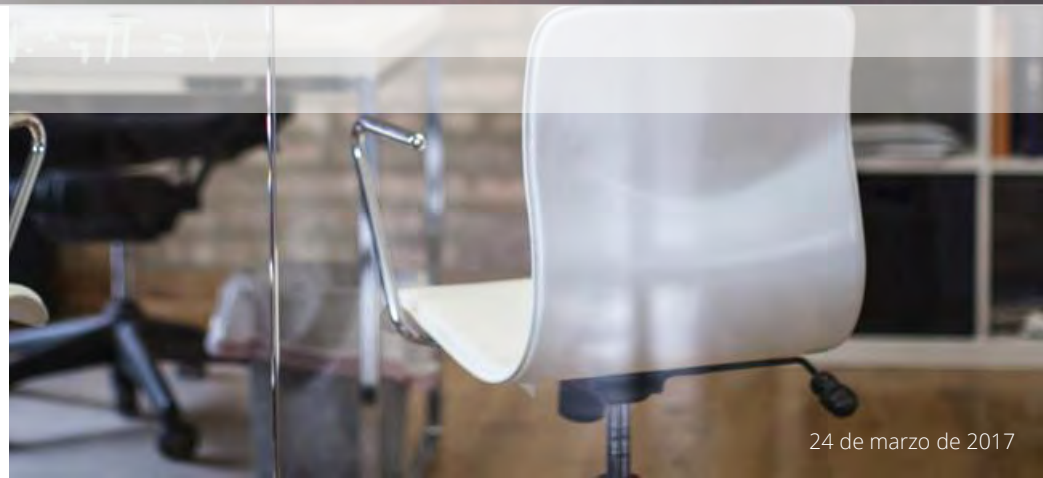
instituto
de ingeniería
del conocimiento

Tendencias actuales en el mundo Big Data

Álvaro Barbero – Chief Data Scientist



www.iic.uam.es

24 de marzo de 2017

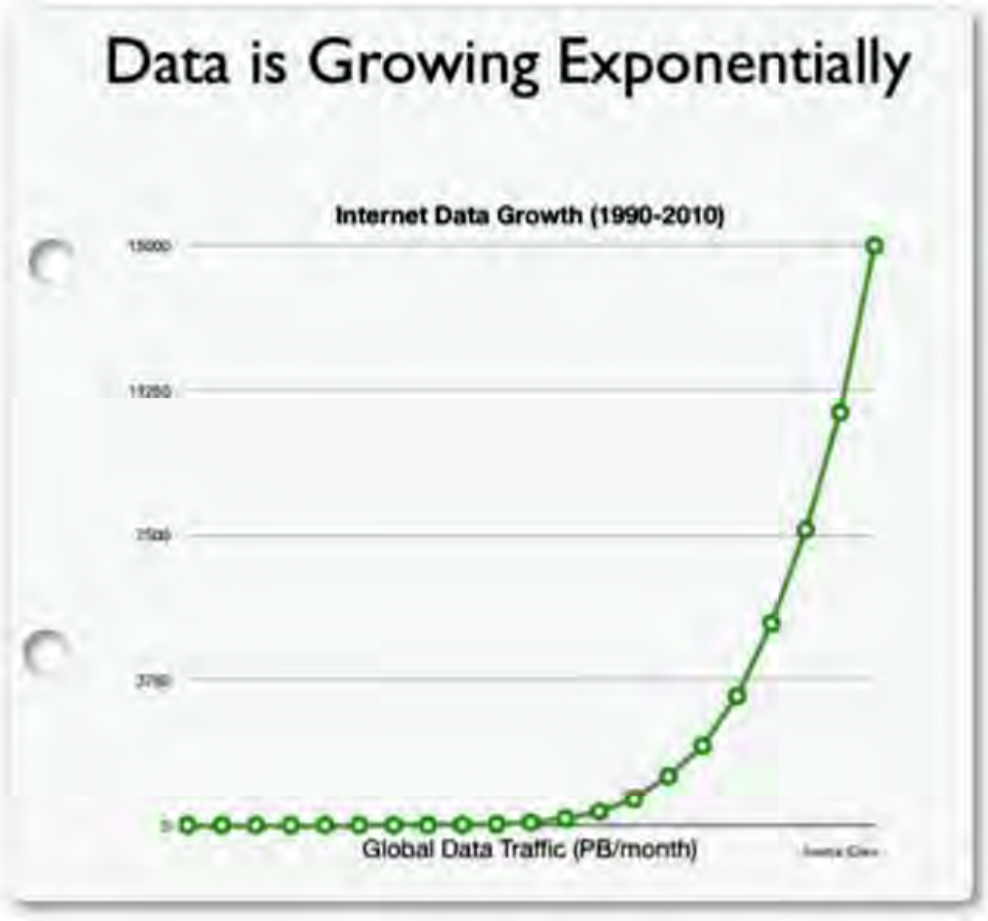
El auge de los datos


Google
Spotify
YouTube
DOMAINS
facebook
2
MILLION
Searches
14
New Songs
Added
72
HOURS
of Video
uploaded
70
New
Registered
41
THOUSAND
posts every
second
15
THOUSAND
Tracks
downloaded
from iTunes
1.8
MILLION
likes
skype
WordPress
1.4
MILLION
350GB
of data
347
New Blog
Posts
Minutes Connecting
with Each other
flickr
20
MILLION
Photo
Views
571
New Created
WEBSITES
278
THOUSAND
Tweets
Transactions
17
THOUSAND
twitter
Walmart
New Tumblr
Photos
20
THOUSAND
104
THOUSAND
Photos
Shared
tumblr.
snapchat
Amazon Sales
$83,000
Professional
searches
11
THOUSAND
Active
Users
11
THOUSAND
emails sent
204
MILLION
photos every
minute
3,600
EMAIL
amazon.com
Linked in
Pinterest
Instagram

En 2012, se crearon cada día 2,5 quintillones de bytes de datos (un 1 seguido de 18 ceros)

El 90% de los datos del mundo se ha creado en los últimos dos años.

Como sociedad, producimos y capturamos cada día más datos de los vistos por todo el mundo desde los comienzos de la Tierra.



Lo que me gustaría

La triste realidad

The 4 V´s of Big Data

Volume	Velocity	Variety	Veracity
Size & Scale	Streaming	Many Forms	Accuracy

‘We are drowning in information, but we are starved for knowledge’.

