

The Value of Machine Learning in Privacy

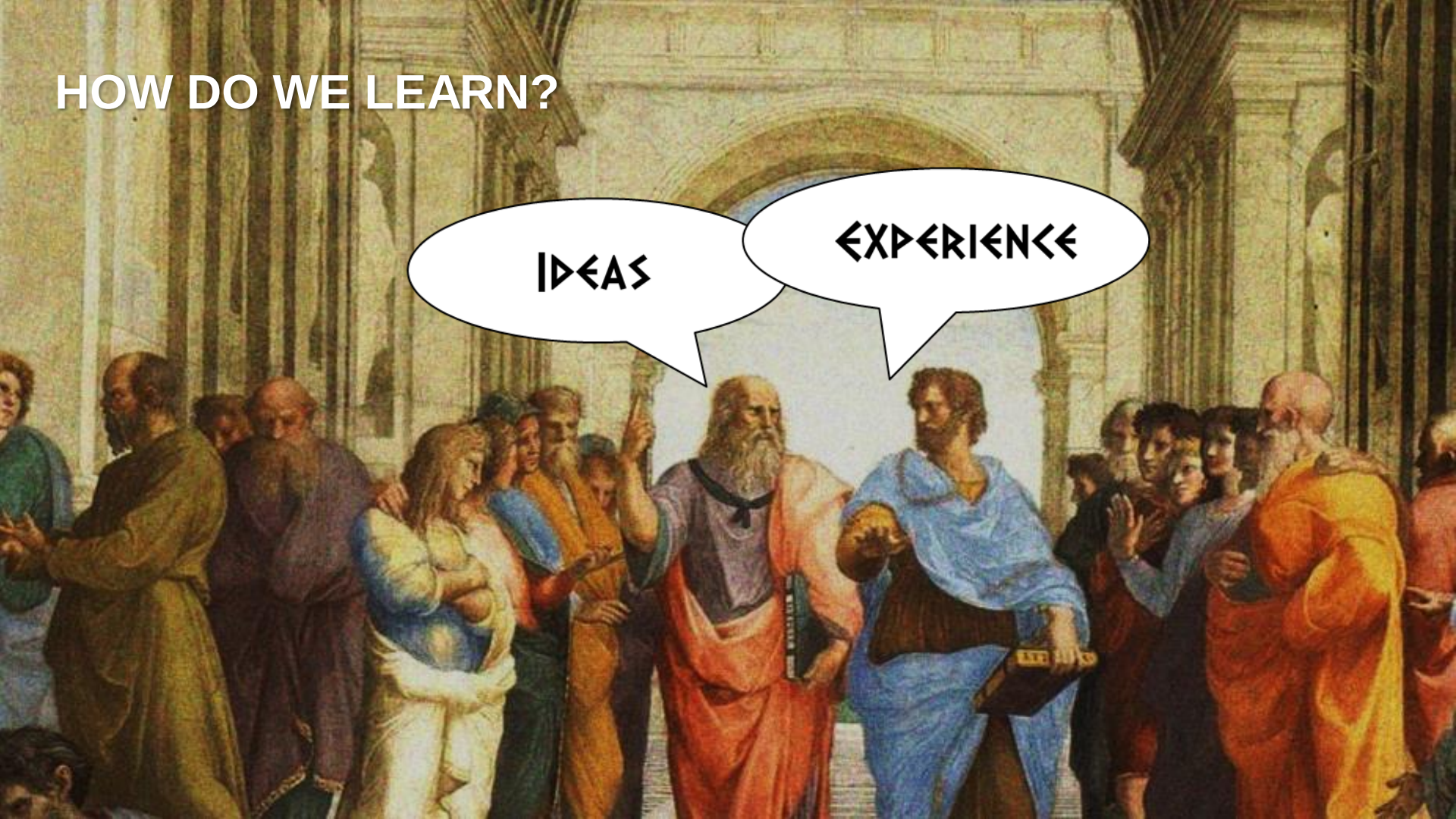
Results-oriented machine learning solution
in securing PII data anonymization



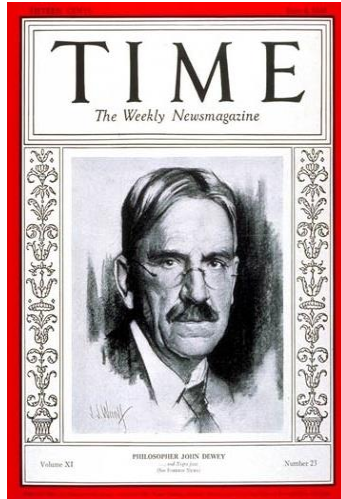
HOW DO WE LEARN?

