simMachines
SIMILARITY SEARCH & PATTERN RECOGNITION

Similarity

One data tool to rule them all

Why Data Science Jobs Are in High Demand

- Analysis by Harvard Extension Hub (link)
  - More data that we can consume
  - Managing it requires skilled individuals
  - By 2018, a shortage of 190,000 data scientists is predicted by McKinsey

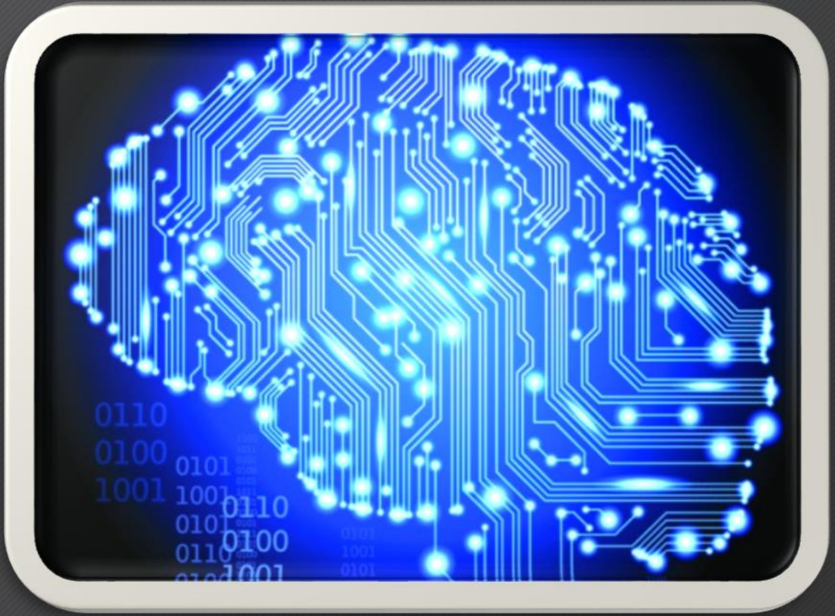
Prasanta Chandra Mahalanobis

"Practitioners with strong programming skills who can build and interpret mathematical models, and communicate the results in a meaningful way have a promising future in any arena."

simMachines
SIMILARITY SEARCH & PATTERN RECOGNITION

Data Science Jobs

- According to NetworkWorld: (link)
  - Data Scientists are elusive unicorns
  - 36,000 openings at 6,000 companies
  - Salaries: $200,000 - $300,000
  - What happens with the rest of the world?
  - What happens with cancer research?
  - What happens with your company?



Data

Similarity at the Supermarket

Discover

Recommend

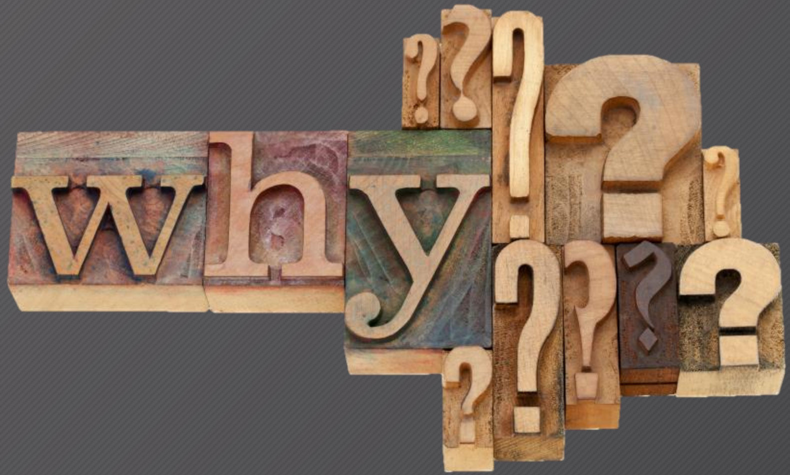
Predict

Tomato

simMachines
SIMILARITY SEARCH & PATTERN RECOGNITION

Why Nearest Neighbor (NN)?

- Why Nearest Neighbor (NN) ?
  - Because it can discover, recommend, predict, classify
  - The more data it has, the better it predicts
  - It is the only machine learning method understandable by the general population
  - It can do much more than traditional Data Mining and BI



About

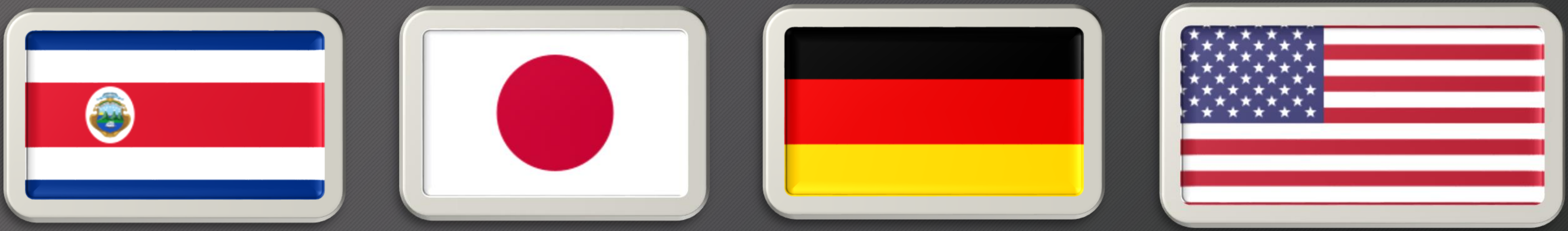
- Founder + CEO + Data Scientist @ simMachines
- Postdoc @ Max Planck Institute for Molecular Biomedicine, Germany
- PhD @ Kyutech（九州工業大学）(Pattern Recognition, ML), Japan
- Engineer @ Intel
- BsC @ Instituto Tecnológico de Costa Rica
- Linkedin profile

simMachines
SIMILARITY SEARCH & PATTERN RECOGNITION

About simMachines

- Funded December 2011
- Support from:
  - E&Y
  - Arch Grants 2012 + Follow-up $150,000
  - CaraoV
  - Plug and Play
  - CONICIT + MICIT
- 6 Full-Time Data Scientists + 2 Bizdevs
- Partners: Prio, Focus IS, Singularities, Safetdoc, JAS Global Advisors
- Offices: St. Louis MO USA, Santo Domingo, Costa Rica
- Customers in: Latin America, US, Germany, UK,

