Corso di Laurea in Informatica

Single e Cross-layer Detection di Siti Web Malevoli: Un Confronto Empirico

Prof. Fabio Palomba

Nicolapio Gagliarde
Mat.: 0512106980

n.gagliarde@studenti.unisa.it
https://github.com/GagliardeNicolapio
https://www.linkedin.com/in/nicolapio-gagliarde-75209018b/

Il problema

170 milioni di siti web malevoli

n.gagliarde@studenti.unisa.it
https://github.com/GagliardeNicolapio
https://www.linkedin.com/in/nicolapio-gagliarde-75209018b/

Il problema

170 milioni di siti web malevoli

650 milioni di attacchi

n.gagliarde@studenti.unisa.it
https://github.com/GagliardeNicolapio
https://www.linkedin.com/in/nicolapio-gagliarde-75209018b/

Il problema

170 milioni di siti web malevoli

650 milioni di attacchi

Milioni di dollari persi al minuto

n.gagliarde@studenti.unisa.it
https://github.com/GagliardeNicolapio
https://www.linkedin.com/in/nicolapio-gagliarde-75209018b/

Alcune soluzioni

Analisi degli URL

n.gagliarde@studenti.unisa.it
https://github.com/GagliardeNicolapio
https://www.linkedin.com/in/nicolapio-gagliarde-75209018b/

Alcune soluzioni

sesa lab
SOFTWARE ENGINEERING SALERNO

- Analisi degli URL
- Analisi dei redirect e delle risorse richieste

n.gagliarde@studenti.unisa.it
https://github.com/GagliardeNicolapio
https://www.linkedin.com/in/nicolapio-gagliarde-75209018b/

Alcune soluzioni

sesa lab
SOFTWARE ENGINEERING SALERNO

Analisi degli URL

Analisi dei redirect e delle risorse richieste

Analisi della pagina web

n.gagliarde@studenti.unisa.it
https://github.com/GagliardeNicolapio
https://www.linkedin.com/in/nicolapio-gagliarde-75209018b/

Alcune soluzioni

sesa lab
SOFTWARE ENGINEERING SALERNO

Analisi degli URL

Analisi dei redirect e delle risorse richieste

Analisi della pagina web

} Tecniche Single-layer

n.gagliarde@studenti.unisa.it
https://github.com/GagliardeNicolapio
https://www.linkedin.com/in/nicolapio-gagliarde-75209018b/

Alcune soluzioni

Potrebbero fallire con:

Analisi degli URL → Indirizzi corti o troppo simili a URL benevoli

Analisi dei redirect e delle risorse richieste

Analisi della pagina web

n.gagliarde@studenti.unisa.it
https://github.com/GagliardeNicolapio
https://www.linkedin.com/in/nicolapio-gagliarde-75209018b/

Alcune soluzioni

Potrebbero fallire con:

Analisi degli URL → Indirizzi corti o troppo simili a URL benevoli

Analisi dei redirect e delle risorse richieste → Siti creati con i CMS

Analisi della pagina web

n.gagliarde@studenti.unisa.it
https://github.com/GagliardeNicolapio
https://www.linkedin.com/in/nicolapio-gagliarde-75209018b/

Alcune soluzioni

Potrebbero fallire con:

Analisi degli URL → Indirizzi corti o troppo simili a URL benevoli

Analisi dei redirect e delle risorse richieste → Siti creati con i CMS

Analisi della pagina web → Siti creati con un interfaccia clonata e tecniche di offuscamento

n.gagliarde@studenti.unisa.it
https://github.com/GagliardeNicolapio
https://www.linkedin.com/in/nicolapio-gagliarde-75209018b/

Single e Cross-layer

Livello rete
Single-layer

n.gagliarde@studenti.unisa.it
https://github.com/GagliardeNicolapio
https://www.linkedin.com/in/nicolapio-gagliarde-75209018b/

