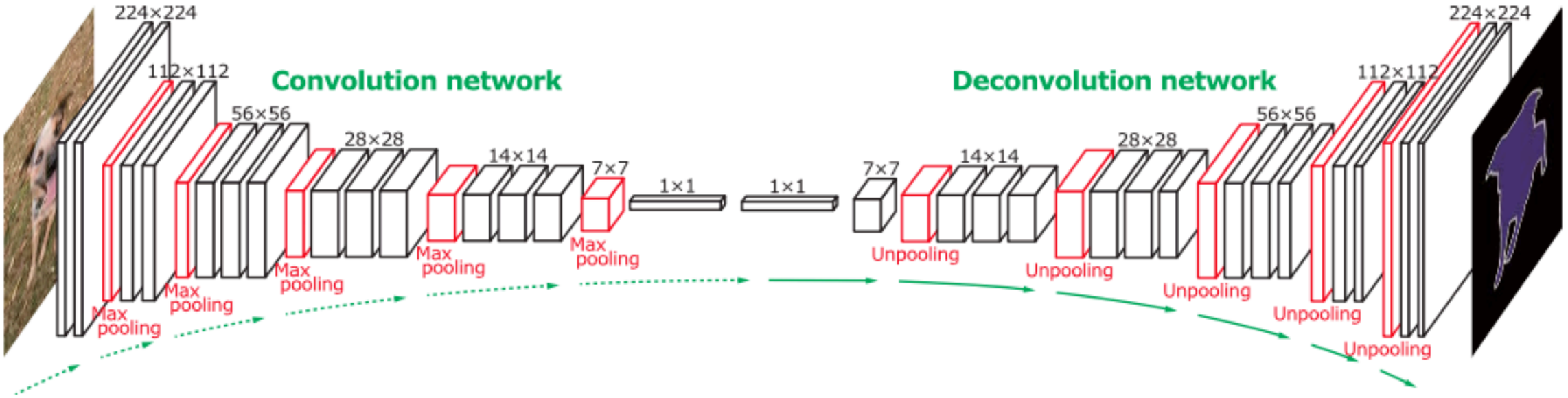


DEEP LEARNING

Anatomia de um projeto de Machine Learning
aplicado à recomendação de notícias

Mar/2018

Fabrício V Matos

linkedin.com/in/fabriciovargasmatos

Vix Data Meetup #7

Recomendação Personalizada de Notícias

29 Estações de TV

Sede em NY


mutual.life
SOUNDS LIKE INSURANCE.
BUT IT'S SHARED PROTECTION.

Pilares

- Mobile
- Inteligência Artificial
- Blockchain


LEFT REARWARD VEHICLE CAMERA
MEDIUM RANGE VEHICLE CAMERA
RIGHT REARWARD VEHICLE CAMERA
MOTION FLOW
LANE LINES
LANE LINES
ROAD FLOW
IN-PATH OBJECTS
ROAD LIGHTS
OBJECTS
ROAD SIGNS

Amazon_rainforest

The Stanford Question Answering Dataset

The Amazon rainforest (Portuguese: Floresta Amazônica or Amazônia; Spanish: Selva Amazónica, Amazonía or usually Amazonia; French: Forêt amazonienne; Dutch: Amazoneregenwoud), also known in English as Amazonia or the Amazon Jungle, is a moist broadleaf forest that covers most of the Amazon basin of South America. This basin encompasses 7,000,000 square kilometres (2,700,000 sq mi), of which 5,500,000 square kilometres (2,100,000 sq mi) are covered by the rainforest. This region includes territory belonging to nine nations. The majority of the forest is contained within Brazil, with 60% of the rainforest, followed by Peru with 13%, Colombia with 10%, and with minor amounts in Venezuela, Ecuador, Bolivia, Guyana, Suriname and French Guiana. States or departments in four nations contain "Amazonas" in their names. The Amazon represents over half of the planet's remaining rainforests, and comprises the largest and most biodiverse tract of tropical rainforest in the world, with an estimated 390 billion individual trees divided into 16,000 species.

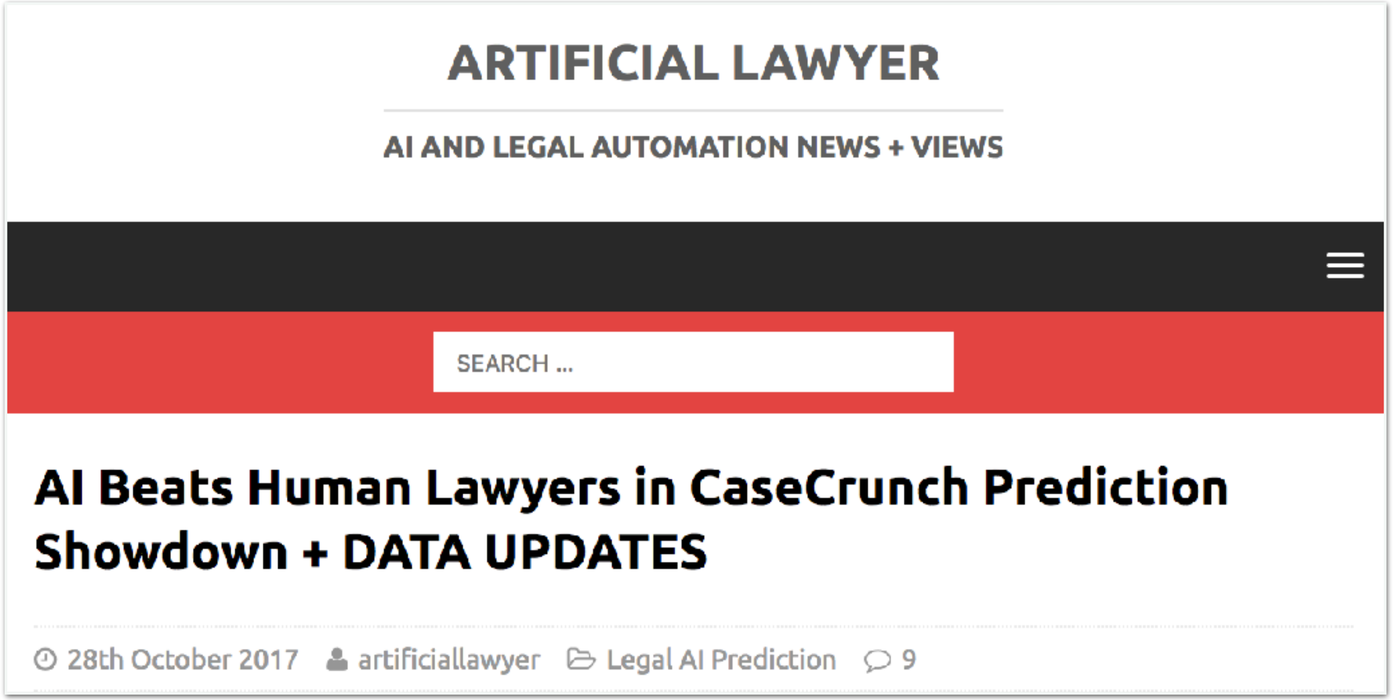
Which name is also used to describe the Amazon rainforest in English?
Ground Truth Answers: also known in English as Amazonia or the Amazon Jungle, | Amazonia or the Amazon Jungle | Amazonia
Prediction: Amazonia

How many square kilometers of rainforest is covered in the basin?
Ground Truth Answers: 5,500,000 square kilometres (2,100,000 sq mi) are covered by the rainforest. | 5,500,000 | 5,500,000
Prediction: 5,500,000

How many nations control this region in total?
Ground Truth Answers: This region includes territory belonging to nine nations. | nine | nine
Prediction: nine

How many nations contain "Amazonas" in their names?
Ground Truth Answers: States or departments in four nations contain "Amazonas" in their names. | four | four
Prediction: four

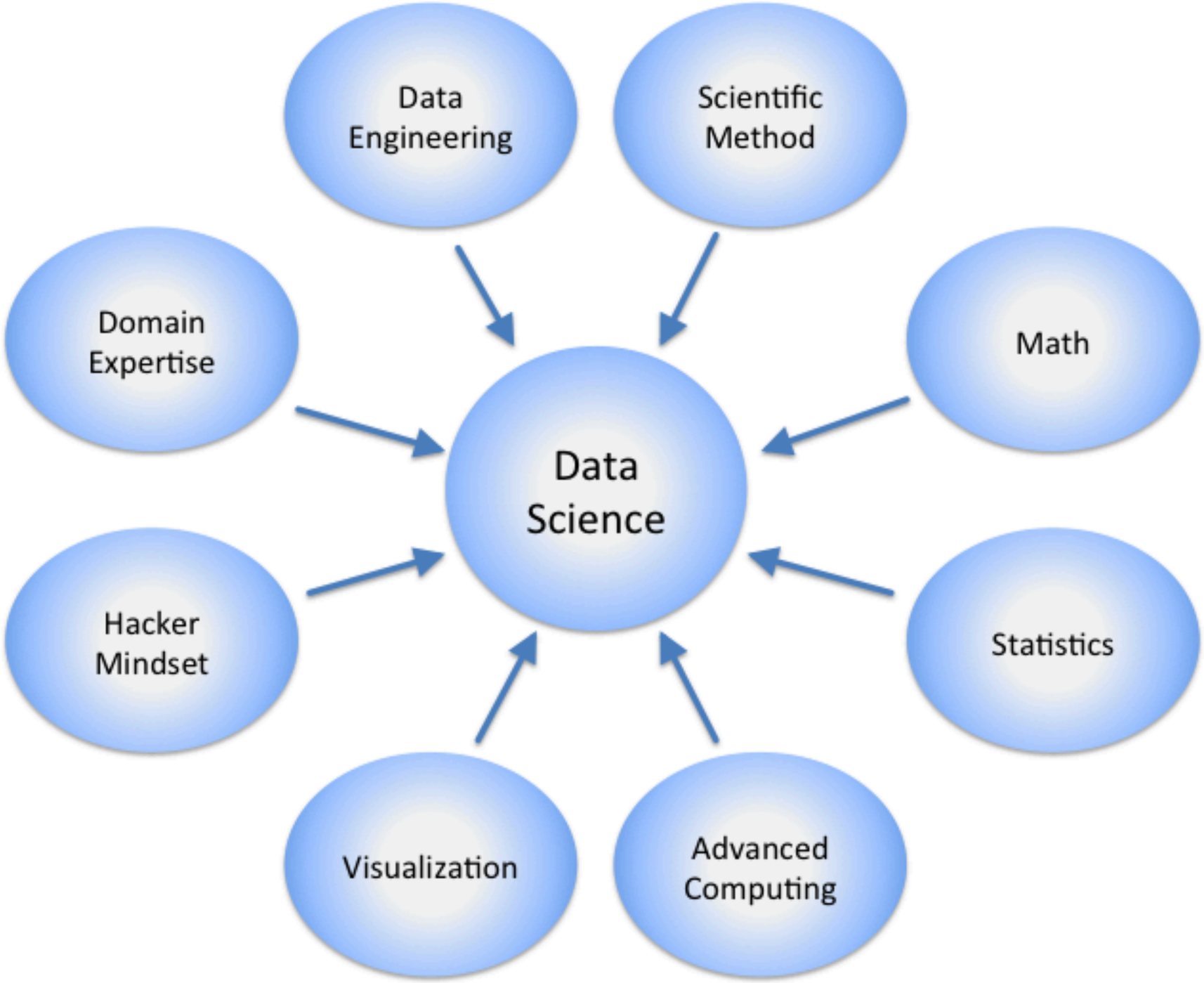
Human Performance: 82.304
AI from Microsoft Research (Fev/18): **82.849**
...
AI from Microsoft Research (Out/17): 76.461



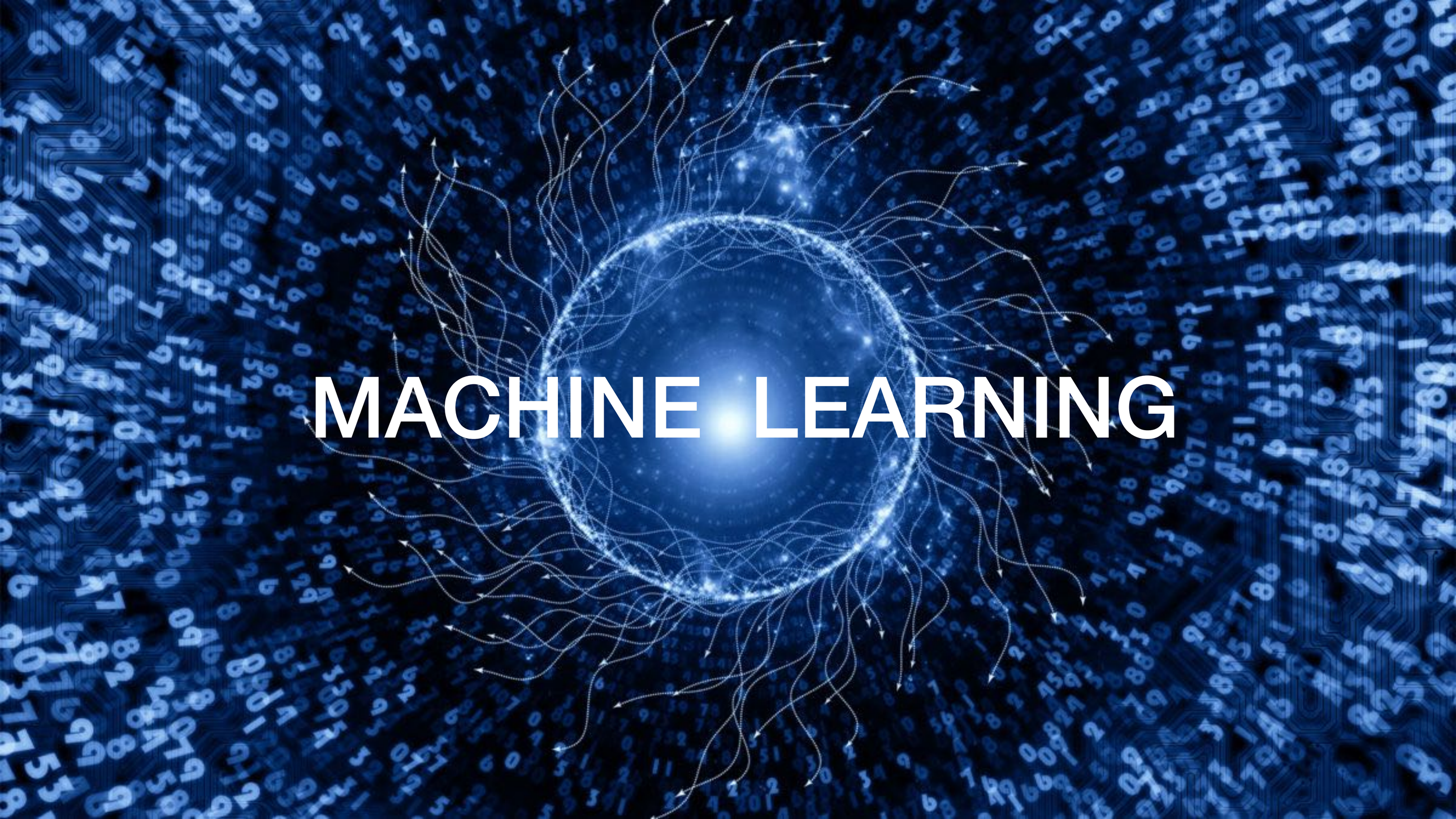
CaseCrunch is proud to announce the results of the lawyer challenge. **CaseCruncher Alpha scored an accuracy of 86.6%. The lawyers scored an accuracy of 62.3%.**

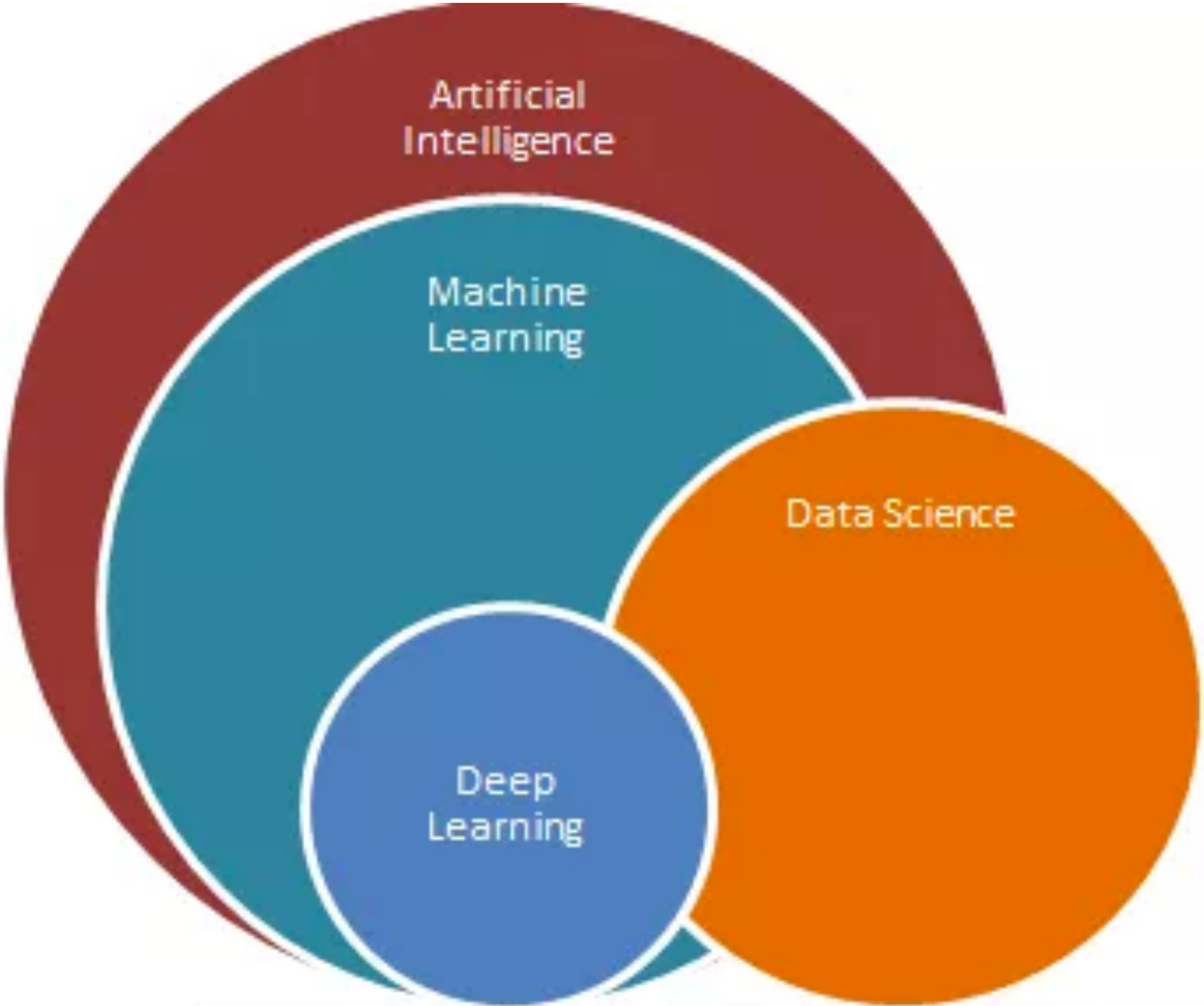
Over 100 commercial London lawyers signed up for the competition and made over 750 predictions over the course of a week in an unsupervised environment.

The problems were real complaints about PPI mis-selling decided and published by the Financial Ombudsman Service (FOS) under the Freedom Of Information Act. (*N.B. Payment Protection Insurance, or PPI, has been a massive issue in the UK, with banks having to pay customers billions of pounds in total refunds for making consumers take on insurance products they never required.*)


Data Engineering
Scientific Method
Domain Expertise
Math
Data Science
Hacker Mindset
Statistics
Visualization
Advanced Computing

Por onde começar?