simMachines: a tale of 4 continents



simMachines
SIMILARITY SEARCH & PATTERN RECOGNITION

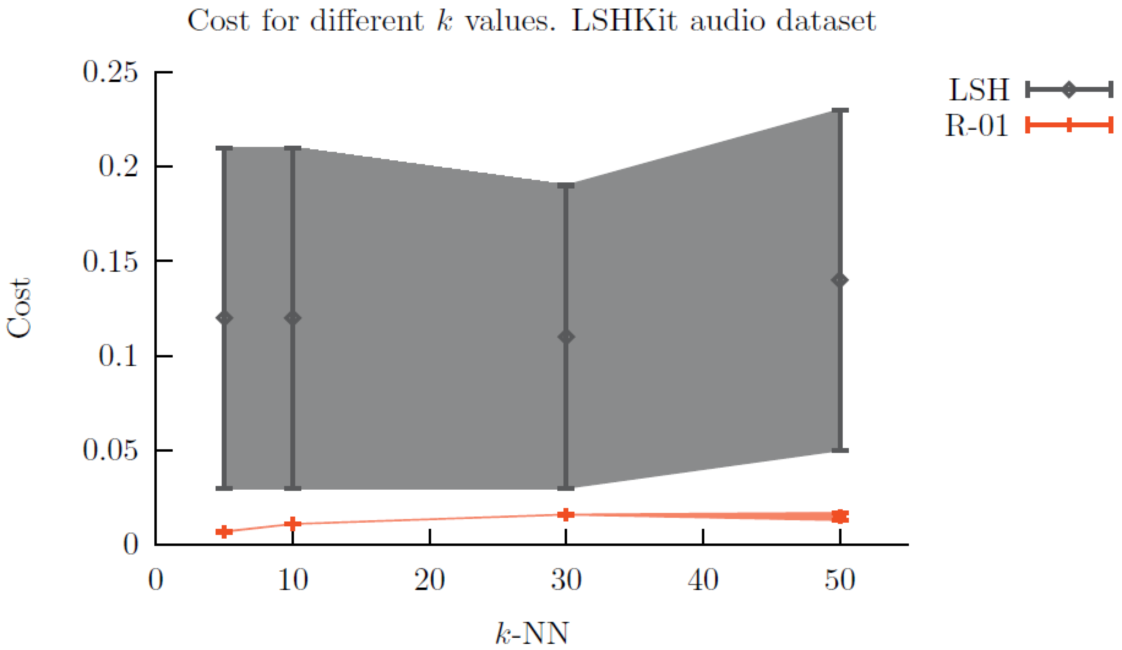
Our Technology, R-01 Similarity Index

Fast Facts:

Our tech is 10X Faster than MIT's LSH (Locality Sensitive Hashing).

LSH is like driving a car that instead of wheels has hexagons.

Our tech is smooth, fluid and fast, a car with proper, racing wheels.



simMachines
SIMILARITY SEARCH & PATTERN RECOGNITION

Scalability (R-01)

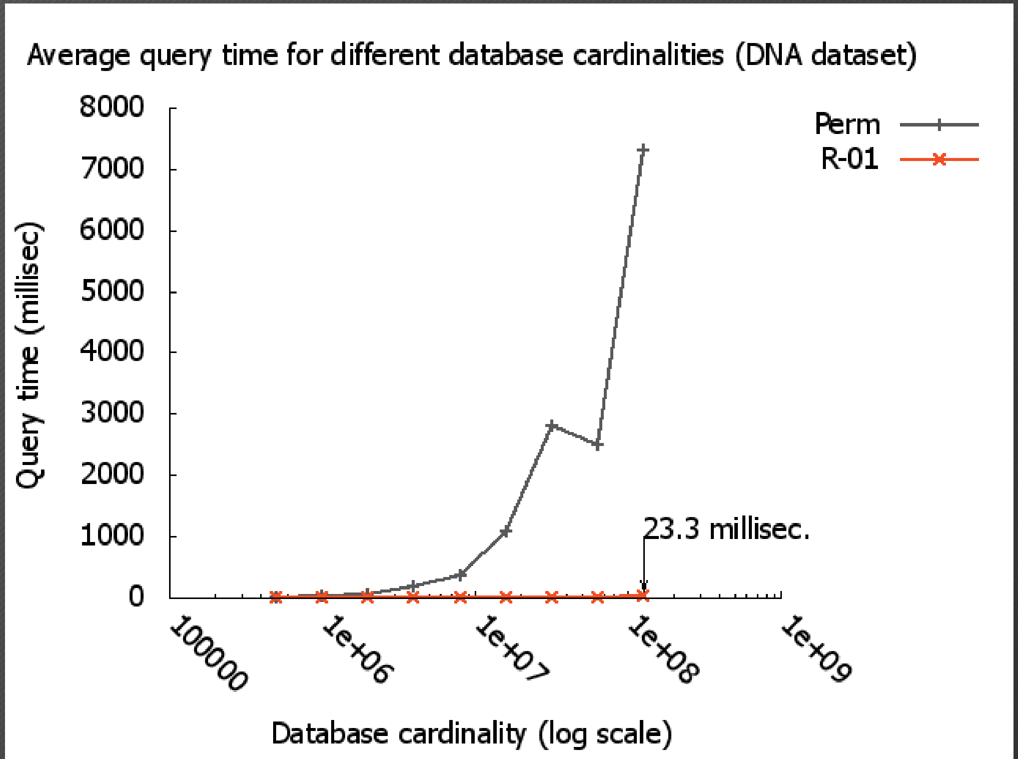
Comparison of our method (R-01) against the Permutation strategy of Amato et al.

120 million strings are inserted and query time is measured and averaged.

Queries remain under 23 millisec. In a nicely flat pattern.

Many more experiments and comparisons here:

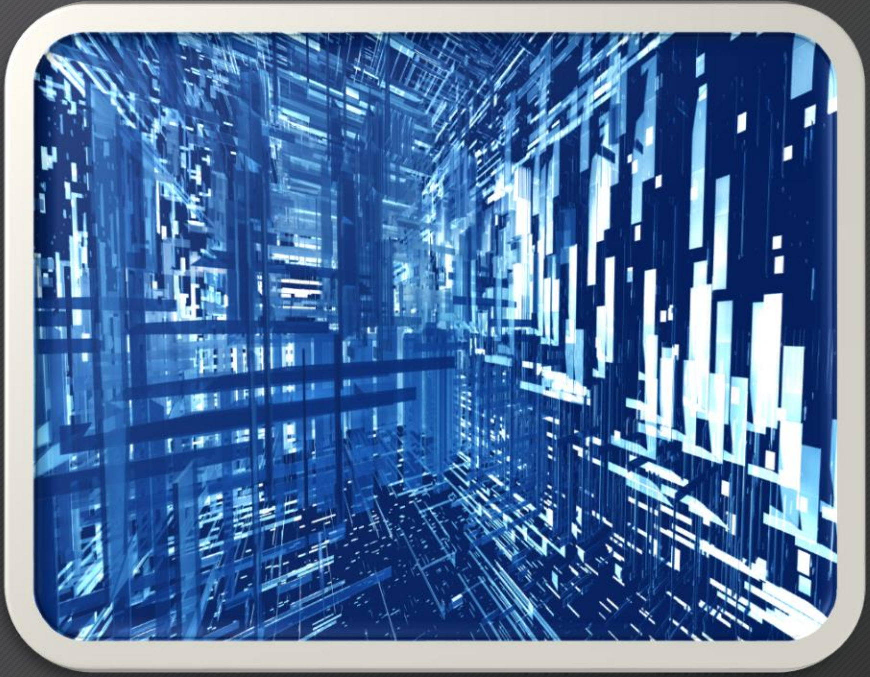
Benchmarks



Experiments executed on a Laptop, 1 CPU

Dense Nearest Neighbor (Dense-NN)

- Distance is a single quantity (uni-dimensional)
- Dense-NN is a precise combination of the following:
  - Measure different distance functions on the whole object to obtain a more clear view of the similarity
  - Dynamic dimensionality reduction: Features weight differently for each object.



