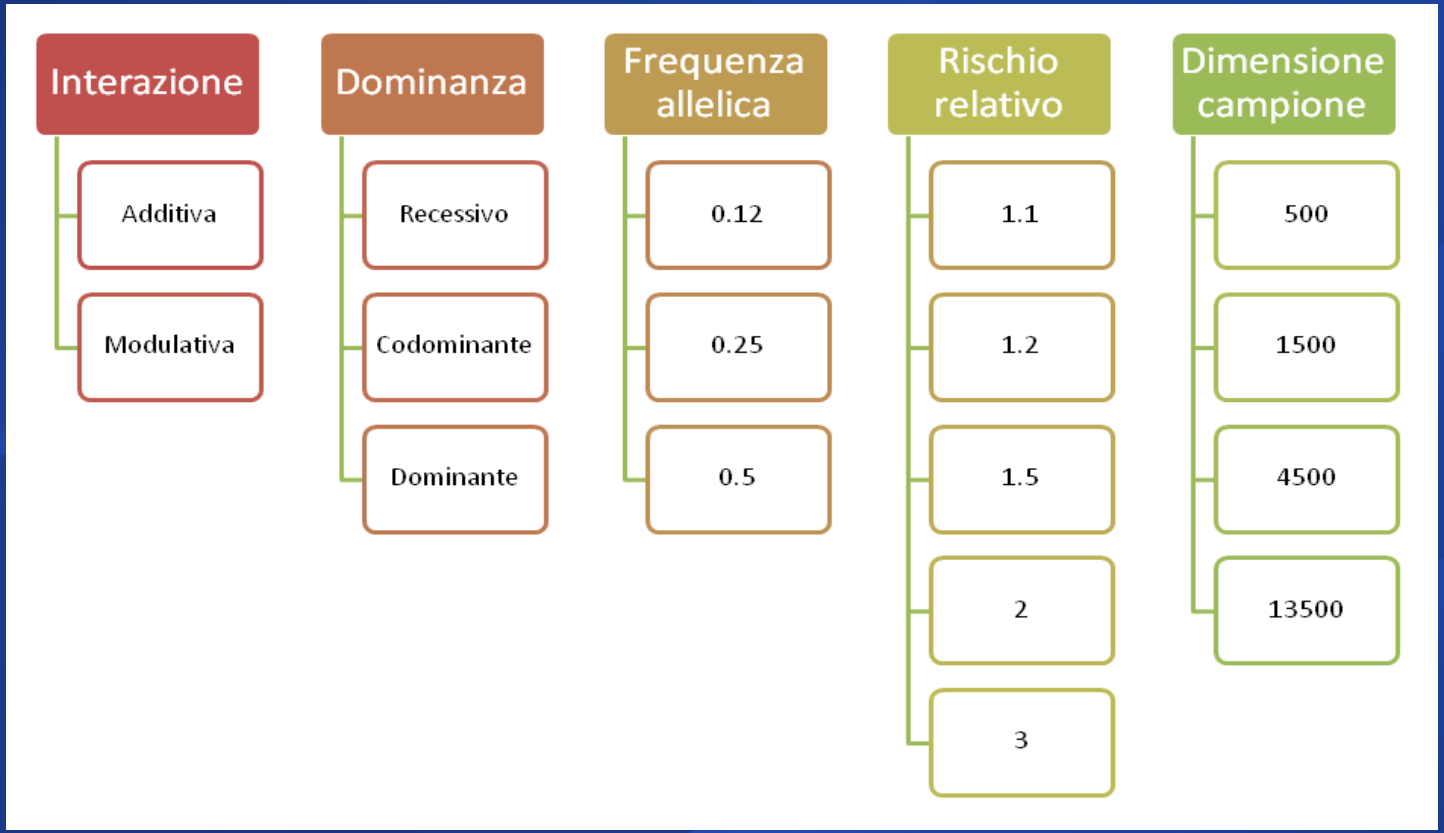
- Modello multilogistico per esprimere il rischio di un individuo:
  - La genetica fissa la forma della funzione di rischio, l'ambiente il valore
  - Tipologie d'interazione esprimibili come relazioni tra le varie funzioni
- E' possibile convertire, in maniera numerica, i parametri epidemiologici in coefficienti per il modello multilogistico