John Naisbitt, 1982

El perfil del Data Scientist

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21th century requires a mixture of multidisciplinary skills ranging fro
intersection of mathematics, statistics, computer science, communication and business. Finding a d
scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a l
cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- Machine learning
- Statistical modeling
- Experiment design
- Bayesian inference
- Supervised learning: decision trees, random forests, logistic regression
- Unsupervised learning: clustering, dimensionality reduction
- Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- Computer science fundamentals
- Scripting language e.g. Python
- Statistical computing packages, e.g. R
- Databases: SQL and NoSQL
- Relational algebra
- Parallel databases and parallel query processing
- MapReduce concepts
- Hadoop and Hive/Pig
- Custom reducers
- Experience with IaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- Passionate about the business
- Curious about data
- Influence without authority
- Hacker mindset
- Problem solver
- Strategic, proactive, creative, innovative and collaborative

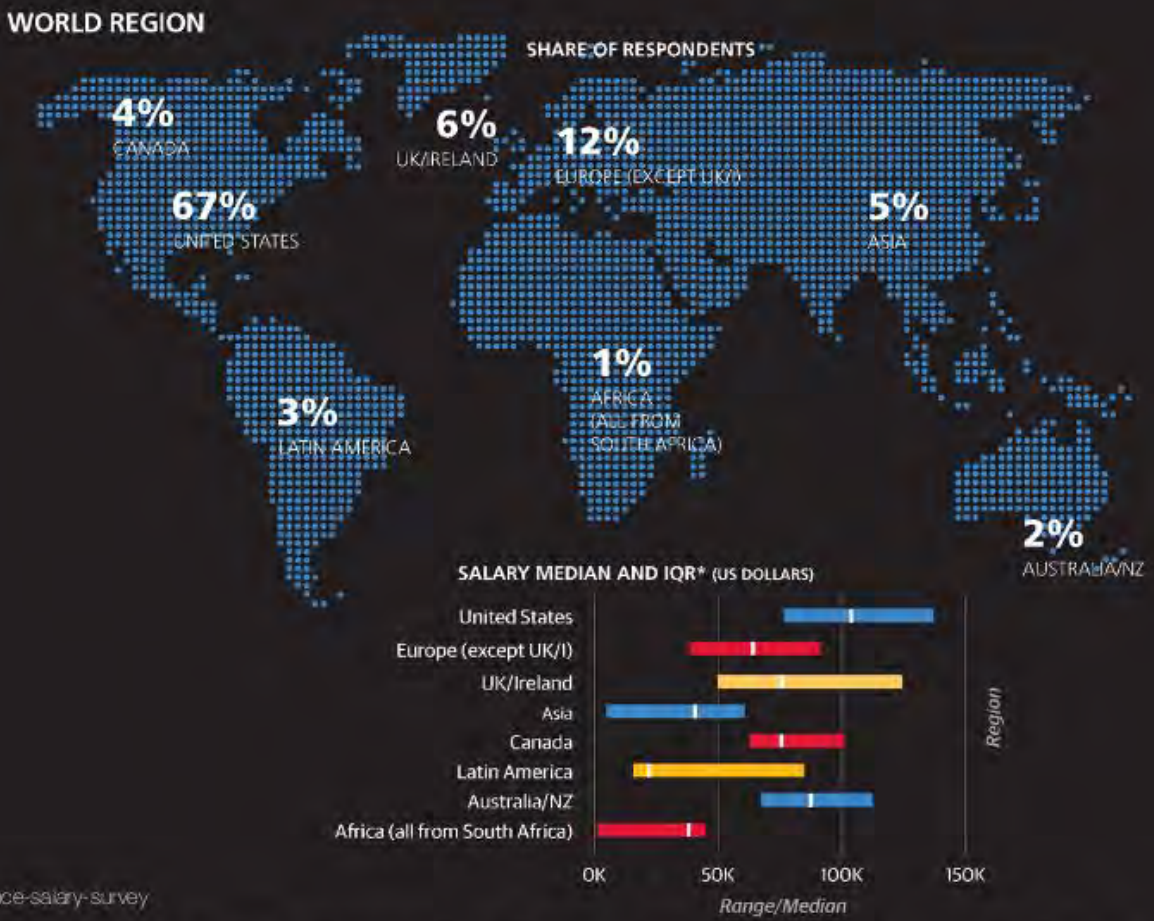
COMMUNICATION & VISUALIZATION

- Able to engage with senior management
- Story telling skills
- Translate data-driven insights into decisions and actions
- Visual art design
- R packages like ggplot or lattice
- Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

MarketingDistillery.com is a group of practitioners in the area of e-commerce marketing. Our fields of expertise include: marketing strategy and optimization, customer tracking and on-site analytics, predictive analytics and econometrics, data warehousing and big data systems, marketing channel insights in Paid Search, SEO, Social, CRM and brand.

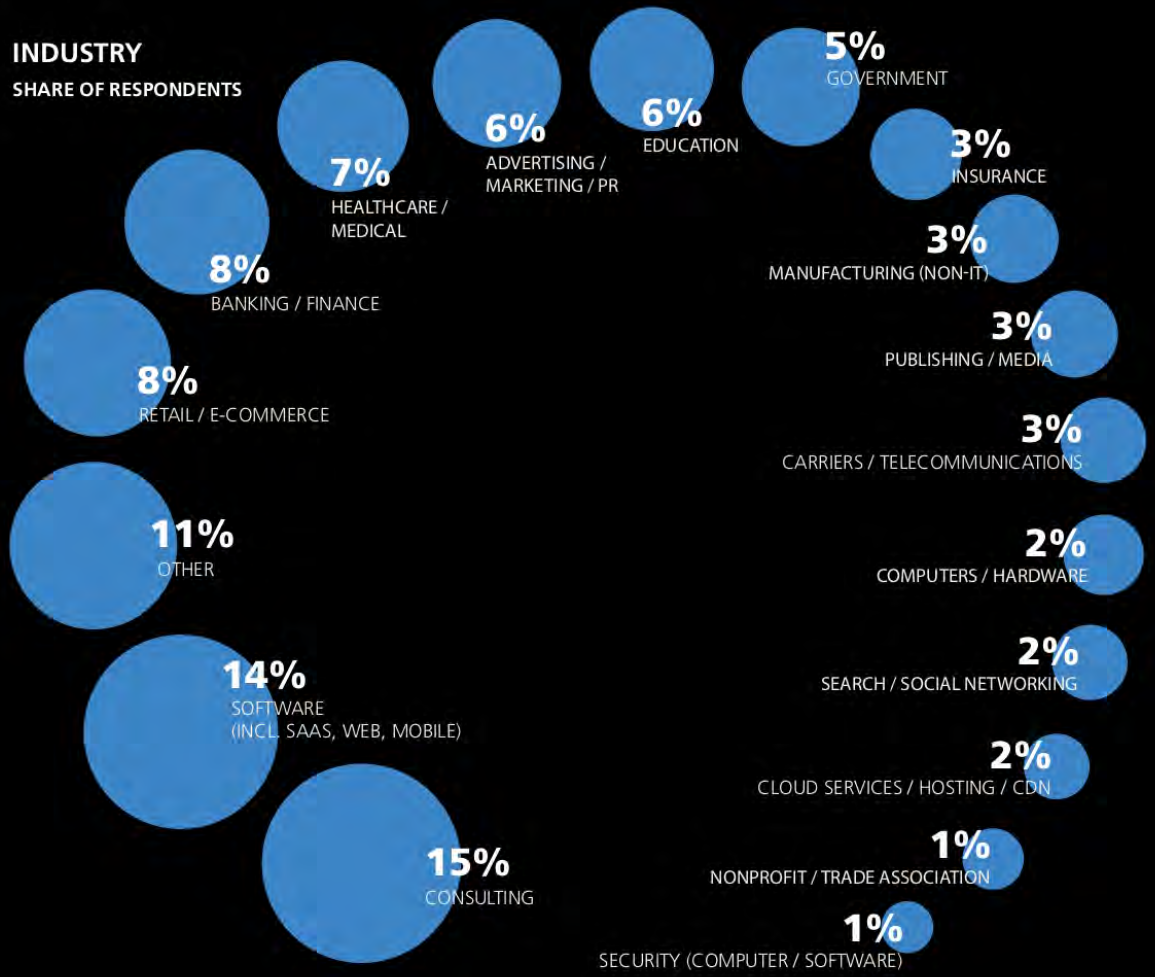
Marketing DISTILLERY
© Krzysztof Zawadzki