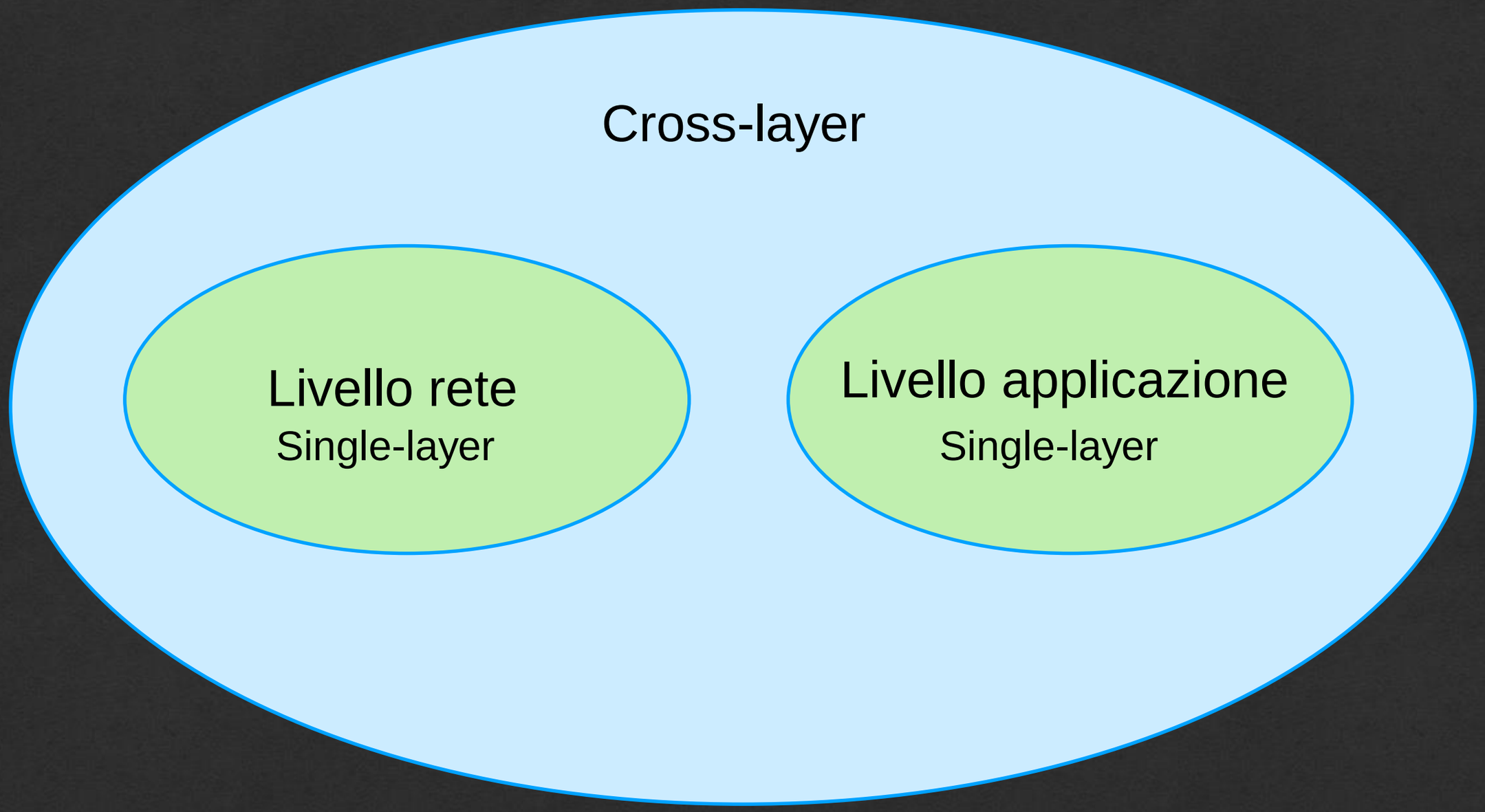
Single e Cross-layer

Livello rete
Single-layer

Livello applicazione
Single-layer

n.gagliarde@studenti.unisa.it
https://github.com/GagliardeNicolapio
https://www.linkedin.com/in/nicolapio-gagliarde-75209018b/

Single e Cross-layer



n.gagliarde@studenti.unisa.it
https://github.com/GagliardeNicolapio
https://www.linkedin.com/in/nicolapio-gagliarde-75209018b/

Lo scopo e le differenze

Lo scopo: confrontare i risultati ottenuti con i risultati di Xu[2014][1]

1 Li Xu. Detecting and characterizing malicious websites. The University of Texas at San Antonio, 2014.

n.gagliarde@studenti.unisa.it
https://github.com/GagliardeNicolapio
https://www.linkedin.com/in/nicolapio-gagliarde-75209018b/

Lo scopo e le differenze

Lo scopo: confrontare i risultati ottenuti con i risultati di Xu[2014][1]

- medesimi algoritmi

1 Li Xu. Detecting and characterizing malicious websites. The University of Texas at San Antonio, 2014.

n.gagliarde@studenti.unisa.it
https://github.com/GagliardeNicolapio
https://www.linkedin.com/in/nicolapio-gagliarde-75209018b/

Lo scopo e le differenze

Lo scopo: confrontare i risultati ottenuti con i risultati di Xu[2014][1]

- medesimi algoritmi
- medesime tecniche

1 Li Xu. Detecting and characterizing malicious websites. The University of Texas at San Antonio, 2014.

n.gagliarde@studenti.unisa.it
https://github.com/GagliardeNicolapio
https://www.linkedin.com/in/nicolapio-gagliarde-75209018b/

Lo scopo e le differenze

Lo scopo: confrontare i risultati ottenuti con i risultati di Xu[2014][1]

- medesimi algoritmi
- medesime tecniche
- dataset diverso!

1 Li Xu. Detecting and characterizing malicious websites. The University of Texas at San Antonio, 2014.

n.gagliarde@studenti.unisa.it
https://github.com/GagliardeNicolapio
https://www.linkedin.com/in/nicolapio-gagliarde-75209018b/