Definizione dello status

$$P(affected \mid g_i, a) = \frac{1}{1 + e^{\alpha_i + \beta_i x}}$$



Scelta dei parametri



- Ogni popolazione contiene 20 variabili (10 genetiche e 10 ambientali) di cui solo 2 (1+1) effettivamente correlate allo status
- 36000 popolazioni analizzate
  - 360 possibili combinazioni dei parametri
  - 100 repliche per ogni combinazione

Complessità computazionale

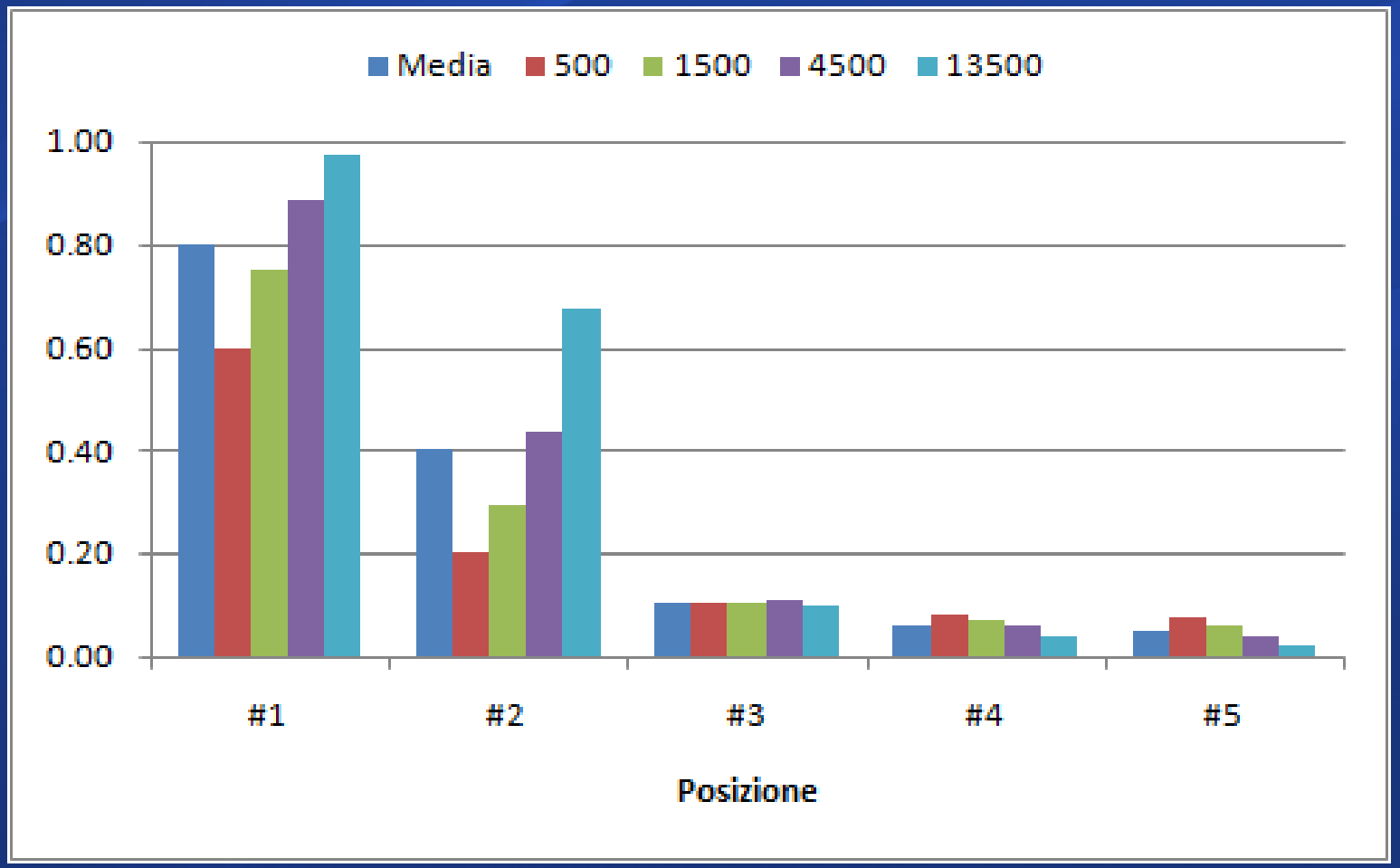
Dato il costo computazionale di alcune tecniche di analisi (ad esempio MDR ha un costo esponenziale nel numero di variabili), il software che esegue le analisi con i quattro metodi è stato concepito fin da principio per essere modulare e funzionare in ambiente distribuito (GRID).