Density

Our Cloud

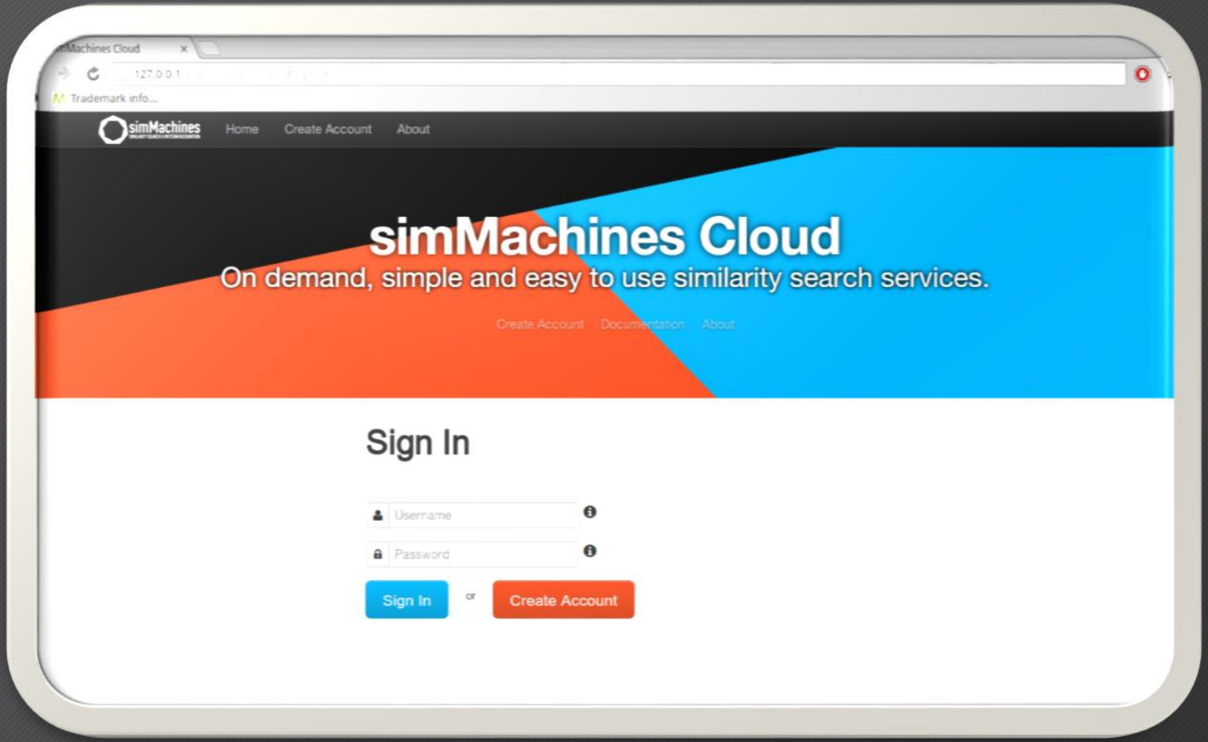
First similarity engine on the cloud

Elastic, failover, parallel.

Trivial to use by even entry-level developers.

1) Load file
2) Start an angel (prediction service)
3) Make predictions

Easier than R, Hadoop and Pivot tables and pretty much every package out there.



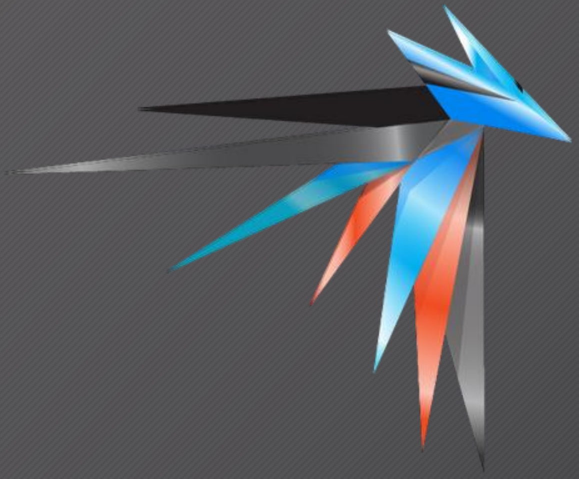
simMachines Cloud

On demand, simple and easy to use similarity search services.

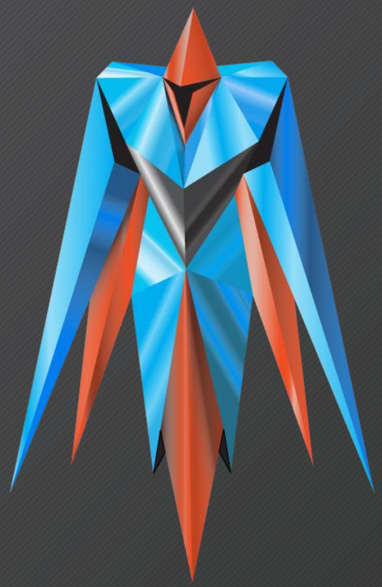
Sign In

Multiple Cloud Products (more coming in)

Gaghiel:
Dense NN
Recommendation engine
Tutorial



Ramiel:
Next generation similarity search
Tutorial



Leliel:
Dense NN classifier
Tutorial

Demo site: http://23.253.135.216:8080/Cloud-1.0.0.1

More products in the pipeline!

simMachines
SIMILARITY SEARCH & PATTERN RECOGNITION

Cloud Benefits

- Solve one problem at a time
- No need to pay for a data scientist
- A young developer is able to:
  - Use our cloud
  - Integrate
  - Discover, recommend, predict