IDEAS

EXPERIENCE



A RADICAL PRAGMATIC VISION



“

Reality is constantly changing and we learn best through applying our experiences and thoughts to problems, as they arise.

There is no absolute and unchanging truth. **Truth is what works.** Only those things that are experienced or observed are real.

”

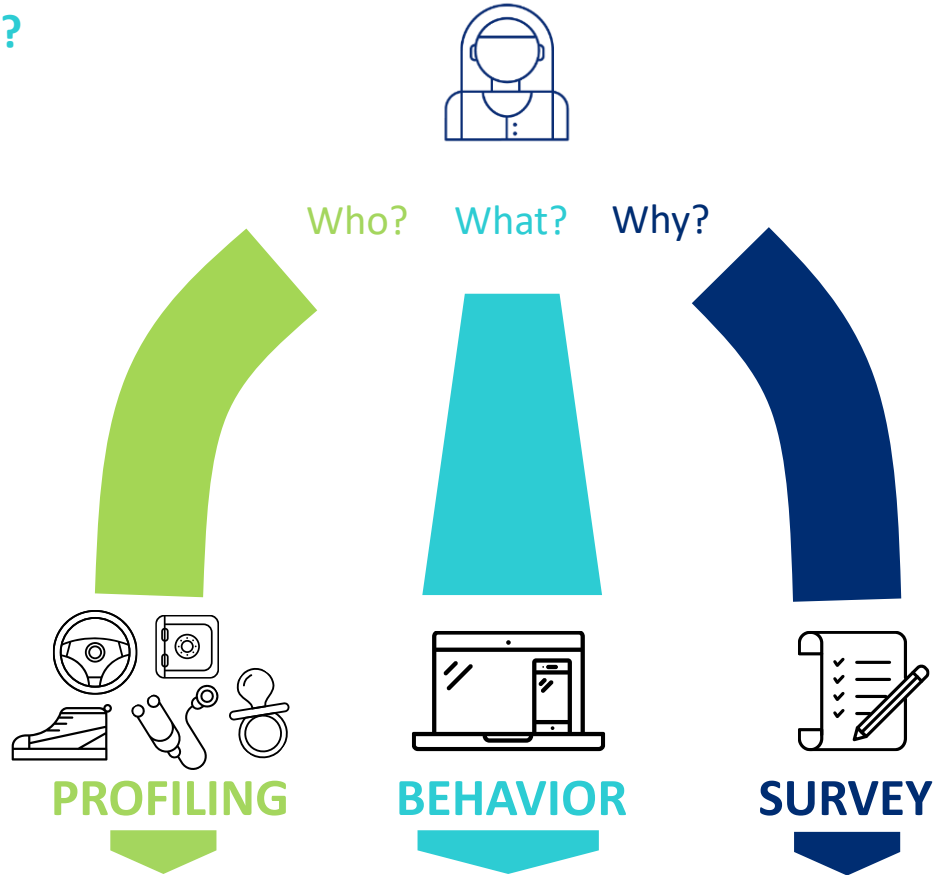
John Dewey

THE PROBLEM

Filtering PII from Online Behavioral Data

WHAT IS ONLINE BEHAVIORAL DATA?

Data on how people behave online,
collected through passive measurement
on an online panel: visited URL's, search
terms, app usage...



PRIVACY AND PERSONAL DATA

Personally identifiable information (PII)

“Is information that can be used on its own or with other information to identify, contact, or locate a single person, or to identify an individual in context”.

+ GDPR

+ ETHICS

UNIVERSAL DECLARATION OF HUMAN RIGHTS

(Article 12 of the 1948)

No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honor and reputation. Everyone has the right to the protection of the law against such interference or attacks.

BEHAVIORAL DATA IS PII

<https://es-la.facebook.com/John.Williams.friends/>

<https://www.google.com>

<http://greatoffers.com/campaign=38/ref=j.williams.88@gmail.com/en>

<https://www.theguardian.com/uk/commentisfree>

<https://twitter.com/jw88?lang=en>

<https://www.linkedin.com/in/1023223/>

...