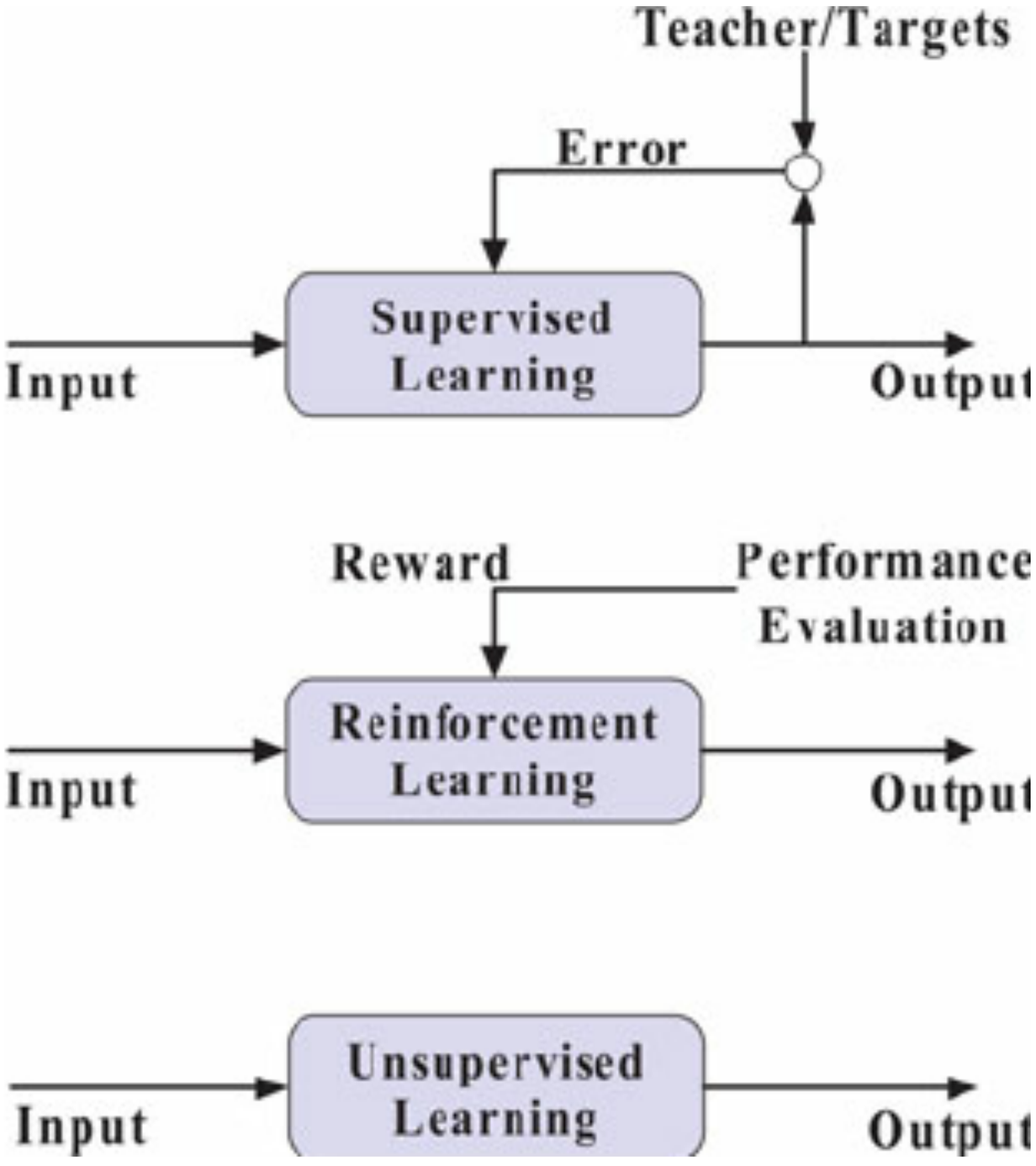

MACHINE LEARNING


Artificial
Intelligence
Machine
Learning
Data Science
Deep
Learning

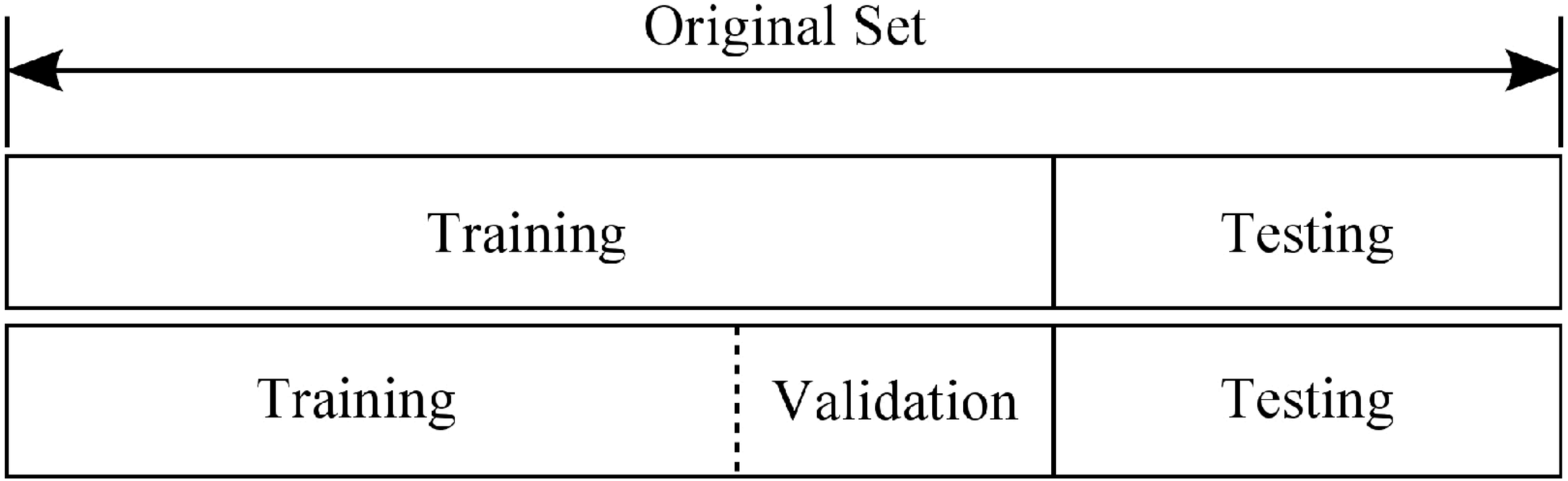
ML: Novo Paradigma

Traditional Programming

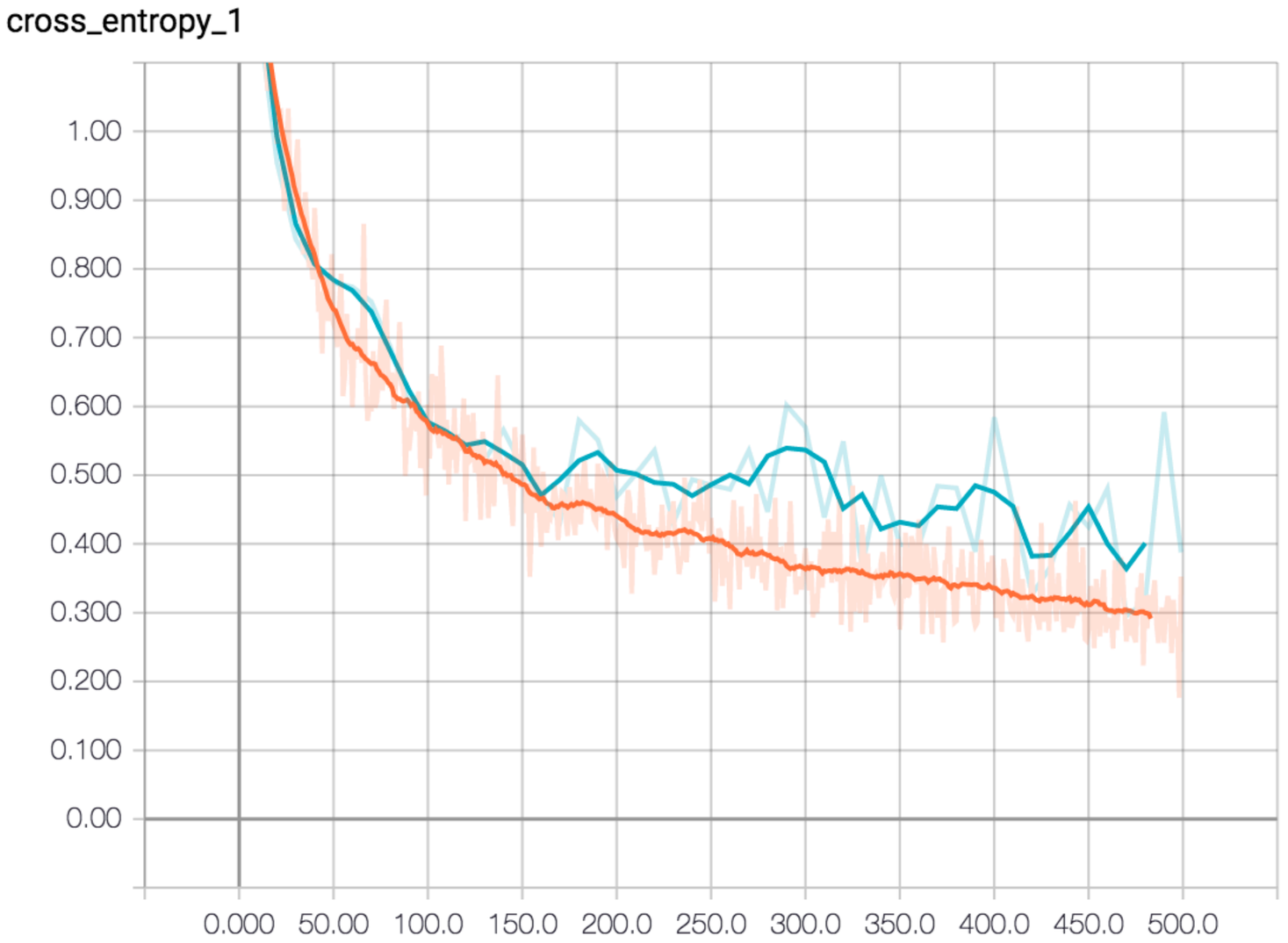
Machine Learning


Teacher/Targets
Error
Supervised Learning
Input
Output
Reward
Performance Evaluation
Reinforcement Learning
Input
Output
Unsupervised Learning
Input
Output

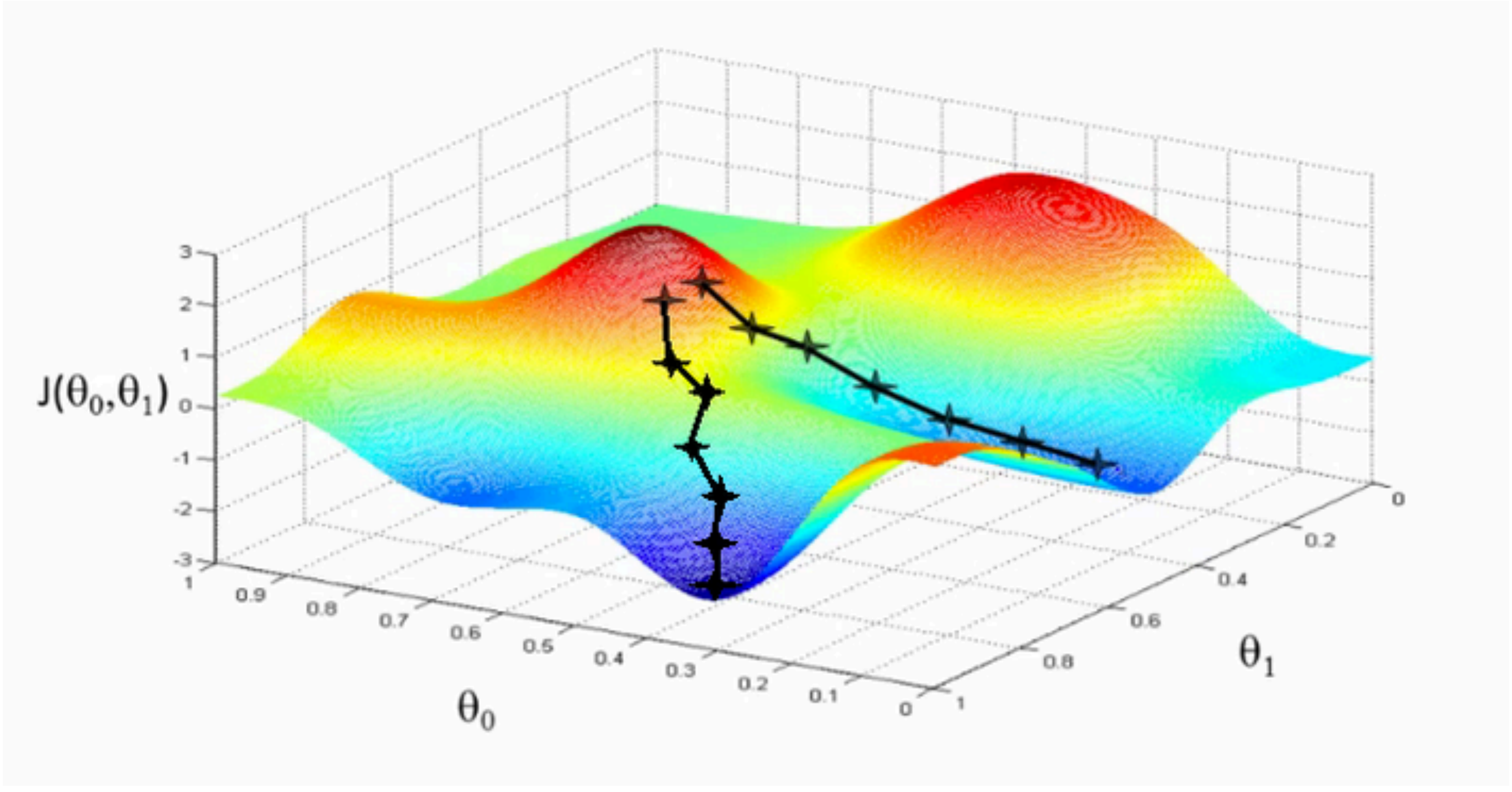
Treino supervisionado: Datasets



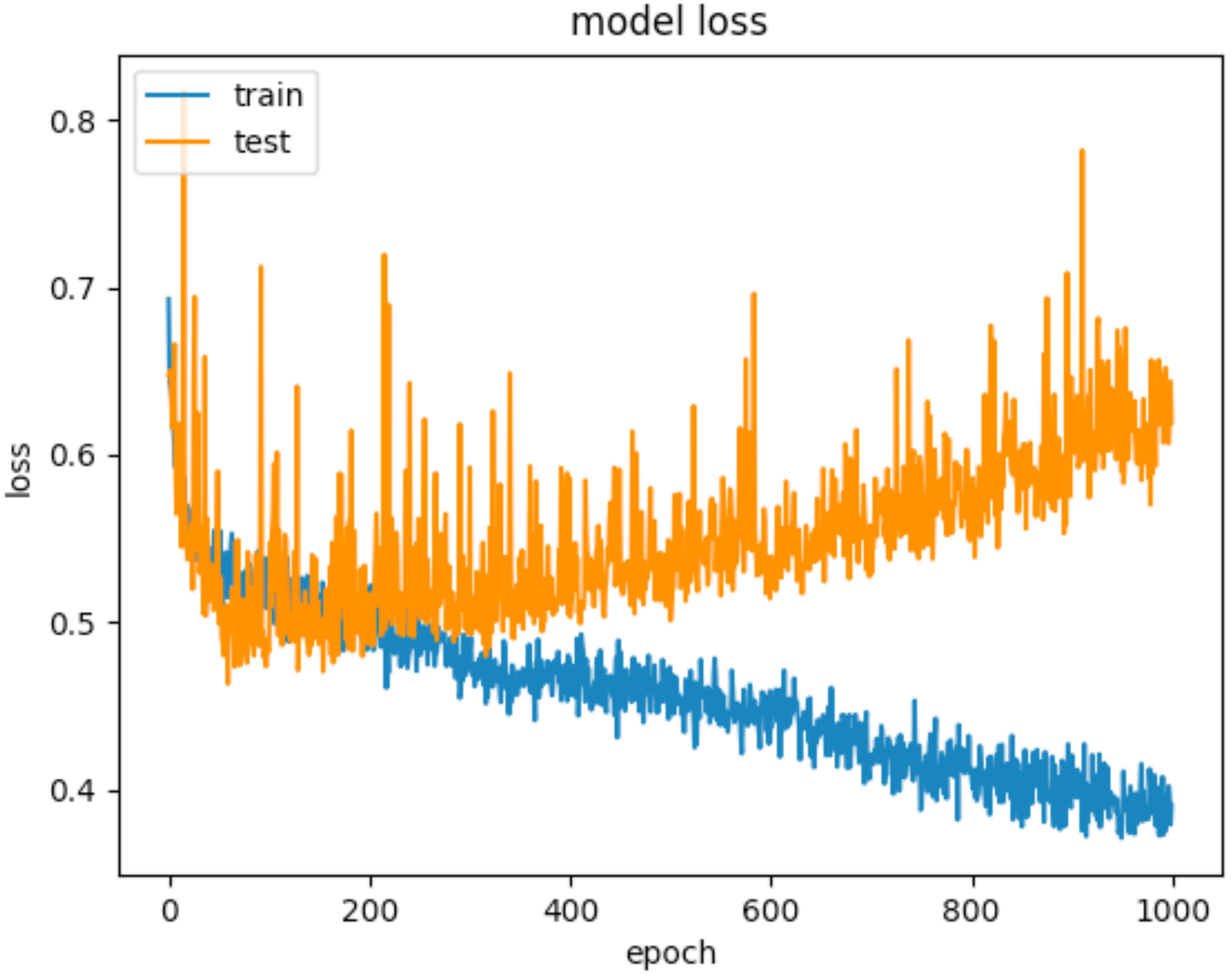
Treino supervisionado: Cost function



Otimizador: Stochastic Gradient Descent



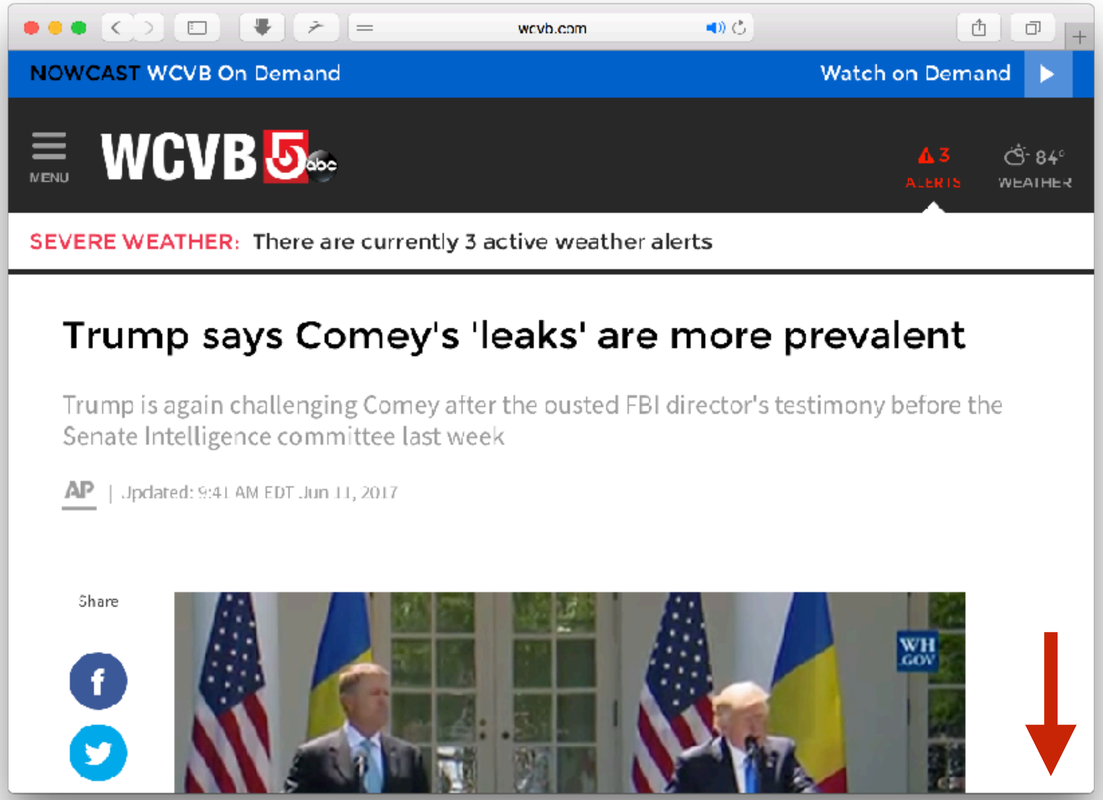
Overfit (“mente fechada”)

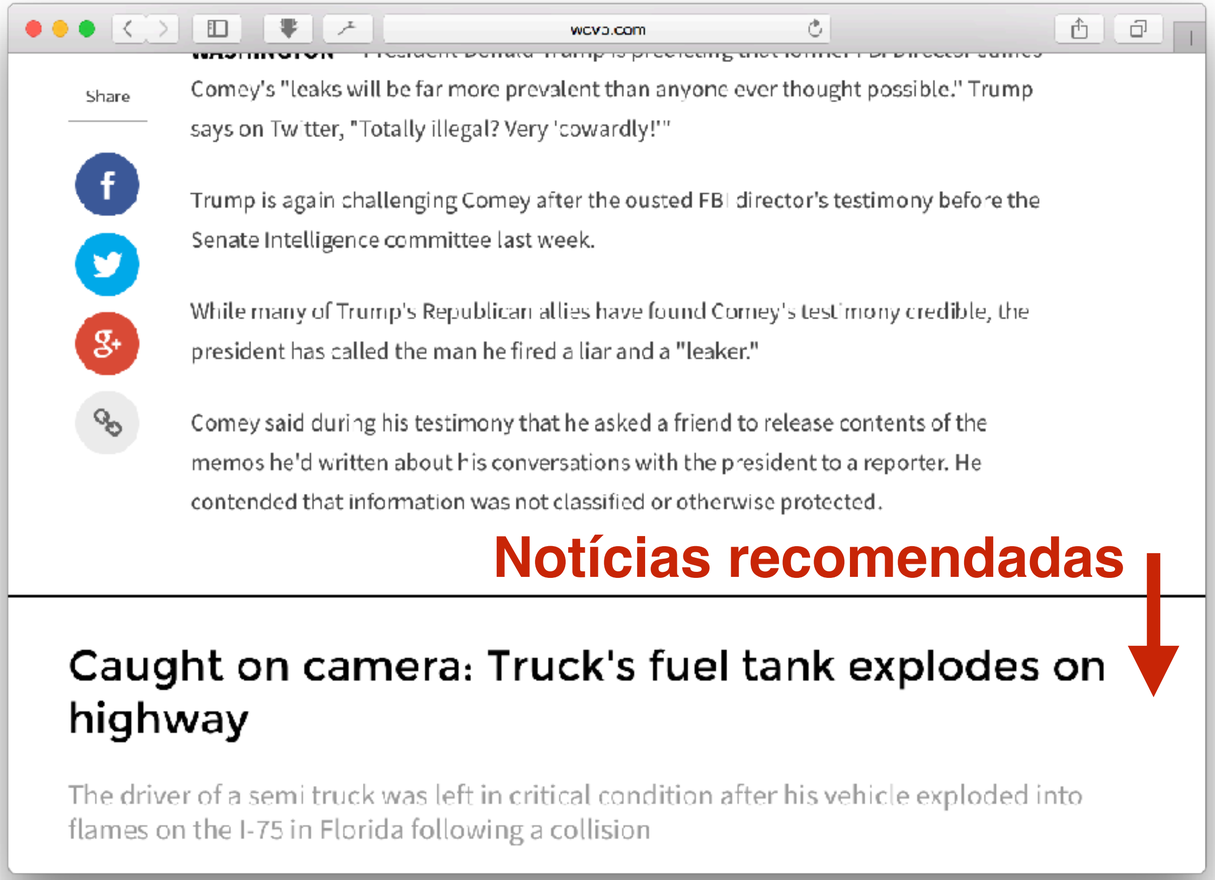


Estudo de caso:

Recomendação personalizada de notícias

Aplicação


wcvb.com
NOWCAST WCVB On Demand
Watch on Demand
MENU
WCVB
3
ALERTS
84°
WEATHER
SEVERE WEATHER: There are currently 3 active weather alerts
Trump says Comey's 'leaks' are more prevalent
Trump is again challenging Comey after the ousted FBI director's testimony before the Senate Intelligence committee last week
AP | Updated: 9:41 AM EDT Jun 11, 2017
Share
WH GOV


wcvb.com
Share
Comey's "leaks will be far more prevalent than anyone ever thought possible." Trump says on Twitter, "Totally illegal? Very 'cowardly!'"
Trump is again challenging Comey after the ousted FBI director's testimony before the Senate Intelligence committee last week.
While many of Trump's Republican allies have found Comey's testimony credible, the president has called the man he fired a liar and a "leaker."
Comey said during his testimony that he asked a friend to release contents of the memos he'd written about his conversations with the president to a reporter. He contended that information was not classified or otherwise protected.
Notícias recomendadas
Caught on camera: Truck's fuel tank explodes on highway
The driver of a semi truck was left in critical condition after his vehicle exploded into flames on the I-75 in Florida following a collision

Desafios específicos

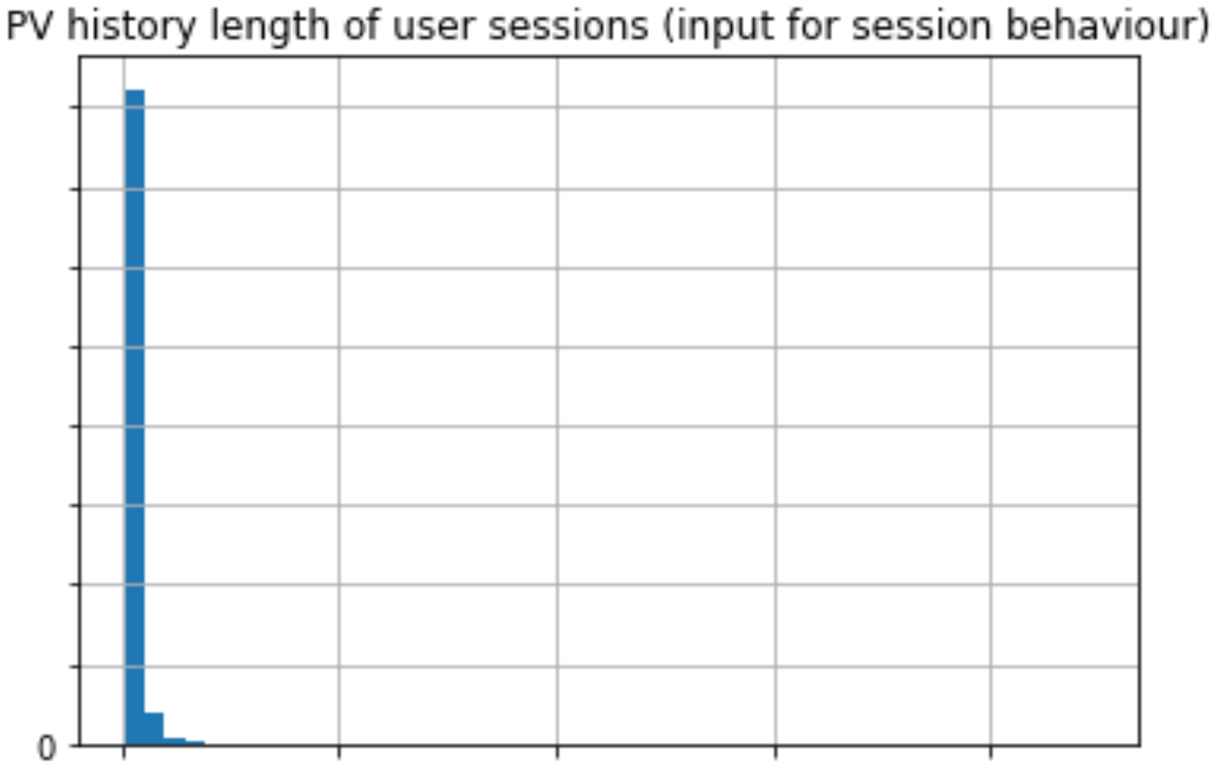
- Grande volume de novos items todos os dias
- Notícias perdem relevância muito rapidamente
- Usuários não se autenticam (escassez de dados históricos)
- Conteúdos devem ser atualizados pelo menos 50 vezes/dia
- API de recomendação escalável
- Implementar v1.0 em 3 meses