Cloud

simMachines
SIMILARITY SEARCH & PATTERN RECOGNITION

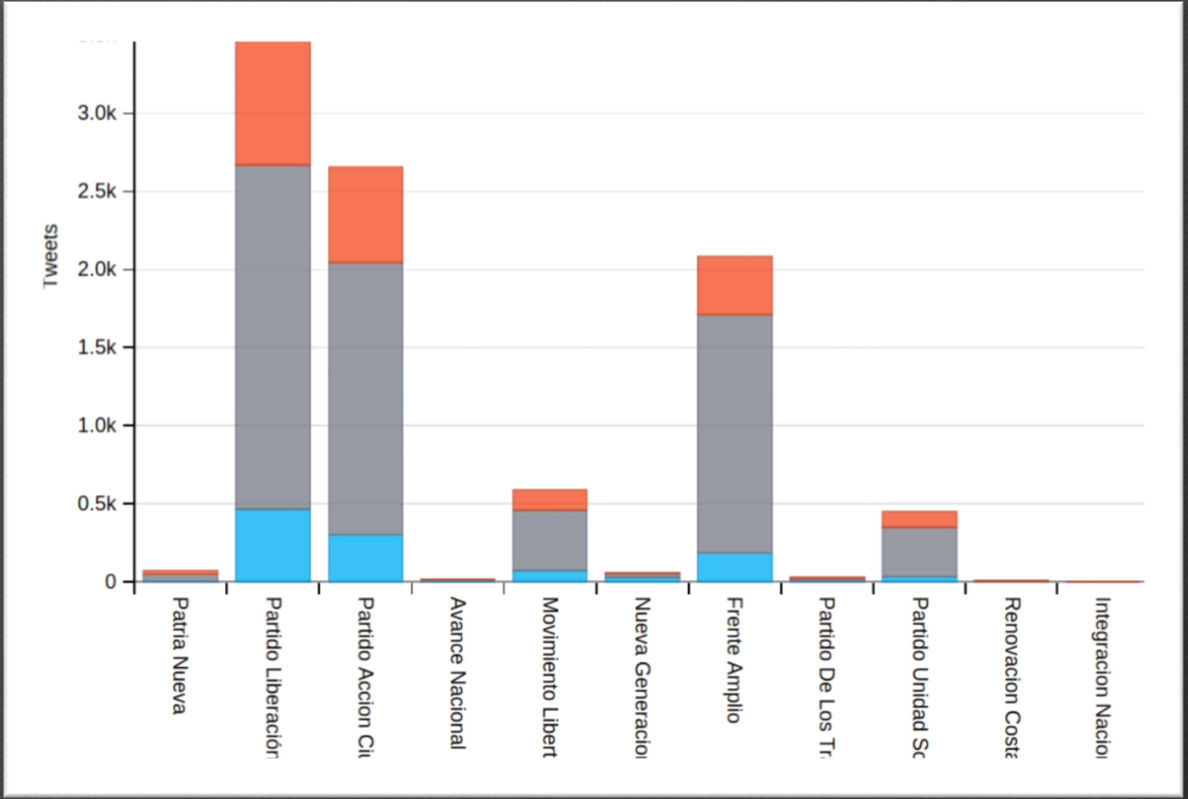
Discover

Ship

simMachines
SIMILARITY SEARCH & PATTERN RECOGNITION

Sentiment Detection

- simMachines successfully predicted the congress layout before election day
- First time somebody predicts an election result before it happens
- A top political party leased our technology



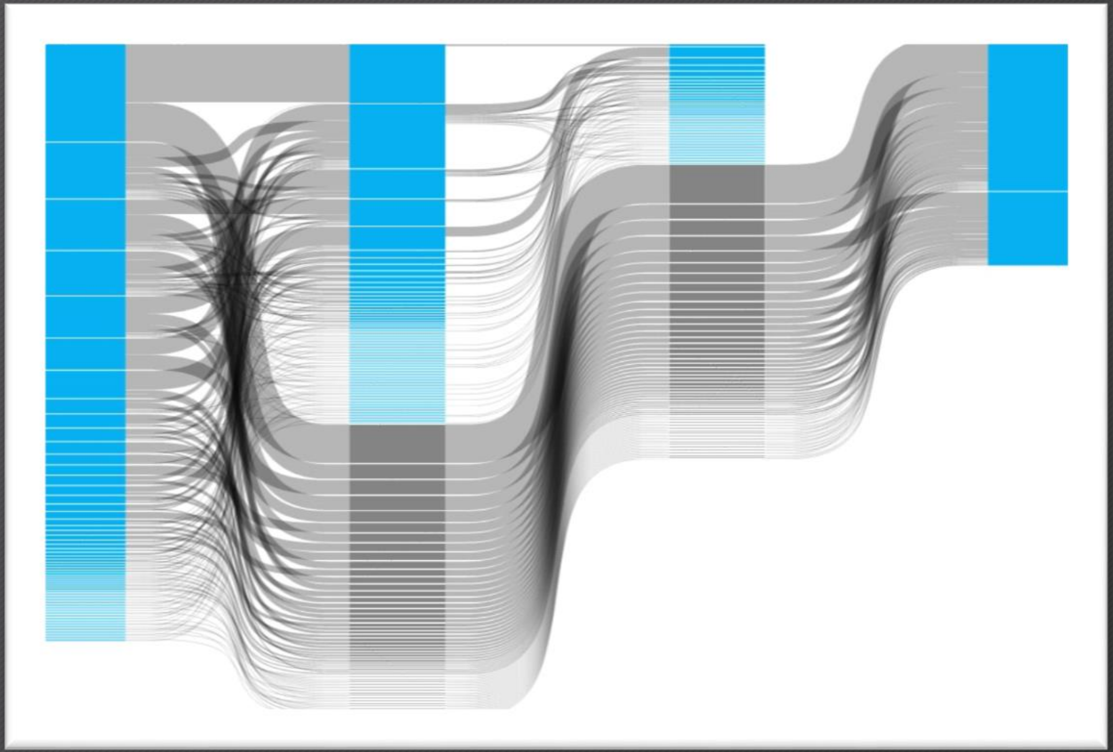
Auditing

Made sure the gTLD application process (a $300 million dollar project) was processed according to specifications.

