Identificación de personas mediante patrones de escritura con teclado

Equipo

Luis

Mario

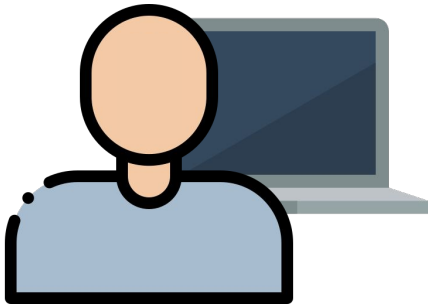
José

Problema

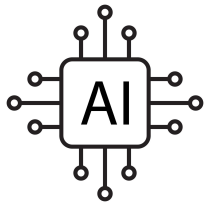
Identificación de personas

Método de identificación complementario…

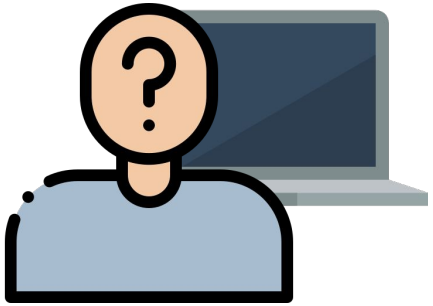
Un alumno entrena al modelo durante todo el ciclo



Inicio de Ciclo

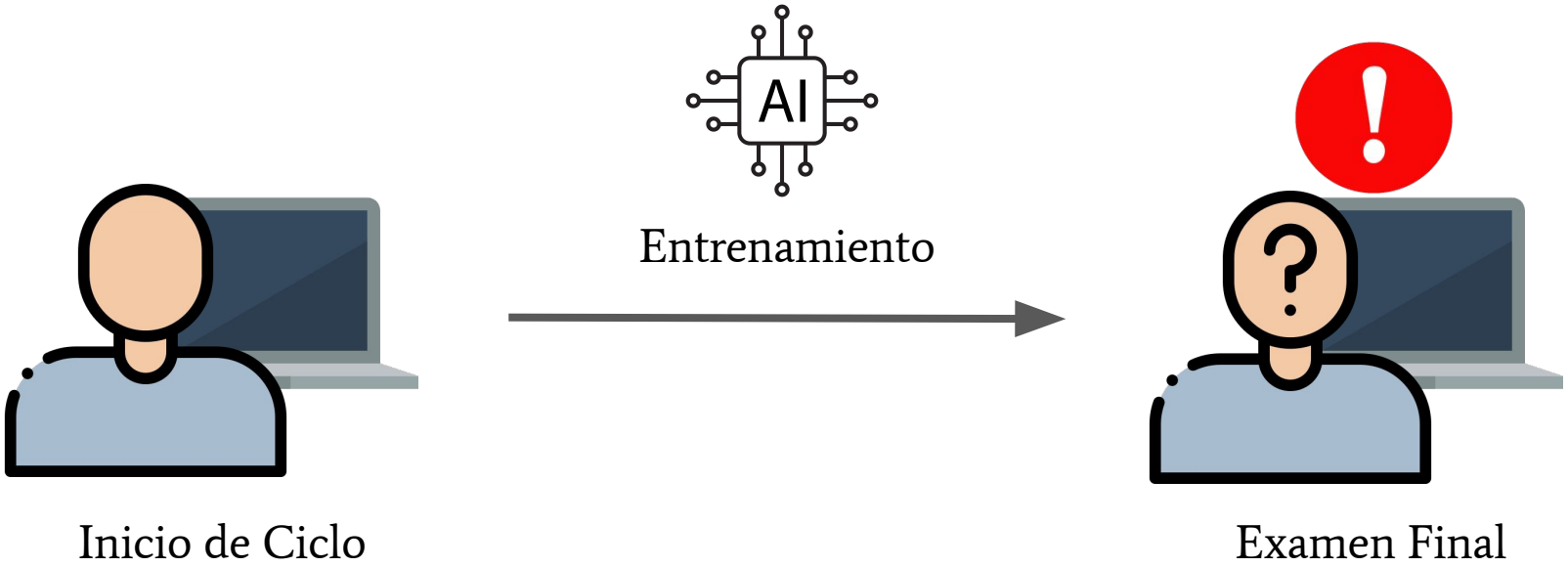


Entrenamiento



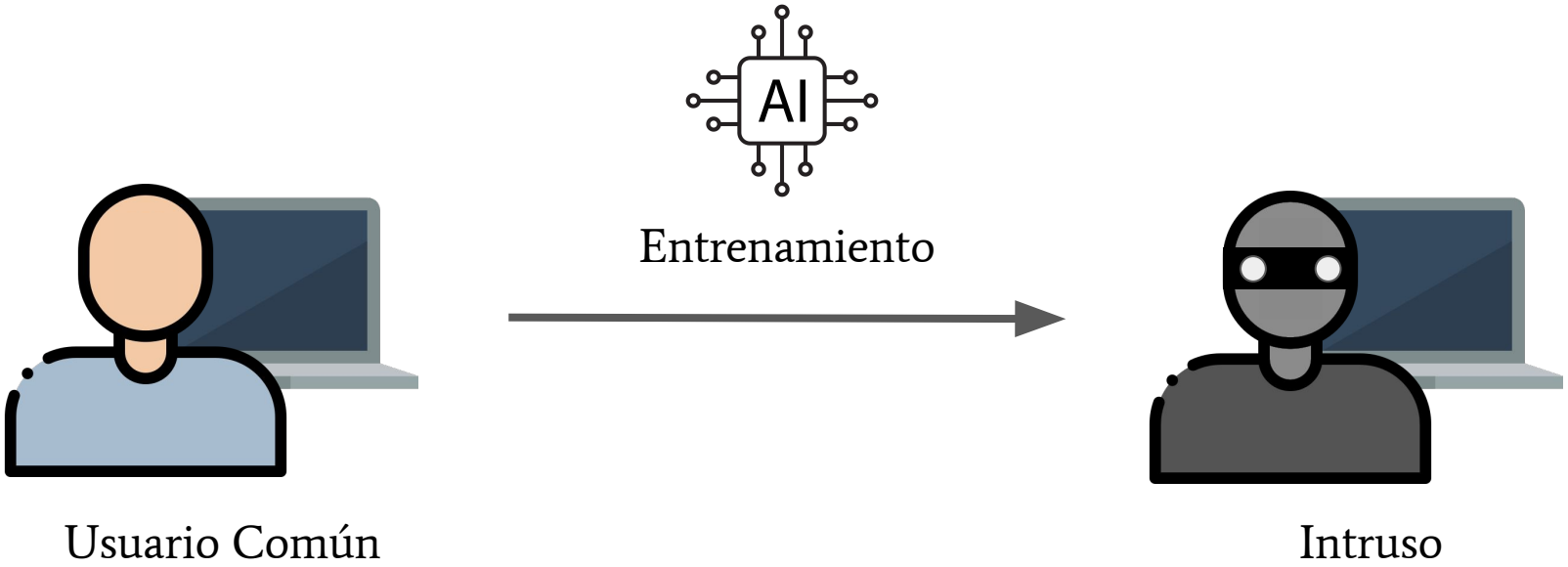
Examen Final

Un alumno entrena al modelo durante todo el ciclo