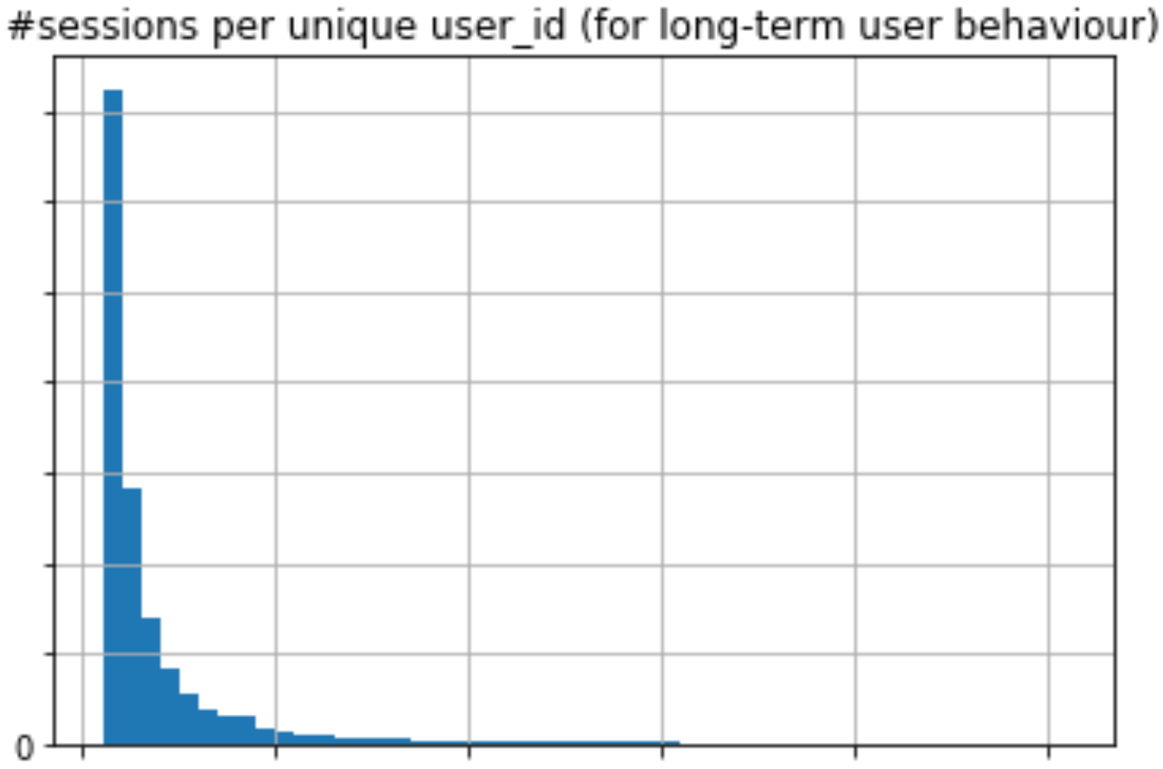
Resultados

- Grupo de teste - um pequeno percentual de usuários de 2 estações de TV: 800 mil usuários (teste + controle)
- Recomendações no rodapé da home: CTR aumentou em 15%
- Recomendações nas páginas de artigo (infinite scroll): Tempo médio de leitura aumentou em 168%
- Sinal verde para versão 2 mais sofisticada (deep learning)

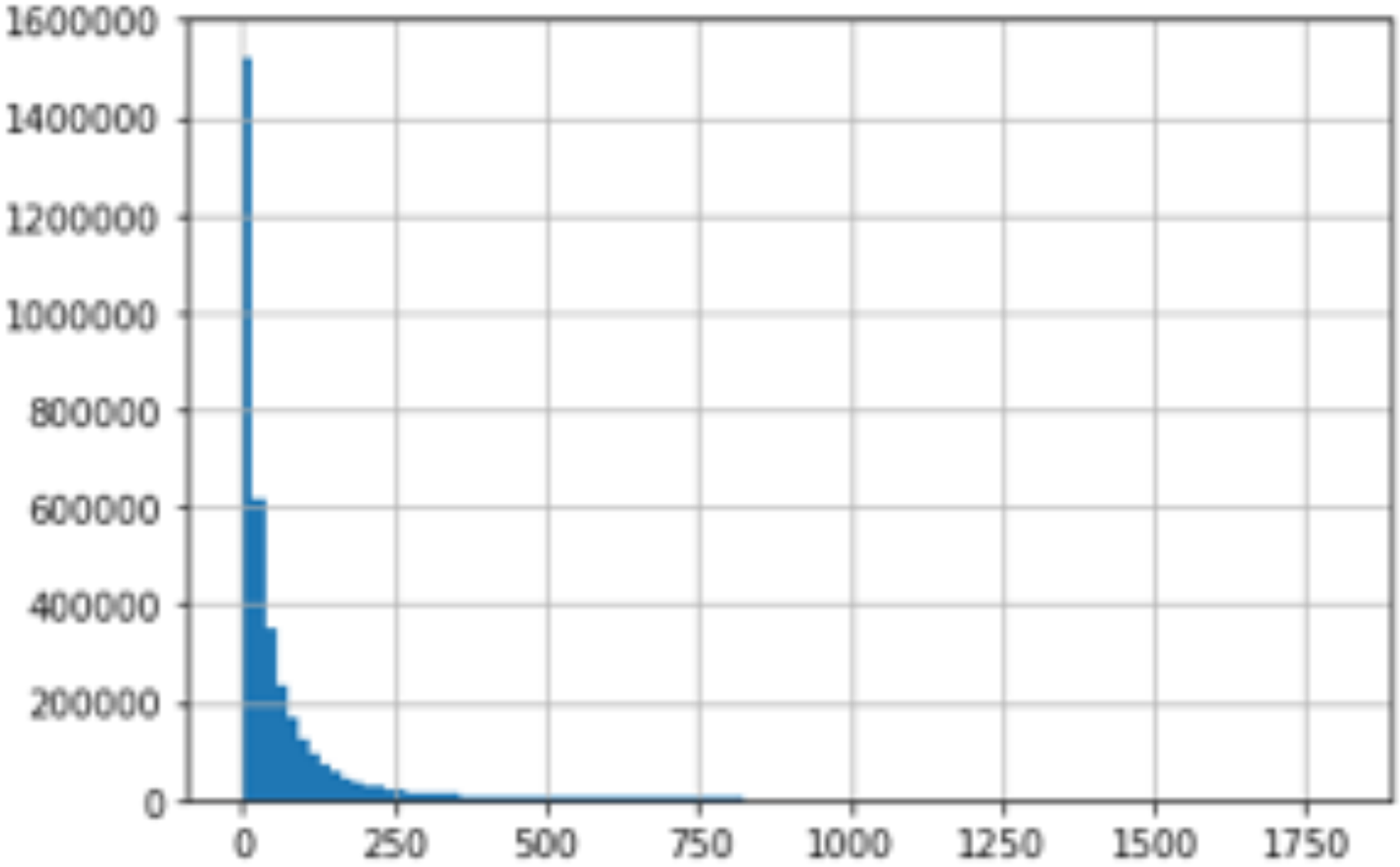
Análise Exploratória dos Dados

Page-views e sessões





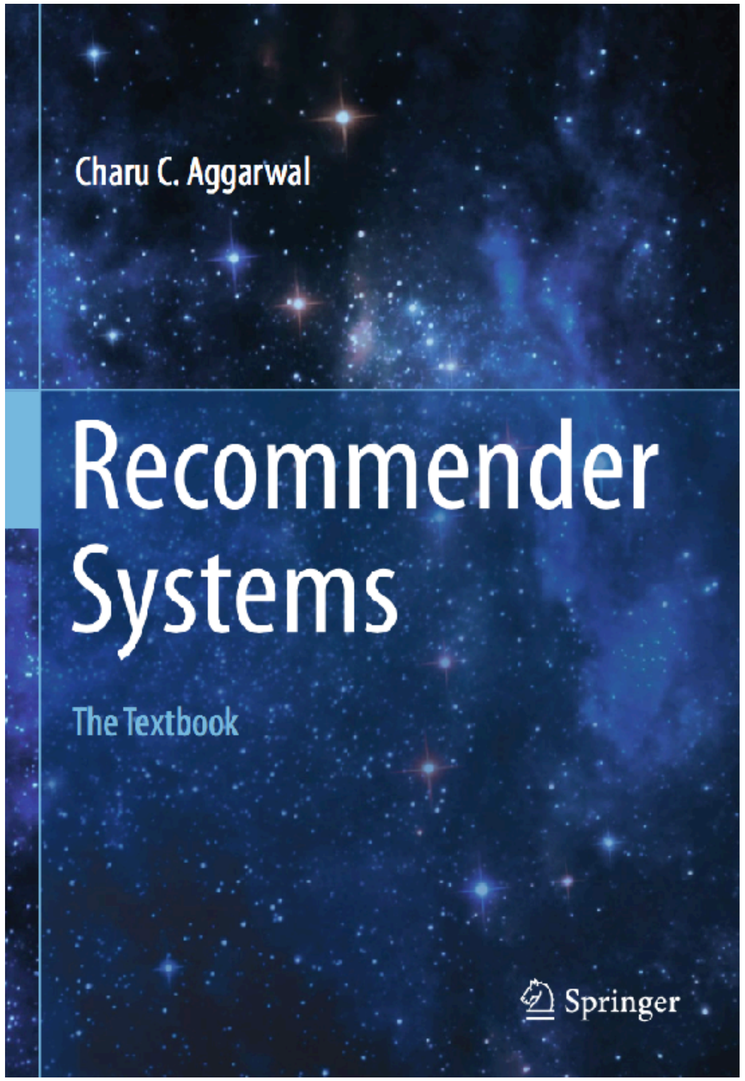
Duração (seg) de cada page-view



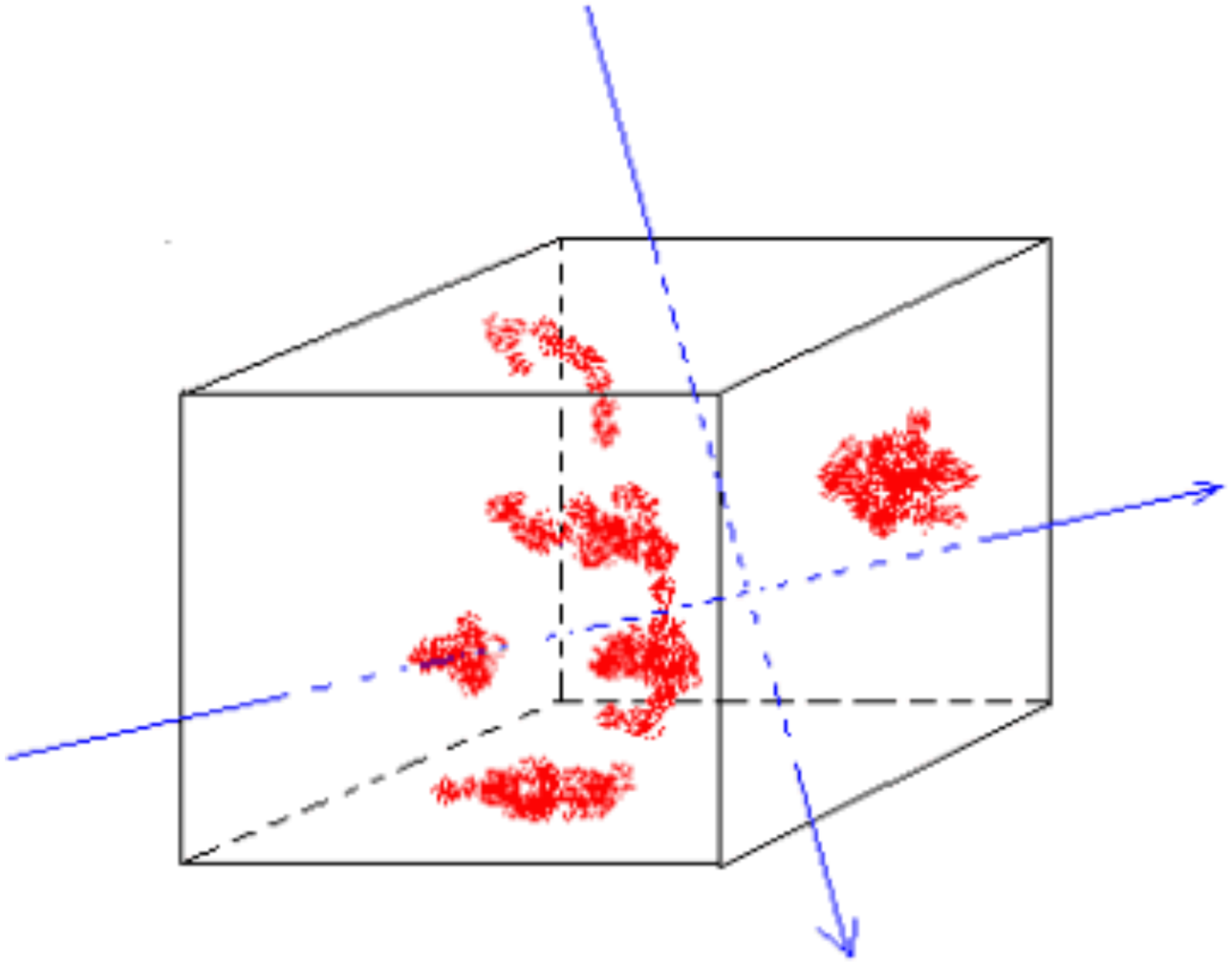
Sparse Matrix: Users x Articles

Dataset	Preenchimento
MovieLens (movies)	1,39%
Netflix (movies)	1,18%
TV Stations (news) 20x menos dados que Netflix	0,06%