Data Science Unicorn



O-Reilly 2015 Data Science Salary Survey:
https://www.oreilly.com/ideas/2015-data-science-salary-survey

*The interquartile range (IQR) is the middle 50% of respondents' salaries. One quarter of respondents have a salary below this range, one quarter have a salary above this range.



Source: O'Reilly Data Science Salary Survey 2016

Desarrollando proyectos de datos

Duda siempre de ti mismo, hasta que los datos no dejen lugar a dudas

Louis Pasteur, 1822-1895


JANUARY
MARCH

Datos

Analítica

Kaizen

改善

Mejora continua


VALOR
sas
learn
SPSS
R
GraphX
ANALIZAR
kafka
NVIDIA CUDA
Spark
STORM
PROCESAR
hadoop
ELASTIC SEARCH
cassandra
mongoDB
APACHE HBASE
HYPERTABLE
ALMACENAR
amazon web services™
IBM Bluemix™
HARDWARE

Niveles de analítica

Analítica prescriptiva

Cuál es la mejor estrategia

Analítica predictiva

Qué va a pasar

Analítica descriptiva

Qué pasa ahora

Analítica descriptiva

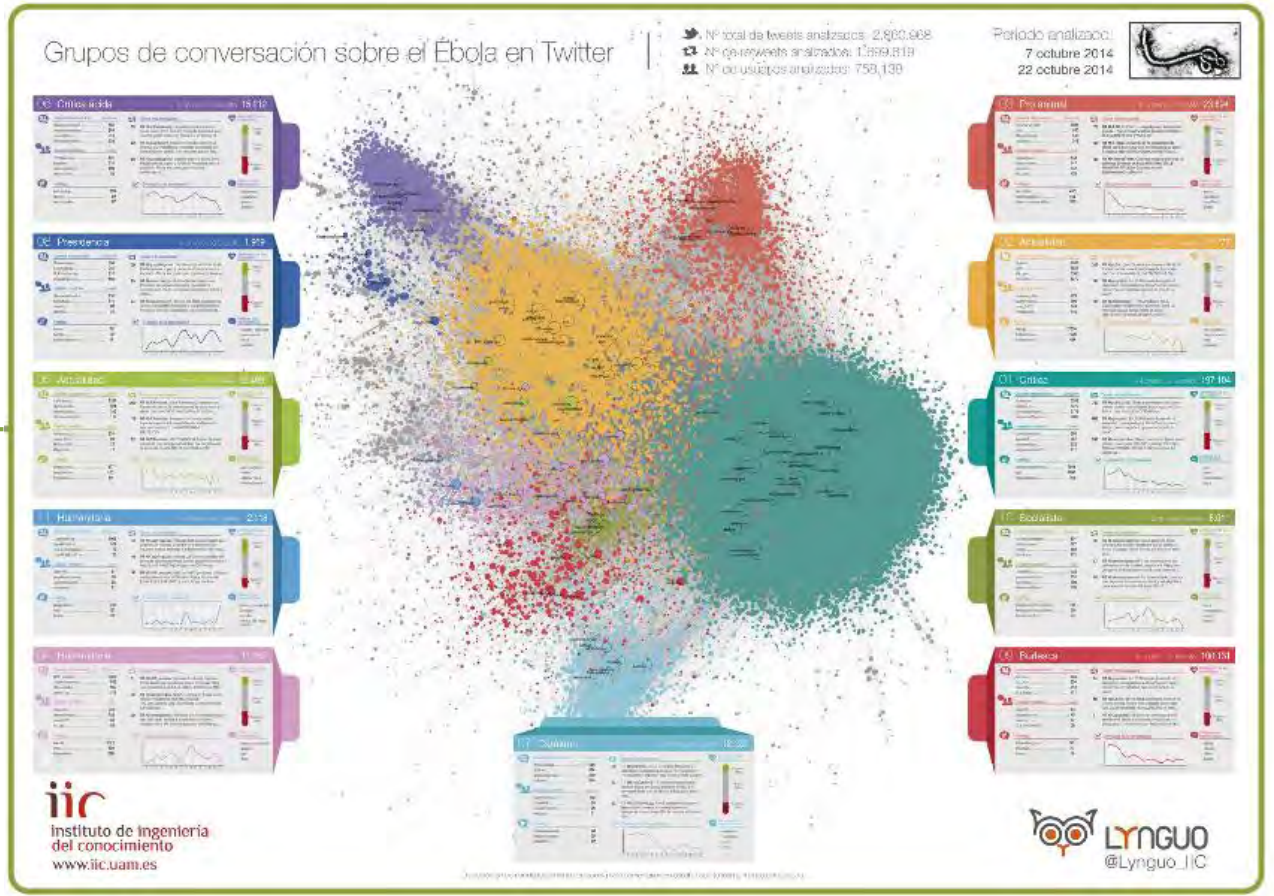
Qué pasa ahora

Describir los datos para sacar información útil

- KPIs
- Dashboards