Our tech audited two large auditing firms:

EY
Building a better working world

KPMG



Commissioned By:

ICANN

simMachines
SIMILARITY SEARCH & PATTERN RECOGNITION

Regular Expression Inference

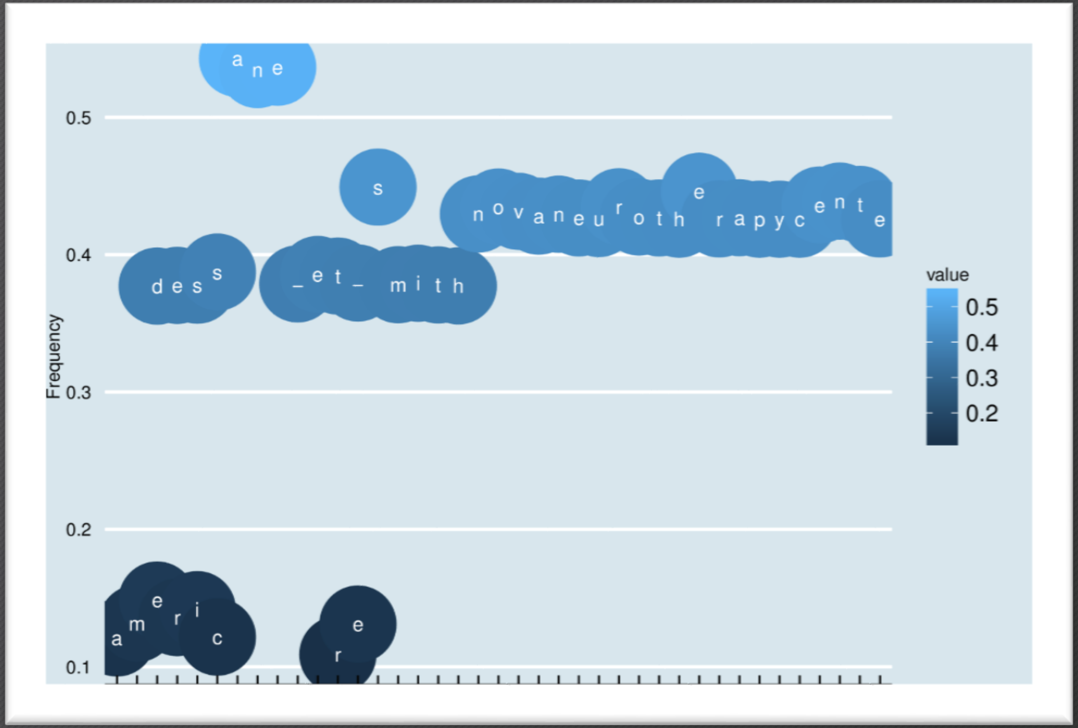
```
.{0,2}+[de]?+[er]?+[is]?+[cs]?+[a][n][e][-]?+[e]?+[rt]?+[-e]?+[s]?+[m]?+.{0,25}+
```

Input: many strings, output: Regexes!

Detected a huge security issue impacting 2.5 billion Internet users (ICANN’s gTLD transition)

OARC Workshop

Paul Vixie congratulated simMachines!



Commissioned by:

ICANN

simMachines
SIMILARITY SEARCH & PATTERN RECOGNITION

Recommend

Key

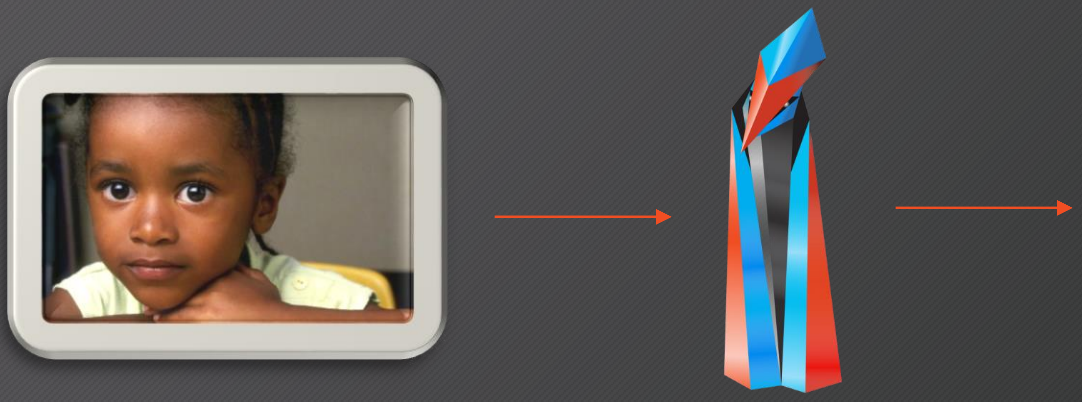
simMachines
SIMILARITY SEARCH & PATTERN RECOGNITION

Recommendations

- Tell me the customers that will buy the following product and associated discounts.
- Tell me the things a customer would like to buy in addition to a base product.
- Prediction success: > 90%
- Joint project with E&Y for a Supermarket

Financial

Predict the kinds of purchases a customer will perform in order to recommend coupons and discounts.