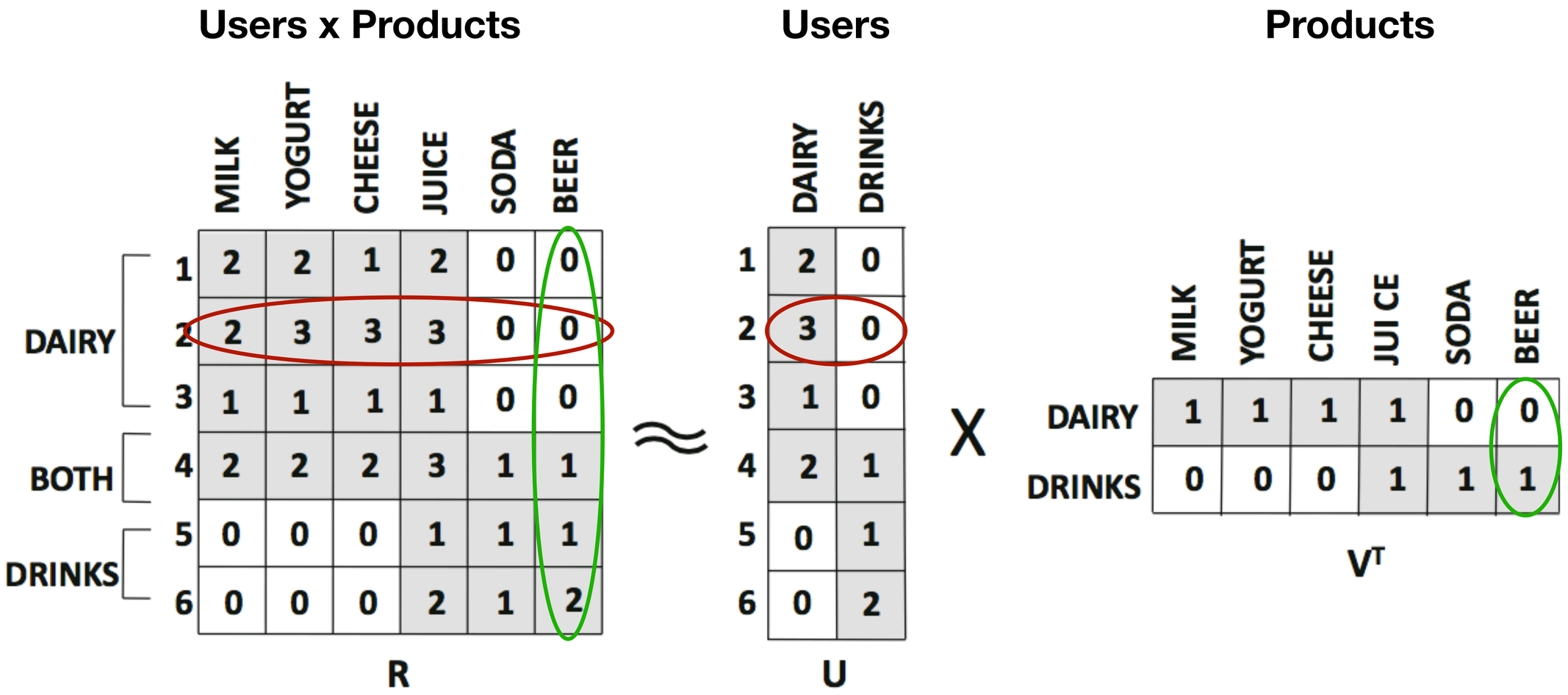
Referências



Espaço de Fatores Latentes

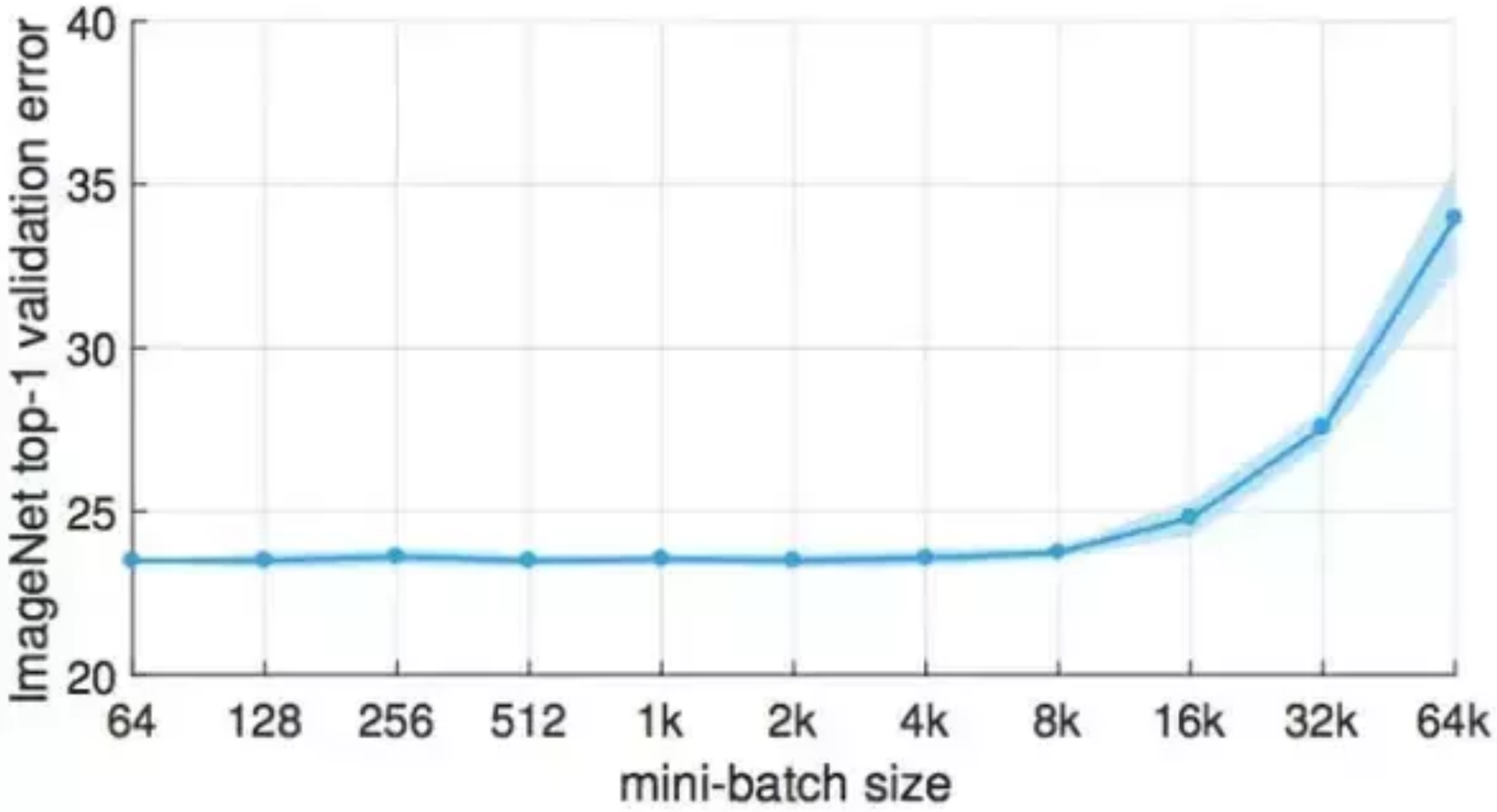


Model-based Collaborative Filtering



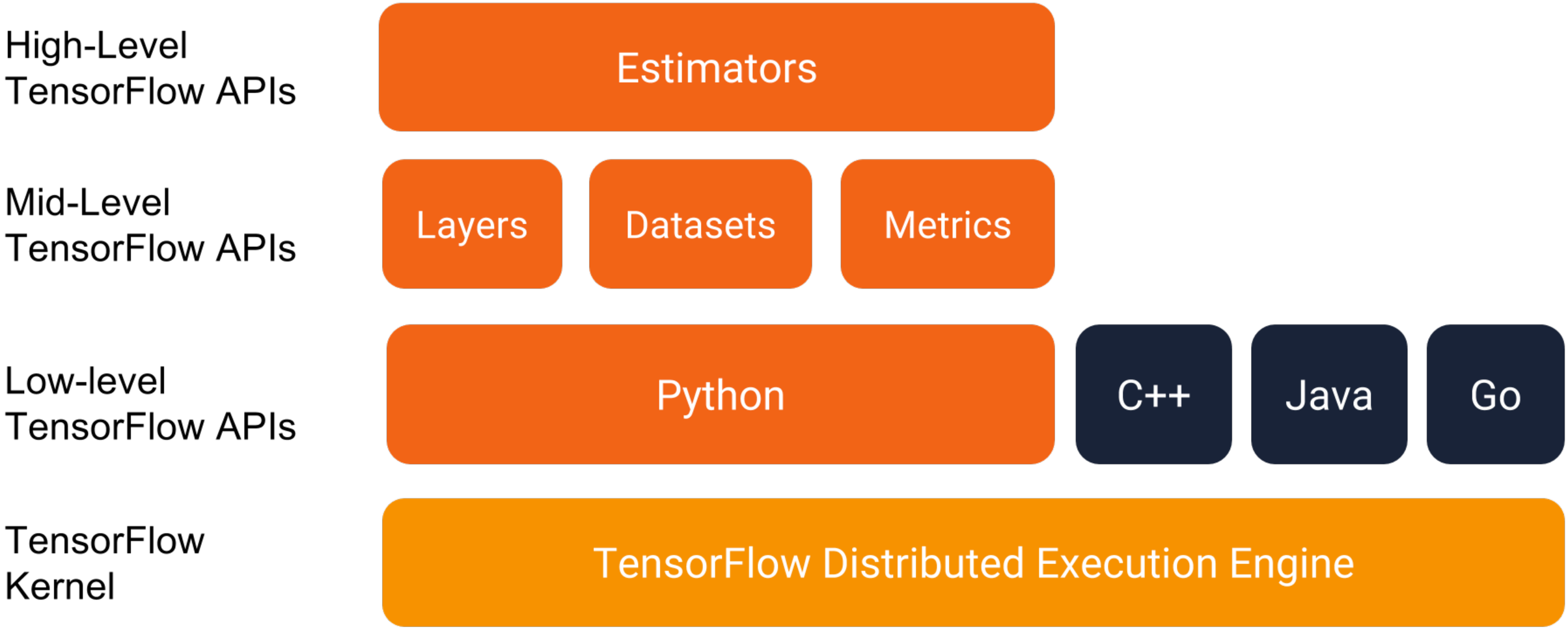
Frameworks e linguagens

Mini-batch size x Validation Error

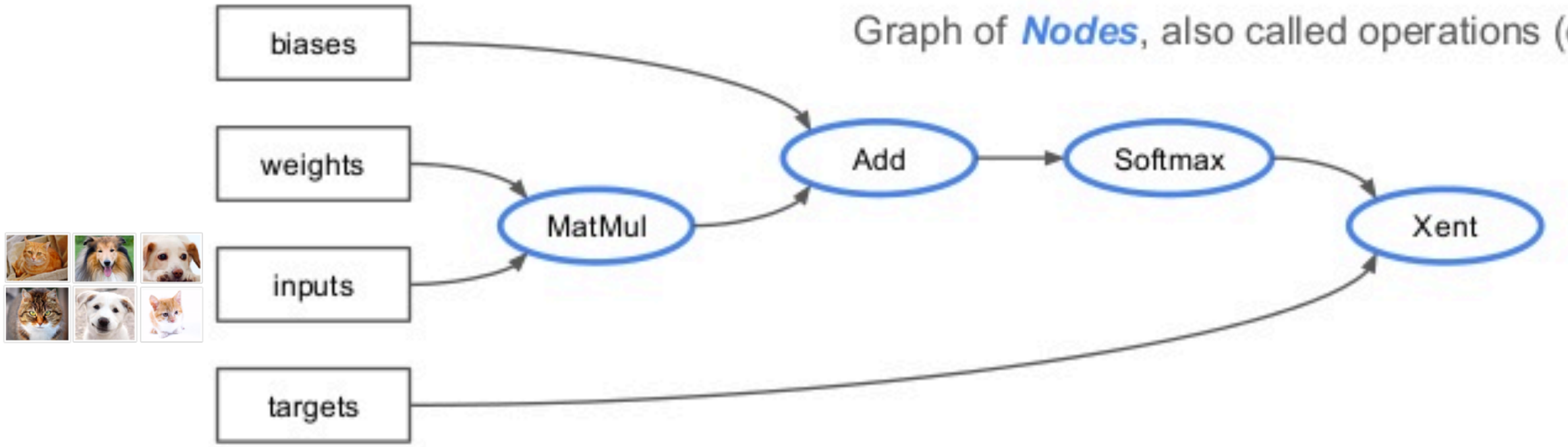


ImageNet dataset: 1h de treino com mini-batches de 8192 imagens em 256 GPUs.

Tensorflow Framework



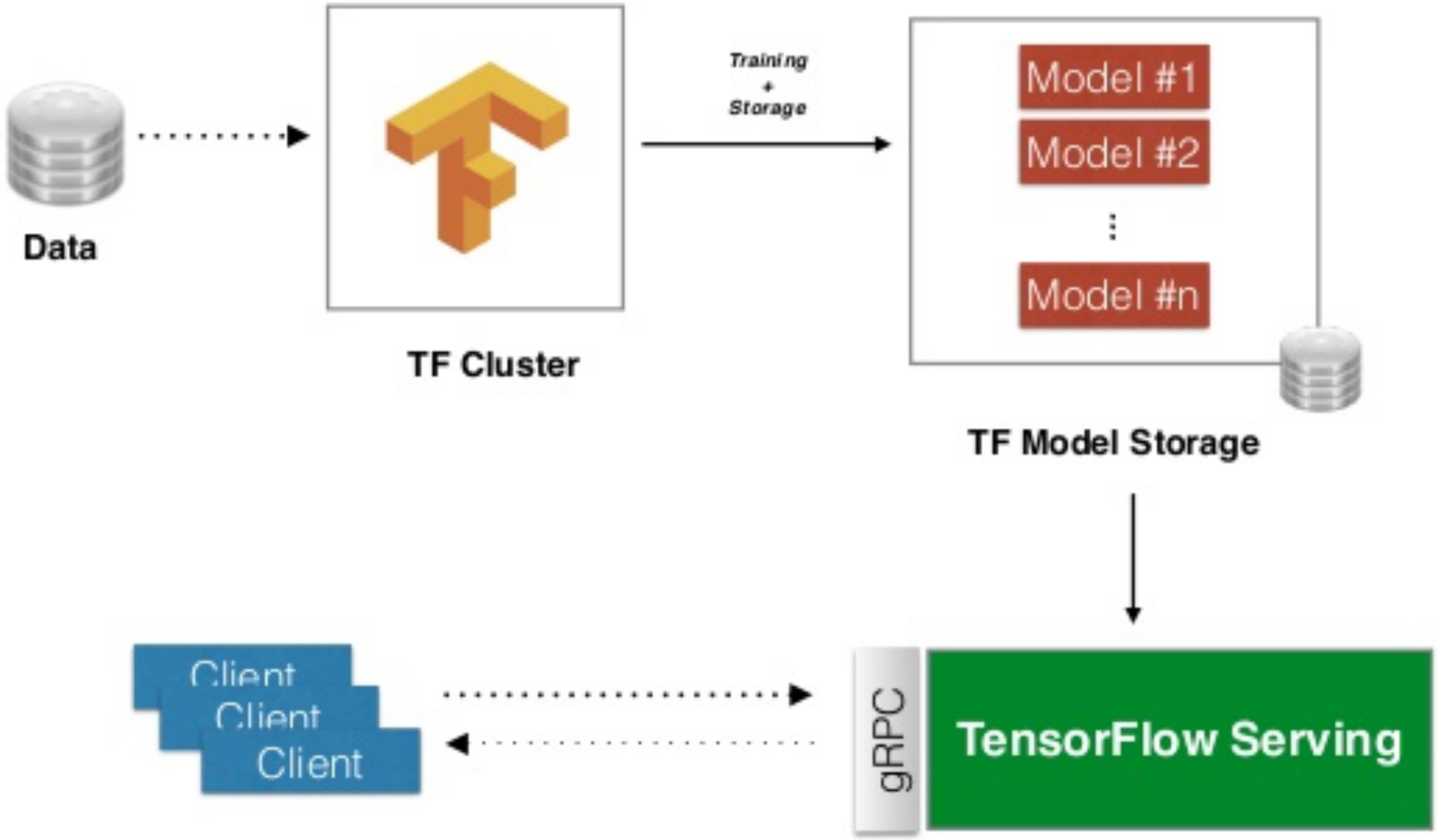
Computation as a Dataflow Graph



classifier = **softmax**(inputs*weights + biases)

cost = **cross_entropy**(classifier, targets)

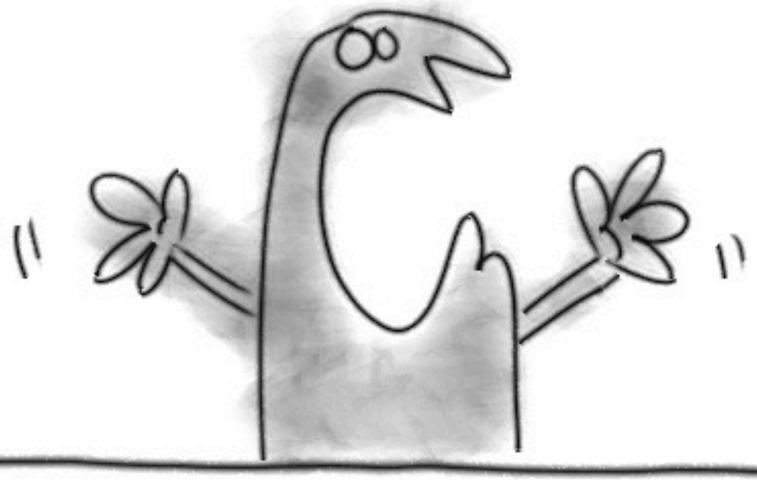
Tensorflow Serving



Como testar o modelo?

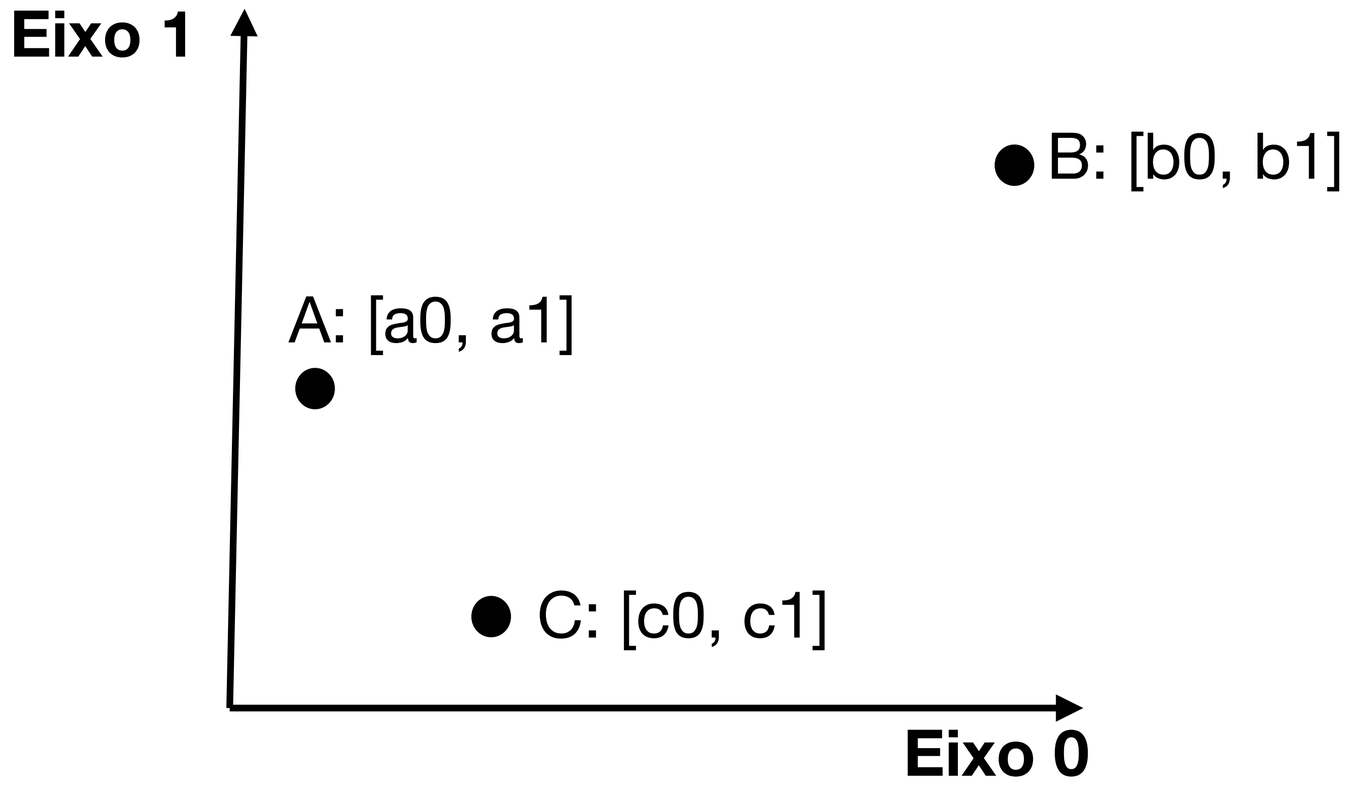
- Benchmark com dados do Movie Lens - Ok
- Apoio Google: Code review e comparação com algoritmo deles - Ok
- "Surrogate problem”
- Personas com histórico fake para testes offline
- Plugin Chrome para teste online
- **Resultado**: Recomendações pouco relevantes/correlacionadas ao histórico das personas

Now What?!!

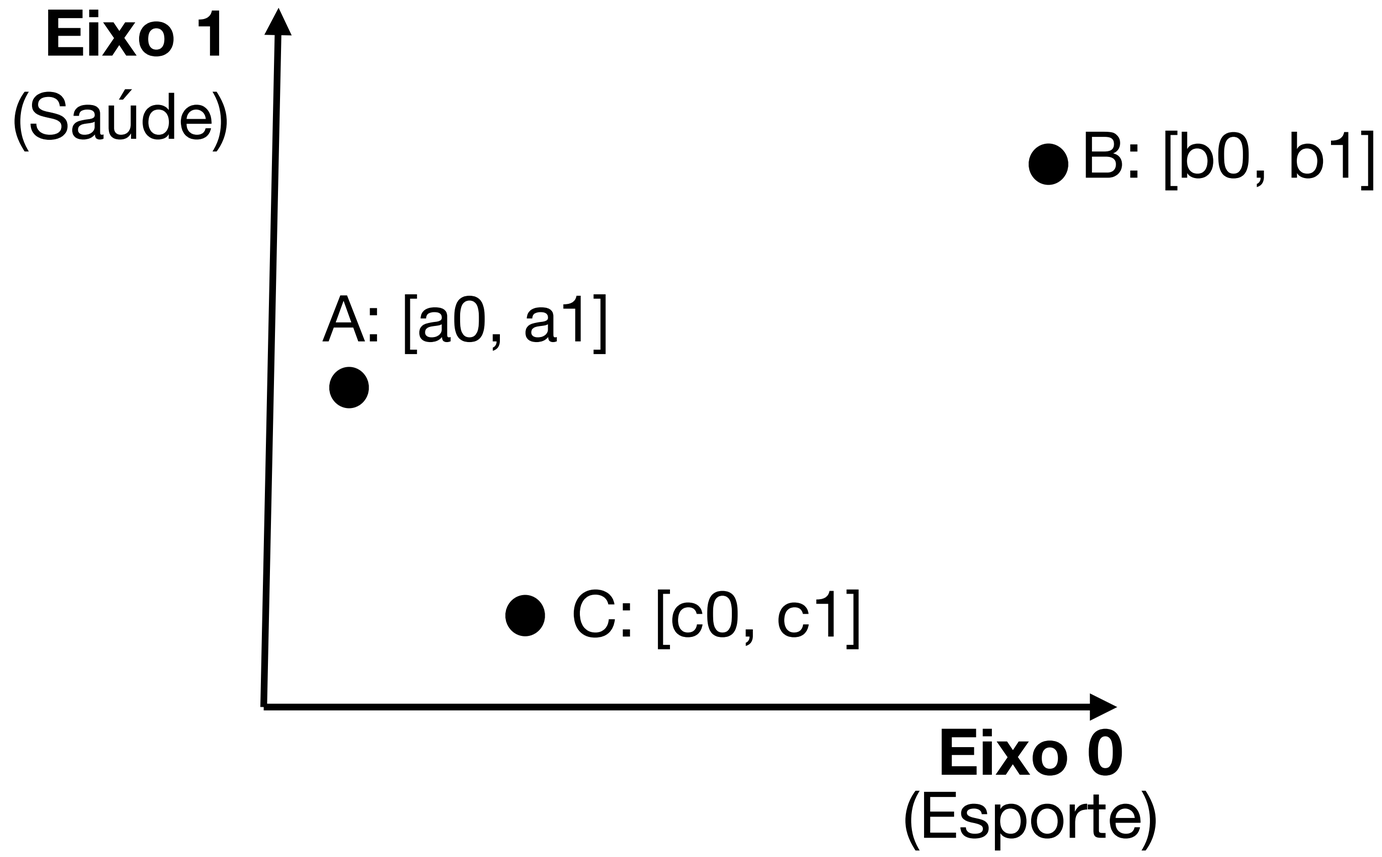


Embeddings: Notícias devem virar vetores em um espaço n-dimensional de assuntos

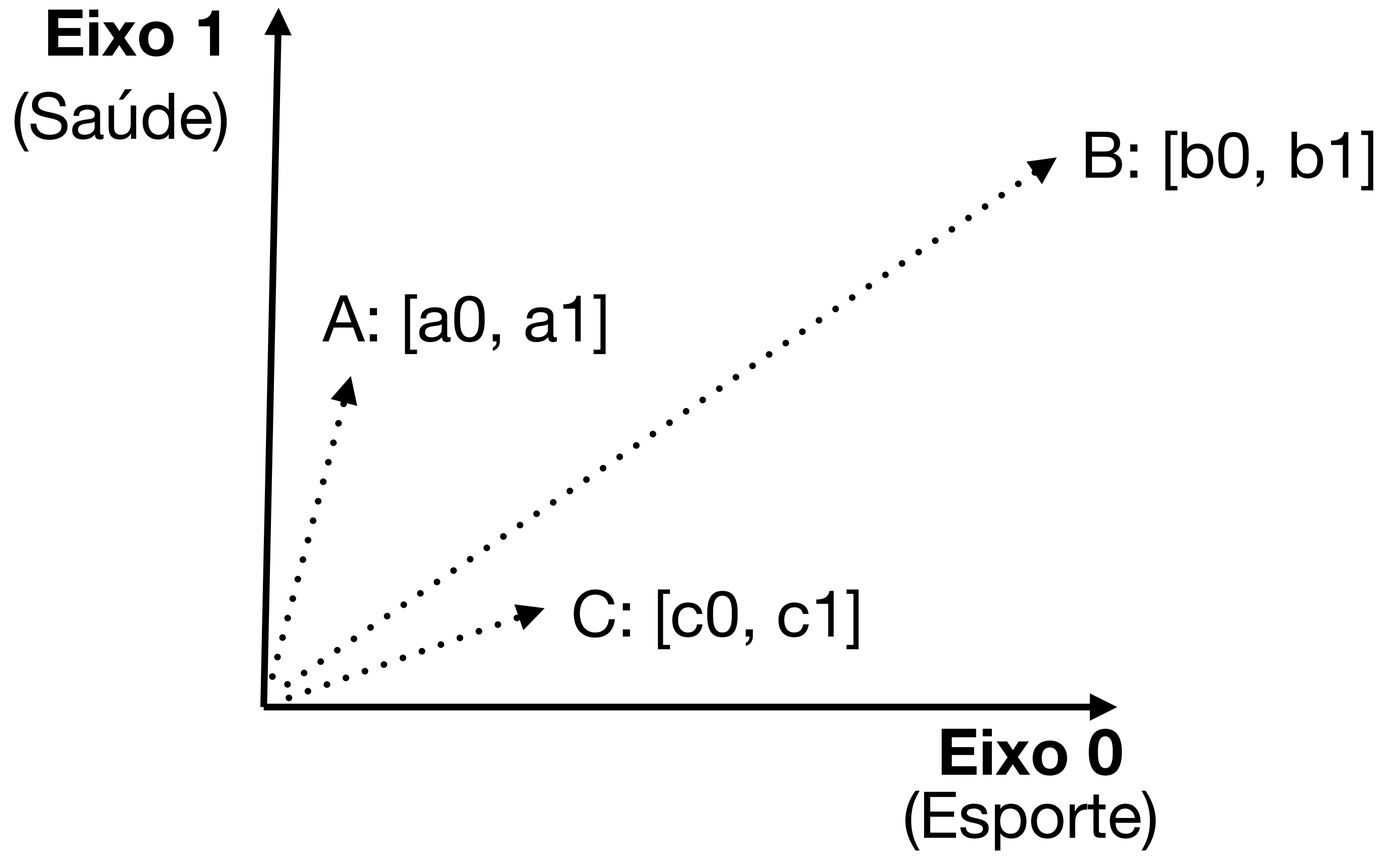
Embeddings (Intuição)



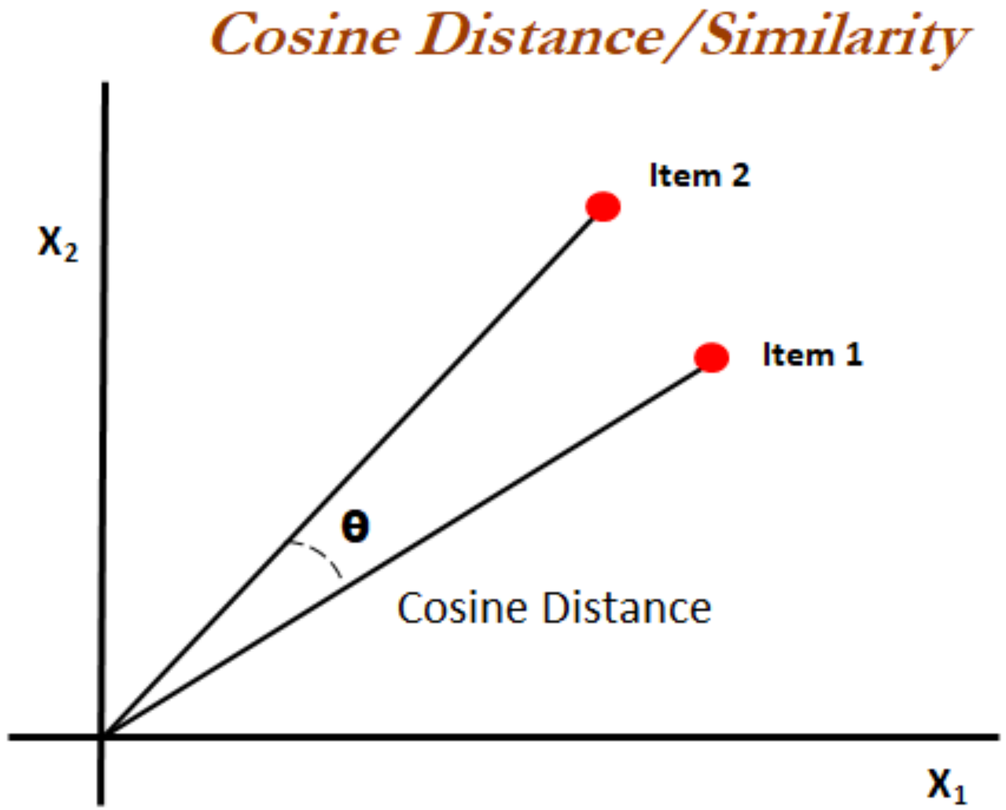
Embeddings (Intuição)



Embeddings (Intuição)