In particolare, le 36000 popolazioni sono state suddivise in vari gruppi e distribuite su due griglie computazionali (farming): Scope e Theophys per un totale complessivo di oltre 1000 ore di calcolo.

Risultati

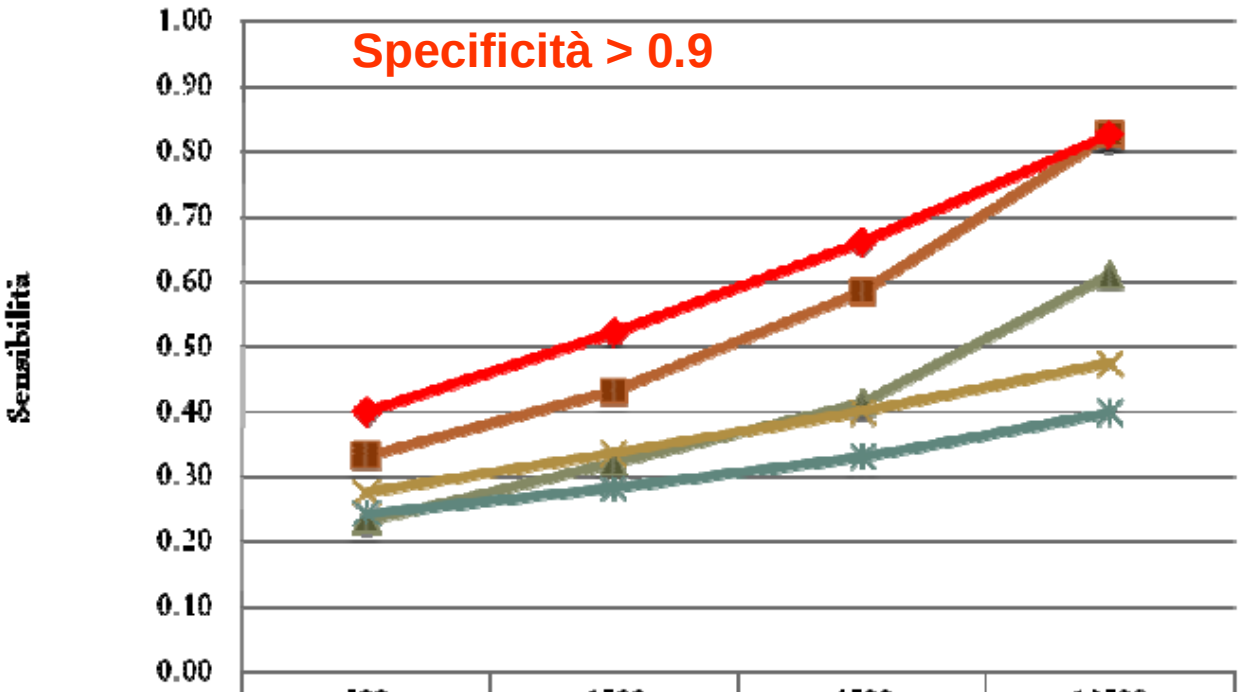
Risultati

Frequenza con la quale in una data posizione appare una delle variabili effettivamente correlate con lo status, in media ed al variare della dimensione del campione