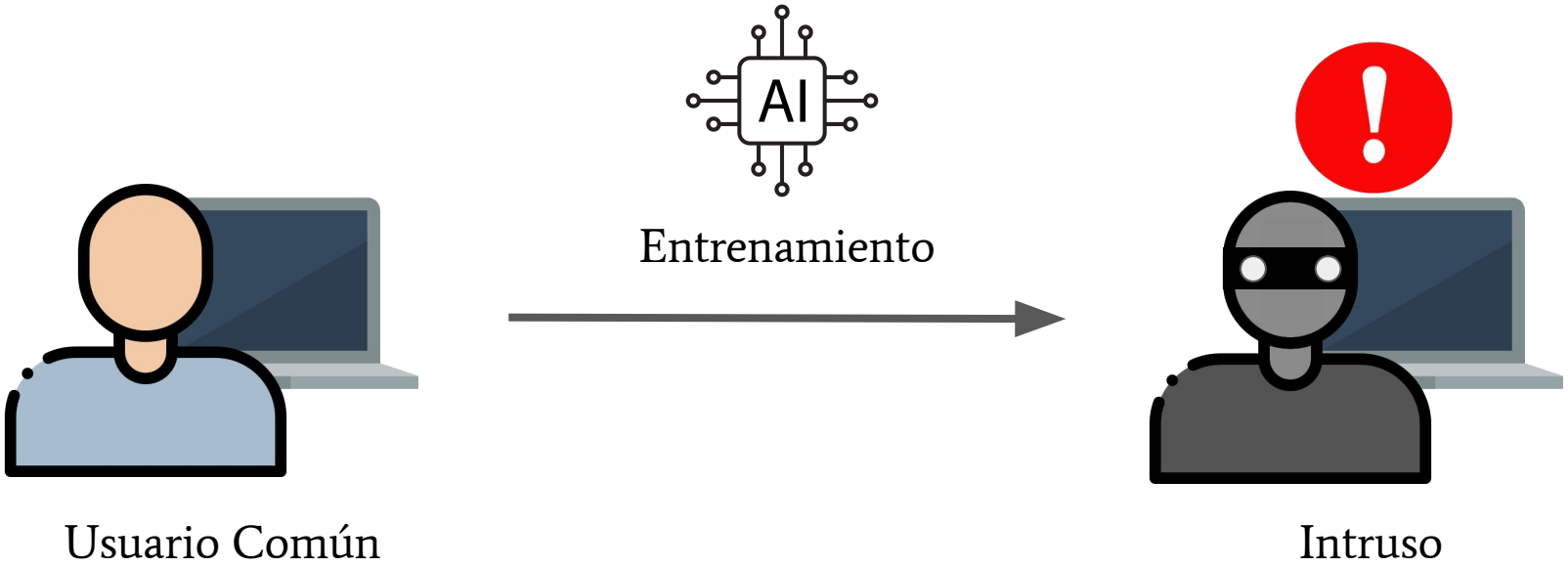


Podría sospechar suplantación

Un usuario común entrena al modelo con el uso diario



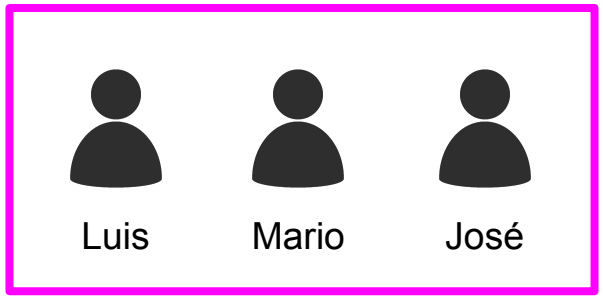
Un usuario común entrena al modelo con el uso diario



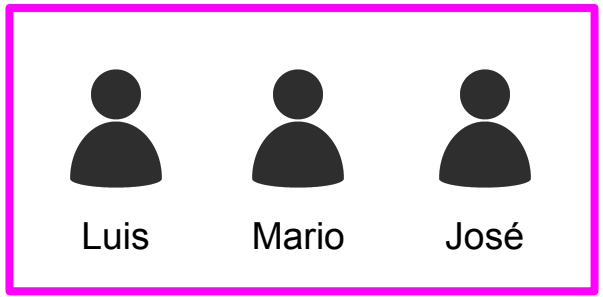
Podría sospechar de intrusos

Datos

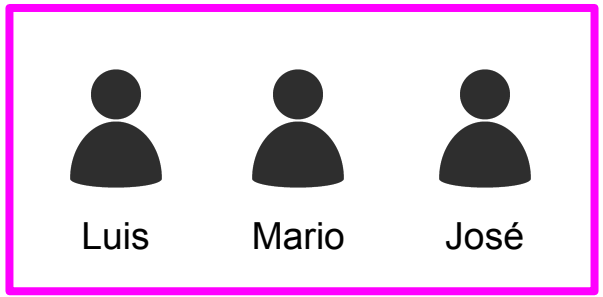
Inexorablemente, se parte de un conjunto definido de **personas**



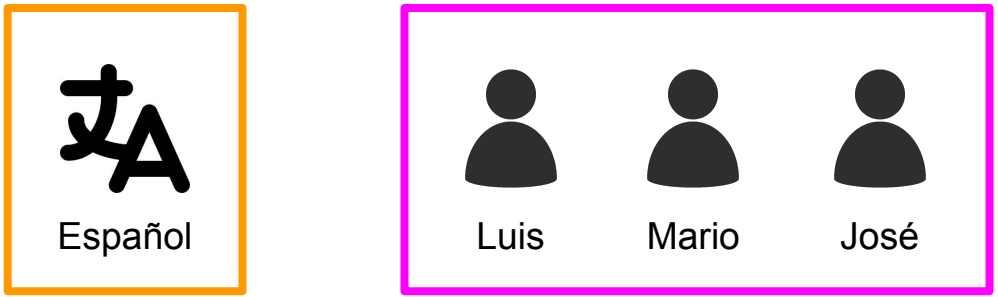
En el peor de los casos, **el contenido** de su producción de texto **no las delata**.