Semelhança entre notícias

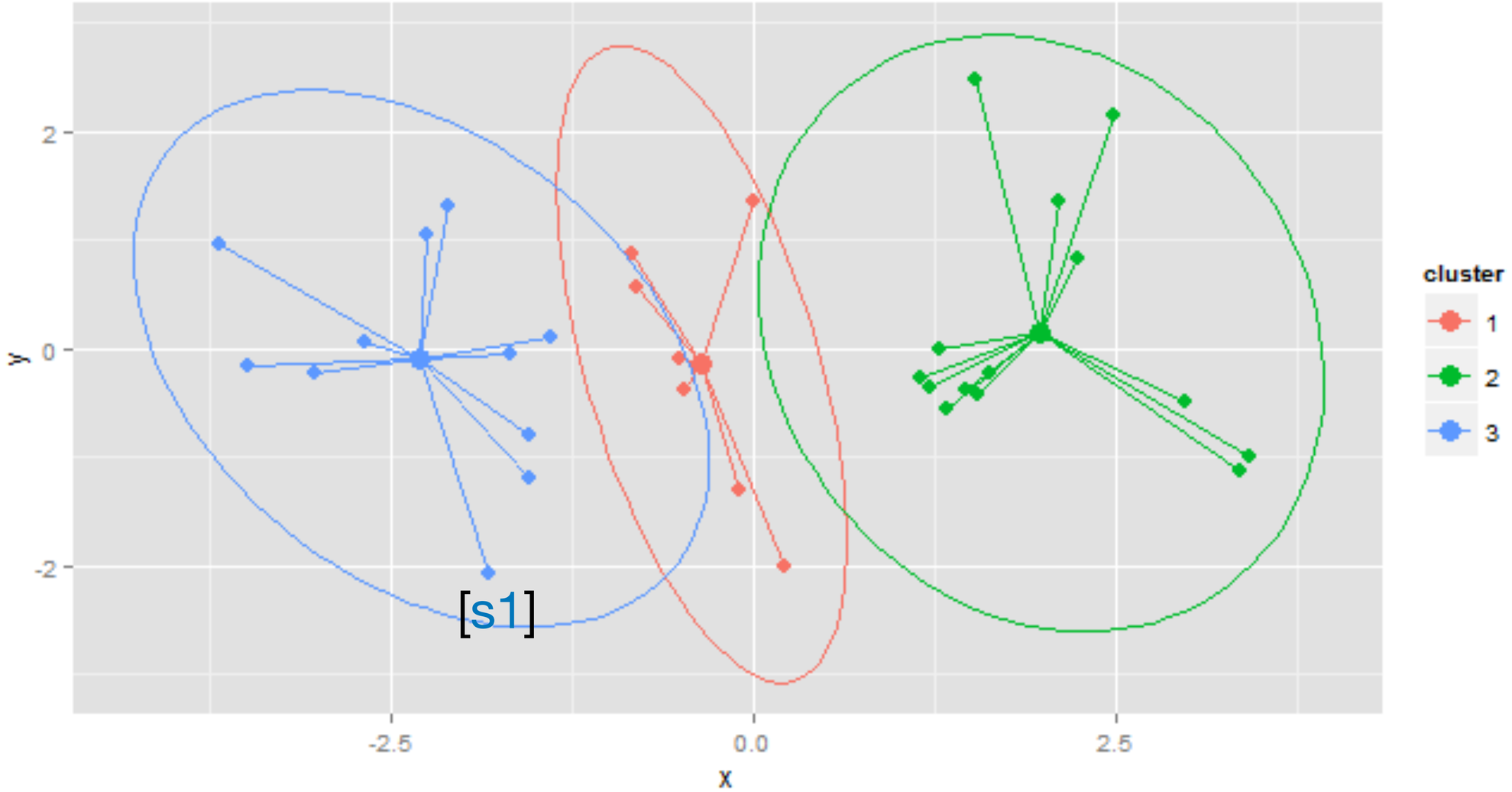


$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

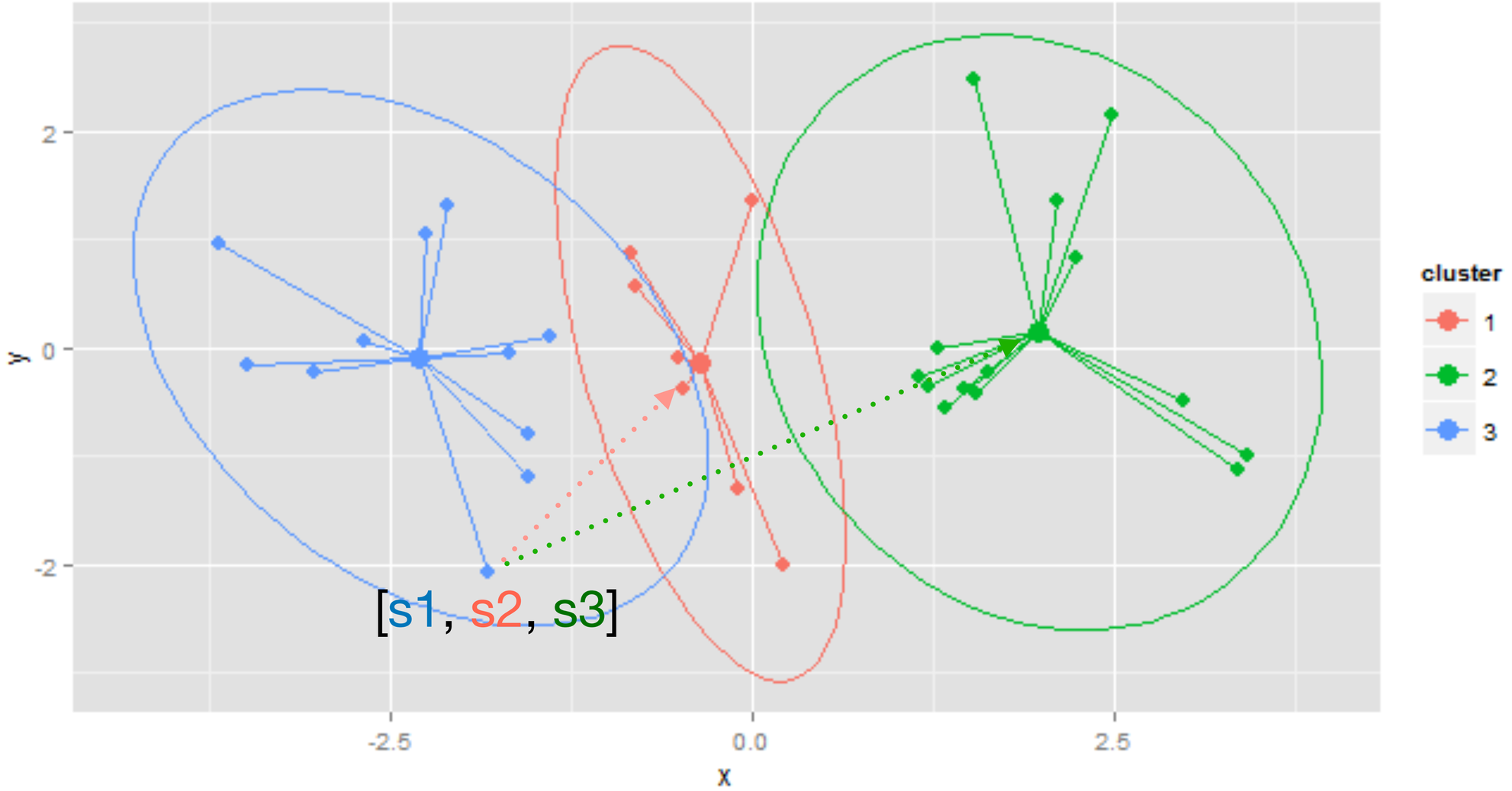
Content (text) Embeddings

- Objetivo: Transformar texto da notícia em um array de tamanho fixo
- Tags
- BOW (bag of words)
- TF-IDF (term-frequency x inverse document frequency)
- Word Embeddings (word2vec, FastText) + PV-DM (paragraph vectors distributed memory)
- Seq2Seq Autoencoder
- HANs (Hierarchical Attention Network)

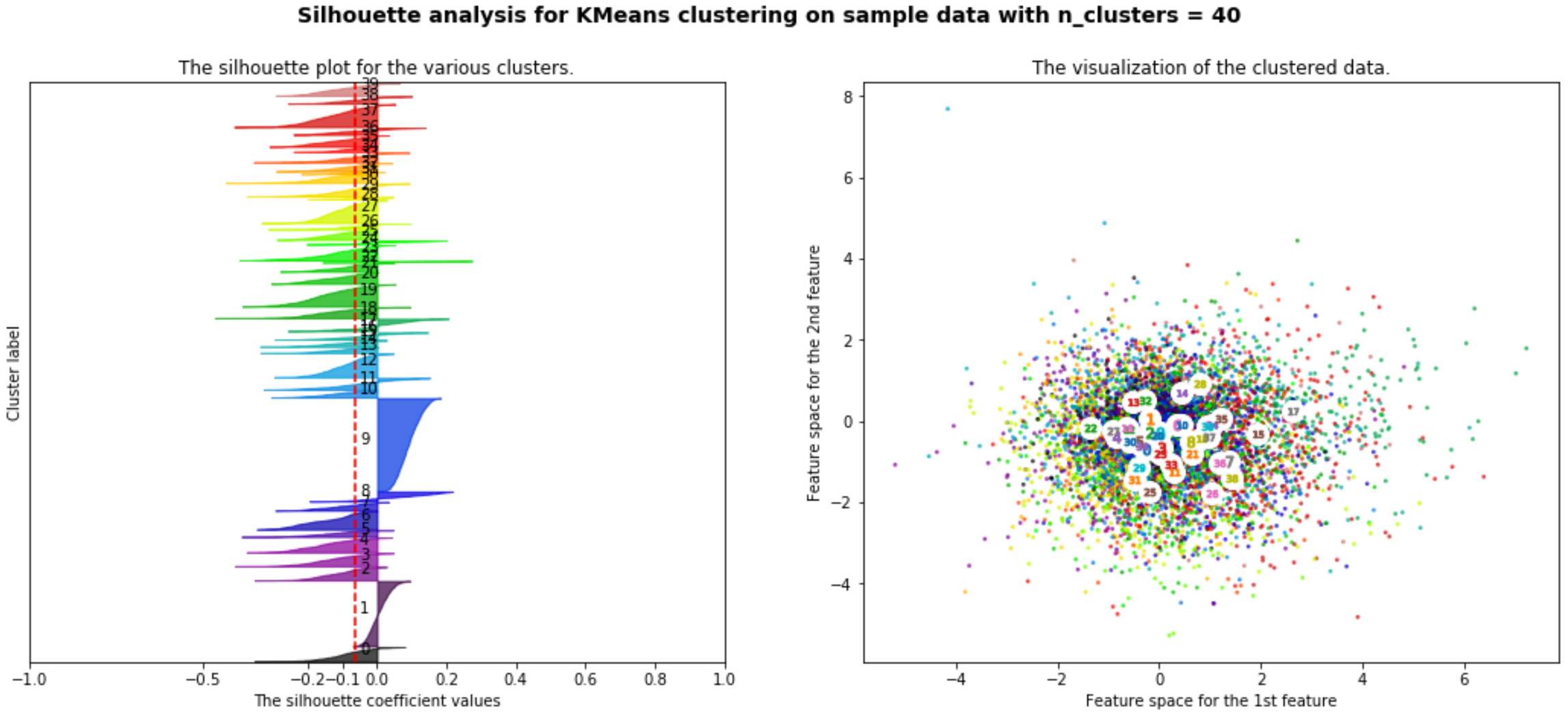
Tópicos via clustering (KMeans)



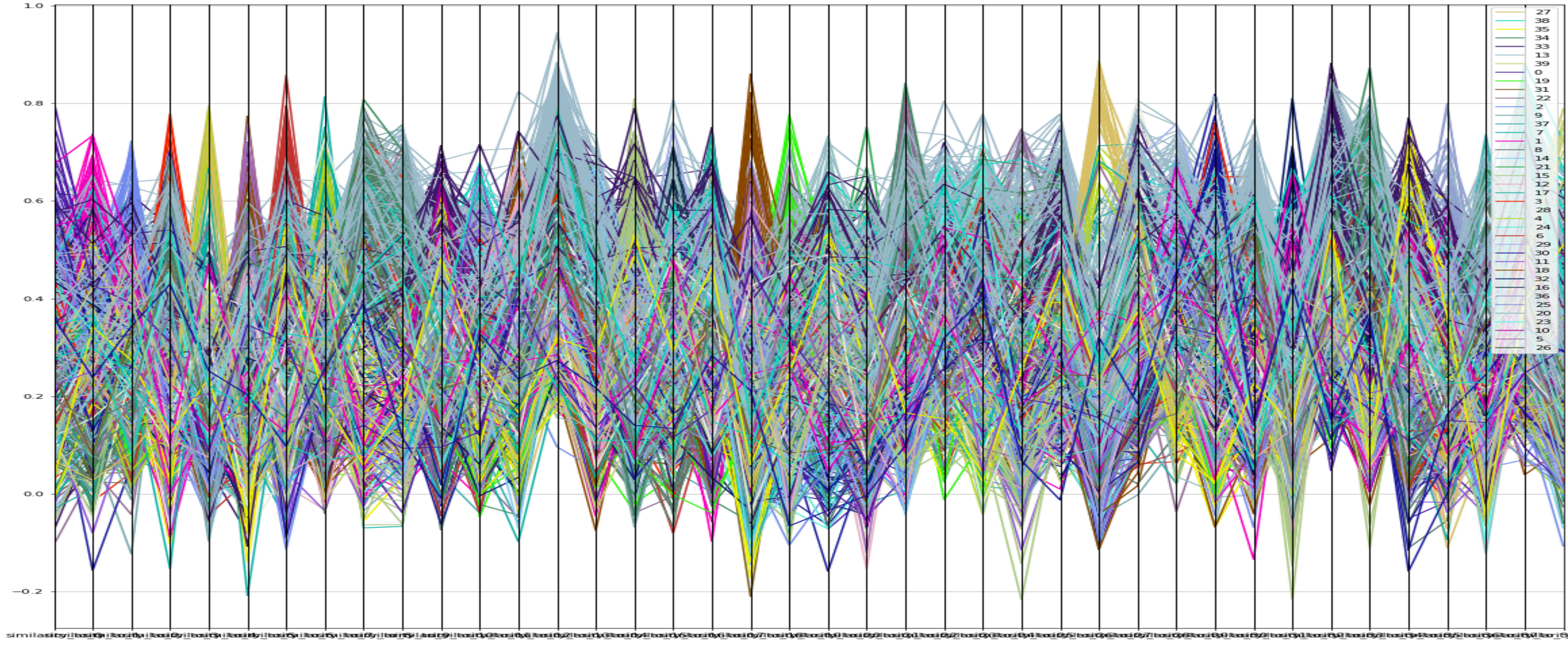
Tópicos via clustering (KMeans)



Qtd de clusters (assuntos latentes)



10k notícias em 40 dimensões

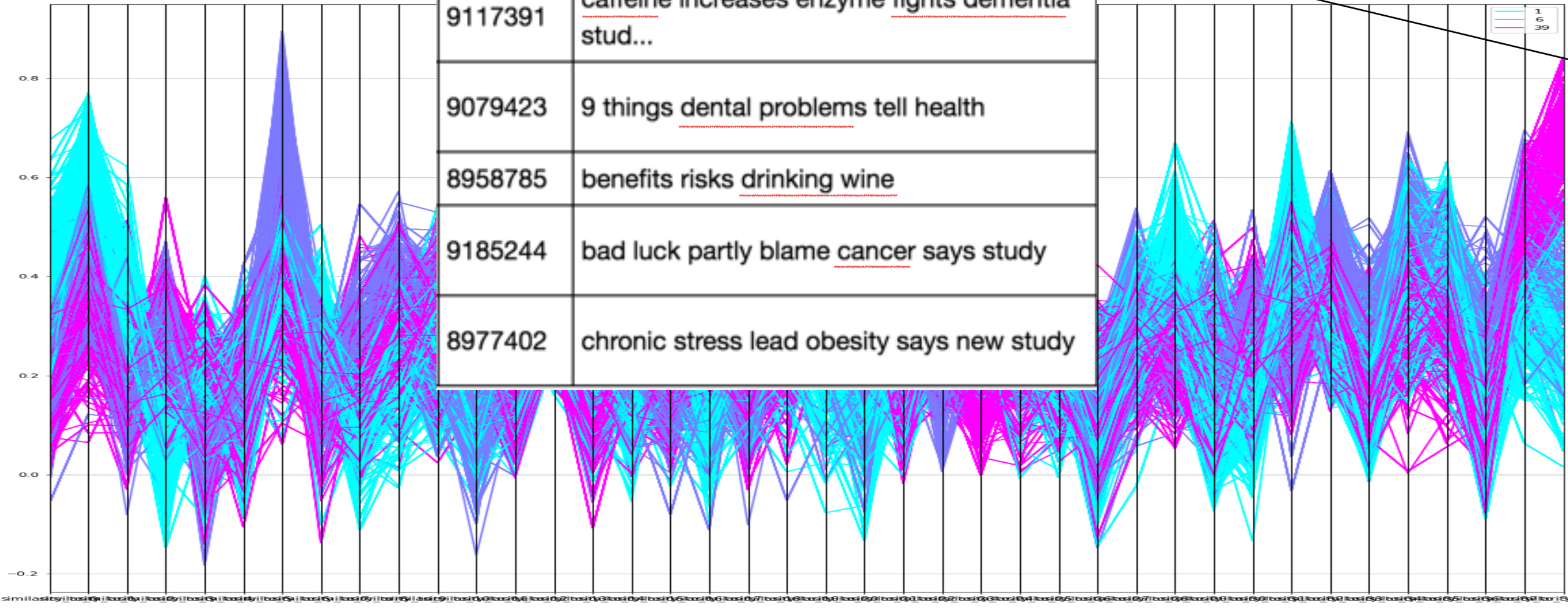


Dimensão 0: “economia/desemprego”

contentId	title
9092218	new bill could help companies pay employee stu...
9116669	us adds 235k jobs unemployment rate falls 4 7 ...
9083135	us jobless claims drop lowest level since 1973
9083100	us applications unemployment benefits drop low...
9136106	fed raises rate sees hikes us economy improves

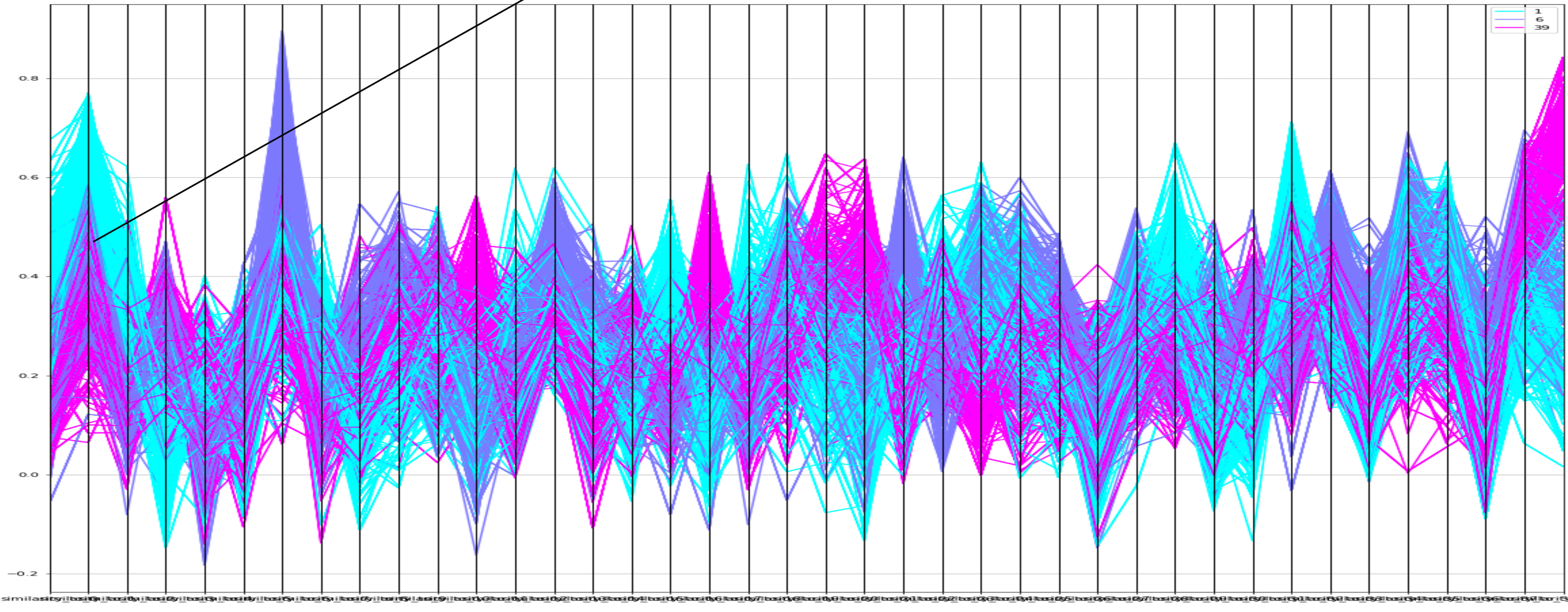
Dimensão 39: “dicas de saúde”

contentId	title
9117391	caffeine increases enzyme fights dementia stud...
9079423	9 things dental problems tell health
8958785	benefits risks drinking wine
9185244	bad luck partly blame cancer says study
8977402	chronic stress lead obesity says new study



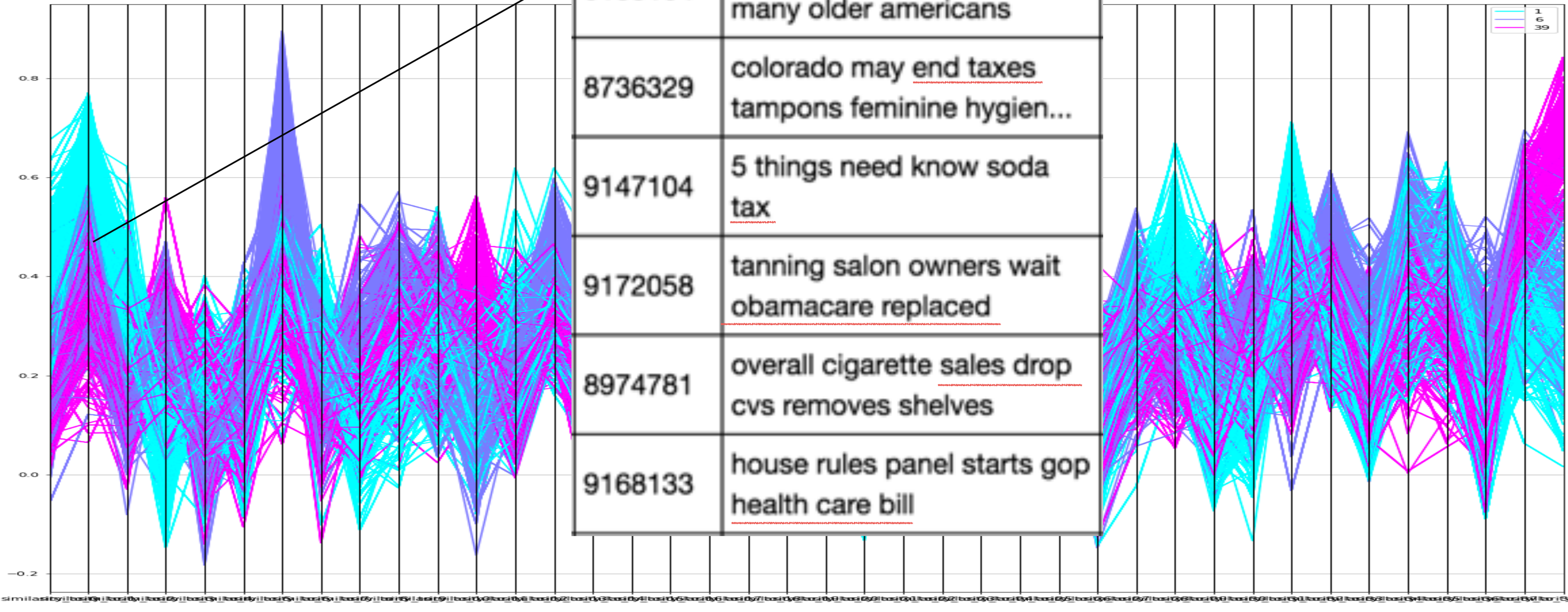
Artigos “magenta” próximo à “economia”?