Confronto con i metodi componenti - 1

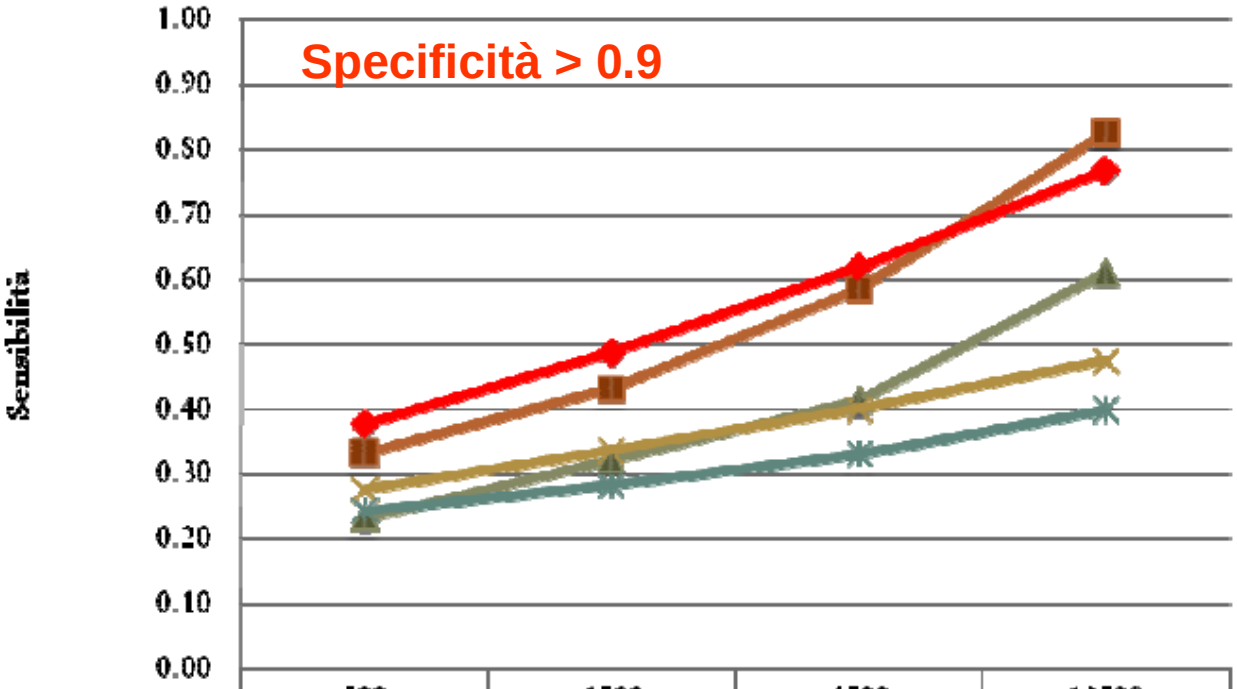
Confronto tra la sensibilità dell'ensemble e quella dei metodi componenti al variare della dimensione del campione



	500	1500	4500	13500
BLR	0.33	0.43	0.58	0.83
UNI	0.23	0.32	0.41	0.61
MDR	0.28	0.34	0.40	0.47
LDA	0.24	0.28	0.33	0.40
ENSEMBLE	0.40	0.52	0.66	0.82

Confronto con i metodi componenti - 2

Confronto tra la sensibilità dell'**ensemble costruito senza la regressione logistica** e quella degli altri metodi al variare della dimensione del campione