BEHAVIORAL DATA IS PII

<https://es-la.facebook.com/John.Williams.friends/>

<https://www.google.com>

<http://greatoffers.com/campaign=38/ref=j.williams.88@gmail.com/en>

<https://www.theguardian.com/uk/commentisfree>

<https://twitter.com/jw88?lang=en>

<https://www.linkedin.com/in/1023223/>

...

<facebook.com/johnny.caldwell> (not anonymised) vs. <facebook.com/XXXXXXXXXX> (anonymised)

BEHAVIORAL DATA IS PII

<https://es-la.facebook.com/John.Williams.friends/>

<https://www.google.com>

<http://greatoffers.com/campaign=38/ref=j.williams.88@gmail.com/en>

<https://www.theguardian.com/uk/commentisfree>

<https://twitter.com/jw88?lang=en>

<https://www.linkedin.com/in/1023223/>

...

[https://www.amazon.es/ECOVACS-Robotics-DEEBOT-N79S-](https://www.amazon.es/ECOVACS-Robotics-DEEBOT-N79S-Smartphone/dp/B077HW9XM7/ref=gbph_img_s-3_3e42_df362a0f?smid=A1AT7YVPFBWXBL&pf_rd_p=12028e22-b41e-4d6b-b6ca-9e691c4e3e42&pf_rd_s=slot-3&pf_rd_t=701&pf_rd_i=gb_main&pf_rd_m=A1AT7YVPFBWXBL&pf_rd_r=XG85Q8REGC8K4RXN0AKC)

[Smartphone/dp/B077HW9XM7/ref=gbph_img_s-](https://www.amazon.es/ECOVACS-Robotics-DEEBOT-N79S-Smartphone/dp/B077HW9XM7/ref=gbph_img_s-3_3e42_df362a0f?smid=A1AT7YVPFBWXBL&pf_rd_p=12028e22-b41e-4d6b-b6ca-9e691c4e3e42&pf_rd_s=slot-3&pf_rd_t=701&pf_rd_i=gb_main&pf_rd_m=A1AT7YVPFBWXBL&pf_rd_r=XG85Q8REGC8K4RXN0AKC)

[3_3e42_df362a0f?smid=A1AT7YVPFBWXBL&pf_rd_p=12028e22-b41e-4d6b-](https://www.amazon.es/ECOVACS-Robotics-DEEBOT-N79S-Smartphone/dp/B077HW9XM7/ref=gbph_img_s-3_3e42_df362a0f?smid=A1AT7YVPFBWXBL&pf_rd_p=12028e22-b41e-4d6b-b6ca-9e691c4e3e42&pf_rd_s=slot-3&pf_rd_t=701&pf_rd_i=gb_main&pf_rd_m=A1AT7YVPFBWXBL&pf_rd_r=XG85Q8REGC8K4RXN0AKC)

[b6ca-9e691c4e3e42&pf_rd_s=slot-](https://www.amazon.es/ECOVACS-Robotics-DEEBOT-N79S-Smartphone/dp/B077HW9XM7/ref=gbph_img_s-3_3e42_df362a0f?smid=A1AT7YVPFBWXBL&pf_rd_p=12028e22-b41e-4d6b-b6ca-9e691c4e3e42&pf_rd_s=slot-3&pf_rd_t=701&pf_rd_i=gb_main&pf_rd_m=A1AT7YVPFBWXBL&pf_rd_r=XG85Q8REGC8K4RXN0AKC)

[3&pf_rd_t=701&pf_rd_i=gb_main&pf_rd_m=A1AT7YVPFBWXBL&pf_rd_r=XG85Q8REGC8K4RXN0AKC](https://www.amazon.es/ECOVACS-Robotics-DEEBOT-N79S-Smartphone/dp/B077HW9XM7/ref=gbph_img_s-3_3e42_df362a0f?smid=A1AT7YVPFBWXBL&pf_rd_p=12028e22-b41e-4d6b-b6ca-9e691c4e3e42&pf_rd_s=slot-3&pf_rd_t=701&pf_rd_i=gb_main&pf_rd_m=A1AT7YVPFBWXBL&pf_rd_r=XG85Q8REGC8K4RXN0AKC)

TWO DIFFERENT STRATEGIES

“There is a clear rule that perfectly defines which elements of a URL are PII. We just need to find it!”