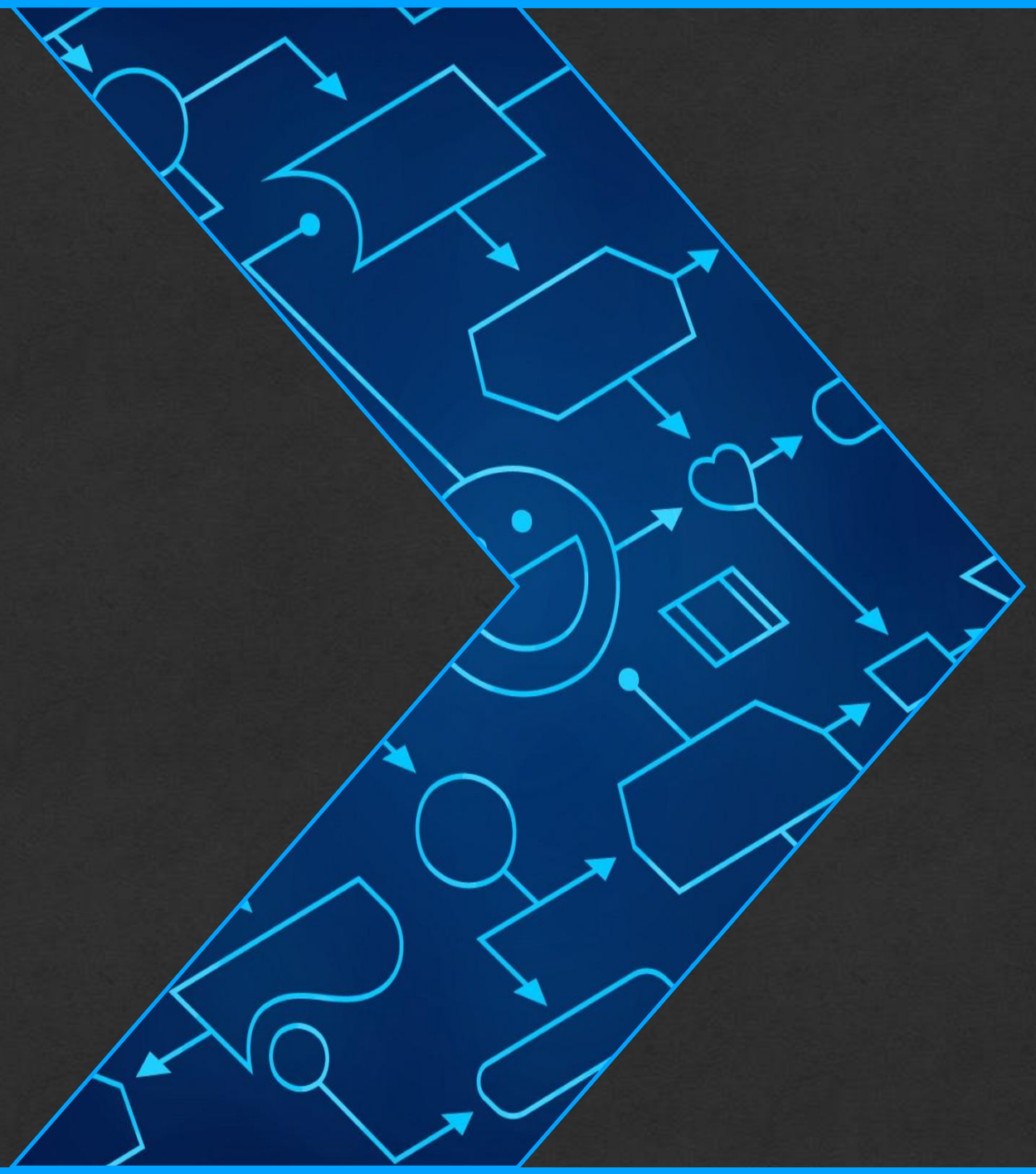
Gli algoritmi

Naive Bayes

Logistic Regression

Support Vector Machine

Decision Tree



n.gagliarde@studenti.unisa.it
https://github.com/GagliardeNicolapio
https://www.linkedin.com/in/nicolapio-gagliarde-75209018b/

Feature selection

Principal Component Analysis

CFS Subset Evaluation

Information Gain

n.gagliarde@studenti.unisa.it
https://github.com/GagliardeNicolapio
https://www.linkedin.com/in/nicolapio-gagliarde-75209018b/