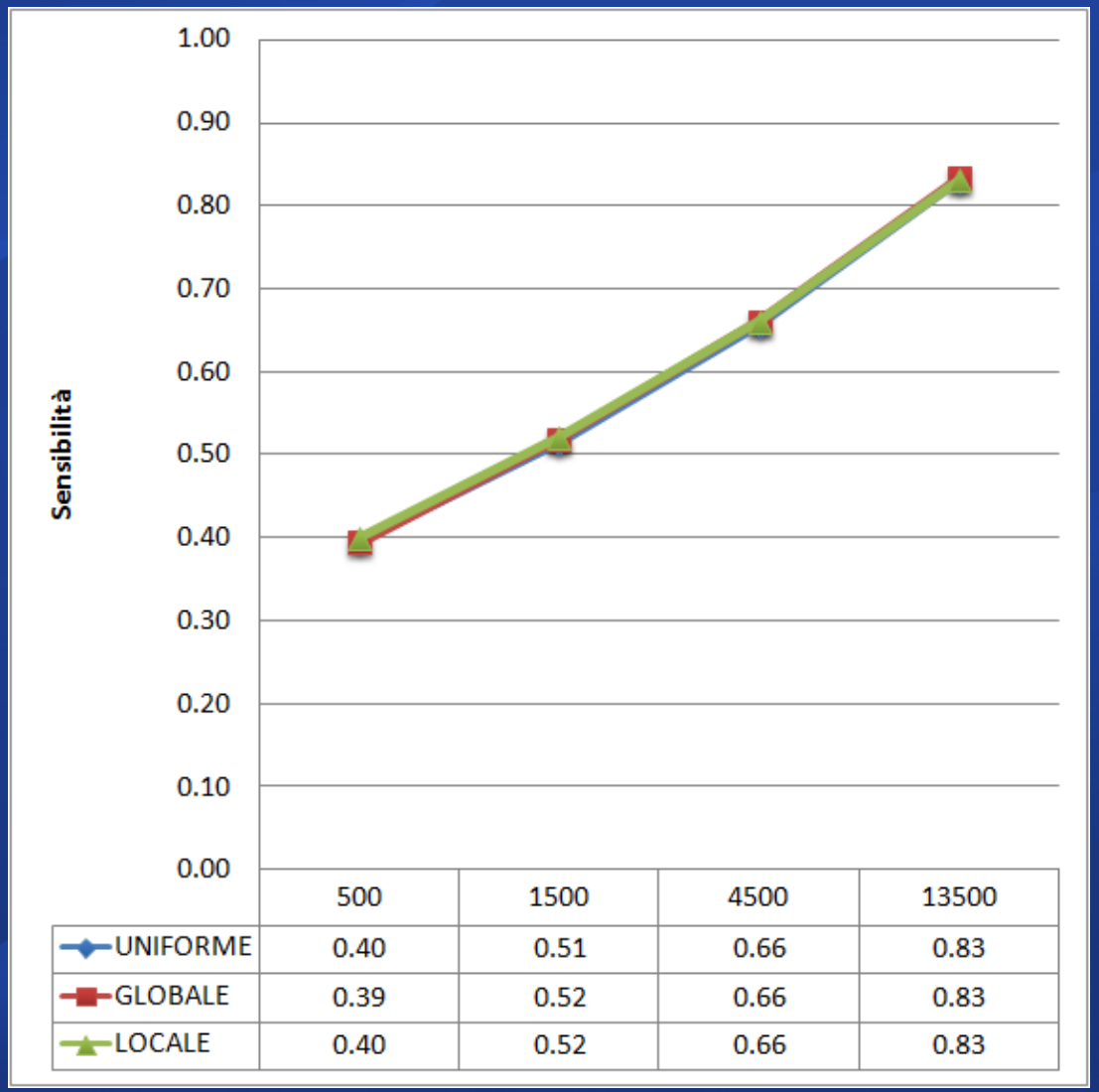


	500	1500	4500	13500
BLR	0.33	0.43	0.55	0.83
UNI	0.23	0.32	0.41	0.61
MDR	0.28	0.34	0.40	0.47
LDA	0.24	0.28	0.33	0.40
ENSEMBLE	0.38	0.49	0.62	0.77

Influenza della scelta dei pesi

Sensibilità, al variare della dimensione del campione, dell'ensemble costruito con 3 differente scelte per i pesi

Comportamenti mediamente simili dei diversi Metodi di Feature Selection $\Longrightarrow$ scarsa dipendenza dalla scelta dei pesi.