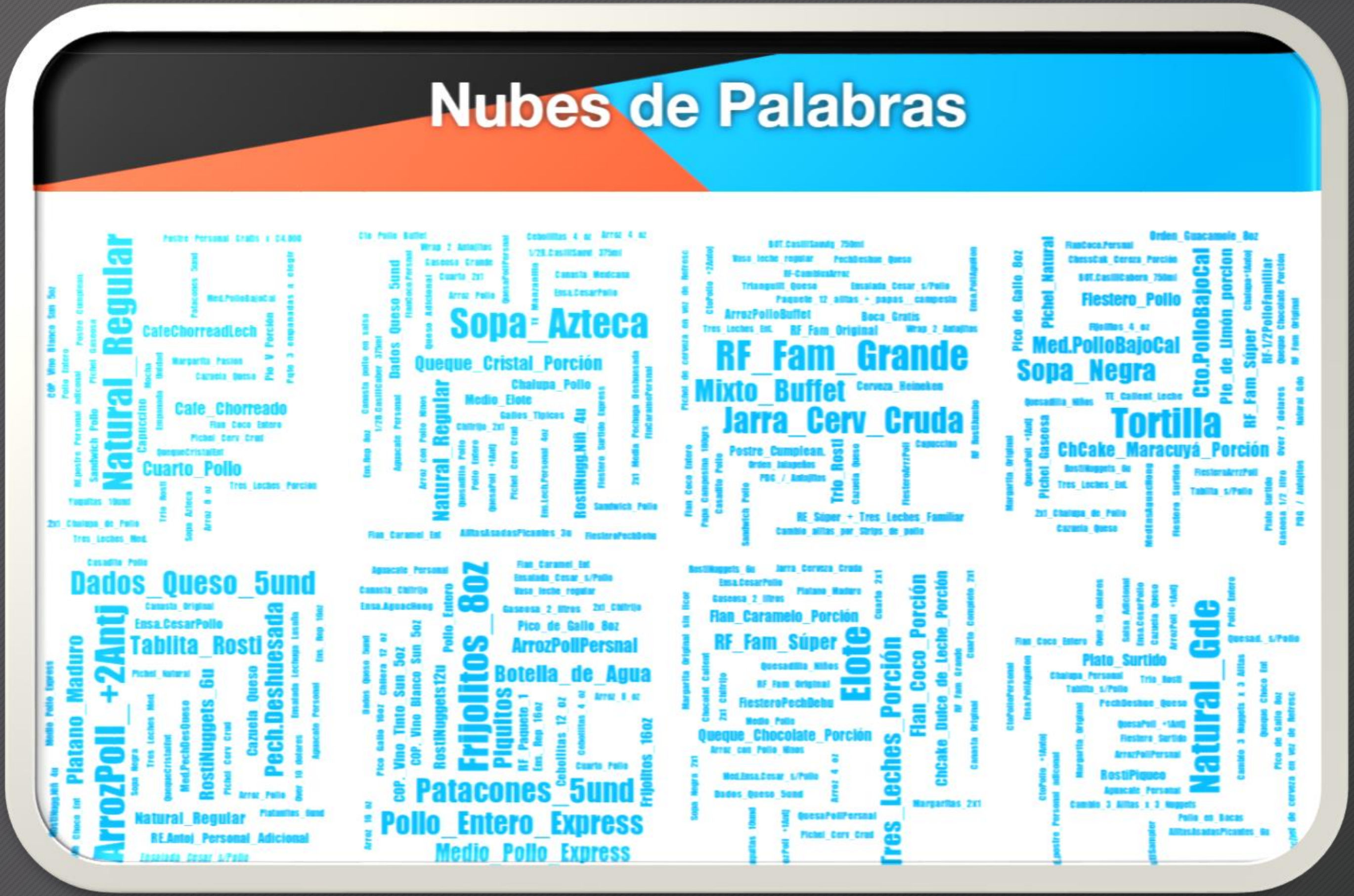
Young Customer

simMachines cloud recommender

Bookstore,
Music store
Phone company

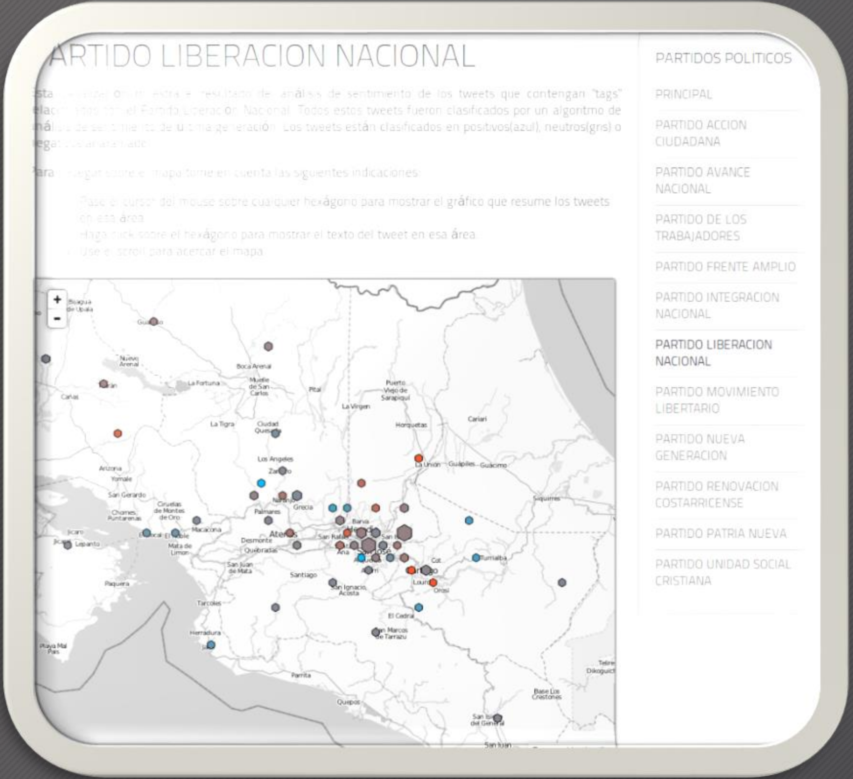
simMachines
SIMILARITY SEARCH & PATTERN RECOGNITION

Combo Recommendation



simMachines
SIMILARITY SEARCH & PATTERN RECOGNITION

Political Messages



ARTIDO LIBERACION NACIONAL

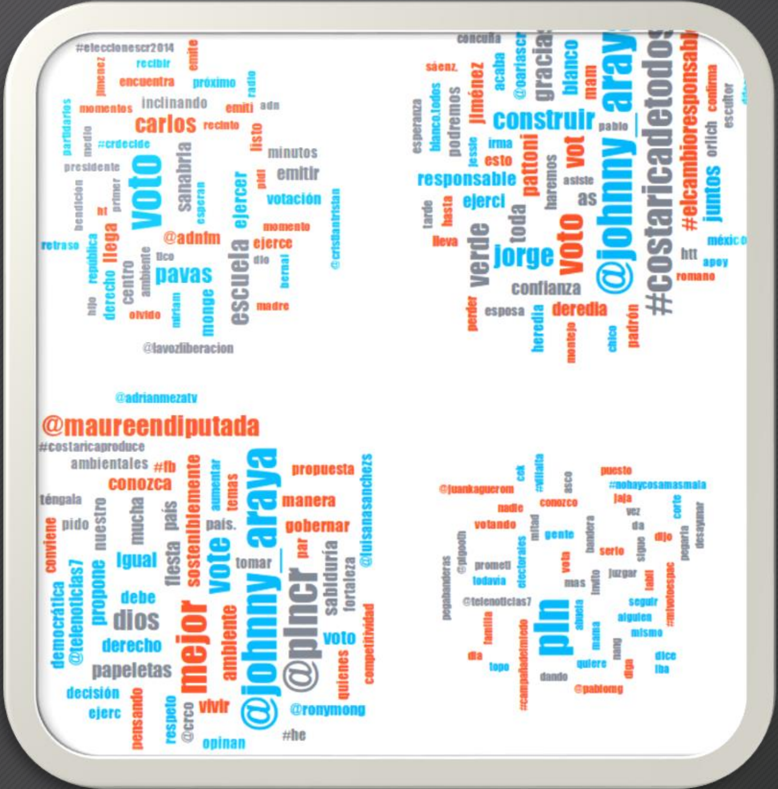
sta visualización muestra el resultado del análisis de sentimiento de los tweets que contengan "tags" elacionados con el Partido Liberación Nacional. Todos estos tweets fueron clasificados por un algoritmo de nálisis de sentimiento de última generación. Los tweets están clasificados en positivos(azul), neutros(gris) o egativos(anaranjado).

ara navegar sobre el mapa tome en cuenta las siguientes indicaciones:

- Pase el cursor del mouse sobre cualquier hexágono para mostrar el gráfico que resume los tweets en esa área
- Haga click sobre el hexágono para mostrar el texto del tweet en esa área
- Use el scroll para acercar el mapa