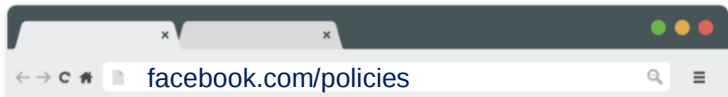
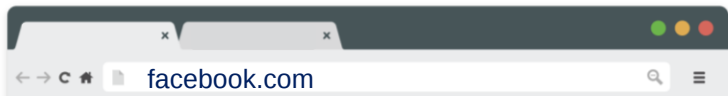


“Forget about simple rules. Learn from experience: look at as many examples as possible, define as many rules as needed”.

THE PLATONIC SOLUTION (Plato)

THE “PLATONIC” RULE

The web was built to share content to many. URLs describing public content must appear in the navigation data set of **several people**.



Contrarily, when a piece of a URL (“token”) is found only in the navigation of **a single individual**, it must be **PII**.



K-ANONYMIZATION FOR TABULAR DATA

“... is a property of certain anonymized data. This property exists if the information of each person cannot be distinguished from at least $k-1$ individuals whose information also appear in the data”.

Sweeney, 2002

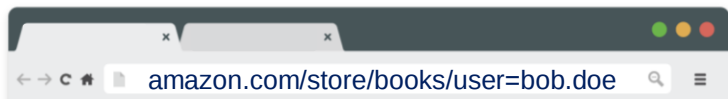
	G	Age	ZIP Code
Dave	M	27	27031
Bob	M	28	27032
Helen	F	30	27134
Charlie	M	29	27121
Alice	F	31	27031
Eve	F	30	27148



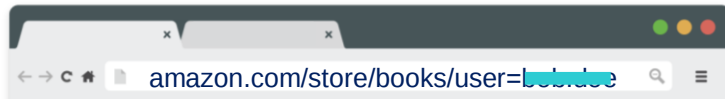
G	Age	ZIP Code
M	27-30	27***
M	27-30	27***
F	30-31	271**
M	27-30	27***
F	30-31	271**
F	30-31	27***

K-ANONYMIZATION FOR URLS

If any part of the URL is personal (a parameter, a name, etc.), it is **very unlikely that this URL has more than one visitor**.



Therefore, our masking procedure consists in generalizing URL components to ensure that **at least k participants** in our data set have visited the generalized URL.



TOO MUCH “PLATONIC”: DRAWBACKS

- 1 Dependent on the size and representativity of the dataset used for comparison.
- 2 If we could compare each URL with all the URLs in the Internet, the algorithm would perfectly identify PII.
- 3 In practice, too much valuable data was masked.

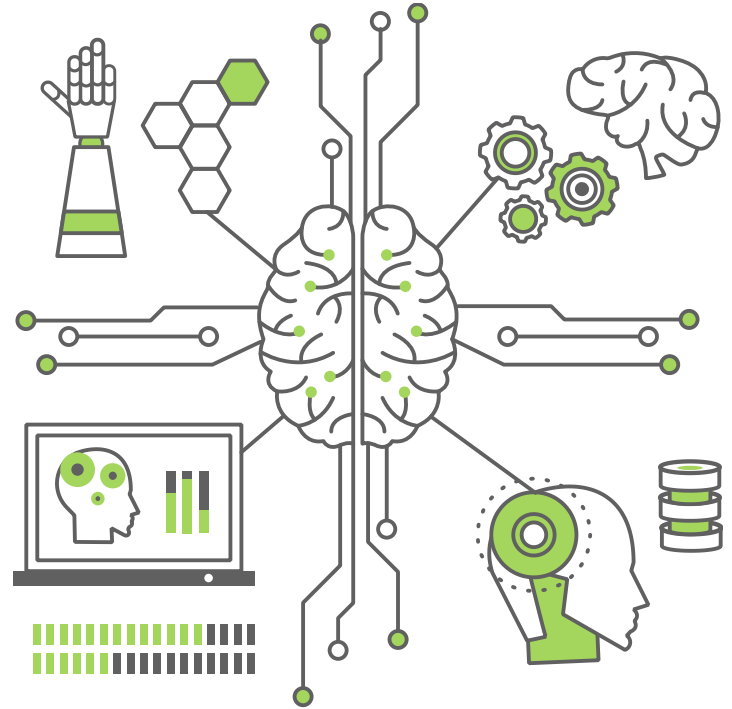
www.google.com/search?source=hp&ei= [redacted] &q= [redacted]
[redacted] &oq= [redacted] &gs_l= [redacted]
www.google.com/search?ei= [redacted] &q= [redacted] &oq= [redacted]
[redacted] &gs_l= [redacted]
www.drogariasapaulo.com.br [redacted] p
www.google.com/search?rlz= [redacted] &ei= [redacted] &q= [redacted]
[redacted] &oq= [redacted] &gs_l= [redacted]
www.clasf.com.br/q/ [redacted]
www.planetabrasileiro.com/salvador/ruas/ [redacted] [redacted]
www.google.com/search?q= [redacted] &sa=X&ved= [redacted] &
biw=1366&bih=637
www.google.com/search?source=hp&ei= [redacted] &q=whats&
oq=whats&gs_l= [redacted]
www.google.com/maps/uv?hl=pt-
BR&pb= [redacted] &imagekey= [redacted] &sa=X&ved= [redacted]
web.whatsapp.com/
www.casasbahia.com.br/BelezaSaude/saude/ProdutosOrtope
dicos/?Filtro= [redacted]
www.google.com/search?rlz= [redacted] &source=hp&ei= [redacted]
[redacted] &q= [redacted] &oq= [redacted] &gs_l= [redacted]