?

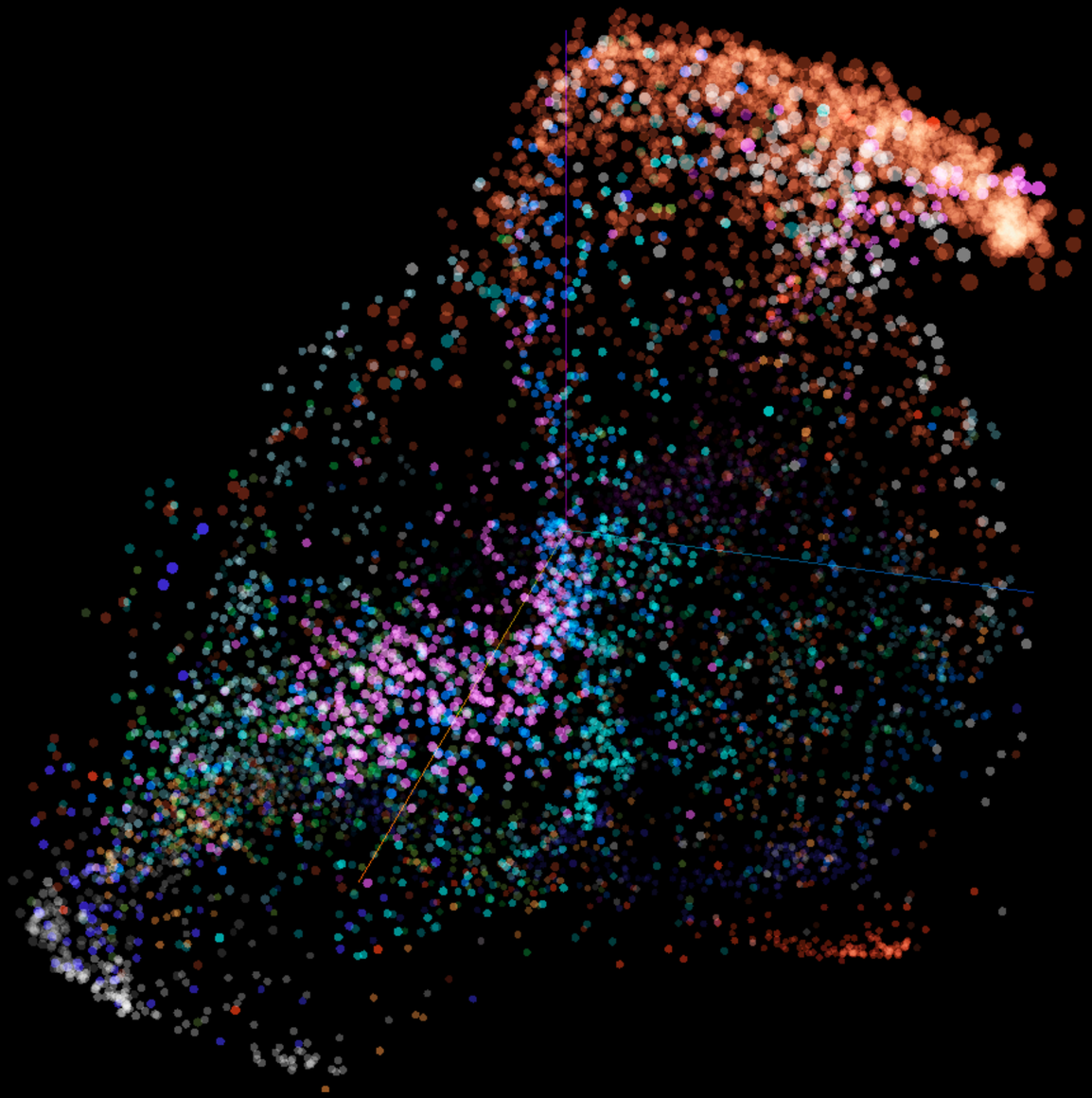


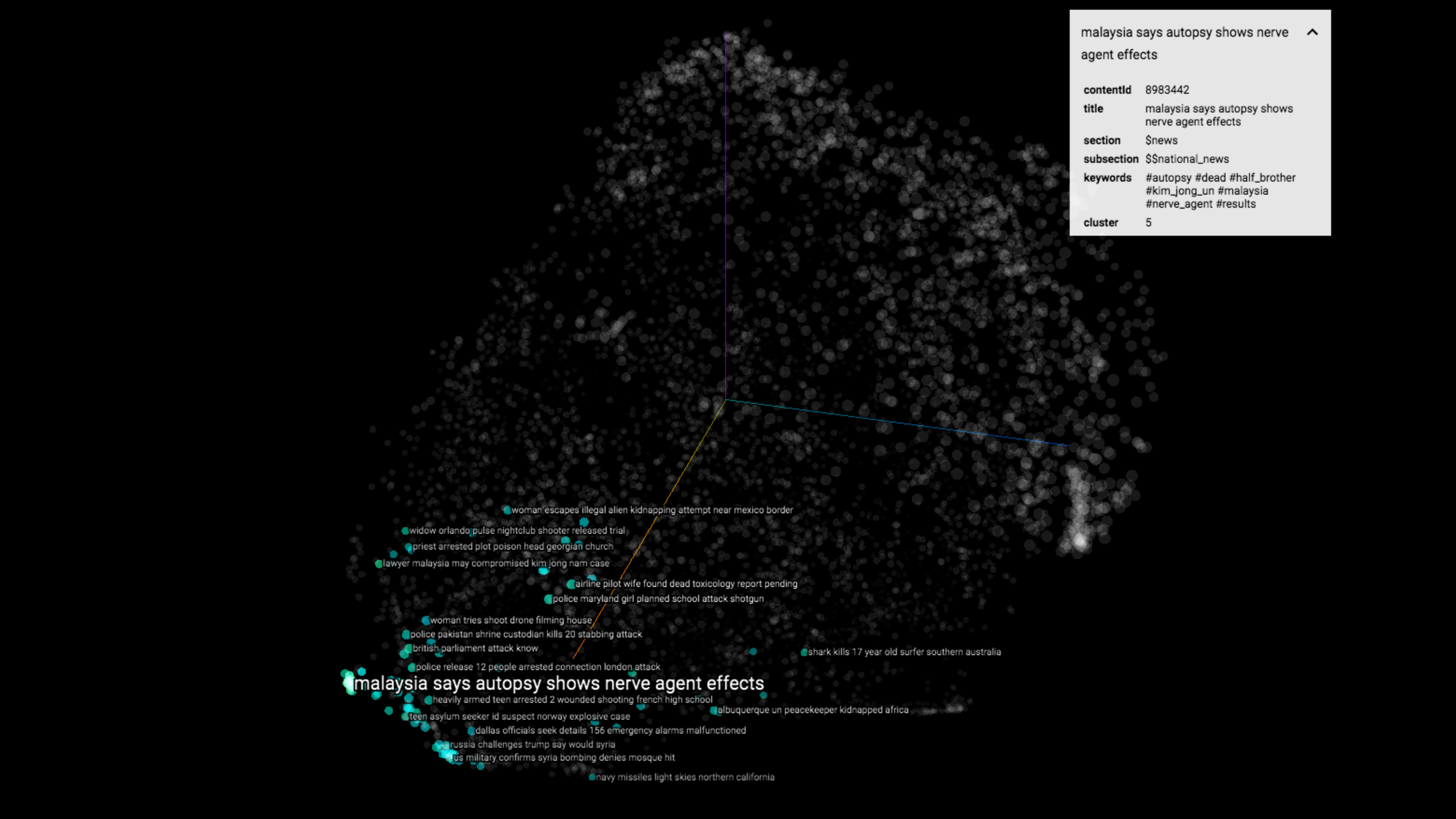
Artigos “magenta” próximo a “economia”?

contentId	title
9153164	gop health plan costs rise many older americans
8736329	colorado may end taxes tampons feminine hygien...
9147104	5 things need know soda tax
9172058	tanning salon owners wait obamacare replaced
8974781	overall cigarette sales drop cvs removes shelves
9168133	house rules panel starts gop health care bill



Visualização com T-SNE




malaysia says autopsy shows nerve agent effects
contentId 8983442
title malaysia says autopsy shows nerve agent effects
section $news
subsection $$national_news
keywords #autopsy #dead #half_brother #kim_jong_un #malaysia #nerve_agent #results
cluster 5
woman escapes illegal alien kidnapping attempt near mexico border
widow orlando pulse nightclub shooter released trial
priest arrested plot poison head georgian church
lawyer malaysia may compromised kim jong nam case
airline pilot wife found dead toxicology report pending
police maryland girl planned school attack shotgun
woman tries shoot drone filming house
police pakistan shrine custodian kills 20 stabbing attack
british parliament attack know
police release 12 people arrested connection london attack
shark kills 17 year old surfer southern australia
malaysia says autopsy shows nerve agent effects
heavily armed teen arrested 2 wounded shooting french high school
albuquerque un peacekeeper kidnapped africa
teen asylum seeker id suspect norway explosive case
dallas officials seek details 156 emergency alarms malfunctioned
russia challenges trump say would syria
us military confirms syria bombing denies mosque hit
navy missiles light skies northern california

Boston News, Weather and Sp ×

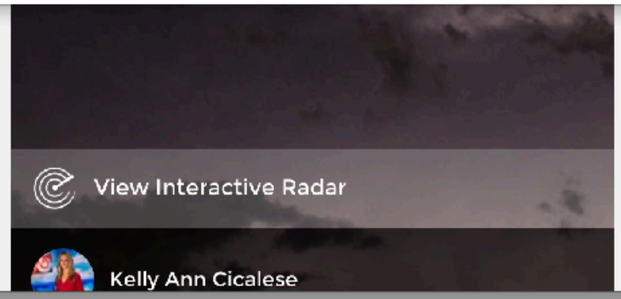
Primeiro usuár...

www.wcvb.com

NOWCAST WCVB On Demand

MENU

WCVB 5 abc



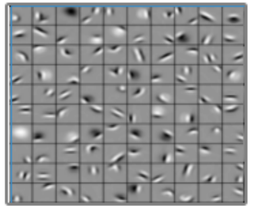
CREam Engine

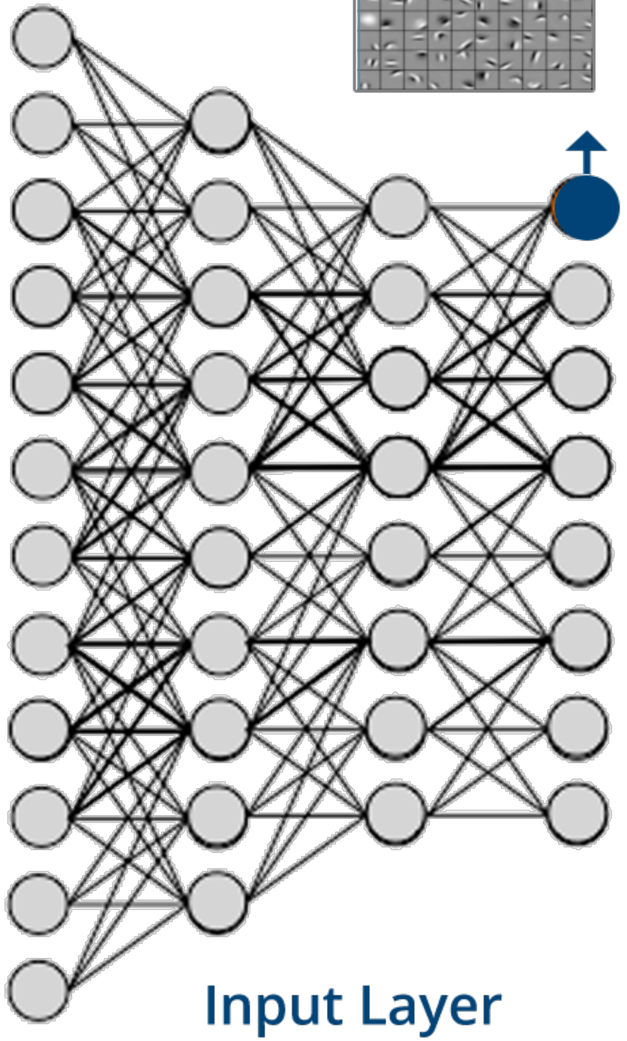
USER ID 06d3a9b57af42c870991347f4bf2fe7b

TITLE	DATE	CONTENT	DISPLAY
Profile in Courage Award: Obama reflects on courage, health care	May 7, 2017 11:21 PM	Video	Standard Article
Obama talks about the courage required to pass the ACA	May 7, 2017 11:37 PM	Video	Standard Article
Trooper chases drunken driving suspect who rear-ended cruiser	May 7, 2017 11:10 PM	Default	Standard Article
Is that Facebook post claiming 'Friends' is returning real?	May 7, 2017 10:11 PM	Default	Standard Article
Twin sisters give birth to baby boys on same day	May 8, 2017 1:24 AM	Video	Standard Article
1 dead, 1 in serious condition after double stabbing in Milford	May 7, 2017 11:42 PM	Default	Standard Article
Police Pull Over Couple To Reveal Pregnancy Announcement	May 7, 2017 2:37 PM	Video	Standard Article
Obama reflects on health care, meaning of courage in award acceptance speech	May 7, 2017 11:32 PM	Default	Standard Article
Teacher sells plasma just to make enough for his family of four	May 7, 2017 4:52 PM	Video	Standard Article

Parte #2 - Deep Learning

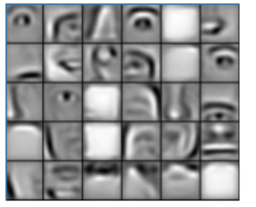
Patterns of Local Contrast

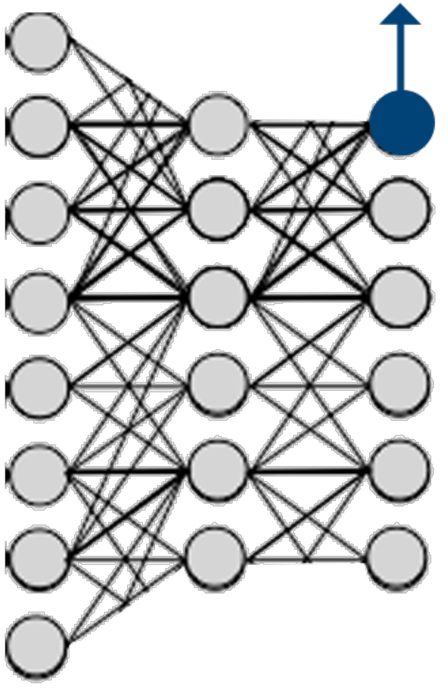




Input Layer

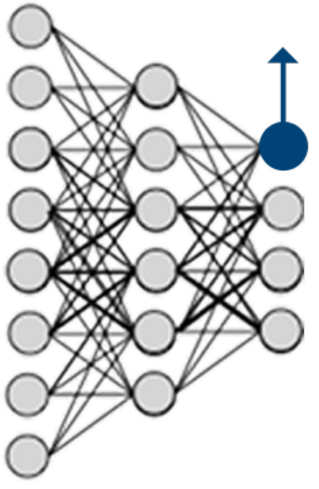
Face Features



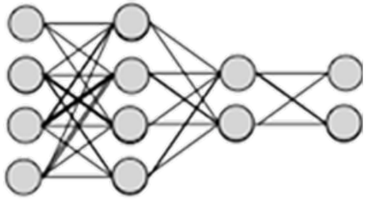


Hidden Layer 1

Face



Hidden Layer 2



Output Layer

Motivações para usar DNN

- V1 captura de cada usuário a intensidade de seu interesse por cada tópico latente
- Queremos capturar a dinâmica comportamental do usuário em diferentes contextos:
  - Temporal: Dia da semana, horário, dia do mês, dia do ano, etc…
  - Assuntos: Sequência dos últimos artigos visitados e seus assuntos.
  - Localização: Em casa? No trabalho? Passeando?

Referência: Survey de Jul/2017

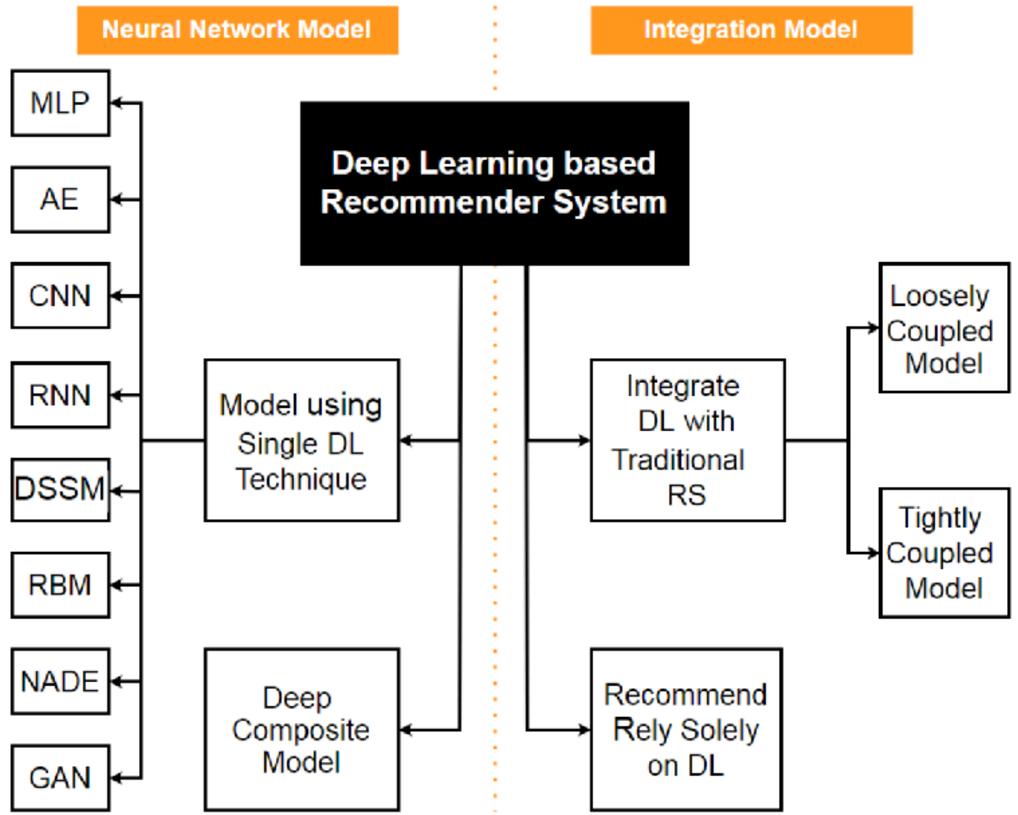
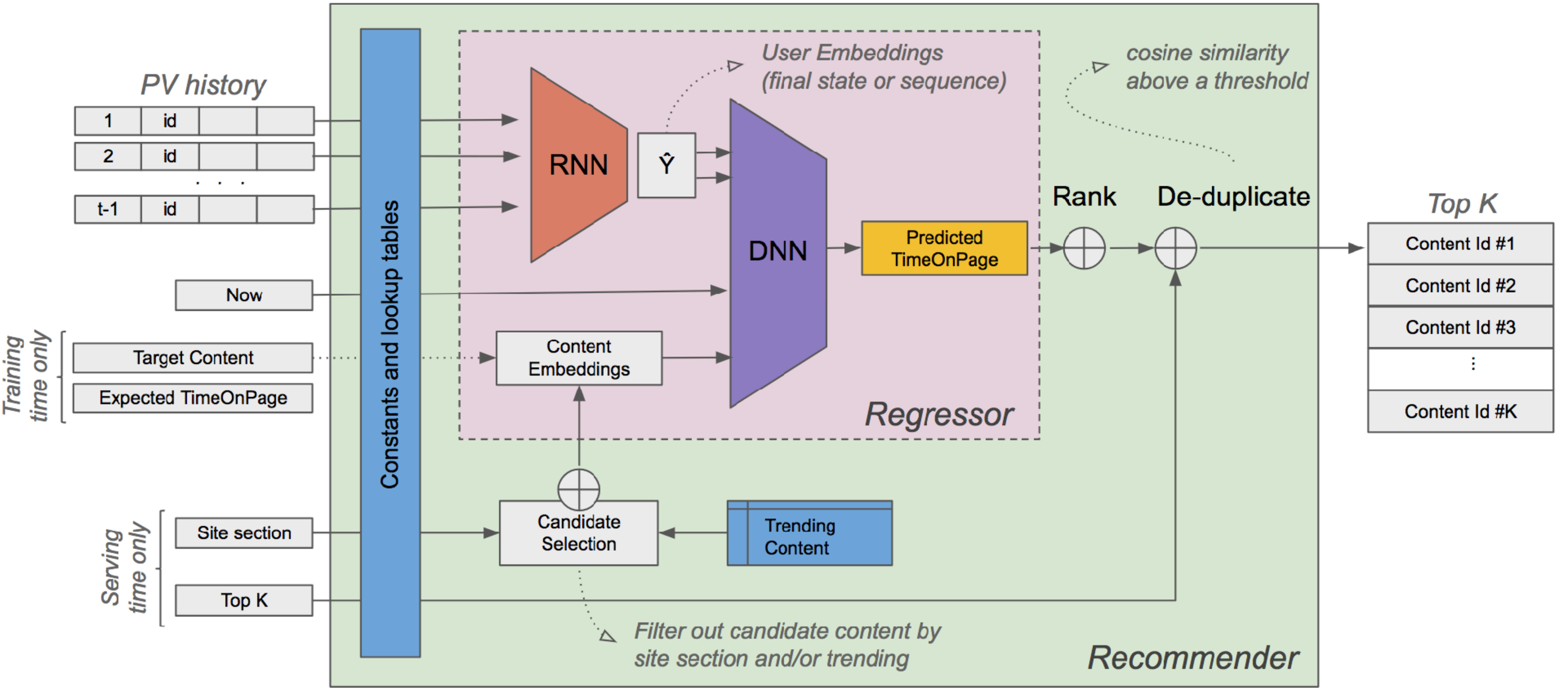


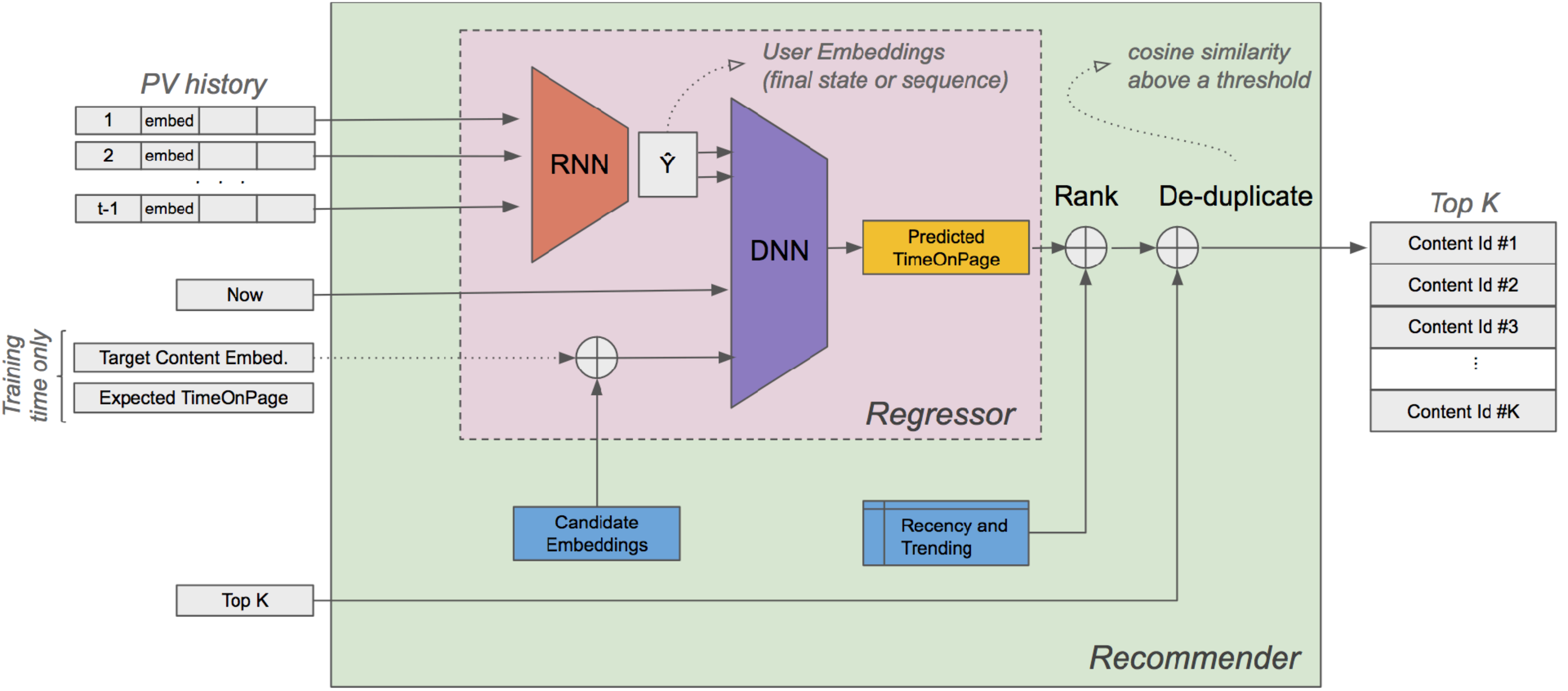
Fig. 1. Two-dimension scheme for classification of deep learning based recommender system. The left part illustrates the neural network models, and the right part illustrates the integration models.