Per il cross-layer

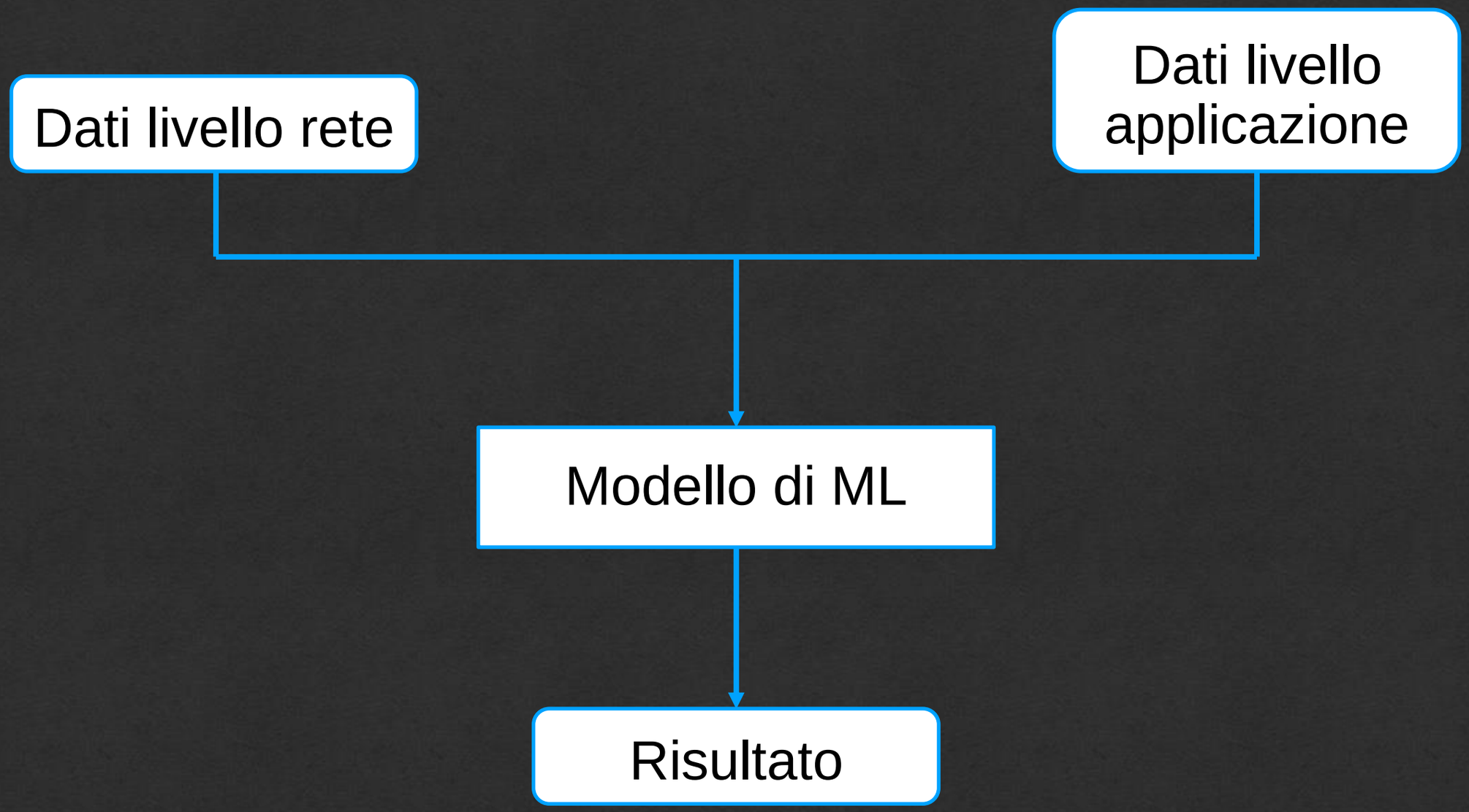
Data-aggregation

OR-aggregation

AND-aggregation

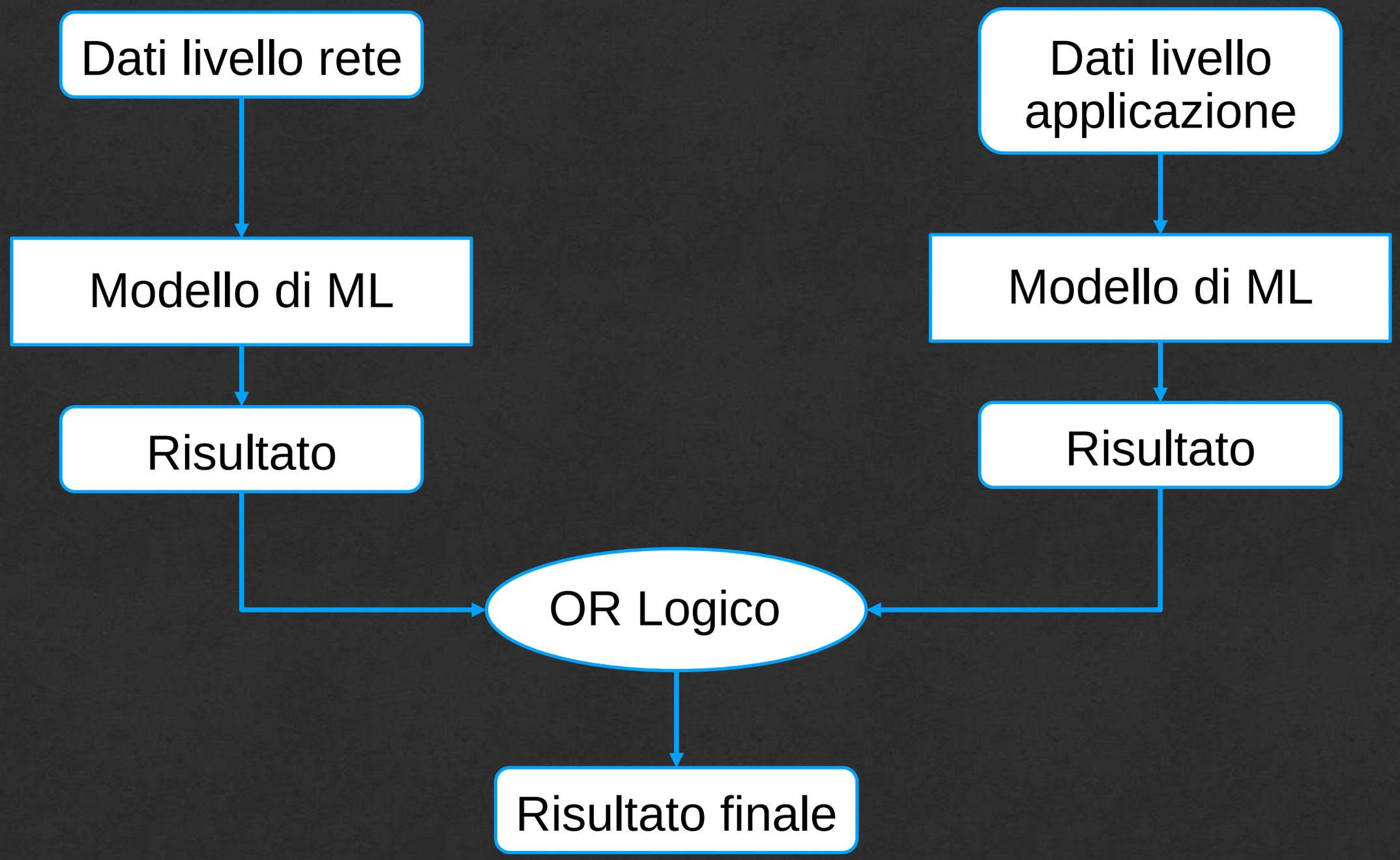
n.gagliarde@studenti.unisa.it
https://github.com/GagliardeNicolapio
https://www.linkedin.com/in/nicolapio-gagliarde-75209018b/