THE PRAGMATIC SOLUTION (Aristo

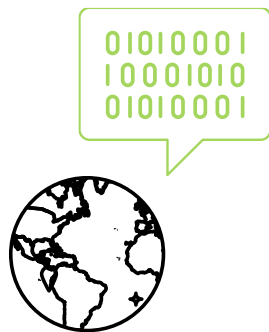
THE “PRAGMATIC” RULE

There is no perfect rule, we can only learn from experience and a lot of imperfect rules.

Machine Learning is about learning from many cases. Train a Machine Learning algorithm to recognize patterns present in PII.



TRAINING DATA



1 day of browsing data from 200 participants from 4 countries:



tokenID	urlID	url	token	PII?
000001	001	twitter.com/jw88?lang=en	jw88	Yes
000002	001	twitter.com/jw88?lang=en	lang	No
000003	001	twitter.com/jw88?lang=en	en	No

Split URLs into tokens (50K URLs -> 200K tokens*)
Each token was manually classified as being PII (or not)

* after removing parameters

Feature engineering

We built 287 features from the data. The ML solution must select the right ones to identify PII.

Example: twitter.com/jw88?lang=en

Relative to the URL of the token

Top 20 most visited domains: **YES**

Category: **INTERNET SERVICE**

Subcategory: **SOCIAL NETWORK**

Relative position: **2**

Absolute position: **14**

Length of URL: **24**

Proportion of symbols: **0.833**

Specific to the token itself

Length of token: **4**

Parameter: **NO**

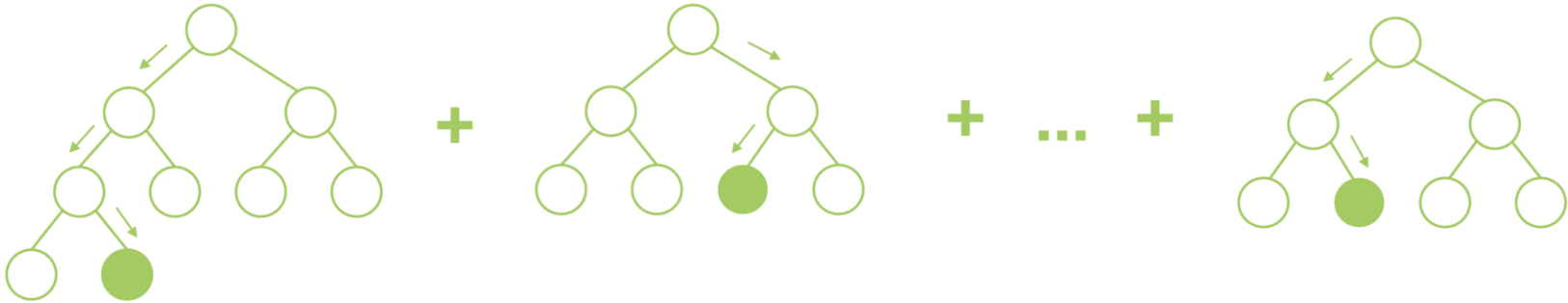
Value: **YES**

Proportion of vowels: **0**

Proportion of uppercase: **0**

Proportion of lowercase: **1**

MODEL



Rule-based methods: complex hierarchy of if-then statements. Each predictor is evaluated through this hierarchy based on its association with the outcome.