PARTIDOS POLITICOS

- PRINCIPAL
- PARTIDO ACCION CIUDADANA
- PARTIDO AVANCE NACIONAL
- PARTIDO DE LOS TRABAJADORES
- PARTIDO FRENTE AMPLIO
- PARTIDO INTEGRACION NACIONAL
- PARTIDO LIBERACION NACIONAL
- PARTIDO MOVIMIENTO LIBERTARIO
- PARTIDO NUEVA GENERACION
- PARTIDO RENOVACION COSTARRICENSE
- PARTIDO PATRIA NUEVA
- PARTIDO UNIDAD SOCIAL CRISTIANA



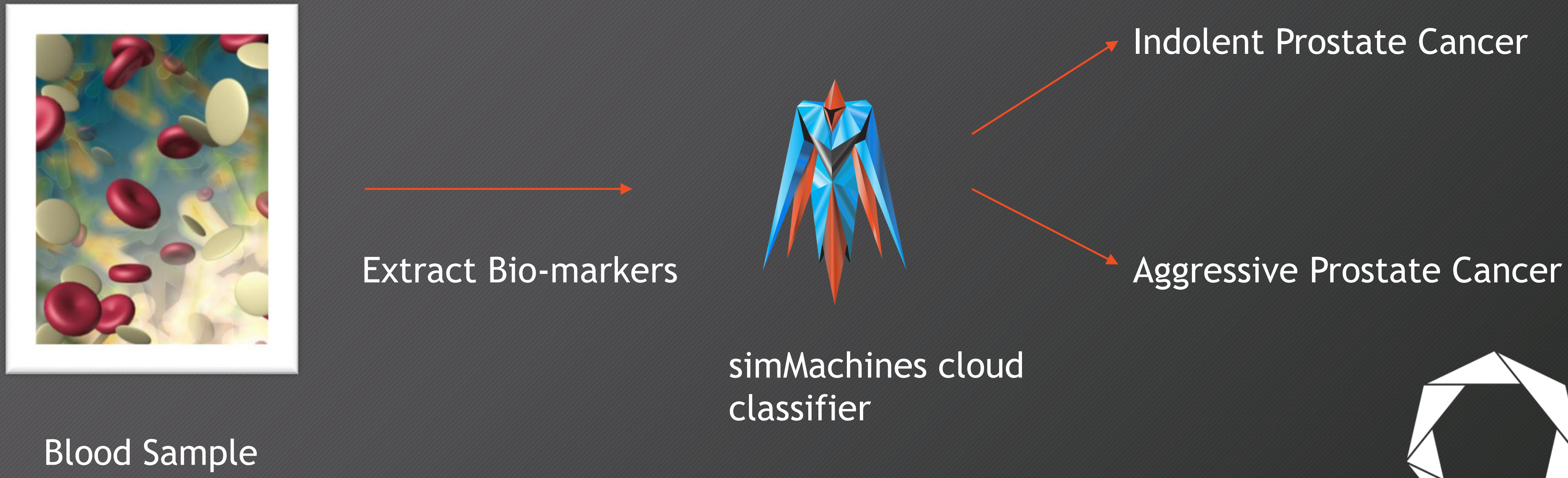
Predict

Crystal Ball

simMachines
SIMILARITY SEARCH & PATTERN RECOGNITION

Healthcare work (I)

Automatically detect the presence of Aggressive or Indolent prostate cancer



Blood Sample

Extract Bio-markers

simMachines cloud classifier

Indolent Prostate Cancer

Aggressive Prostate Cancer

simMachines
SIMILARITY SEARCH & PATTERN RECOGNITION

Healthcare Work (II)

Based on the patient's historical data, predict how many times he or she will receive a CT scan in
The future. Reducing exposure to CT scans is important because they are correlated to Cancer occurrences.

Patient Historical Data

Process-data

simMachines cloud
Regression engine

Average # of CT Scans

simMachines
SIMILARITY SEARCH & PATTERN RECOGNITION

Healthcare Work (III)

- Etherapeutics (based in Oxford, UK)
- Uses simMachines technology in the search of similar chemical interaction patterns.

"We were looking for similarity engines, and we found only you guys"
"It is refreshing to receive the kind of service you provide"
Jonny Wray, Head of Discovery Informatics

e-Therapeutics plc
systems biology drug discovery

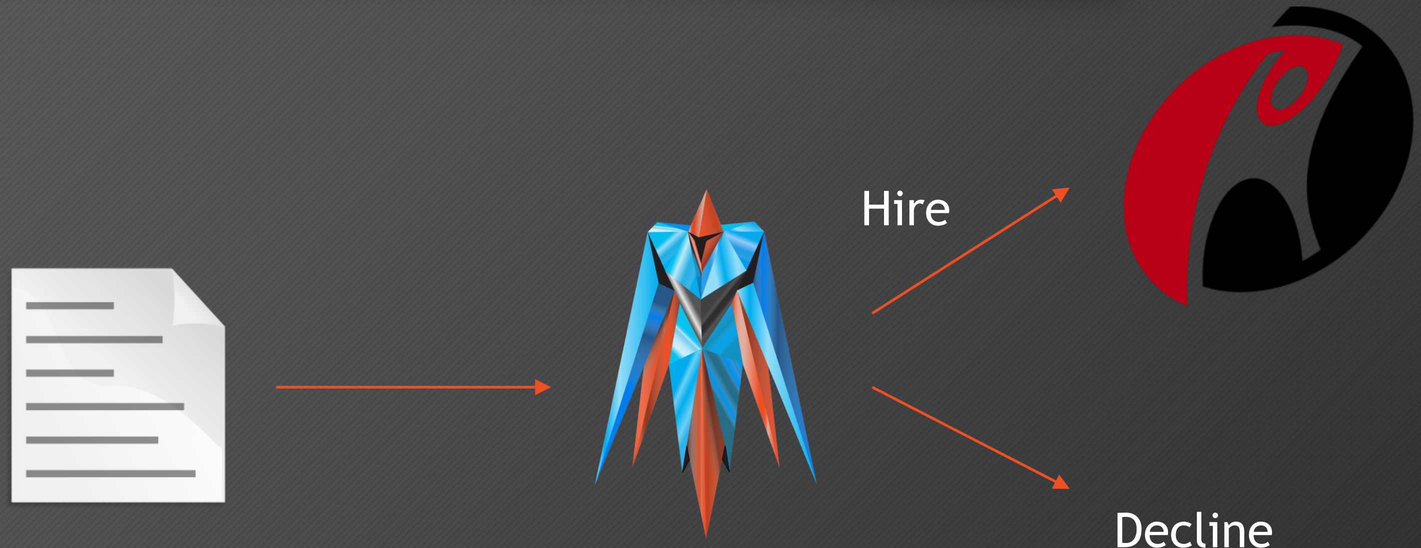
simMachines
SIMILARITY SEARCH & PATTERN RECOGNITION

Prediction & Classification

From 200,000 resumes, find if a resume will be a good fit to an organization.