Data-aggregation



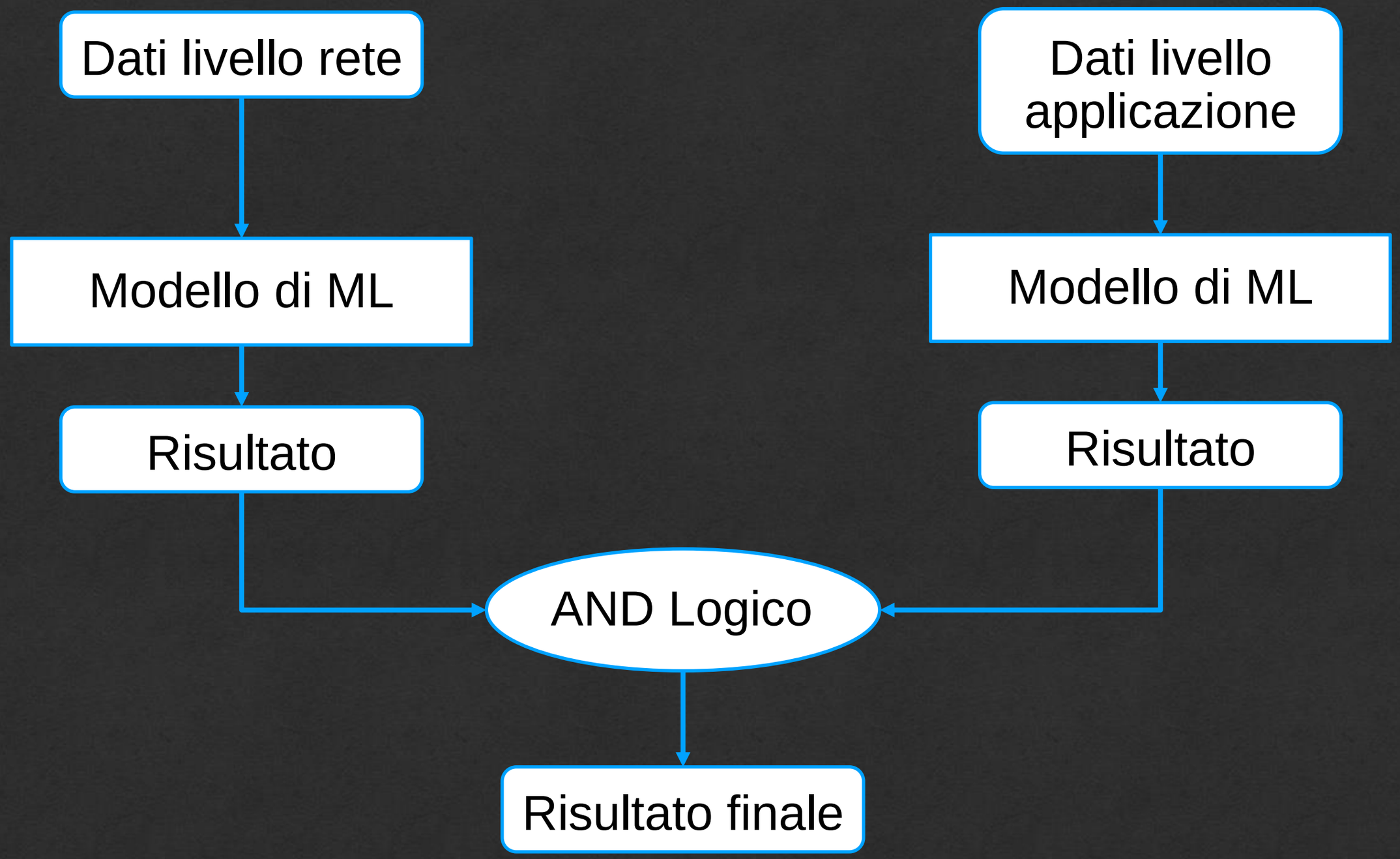
n.gagliarde@studenti.unisa.it
https://github.com/GagliardeNicolapio
https://www.linkedin.com/in/nicolapio-gagliarde-75209018b/

OR-aggregation



n.gagliarde@studenti.unisa.it
https://github.com/GagliardeNicolapio
https://www.linkedin.com/in/nicolapio-gagliarde-75209018b/