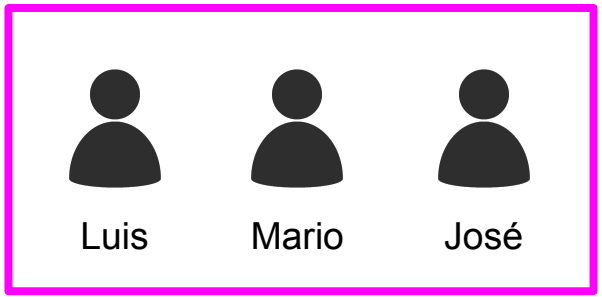
Es decir, son similares en **léxico y sintaxis**.



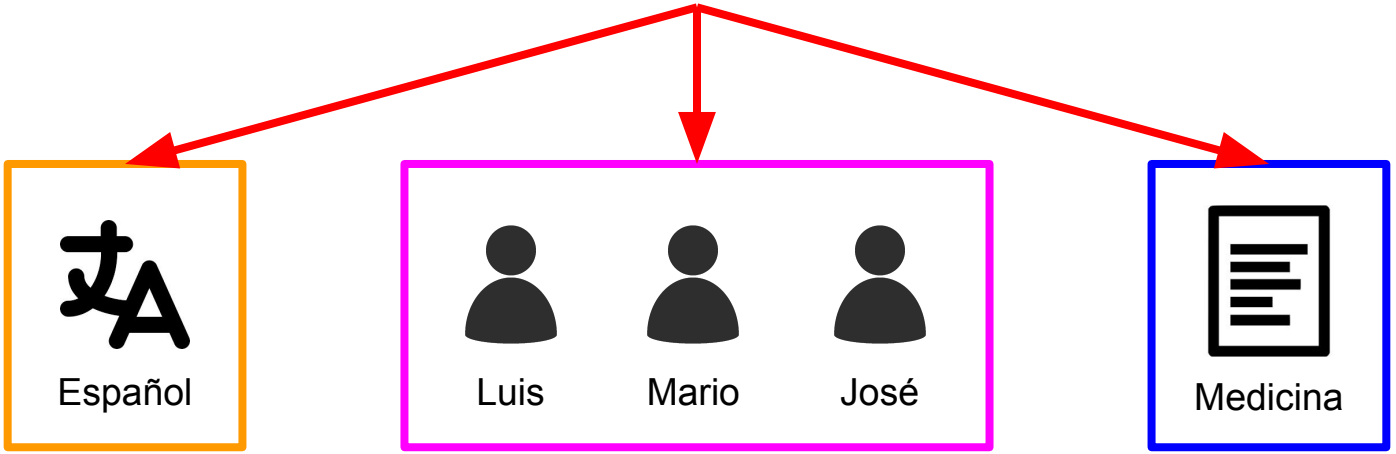
Eso implica: compartir mismos **idiomas**



Y manejar los mismos **registros, temas y estilos**.



En este trabajo, buscamos tener control sobre estas tres variables para tener mayor flexibilidad y capacidad de prueba en nuestro entorno.



¿Qué hicimos como buenos científicos de la computación?

¡Una página web!

¡Una página web!

Identidad

Texto de un conjunto fijo de textos

Espacio para transcribir el texto

¡No hay lugar como el hogar!

Se fomenta la realización de la prueba en las condiciones más **representativas / comunes**

Con su teclado, en su setup, etc.

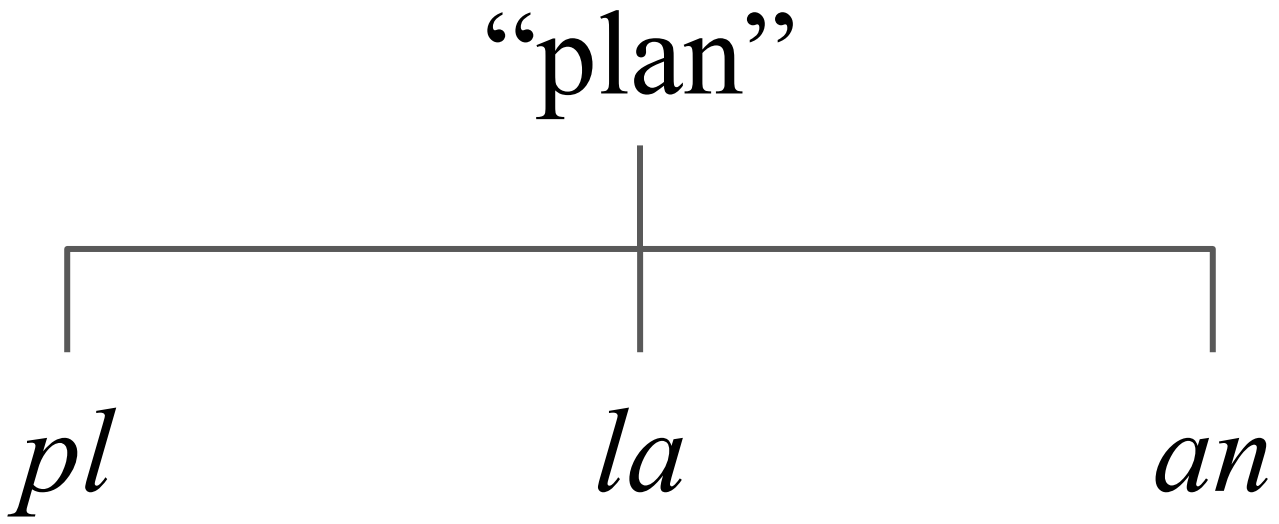
¿Pero qué **información** se puede **extraer**?