- Neural Autoregressive Distribution Estimation (NADE) [57, 108] is an unsupervised neural network built atop autoregressive model and feedforward neural network. It is a tractable and efficient estimator for modelling data distributions and densities.

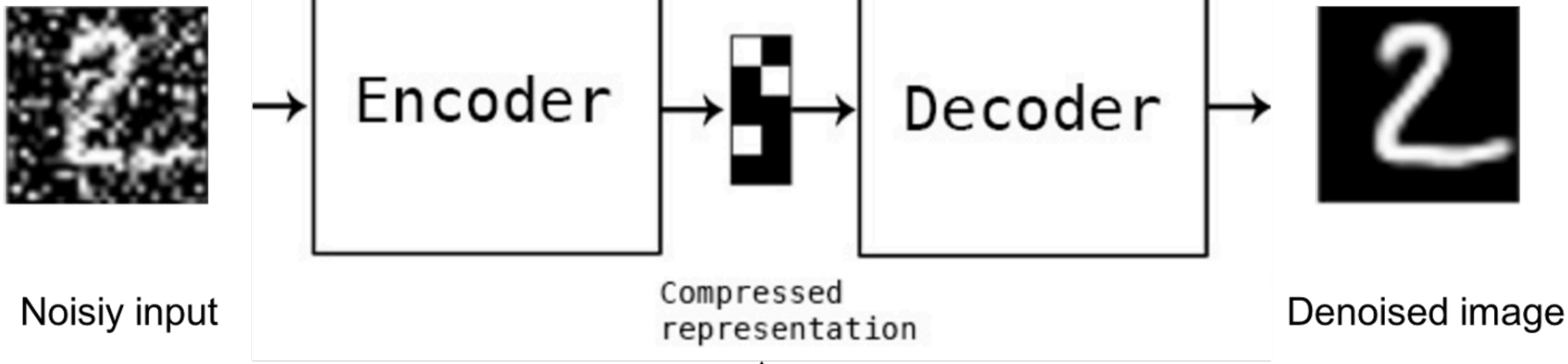
Proposta Inicial



Proposta após feedback do Google

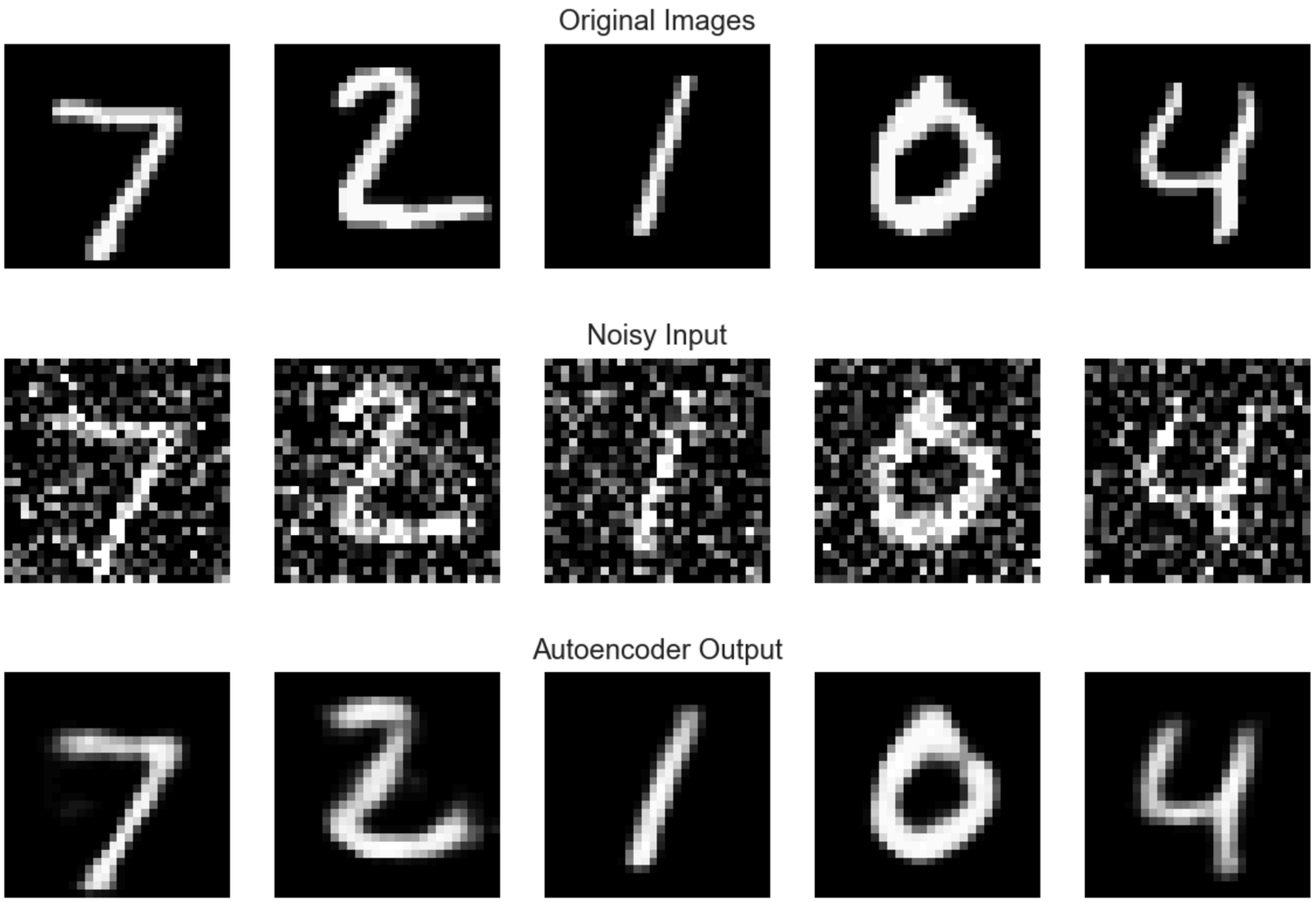


Deep Autoencoder

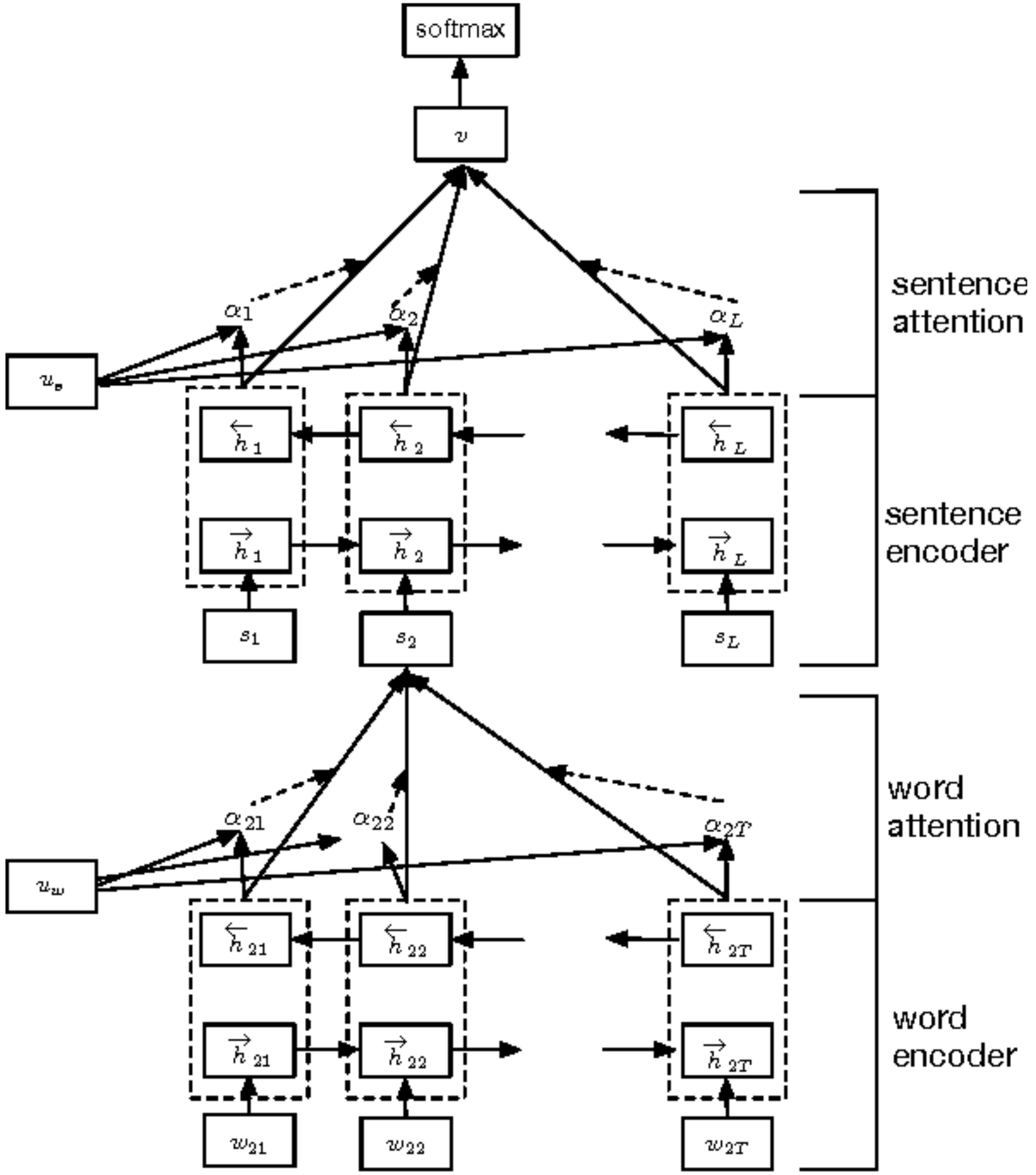


The feature we want to extract from the image

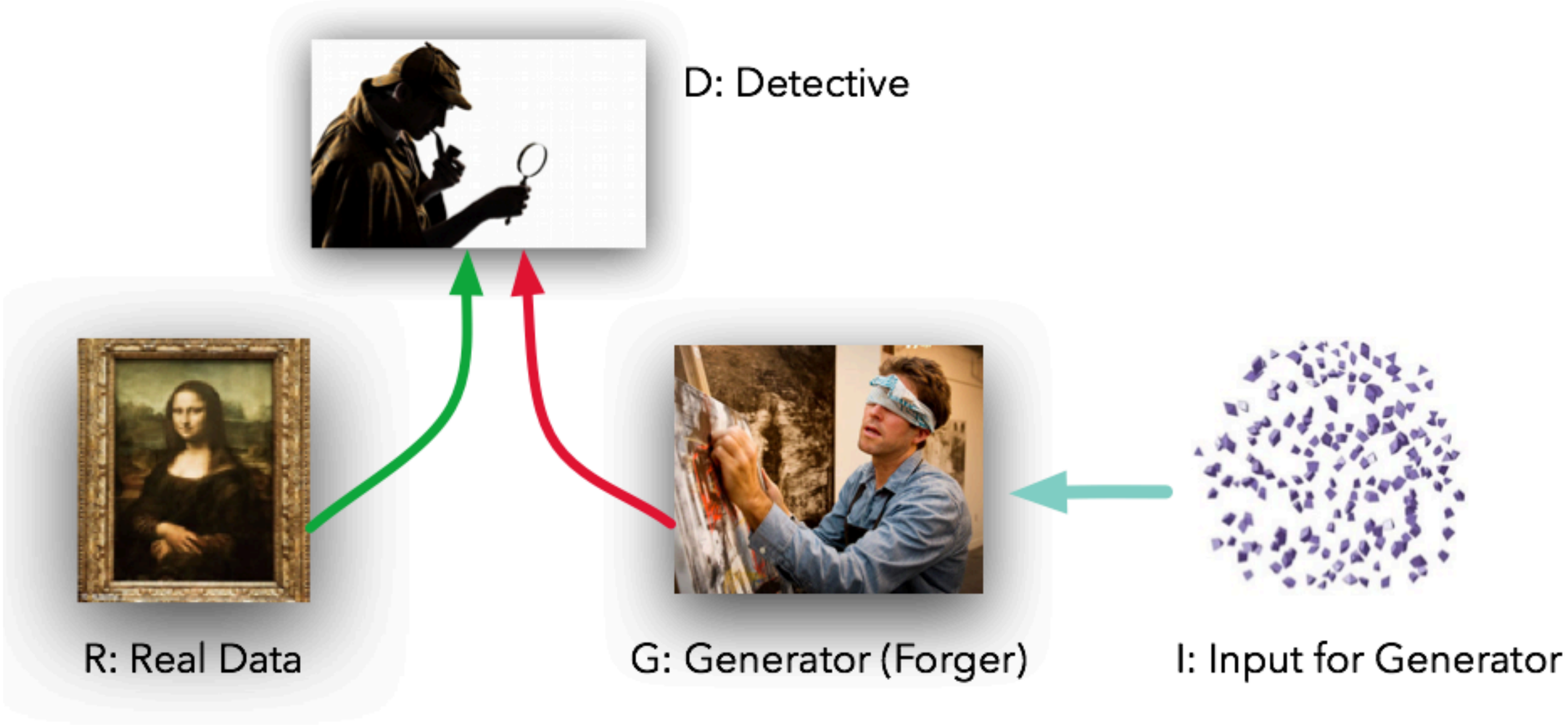
Deep Autoencoder



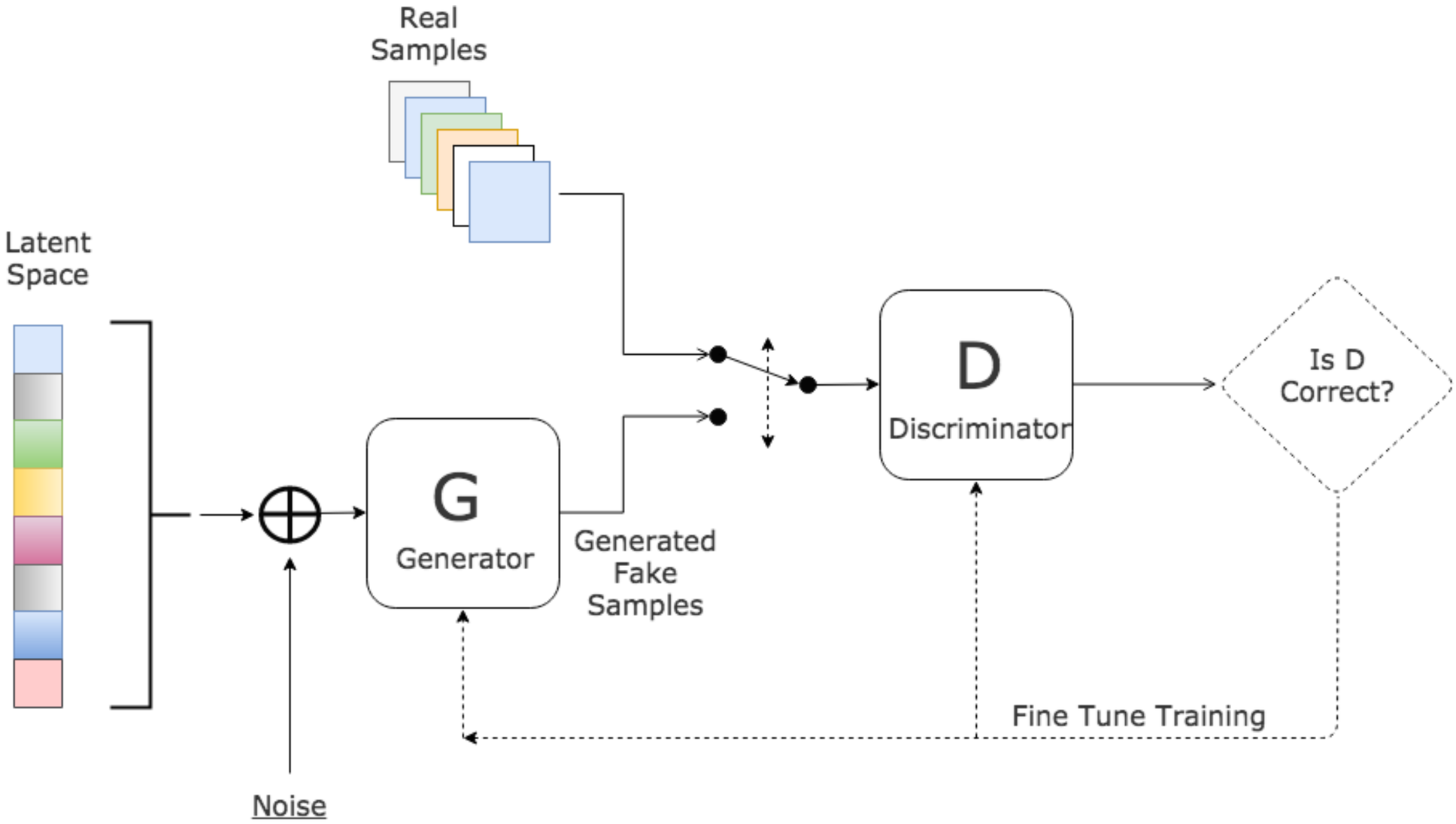
HANs: Hierarchical Attention Network



GAN: Generative Adversarial Network (2013)



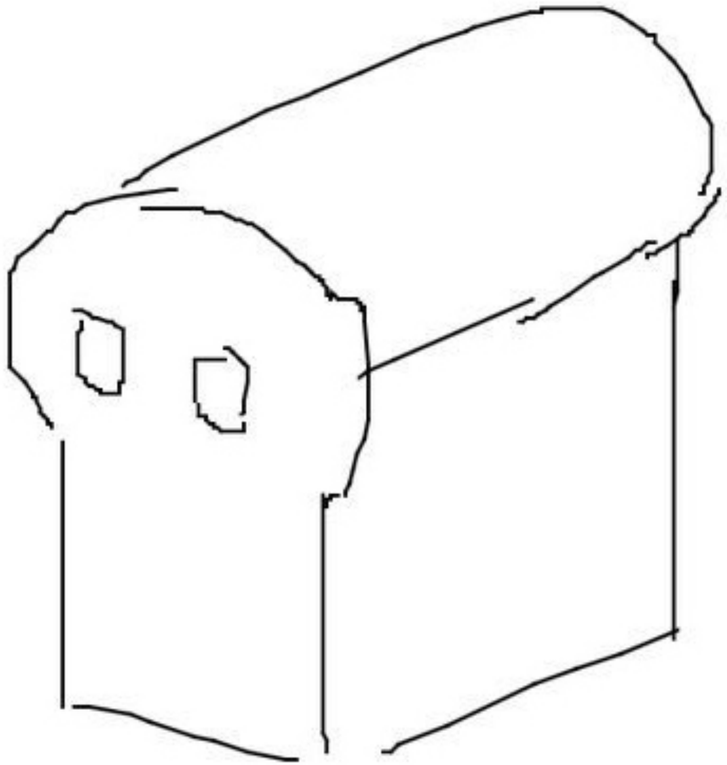
GAN: Generative Adversarial Network (2013)



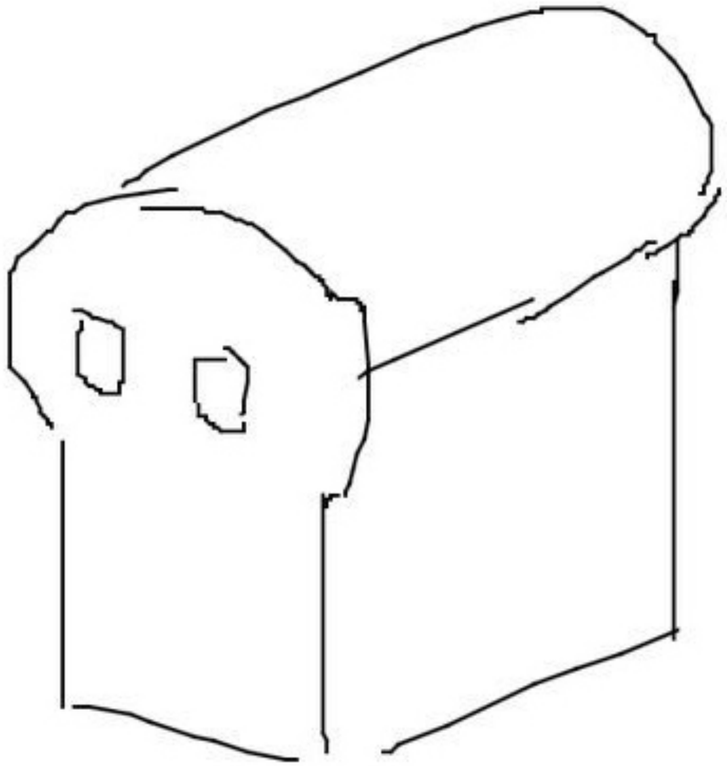
Espaço Latente “Gato”

Gato fake

Pão de forma => ?