AND-aggregation

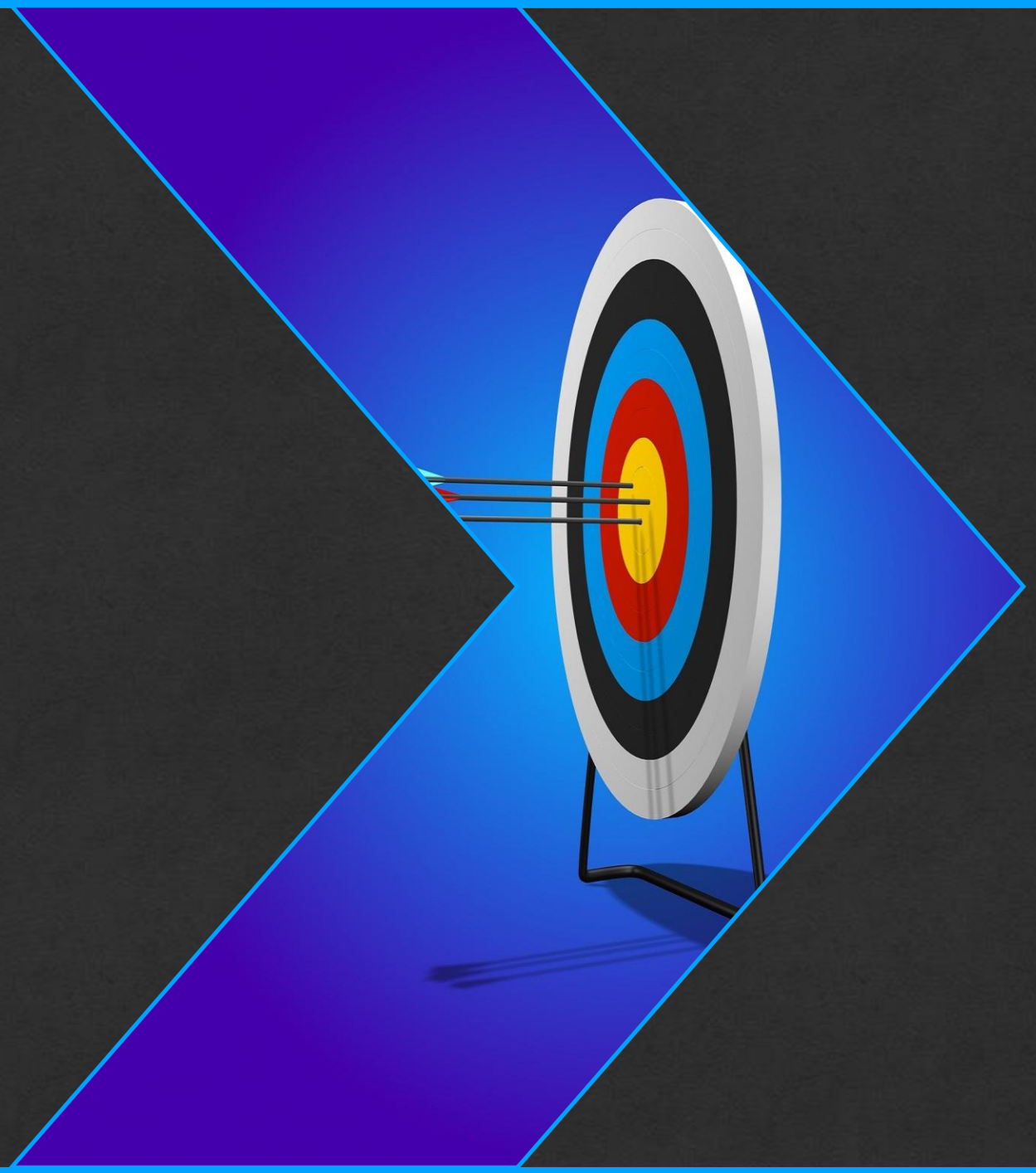


n.gagliarde@studenti.unisa.it
https://github.com/GagliardeNicolapio
https://www.linkedin.com/in/nicolapio-gagliarde-75209018b/

I risultati

Algoritmo migliore: Decision Tree

- Accuracy: 95%
- Falsi negativi: 5%
- Falsi positivi: 7%

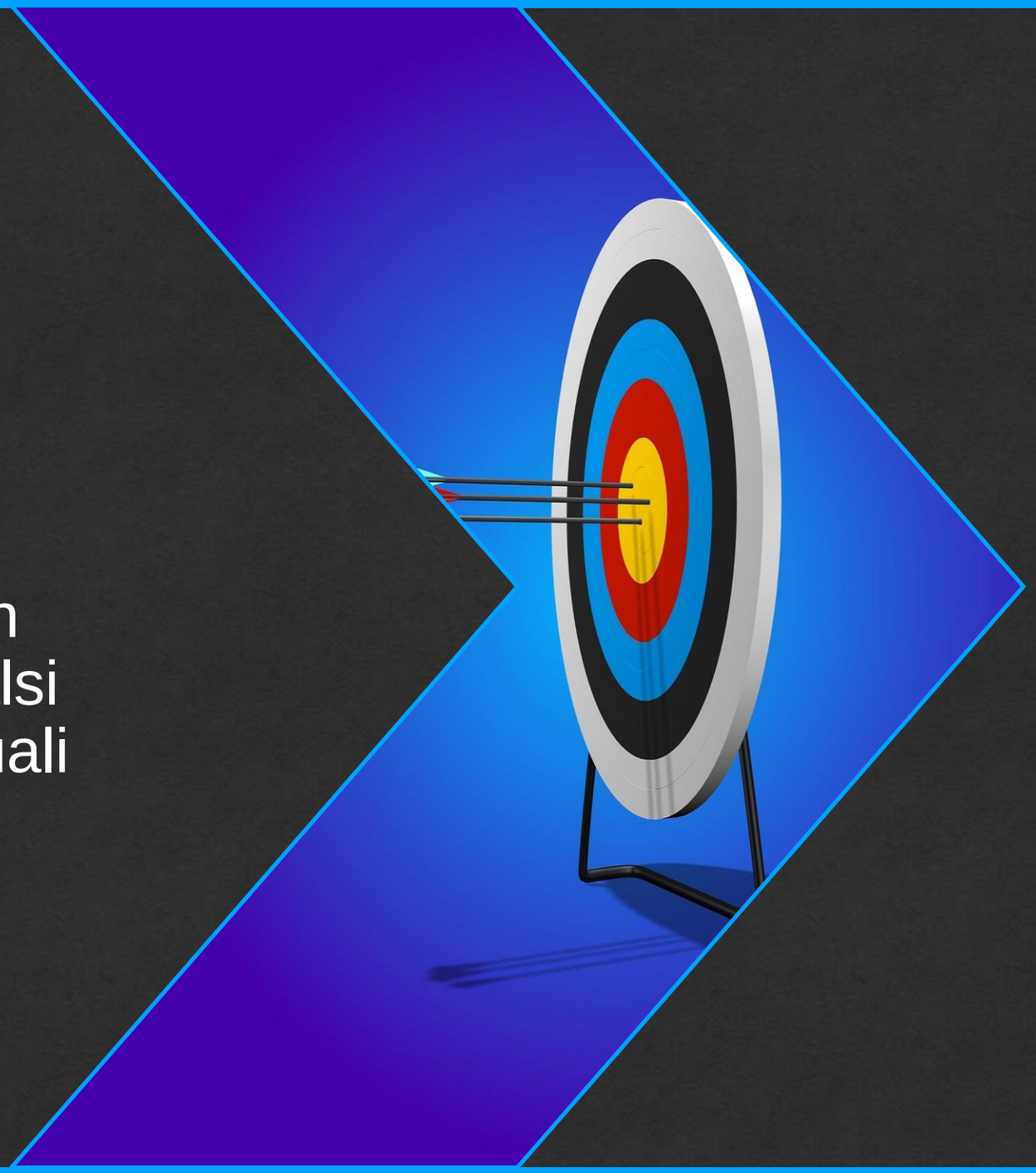


n.gagliarde@studenti.unisa.it
https://github.com/GagliardeNicolapio
https://www.linkedin.com/in/nicolapio-gagliarde-75209018b/