Gradient Boosting: a general method of producing a very accurate prediction rule by combining rough and moderately inaccurate rules of thumb .

EVALUATION METRICS

Select the **metric** to optimize.

Select a **threshold** in the classification criteria: are false positives hurting you as much as false negatives?

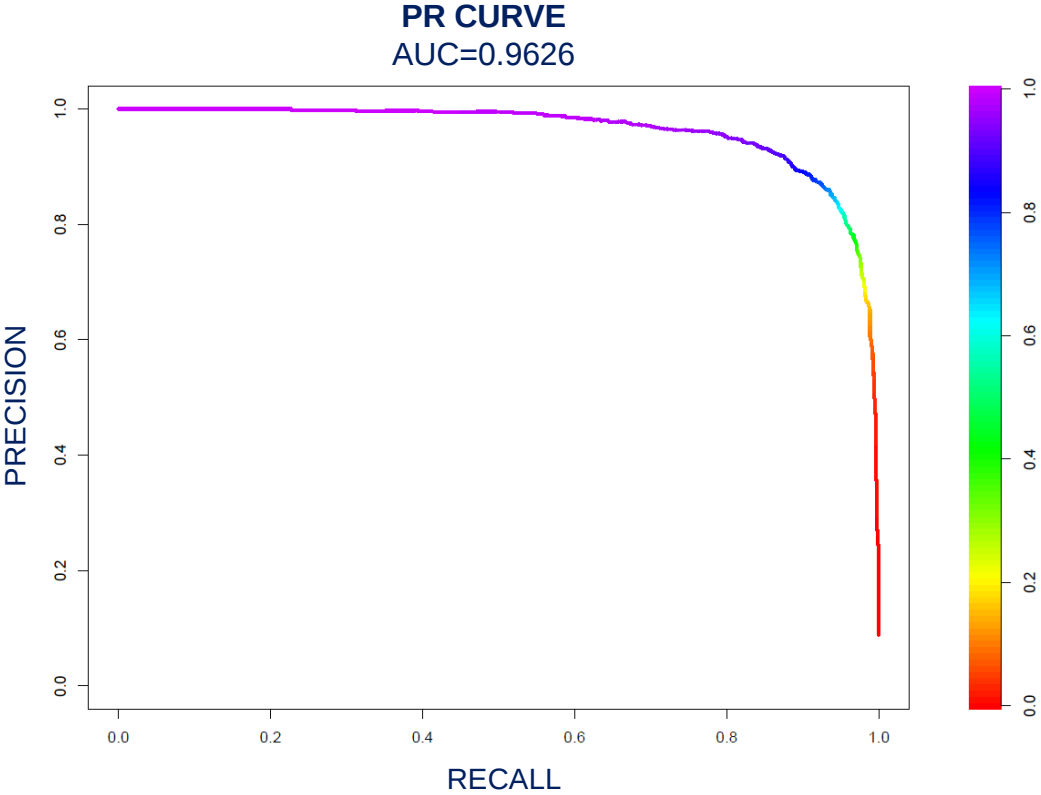
		PREDICTED VALUES	
		0: NOT PII	1: PII
TRUE VALUES	0: NOT PII	TRUE NEGATIVES	FALSE POSITIVES
	1: PII	FALSE NEGATIVES	TRUE POSITIVES