- Visualizaciones

Analítica descriptiva

Qué pasa ahora



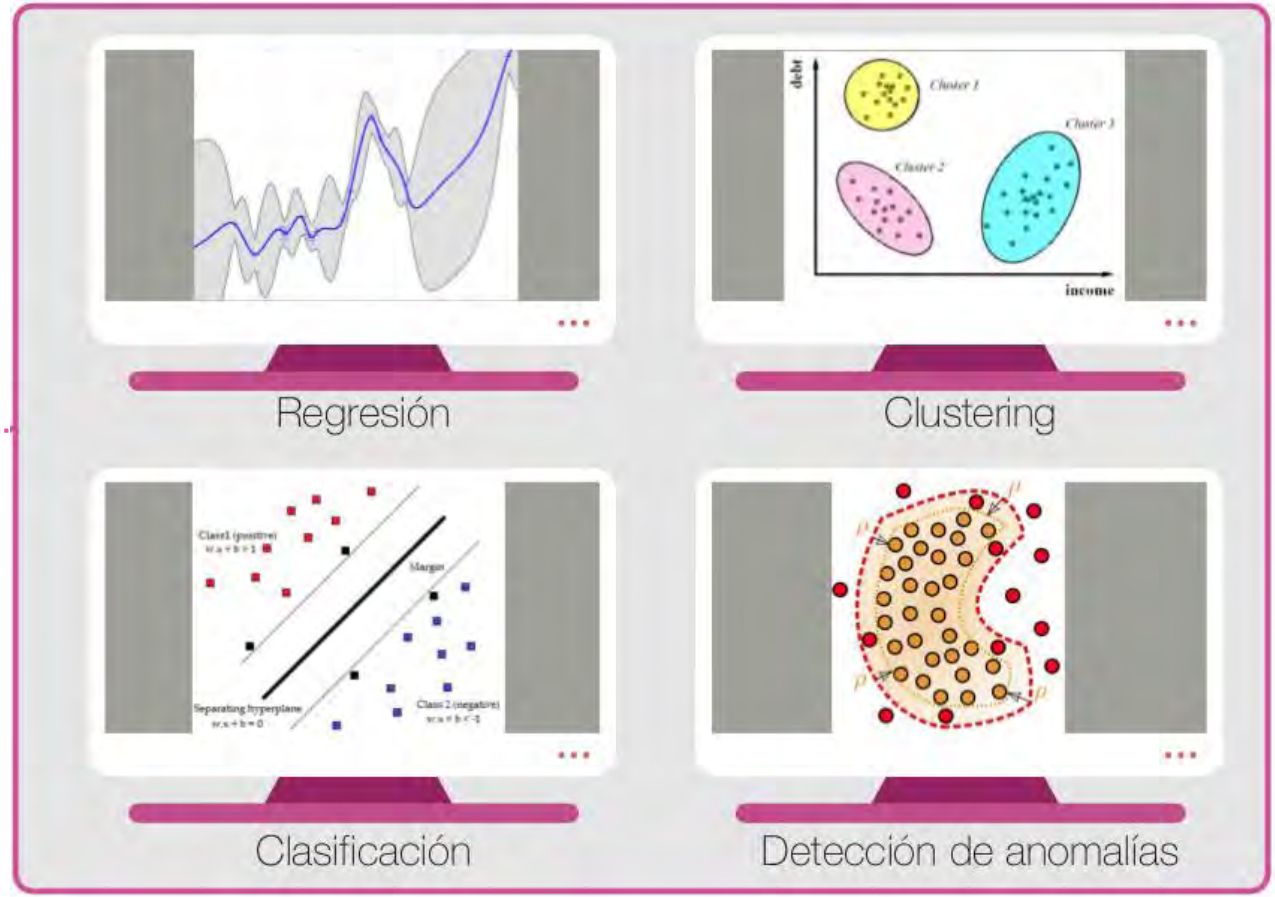
Analítica descriptiva

Qué pasa ahora

Analítica predictiva

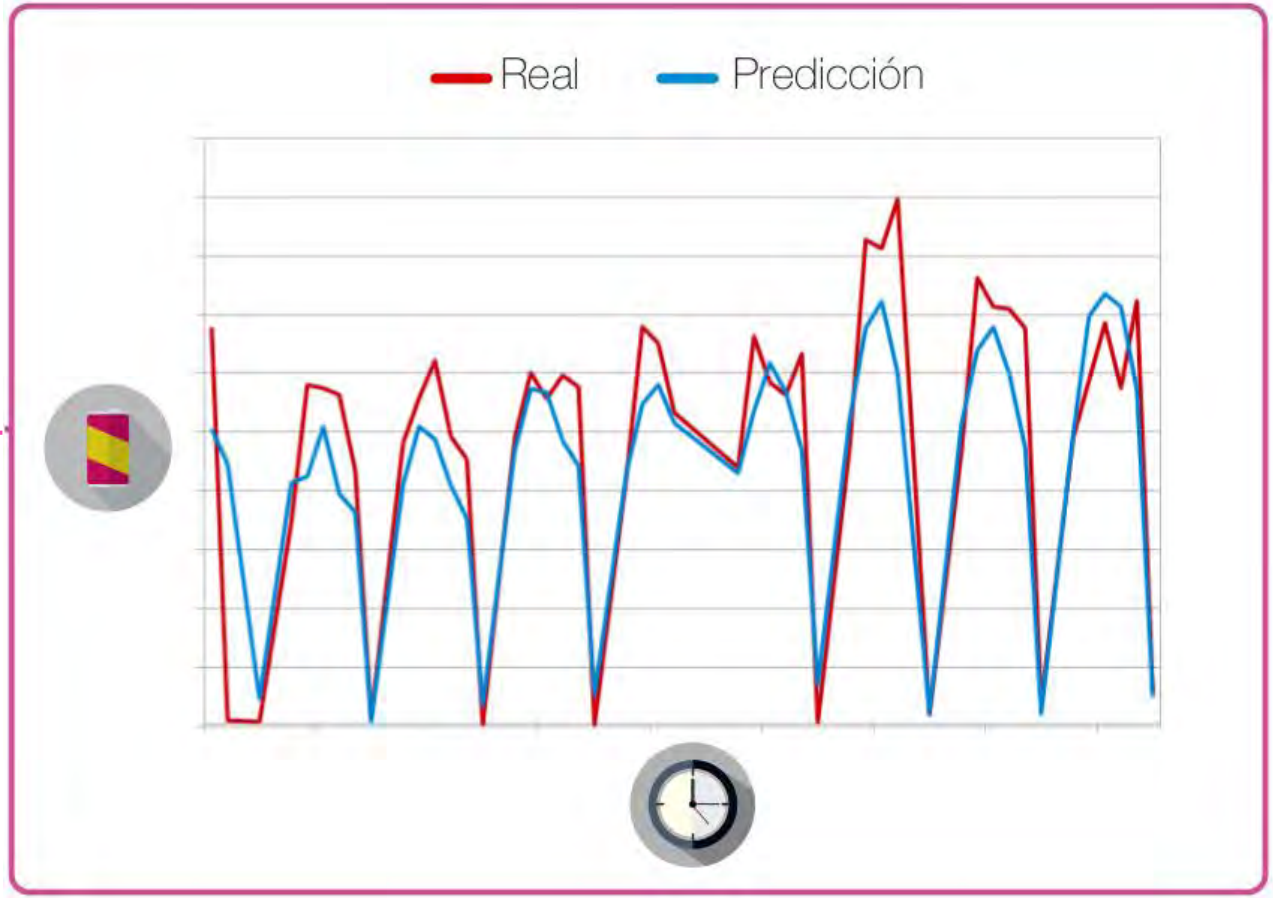
Qué va a pasar

Estimar datos que no tenemos



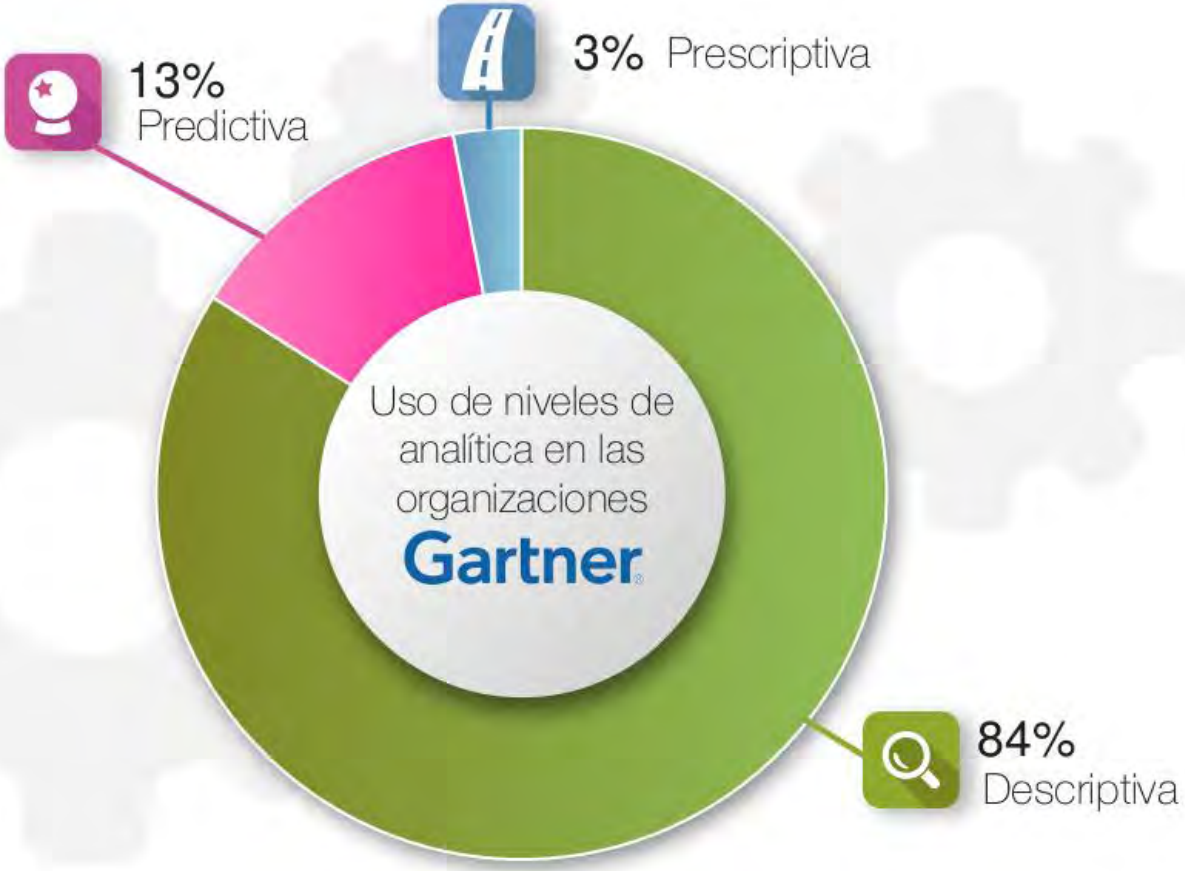
Analítica predictiva

Qué va a pasar



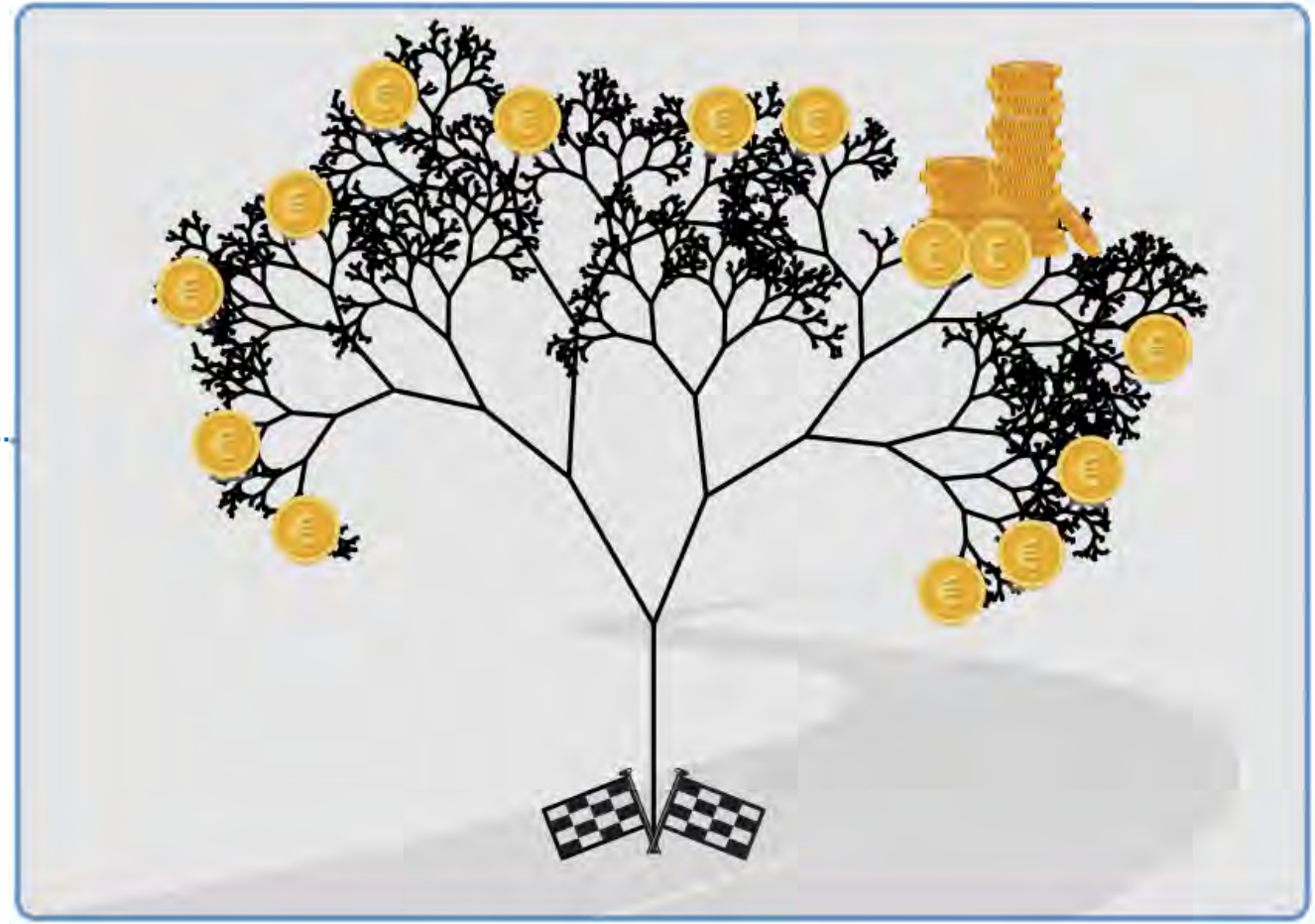
Analítica predictiva

Qué va a pasar


3% Prescriptiva
13% Predictiva
Uso de niveles de analítica en las organizaciones
Gartner
84% Descriptiva