Decidimos concentrarnos en **un aspecto**:

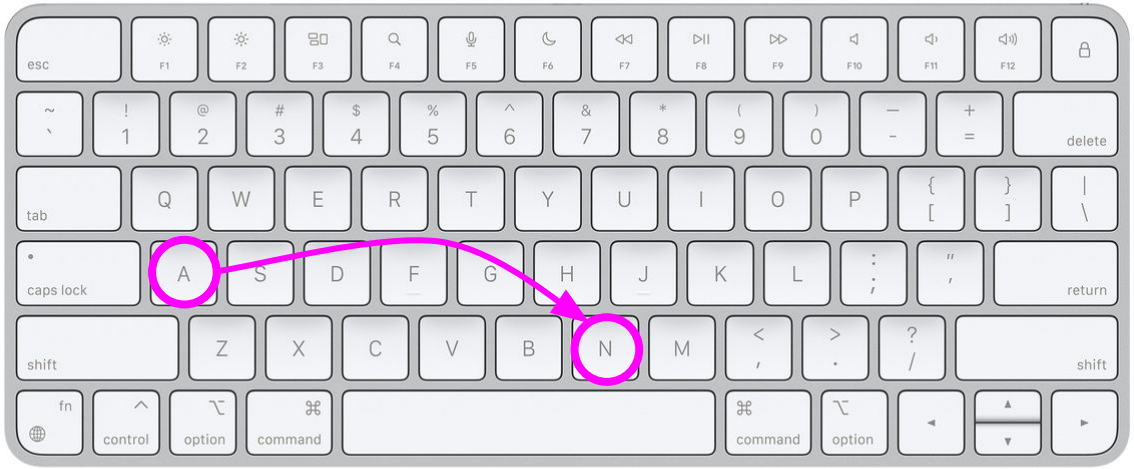
Tiempo de escritura de cada **dígrafo**

¿Un qué?

Dígrafo

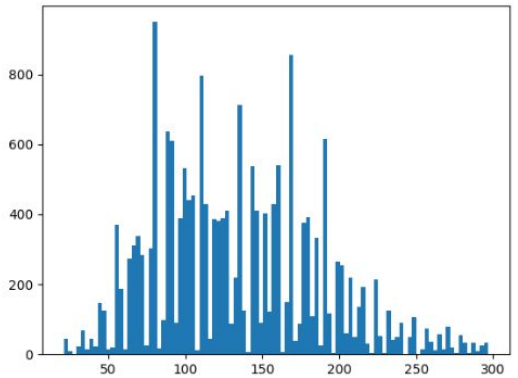


a-n

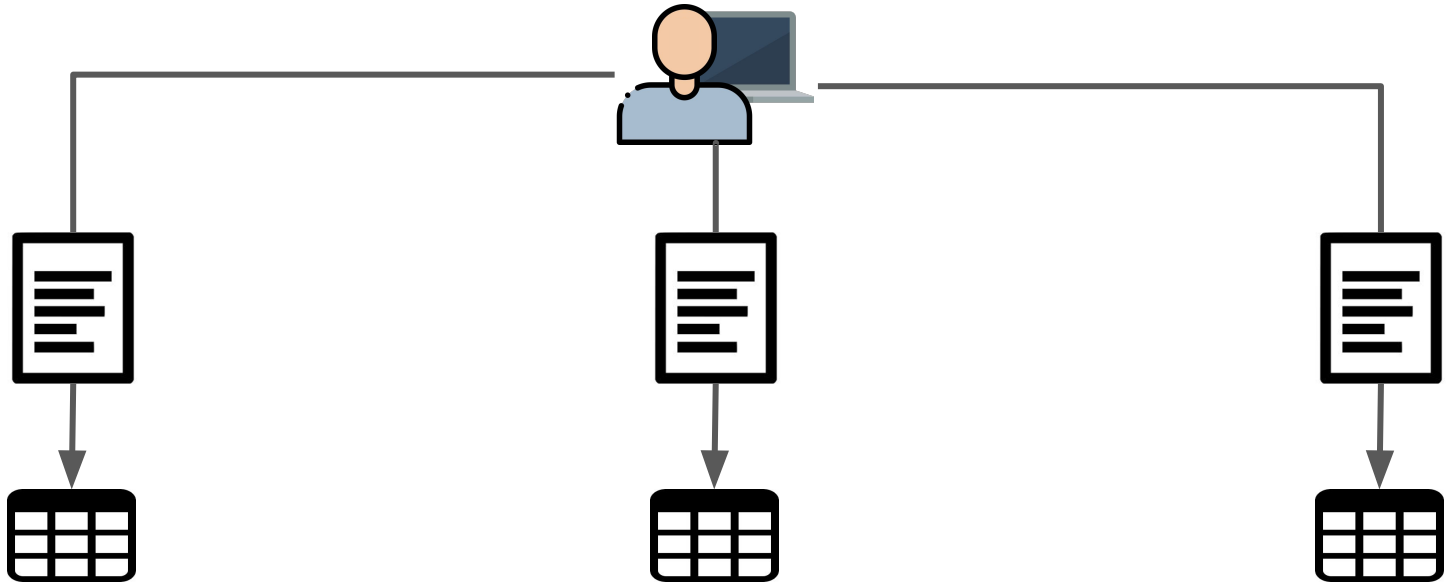

esc
F1
F2
F3
F4
F5
F6
F7
F8
F9
F10
F11
F12
delete
tab
Q
W
E
R
T
Y
U
I
O
P
caps lock
A
S
D
F
G
H
J
K
L
return
shift
Z
X
C
V
B
N
M
shift
fn
control
option
command
command
option

120 ms

	A	B	C	•••
A	120	140	59	•••
B	87	90	47	•••
C	26	102	95	•••
⋮	⋮	⋮	⋮	



Una tabla de tiempos por cada párrafo que se escribe en la web



Así cada registro de nuestro DataSet estaba compuesto por:

Usuario | tiempoAA | tiempoAB | tiempoAC |…

Y con esto entrenaremos a nuestro modelos

Preprocesamiento

Una persona puede escribir a **diferentes velocidades** en diferentes momentos

Un día, uno puede estar más **apurado**
Otro día, uno podría estar más **distraído**
etc …

Si capturamos la **velocidad** absoluta, probablemente tengamos **más información**

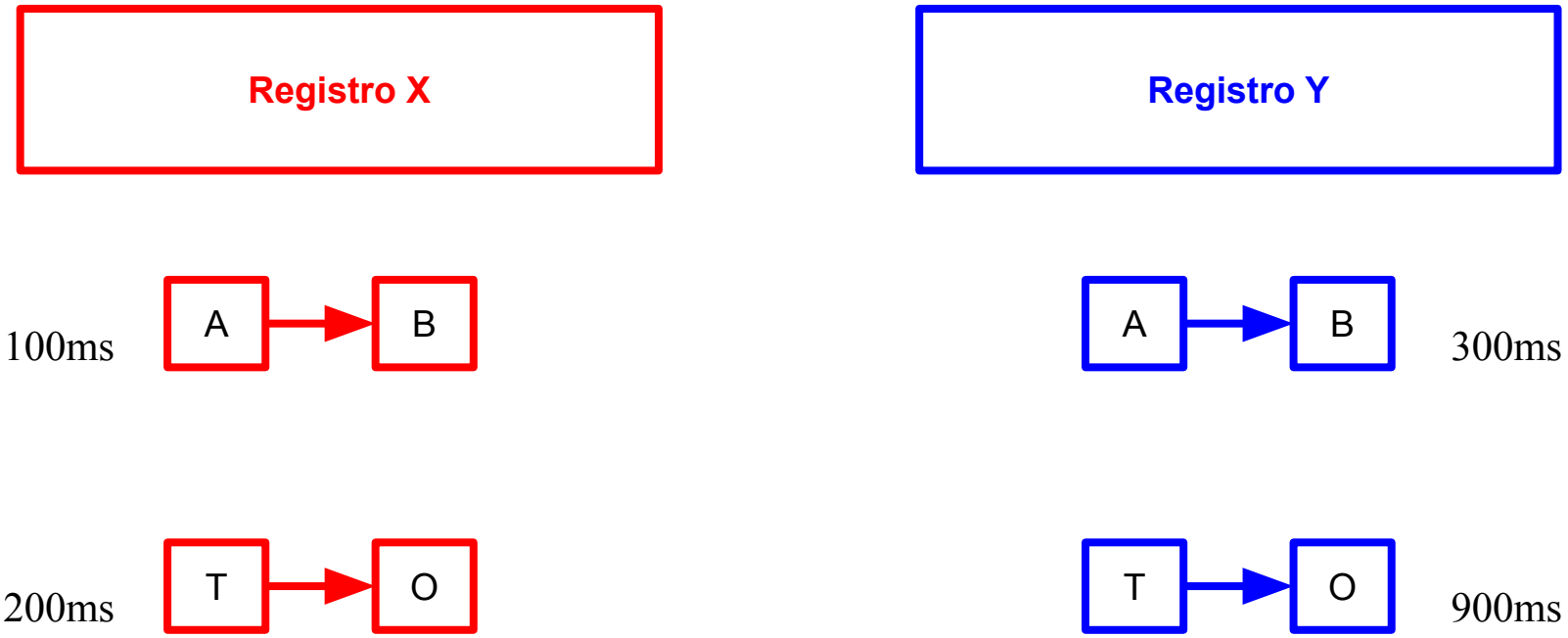
Pero **mayor caos**, requerirá de **más datos** de entrada, **más recursos** y métodos **más sofisticados**

Por ello, proponemos la siguiente hipótesis:

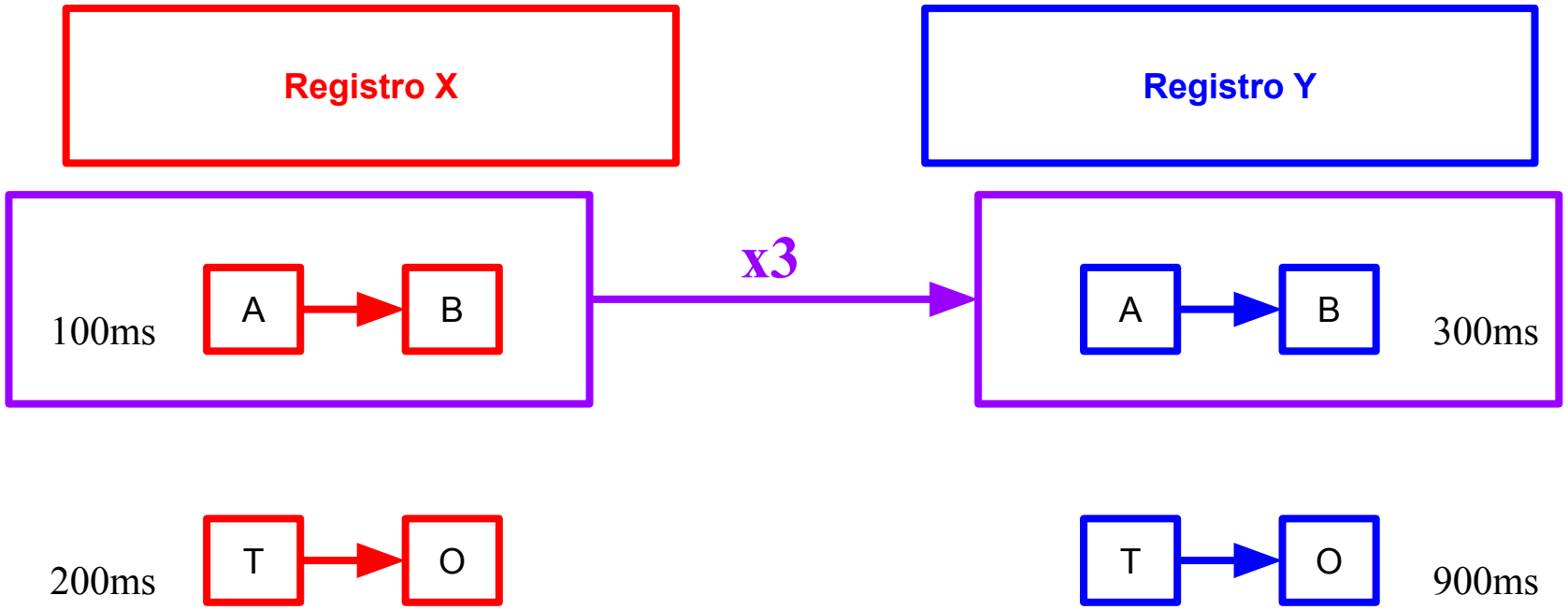
La diferencia de **velocidades** absolutas **no son** el **único patrón** que **revela** la identidad de la **persona**.