$$\text{FNR} = \frac{\text{FN}}{\text{TP} + \text{FN}}$$

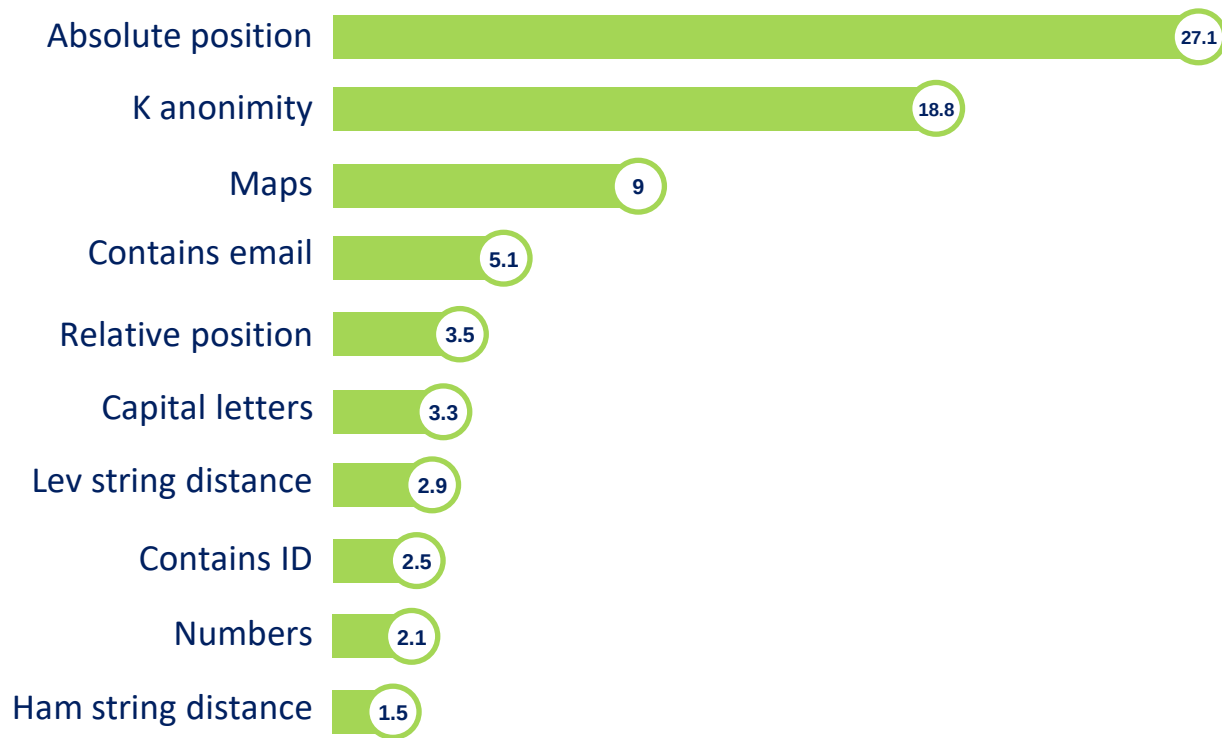
$$\text{PRECISION} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{RECALL} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