(95% success rate)

Our customer: Rackspace



Hire

Decline

Resume

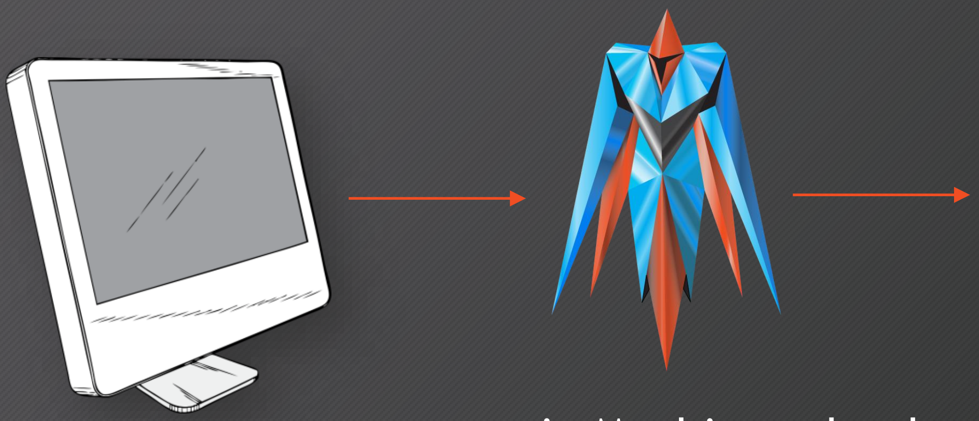
simMachines cloud classifier

simMachines
SIMILARITY SEARCH & PATTERN RECOGNITION

Prediction and Classification

TAX category prediction

Take a product description from Amazon, predict the tax ID the product will have to pay when it enters multiple countries abroad.

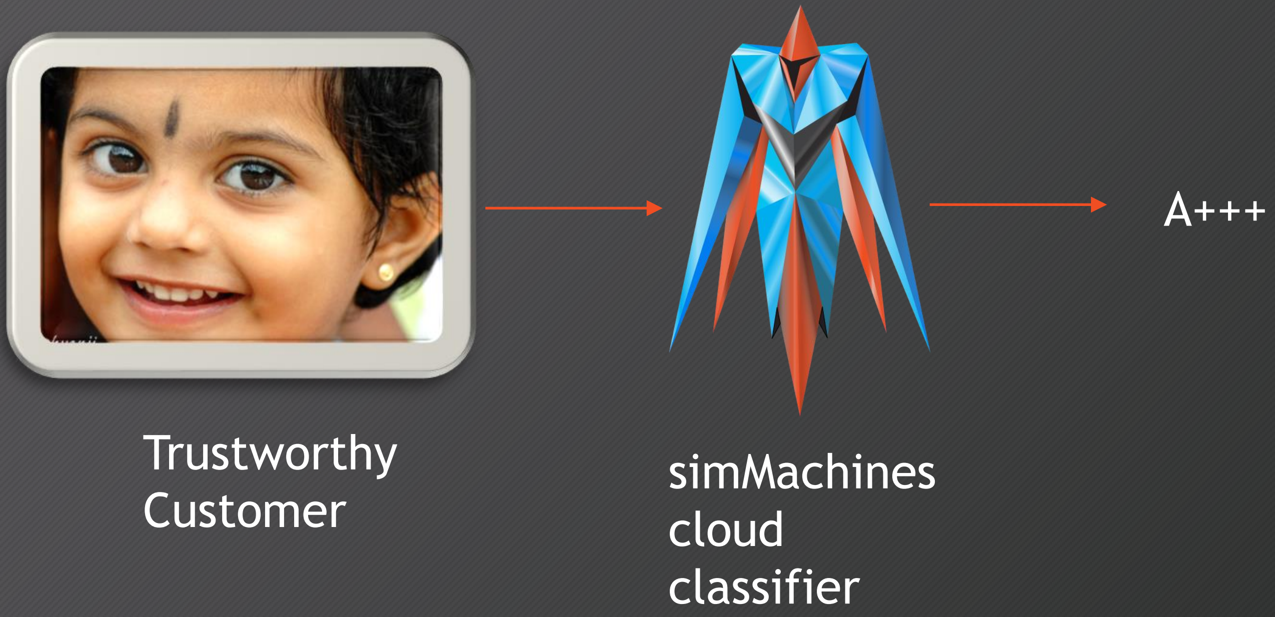


TAX ID 203055

simMachines cloud classifier

Financial

Predict customer Category before issuing a loan



Trustworthy Customer

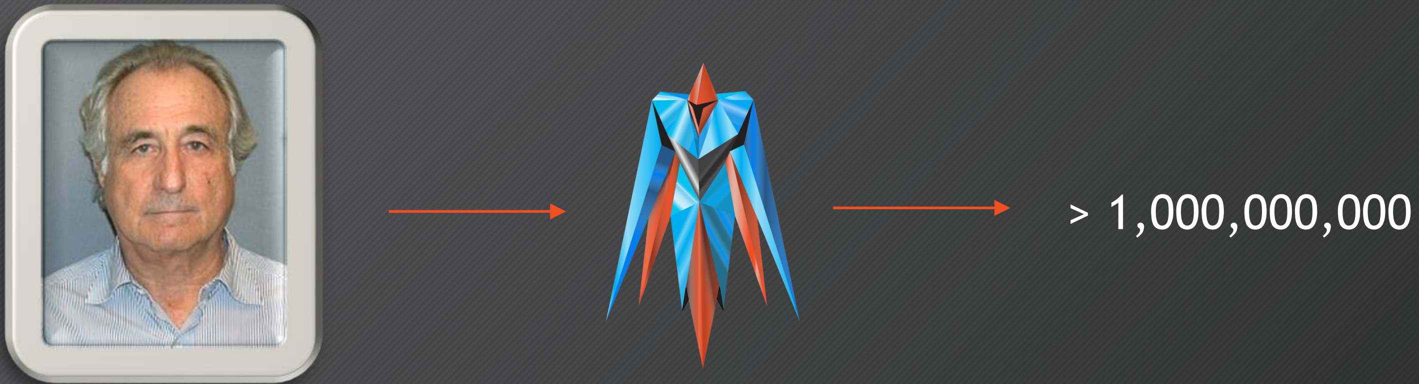
simMachines cloud classifier

A+++

simMachines
SIMILARITY SEARCH & PATTERN RECOGNITION

Financial

Predict number of times I will have to dial to collect payments



Bad Customer

simMachines cloud classifier

> 1,000,000,000

simMachines
SIMILARITY SEARCH & PATTERN RECOGNITION

Summary

- Similarity is:
  - Simple
  - Powerful
  - Modern
  - Effective
  - Scalable Data Science
  - A tool to rule them all

Purpose

Thank you!

simMachines

SIMILARITY SEARCH & PATTERN RECOGNITION