Pongamos un **ejemplo**

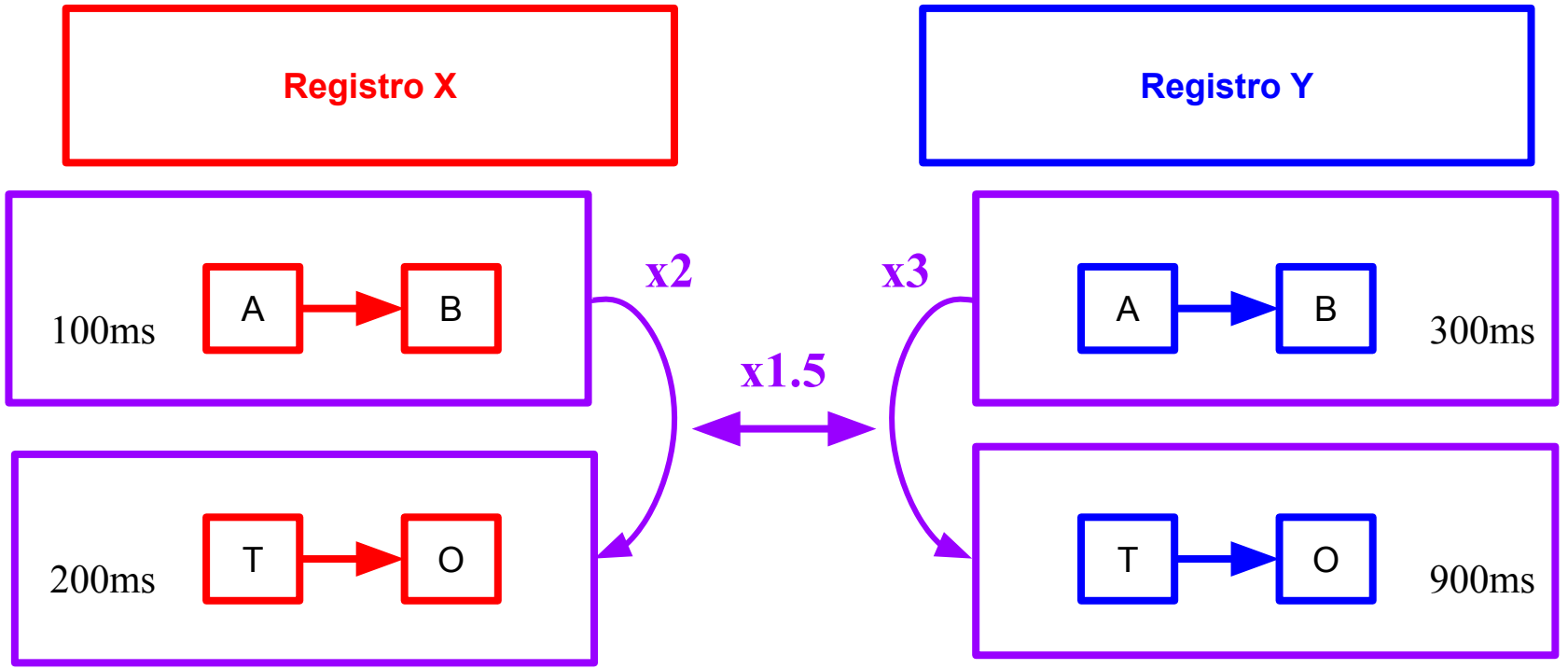
Supongamos dos registros y dos dígrafos



En lugar de potencialmente comparar …

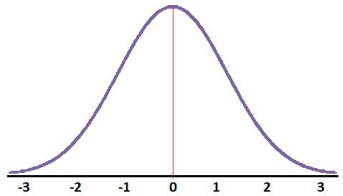


Buscamos comparar los cambios de proporcionalidades



Para lograrlo, aplicamos una **normalización** cada registro en función de **su propia distribución normal**

	A	B	C	...
A	120	140	59	...
B	87	90	47	...
C	26	102	95	...
⋮	⋮	⋮	⋮	



	A	B	C	...
A	1.3	1.5	0.3	...
B	0.7	0.8	0	...
C	-1	1.1	0.9	...
⋮	⋮	⋮	⋮	

Como consecuencia, se **reduce el impacto** de **diferentes velocidades** de escritura por una misma persona