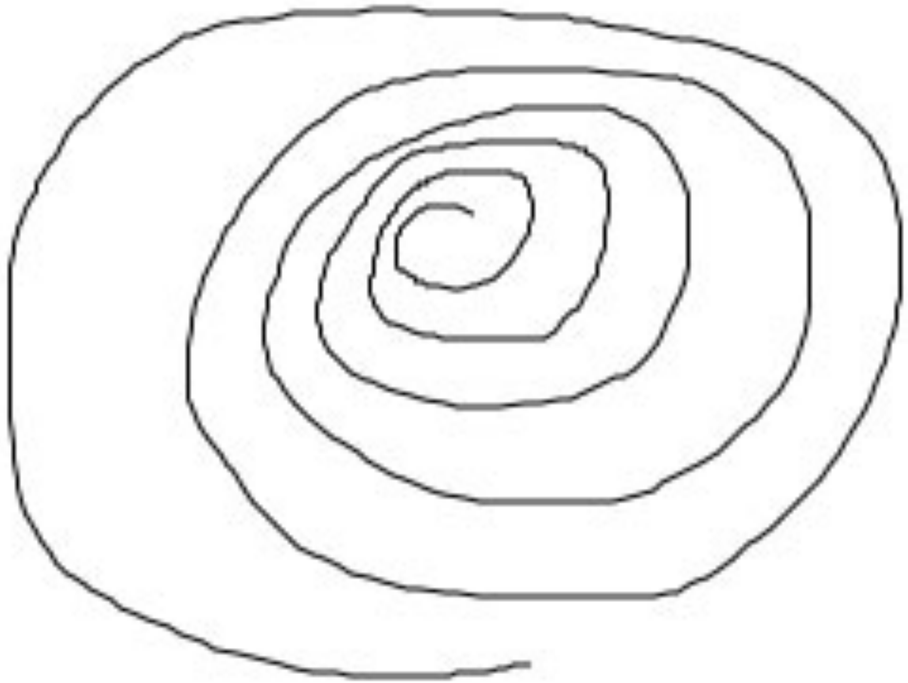
Pão de forma => “Pão Gato” :-)



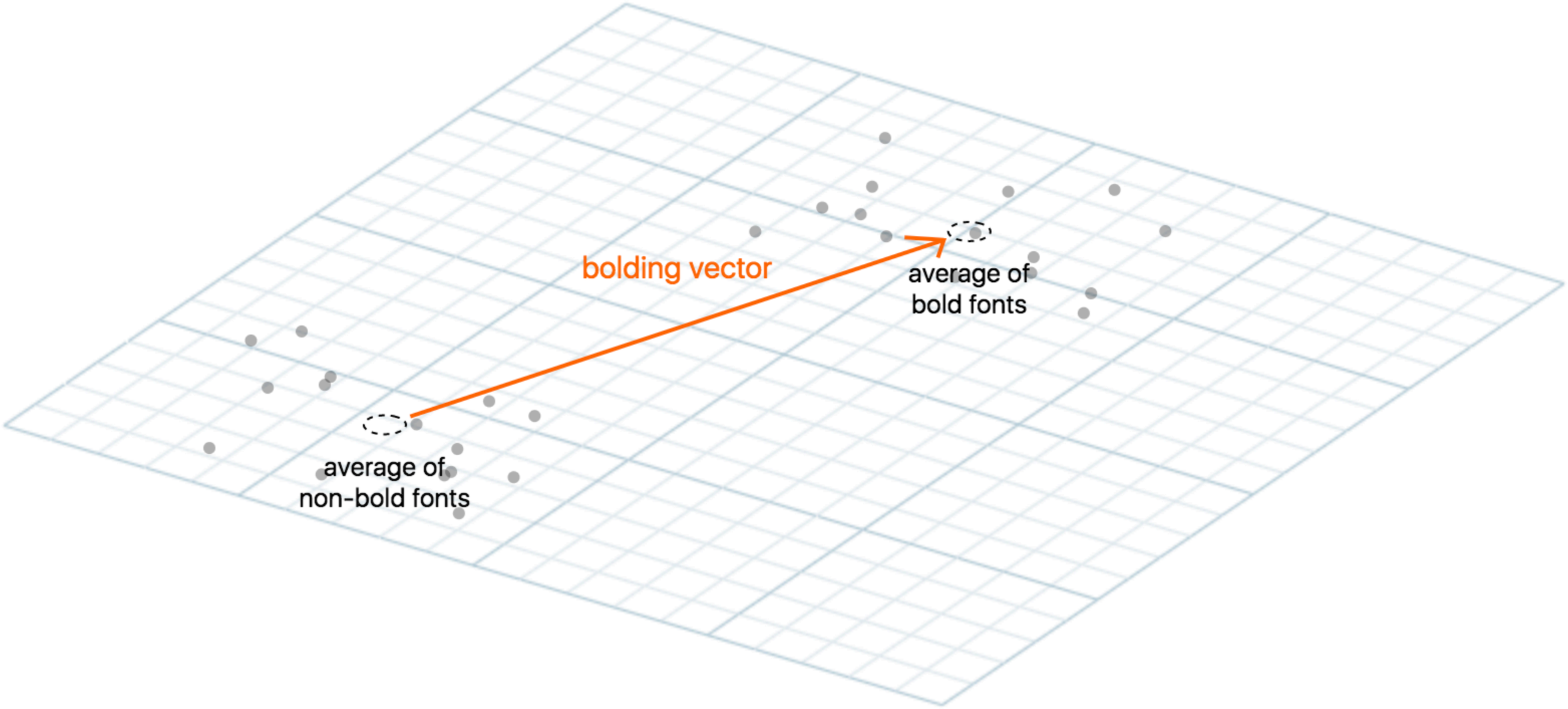
Beholder (Dungeons & Dragons)

Beholder Gato

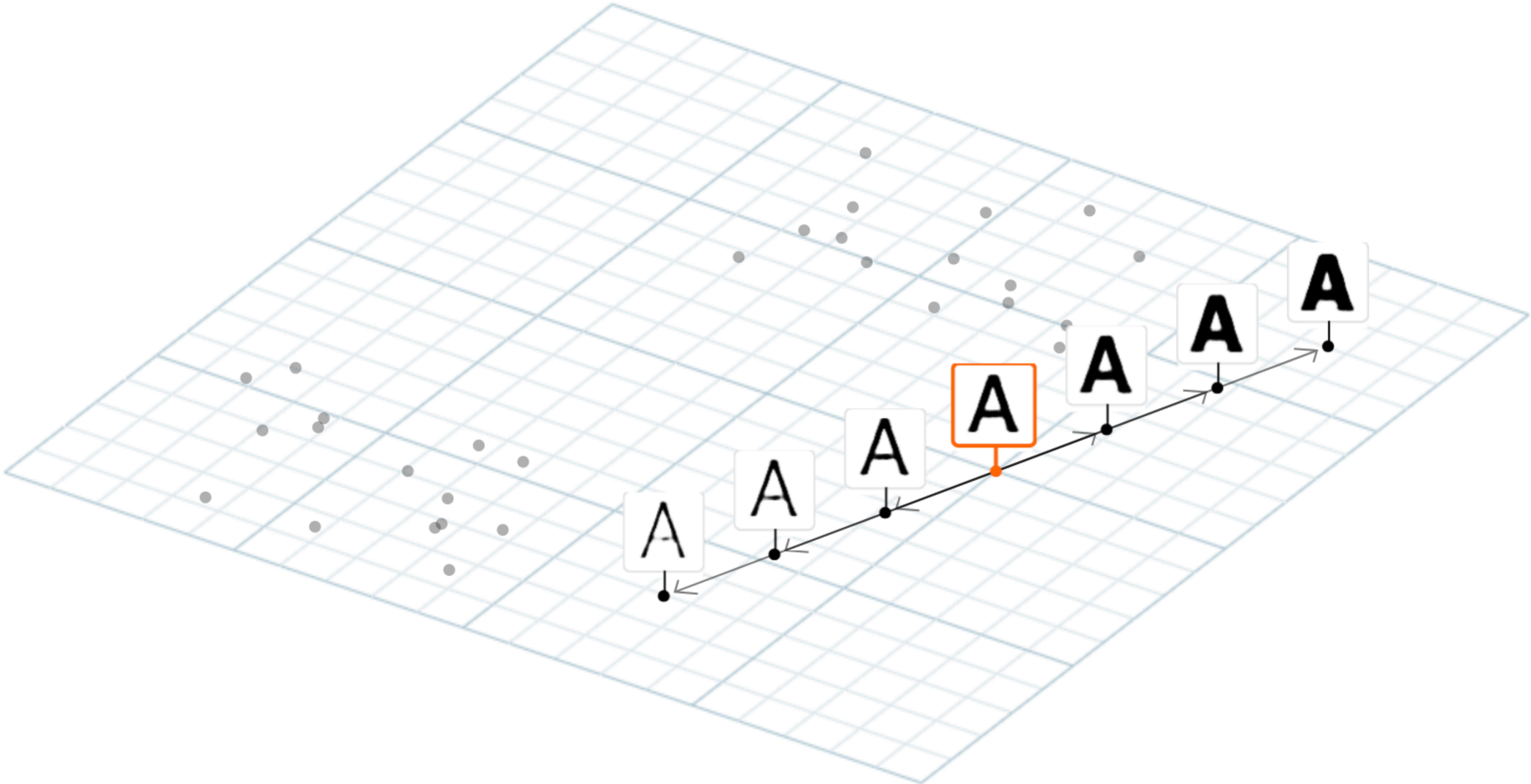
Gato espiral



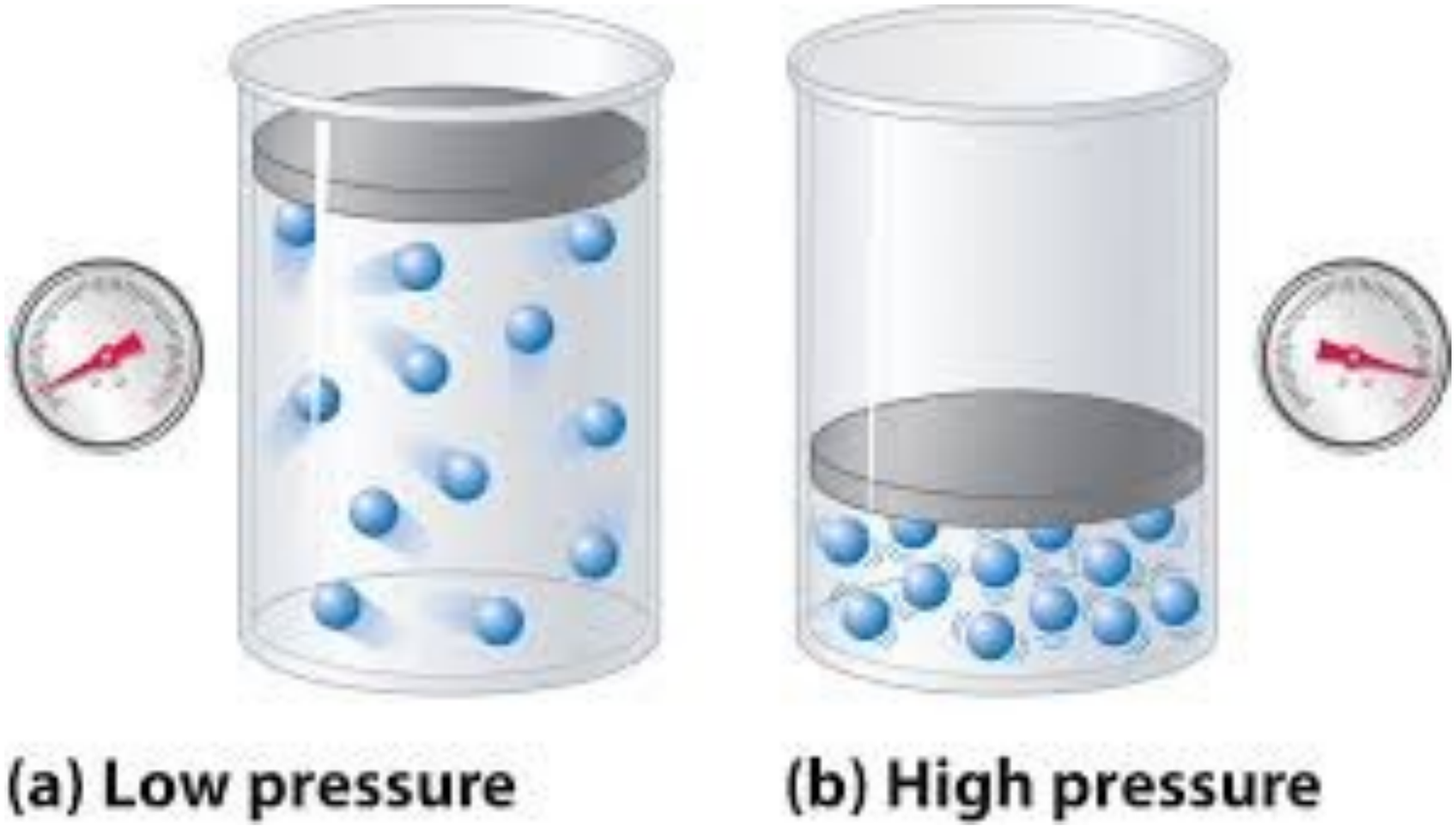
Estilo de Fontes



Estilo de Fontes



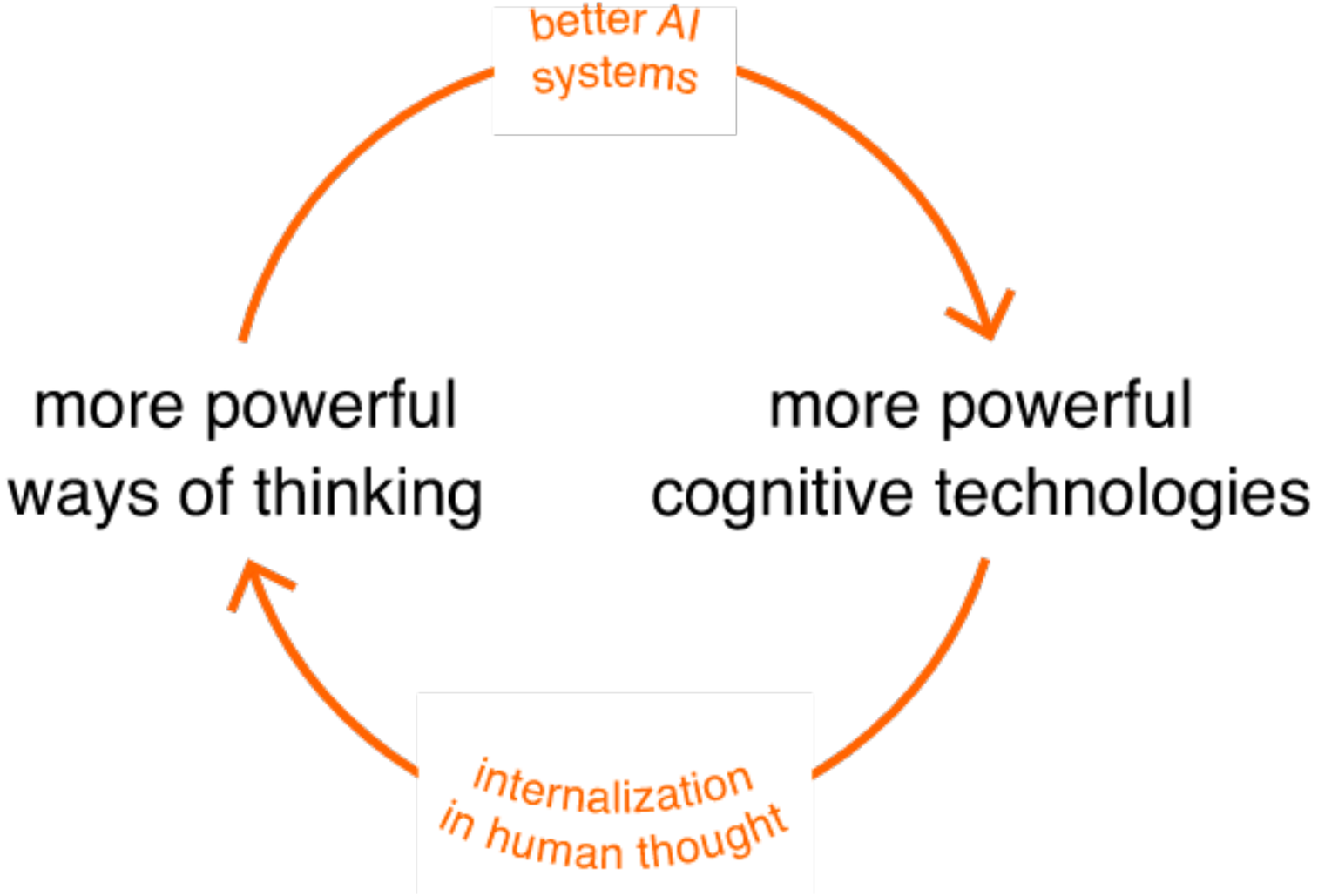
Novas Primitivas



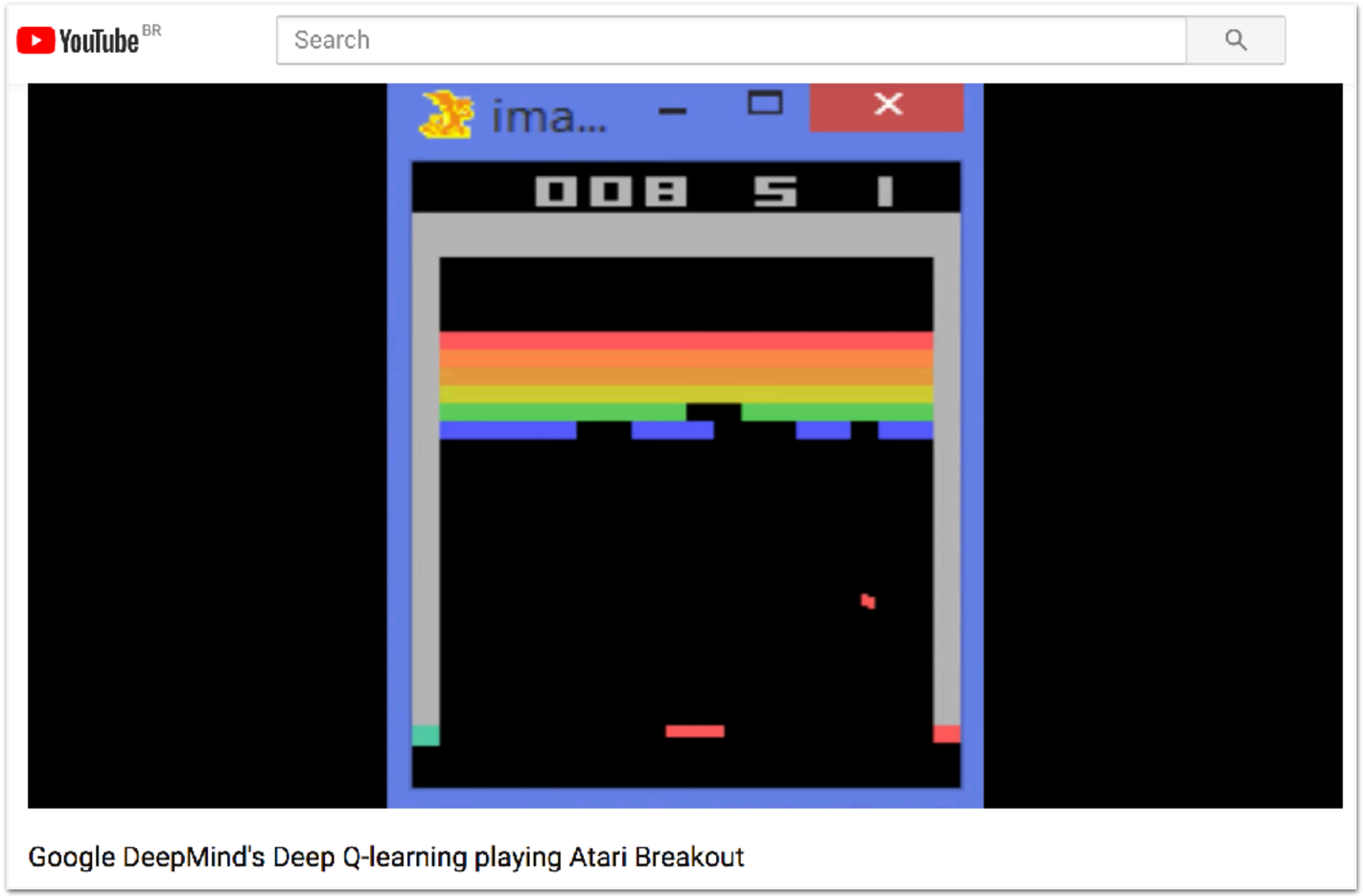
(a) Low pressure

(b) High pressure

Singularidade

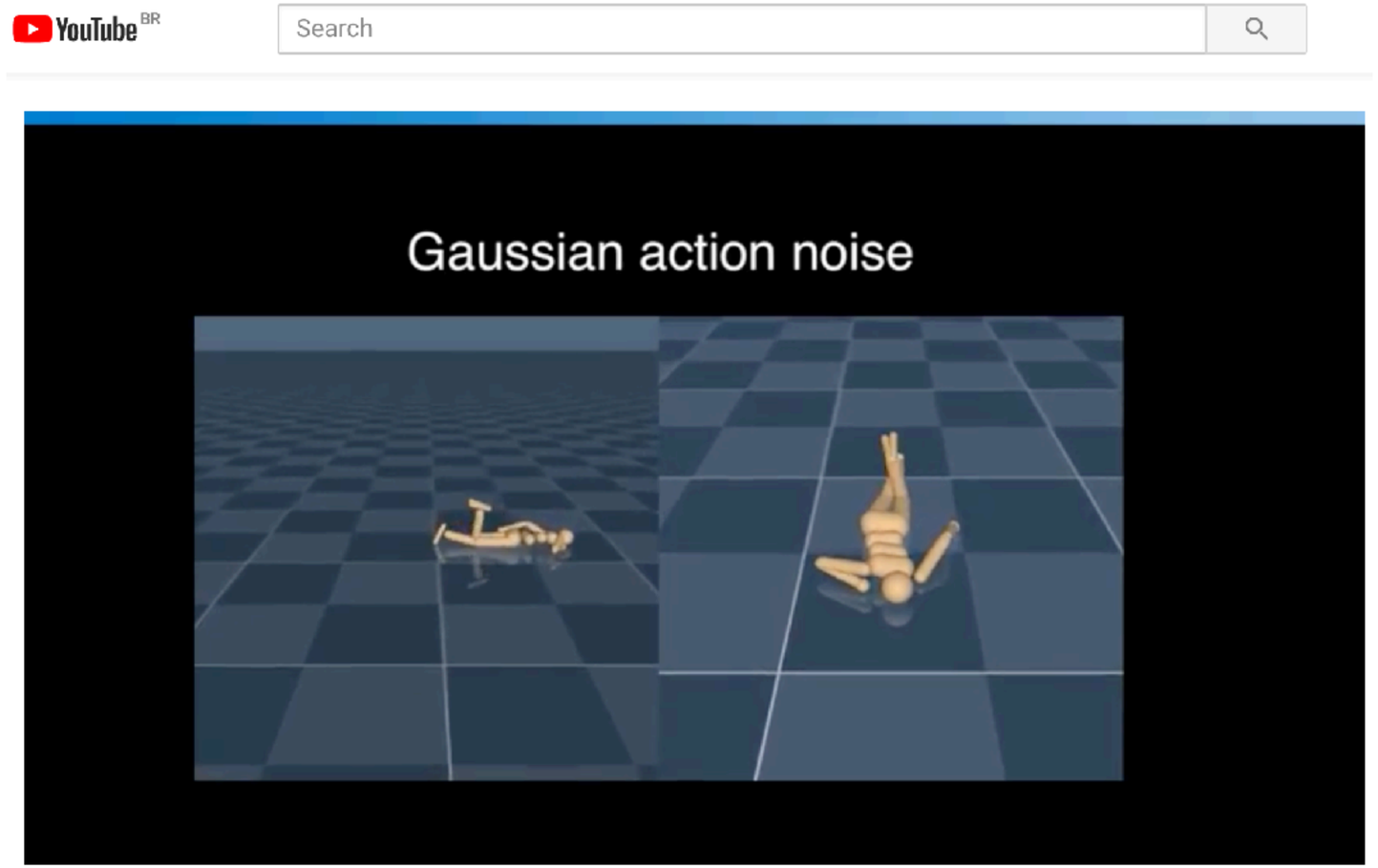


Reinforcement Learning



https://www.youtube.com/watch?v=V1eYniJ0Rnk

Reinforcement Learning



RAAIS 2017 - Raia Hadsell, Senior Research Scientist at DeepMind

https://www.youtube.com/watch?v=0e_uGa7ic74&feature=youtu.be&t=2065

Profissões impactadas: todas!