I risultati

Algoritmo migliore: Decision Tree

La AND-aggregation con Decision Tree implica un incremento dei falsi negativi di circa 20 punti percentuali



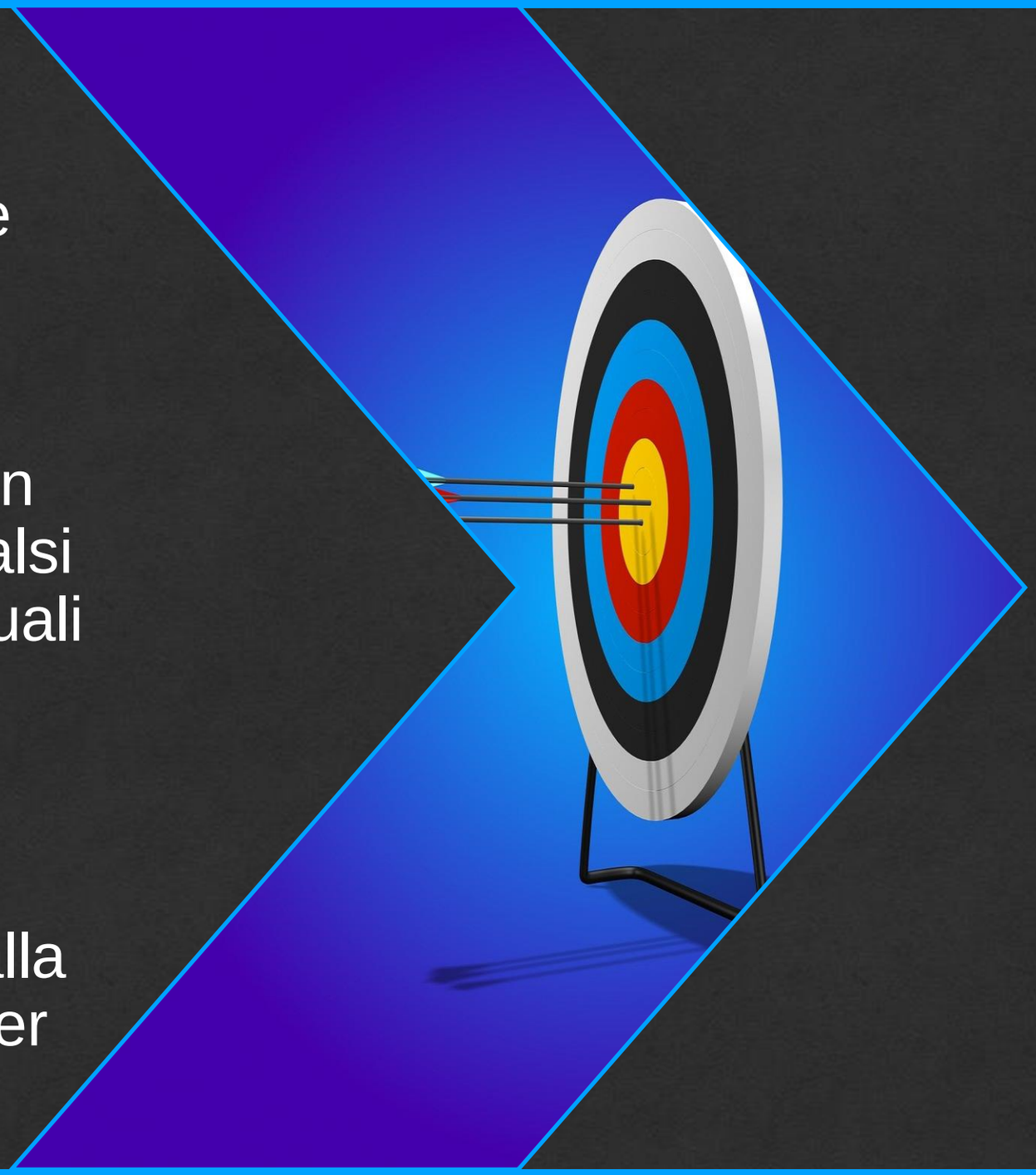
n.gagliarde@studenti.unisa.it
https://github.com/GagliardeNicolapio
https://www.linkedin.com/in/nicolapio-gagliarde-75209018b/