Y se le da **más importancia** al **resto de patrones**

Método de solución

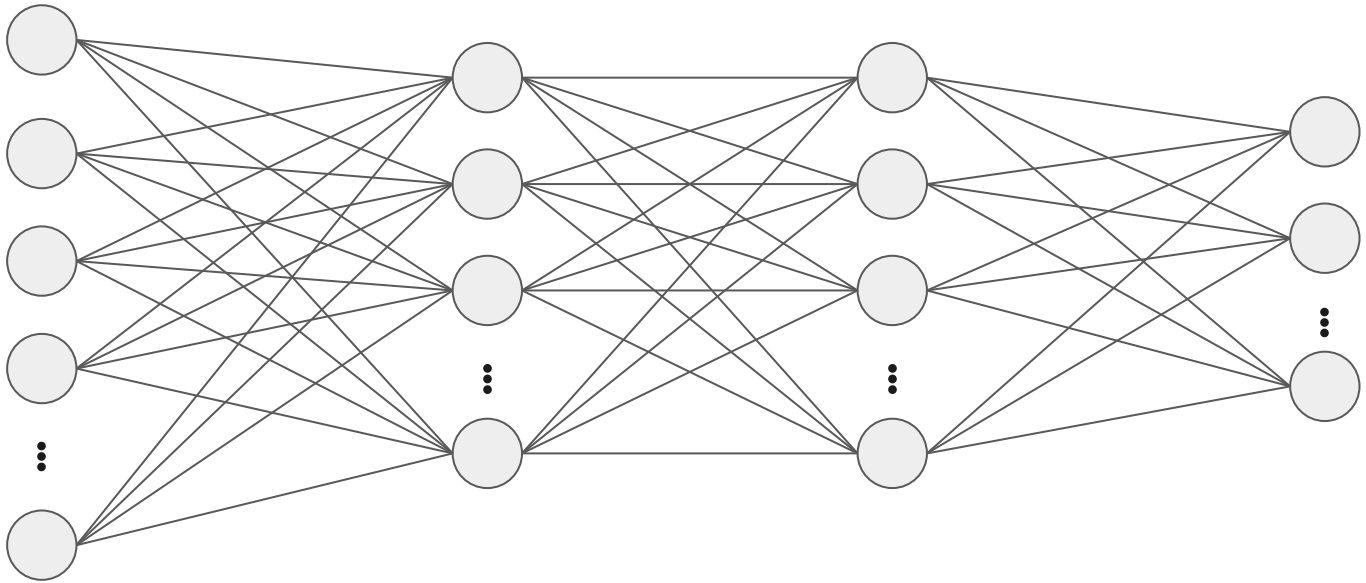
Clasificación con MLP y KNN

Arquitectura MLP

Input Layer

Hidden Layer

Output Layer



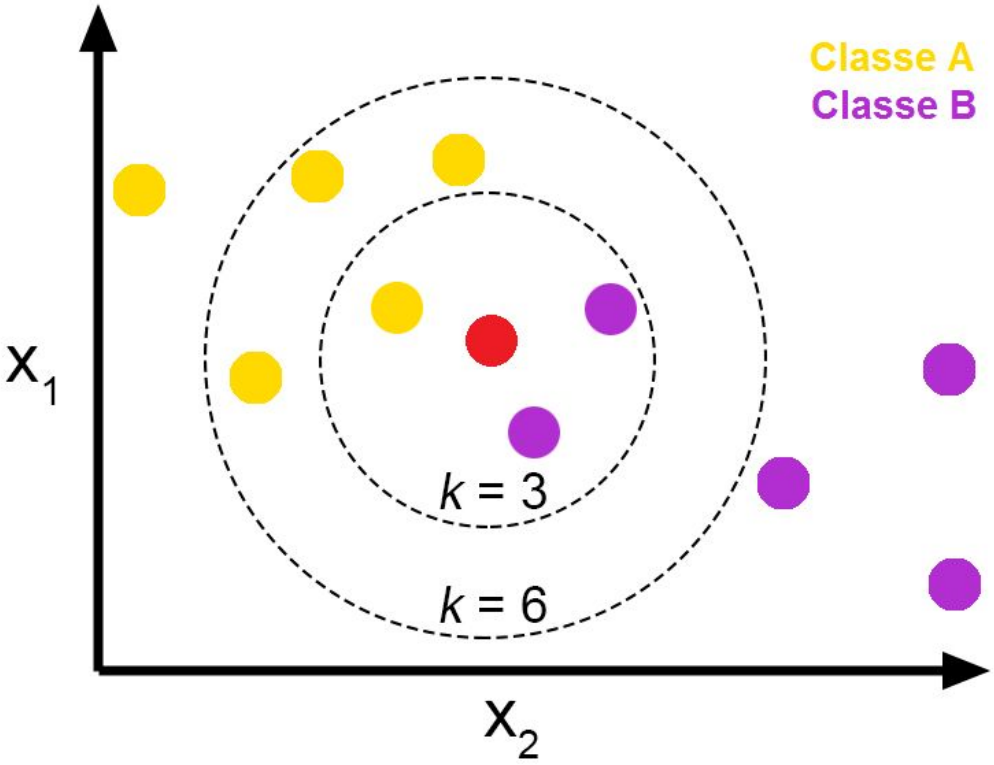
Una por cada dígrafo único existente

1024 neuronas ReLU

128 neuronas ReLU

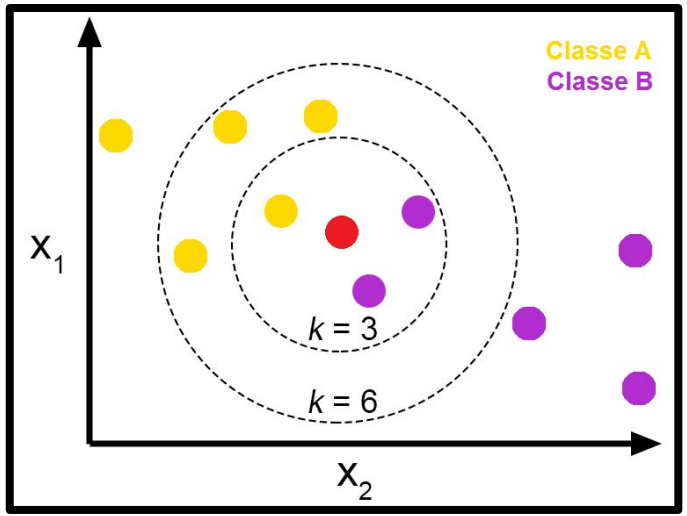
Una por cada label

KNN



Alrededor de 3000 dimensiones!

KNN



Reducción de dimensión

PCA (10 dim)

LDA (#labels-1 dim)

Experimentación y Resultados

Mientras experimentamos, vimos útil aplicar ciertos **filtros**.

Removimos las **tildes**
Removimos los dígrafos del tipo **“aa”** o **“dd”**
Removimos caracteres **no alfabéticos**
Removimos dígrafos con **“ñ”**

REAL
ACADEMIA
ESPAÑOLA

Por cuestiones de tiempo y la cantidad de datos que necesitábamos, se pudo estudiar a **2 personas** con **50 registros cada una**.

Donde cada registro es un párrafo de 3 a 5 oraciones.