RESULTS: CLASSIFICATION PERFORMANCE

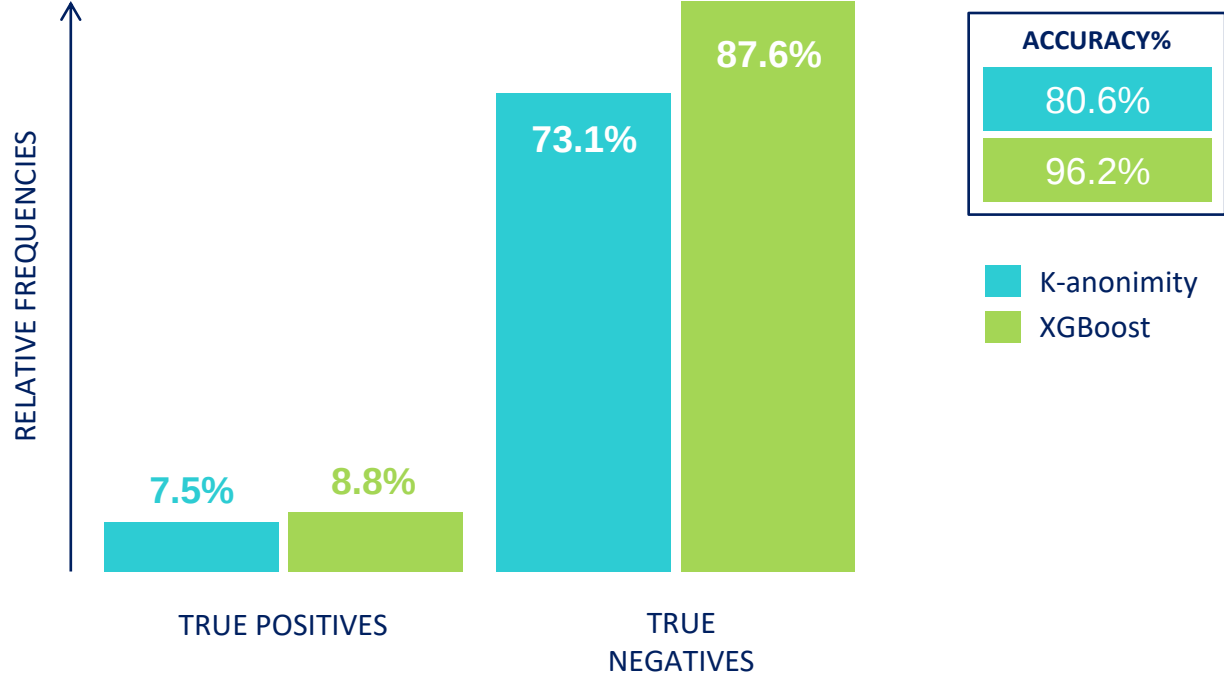


RESULTS: RELEVANT FEATURES

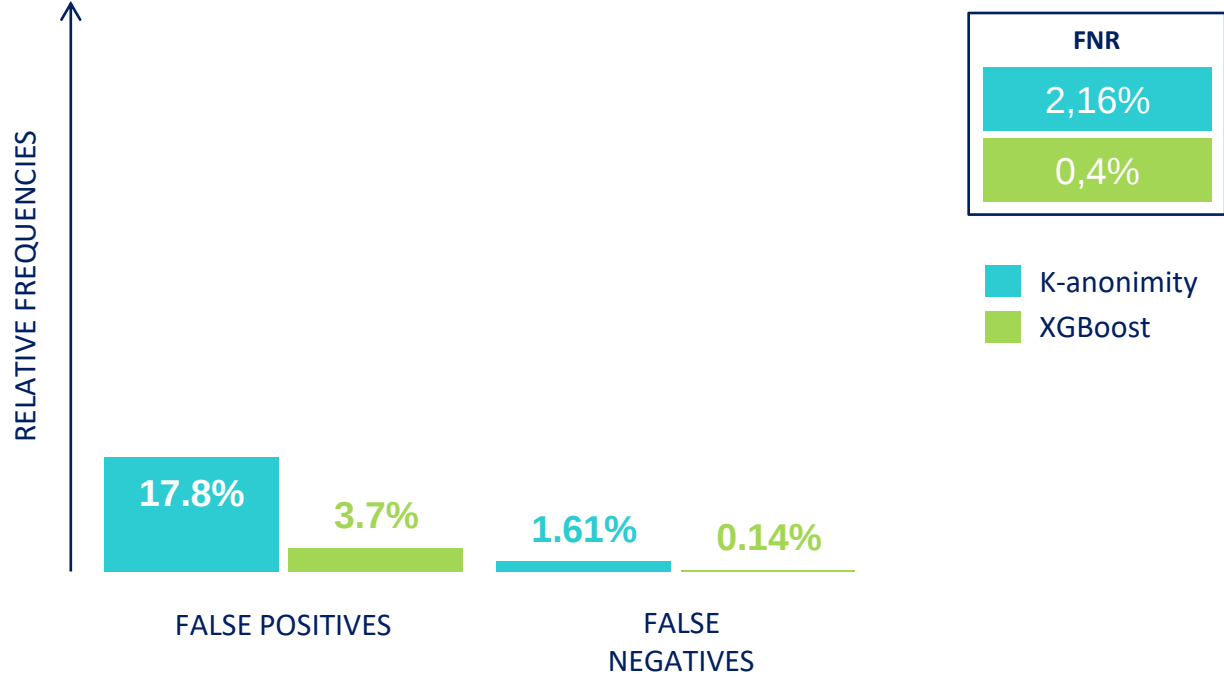


AND THE WINNER IS...

COMPARED PERFORMANCE: SUCCESSES

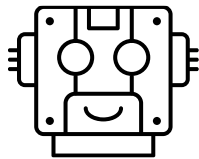


COMPARED PERFORMANCE: FAILURES



KEY LEARNINGS

WHAT DID WE LEARN FROM THIS EXPERIENCE?



1

Artificial Intelligence **is about using** huge computational power to learn from experience = pragmatic approach.



2

“Great ideas” still play a role:
1. Inspiring ML approaches, and;
2. Guiding the development.

Thank you!

**SPAIN**

iberia@netquest.com

BRAZIL

atendimento@netquest.com

EUROPE & ASIA

rfqinternational@netquest.com

SOUTHERN CONE

cs.sales@netquest.com

PORTUGAL

portugal@netquest.com

COLOMBIA

colombia@netquest.com

USA & CANADA

americas@netquest.com

MEXICO

latamnorte@netquest.com

www.netquest.com



genuine data