I risultati

Algoritmo migliore: Decision Tree

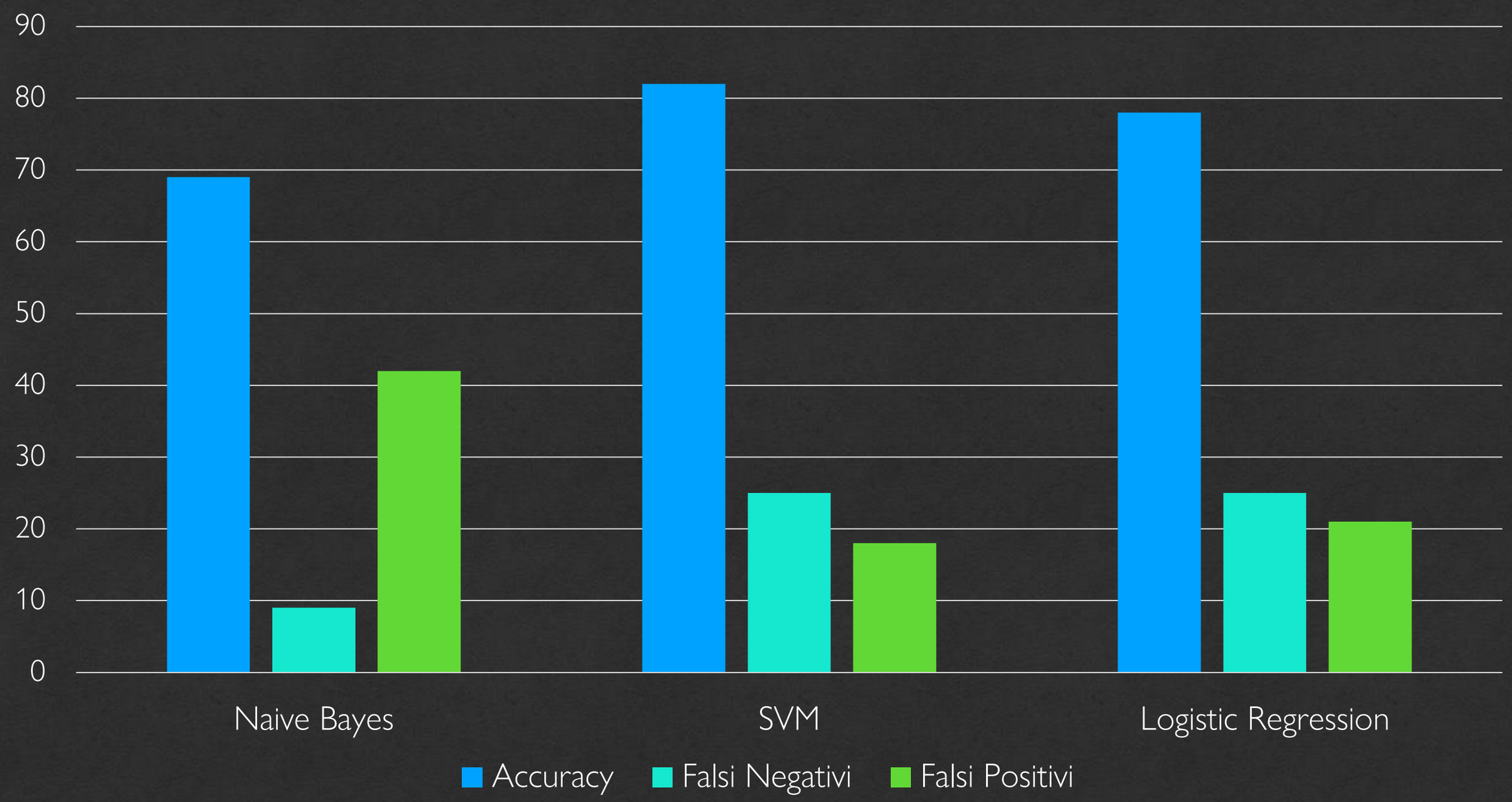
La AND-aggregation con Decision Tree implica un incremento dei falsi negativi di circa 20 punti percentuali

Naive Bayes, SVM e Logistic Regression risultano non adatti alla classificazione single e cross-layer



n.gagliarde@studenti.unisa.it
https://github.com/GagliardeNicolapio
https://www.linkedin.com/in/nicolapio-gagliarde-75209018b/

I risultati



n.gagliarde@studenti.unisa.it
https://github.com/GagliardeNicolapio
https://www.linkedin.com/in/nicolapio-gagliarde-75209018b/

Single e Cross-layer Detection di Siti Web Malevoli: Un Confronto Empirico

Nicolapio Gagliarde
Università degli Studi di Salerno

Il problema

170 milioni di siti web malevoli

650 milioni di attacchi

Milioni di dollari persi al minuto

Alcune soluzioni

Analisi degli URL

Analisi dei redirect e delle risorse richieste

Analisi della pagina web

} Tecniche Single-layer

Alcune soluzioni

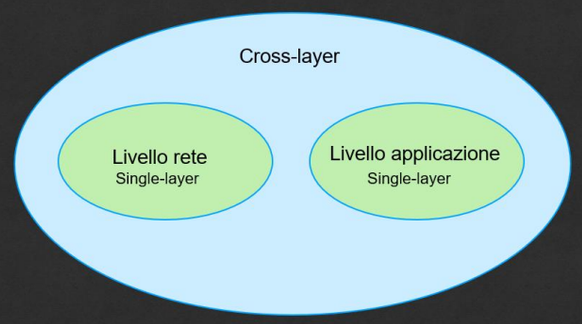
Potrebbero fallire con:

Analisi degli URL → Indirizzi corti o troppo simili a URL benevoli

Analisi dei redirect e delle risorse richieste → Siti creati con i CMS

Analisi della pagina web → Siti creati con un interfaccia clonata e tecniche di offuscamento

Single e Cross-layer



Lo scopo e le differenze

Lo scopo: confrontare i risultati ottenuti con i risultati di Xu[2014][1]

- medesimi algoritmi
- medesime tecniche
- dataset diverso!

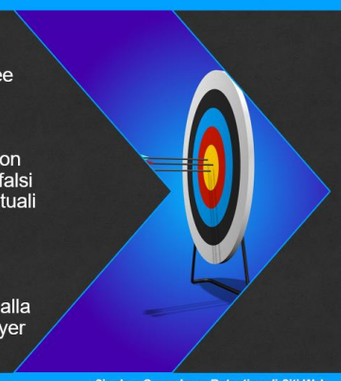
1 Li Xu. Detecting and characterizing malicious websites. The University of Texas at San Antonio, 2014.

I risultati

Algoritmo migliore: Decision Tree

La AND-aggregation con Decision Tree implica un incremento dei falsi negativi di circa 20 punti percentuali

Naive Bayes, SVM e Logistic Regression risultano non adatti alla classificazione single e cross-layer



Single e Cross-layer Detection di Siti Web Malevoli: Un Confronto Empirico

Grazie!

Nicolapio Gagliarde

n.gagliarde@studenti.unisa.it

https://github.com/GagliardeNicolapio

https://www.linkedin.com/in/nicolapio-gagliarde-75209018b/