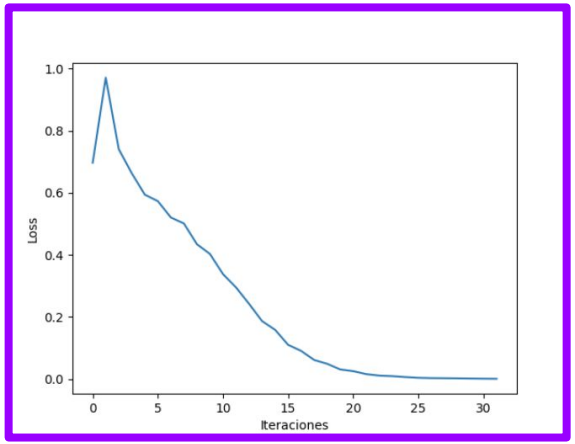
MLP

Con **70%** de data de entrenamiento y **30%** de validación se pudo llegar a un **accuracy de [70-90]%**

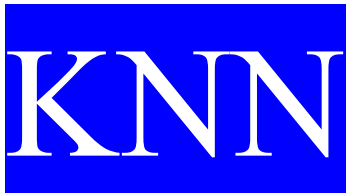
Epochs = 32

Batch size = 1 (SGD)

Learning rate = 0.001



Con esta configuración el loss ya convergía. **No aumentamos** los parámetros, para **evitar el overfitting**.


KNN

K Neighbors = 5

Distance metric = Euclidean

Neighbors weights = Uniform

K Neighbors = 5

Distance metric = Euclidean

Neighbors weights = Uniform

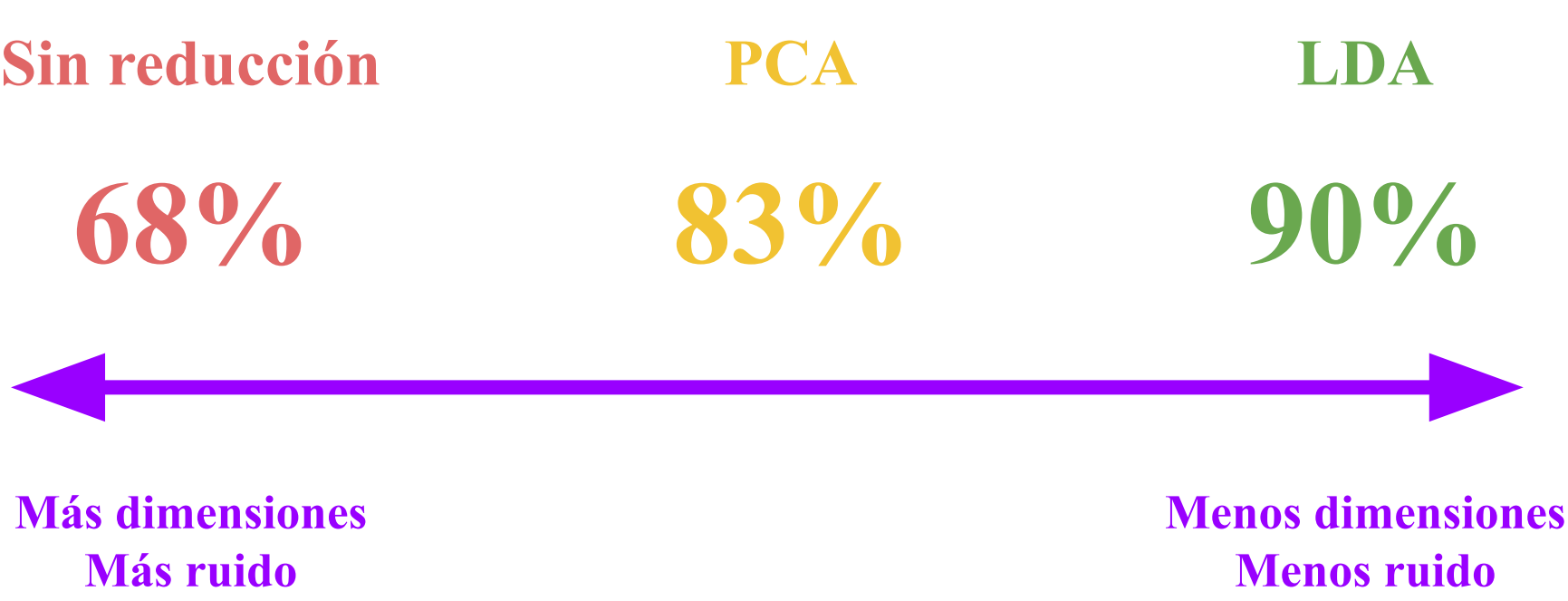
→ **Default**

+

K Fold = 5

Con **70%** de data de entrenamiento y **30%** de validación se pudo llegar a los siguientes resultados:

Sin reducción	PCA	LDA
68%	**83%**	**90%**


Sin reducción
PCA
LDA
68%
83%
90%
Más dimensiones
Más ruido
Menos dimensiones
Menos ruido

Sin reducción **PCA** **LDA**

68% **83%** **90%**

Más dimensiones
Más ruido

Menos dimensiones
Menos ruido

Existe un conjunto pequeño de digrafos, cuya variabilidad es tan alta que contiene suficiente información

Conclusiones

Pudimos llegar a una **precisión de 90%** para la identificación de **2 personas**.

Es un **punto de partida** para algo más interesante. Podría probarse en un **salón de clase**, con 20 o 30 personas y alertar posibles casos de **suplantación**.

En este caso, un método de machine learning **tradicional (KNN)** resultó más preciso que uno **más moderno (MLP).**

En este caso, un método de machine learning **tradicional (KNN)** resultó más preciso que uno **más moderno (MLP).**

En este caso, un método de machine learning **tradicional (KNN)** resultó más preciso que uno **más moderno (MLP).**

Guarda consistencia con el bajo **volúmen** de datos y su simple **estructura**

Guarda consistencia con el estado del arte

Otros intentos lograron **98% de accuracy** utilizando distancia **euclidiana** con un vector con los **25 dígrafos** más comunes

Identification Based on Typing Patterns Between Programming and Free Text

Petrus Peltola, Vilma Kangas, Nea Pirttinen, Henrik Nygren, Juho Leinonen
University of Helsinki
Helsinki, Finland
{petrus.peltola,vilma.l.kangas,nea.pirttinen,henrik.nygren,juho.leinonen}@helsinki.fi

Gracias