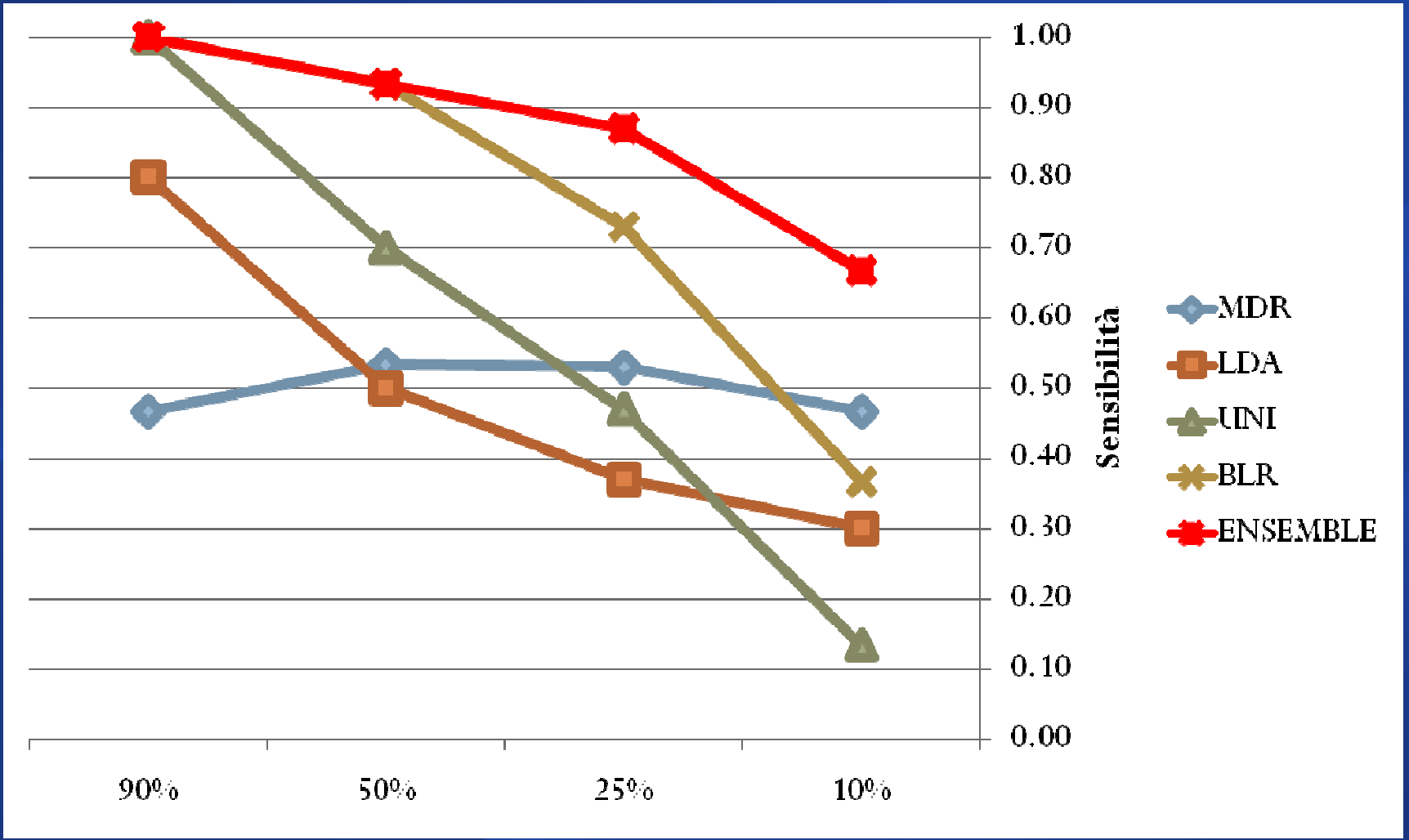
	500	1500	4500	13500
UNIFORME	0.40	0.51	0.66	0.83
GLOBALE	0.39	0.52	0.66	0.83
LOCALE	0.40	0.52	0.66	0.83

Caso di studio Reale

Materiali e metodi

- Dataset di *diabetici napoletani*
- 591 pazienti
  - 369 casi
  - 222 controlli
- 7 variabili
  - 3 ambientali (Età, Sesso, BMI)
  - 4 marker genetici (TCF7L2, UCP3, PPARg, FTO)
- I fattori più frequentemente ritenuti coinvolti dai 5 metodi sono età, BMI, TCF7L2
  - Questo insieme è considerato la risposta di riferimento
- Dal dataset completo vengono estratti 10 subsample alla volta di dimensione via via più piccola

Risultati



E’ da notare l’andamento simile ai dati simulati dei 4 Metodi in funzione del sample size.

Conclusioni

- Un problema "complesso" ma di grande importanza per la salute pubblica
  - Difficoltà nel reperire datasets adeguati
- Molti approcci disponibili ma nessuno universalmente accettato o significativamente migliore degli altri.
- Una possibile soluzione combinando opportunamente le risposte di più metodi di analisi (*ensemble*).
- Un semplice sistema a voto di maggioranza migliora significativamente le prestazioni rispetto ai singoli metodi.