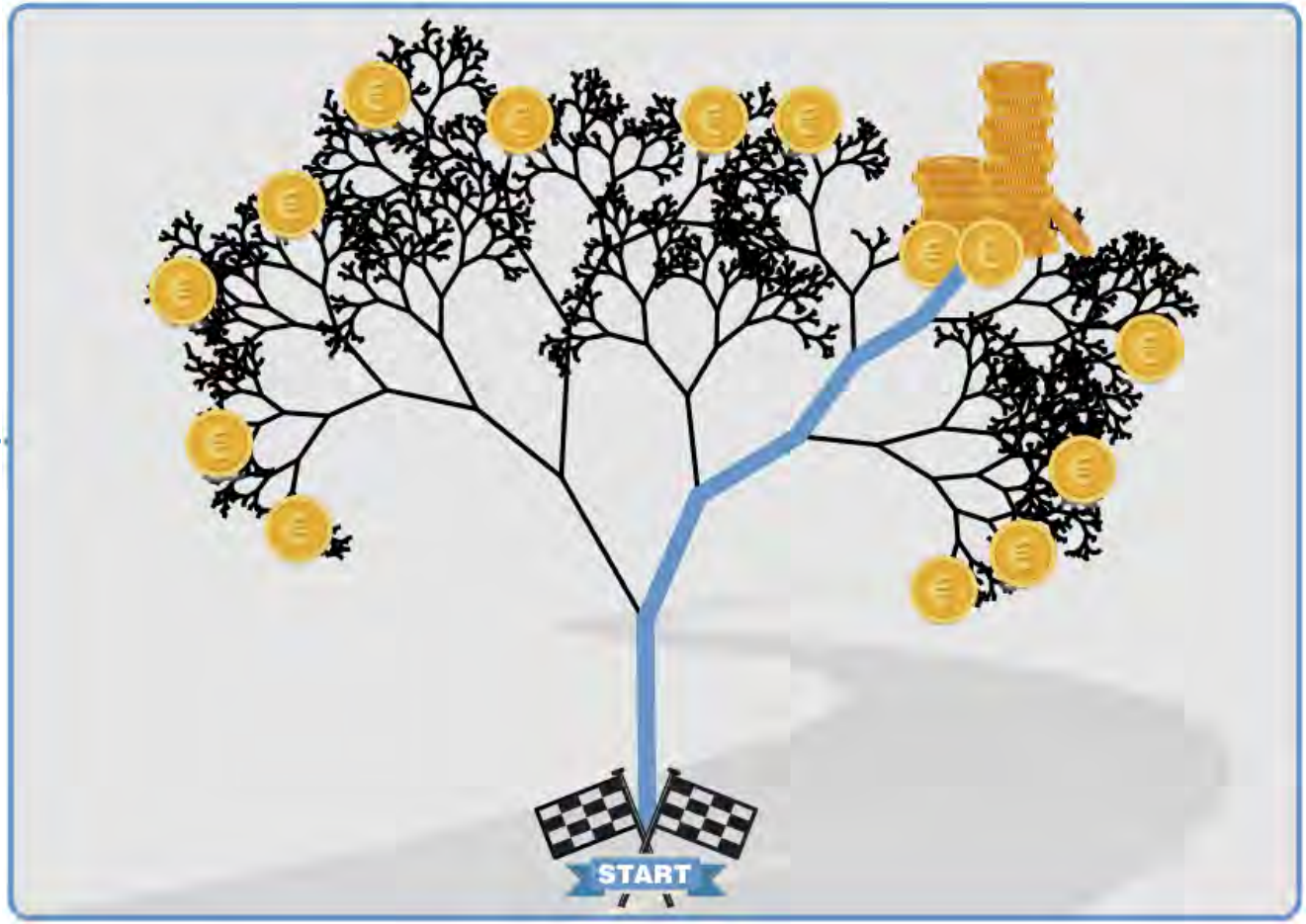
Analítica prescriptica:

Cuál es la mejor estrategia



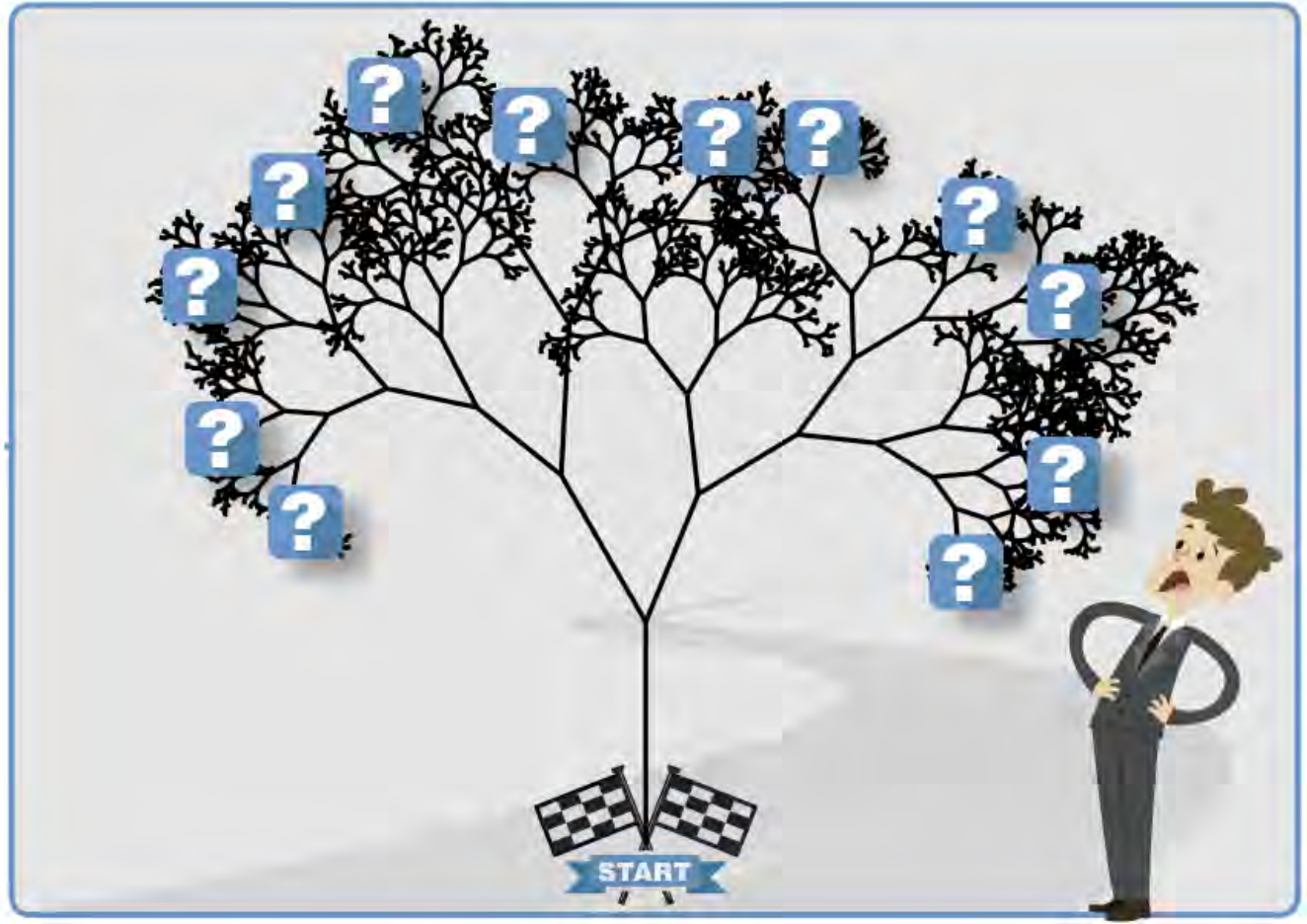
Analítica prescriptica:

Cuál es la mejor estrategia



Analítica prescriptica:

Cuál es la mejor estrategia



Analítica prescriptica:

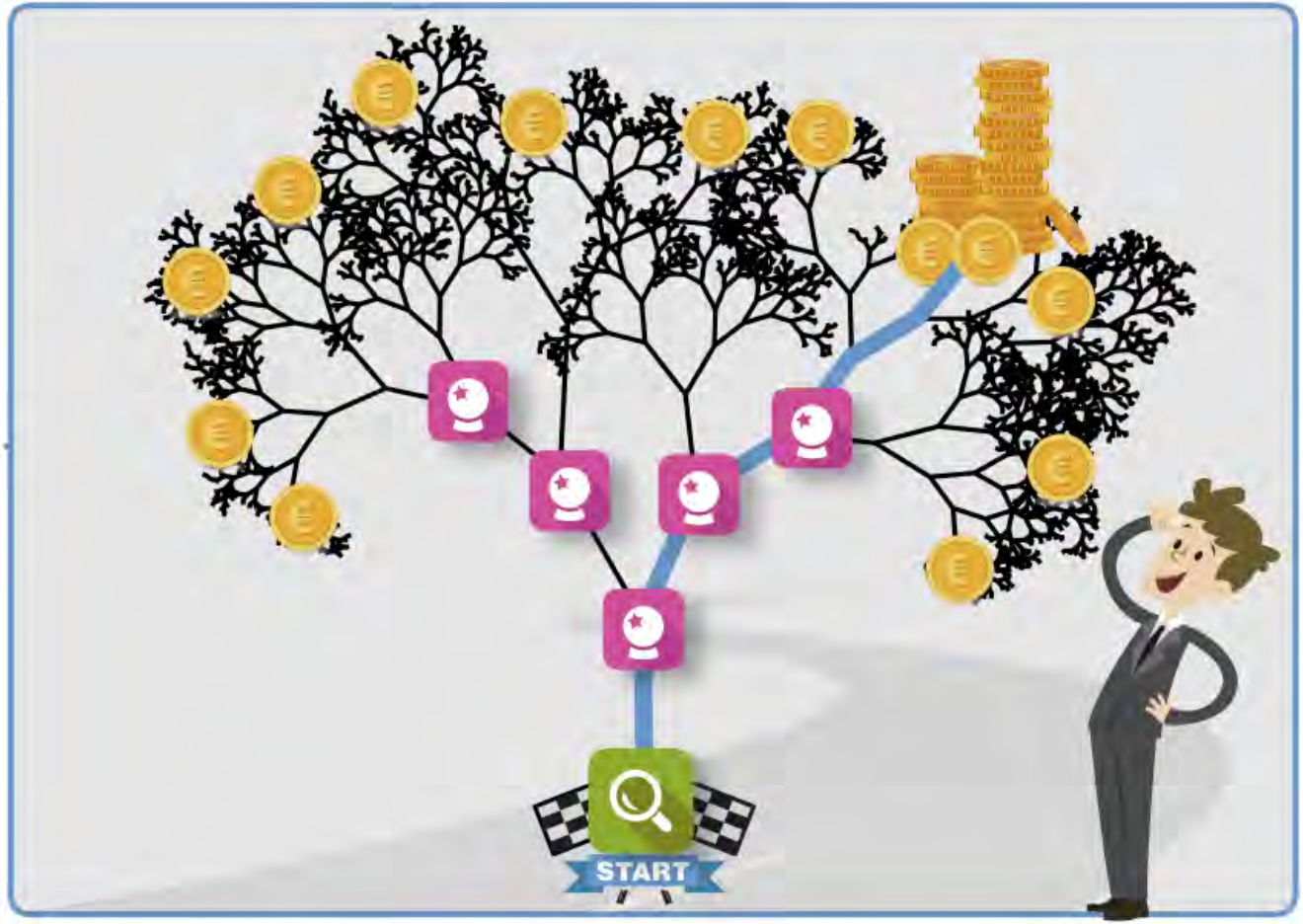
Cuál es la mejor estrategia

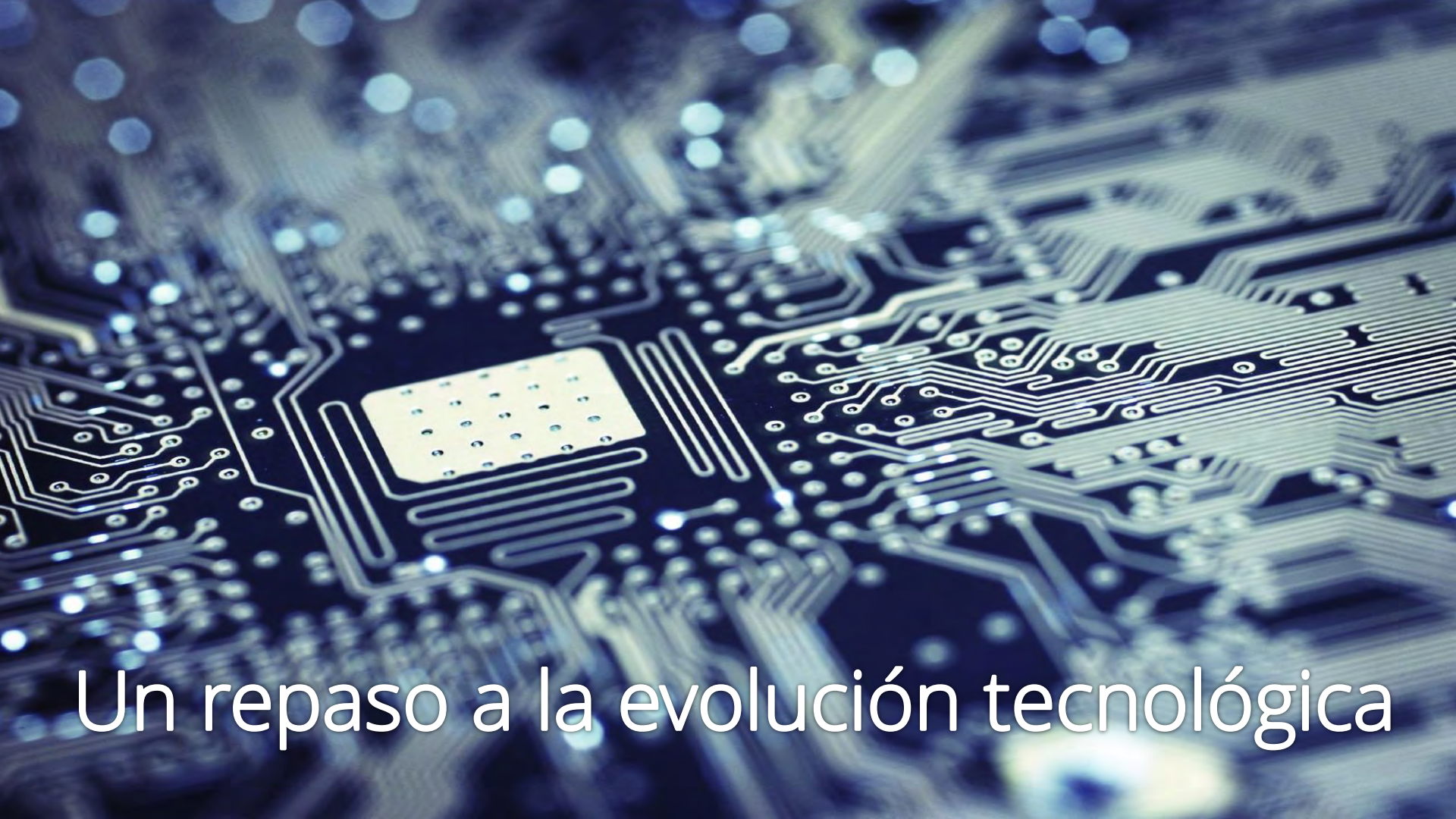
Analítica prescriptica:

Cuál es la mejor estrategia

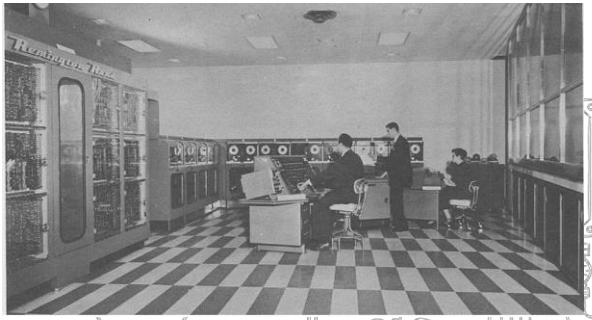
Analítica prescriptica:

Cuál es la mejor estrategia





Un repaso a la evolución tecnológica



(1951)

UNIVAC I
5000 tubos de vacío

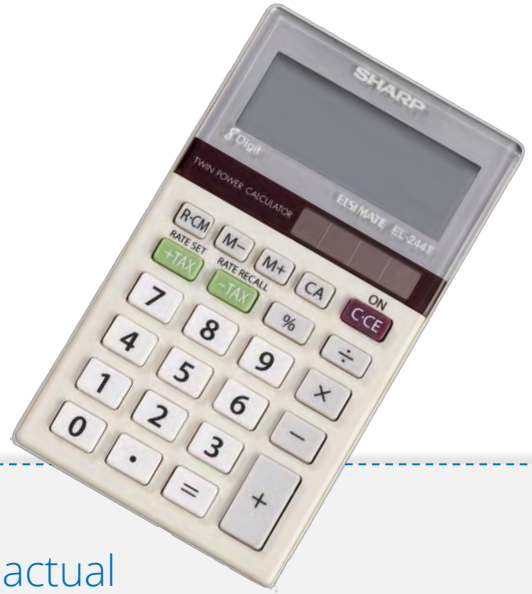
(1985)

Procesador 30386
275.000 transistors
(x55)

(2008)

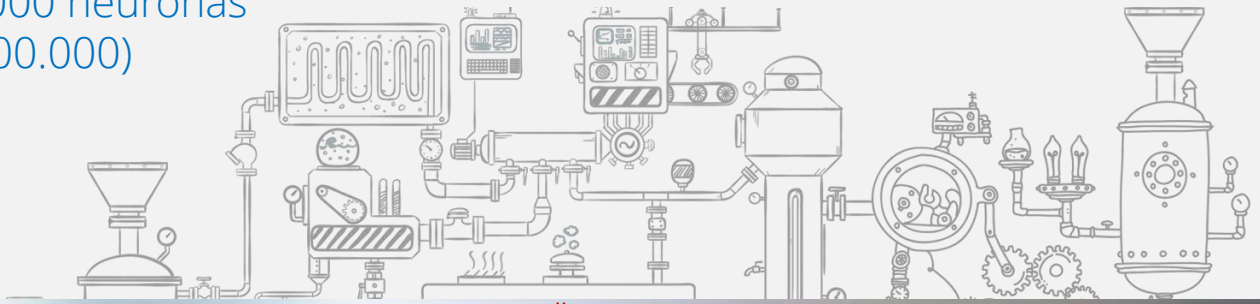
Core i7
731.000.000 transistores
(x146.200)

(195.000 A.C)

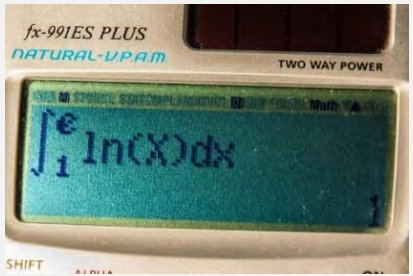


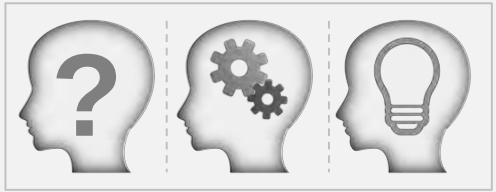
Cerebro Homo Sapiens Sapiens
50.000.000.000 neuronas
(x10.000.000)

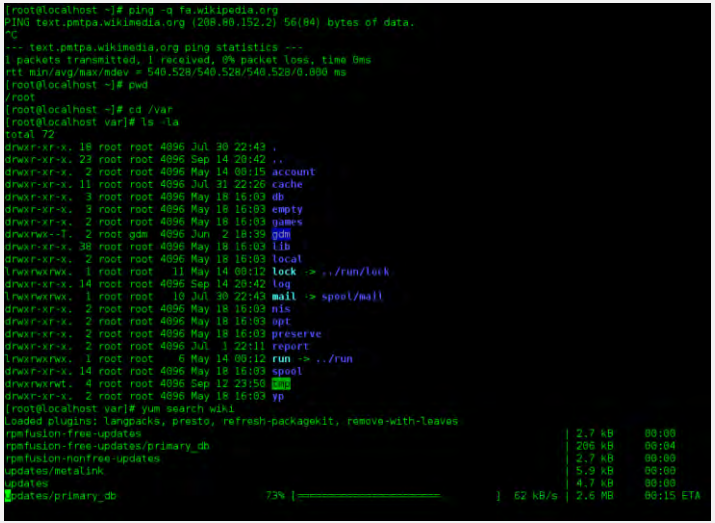
Computadora actual

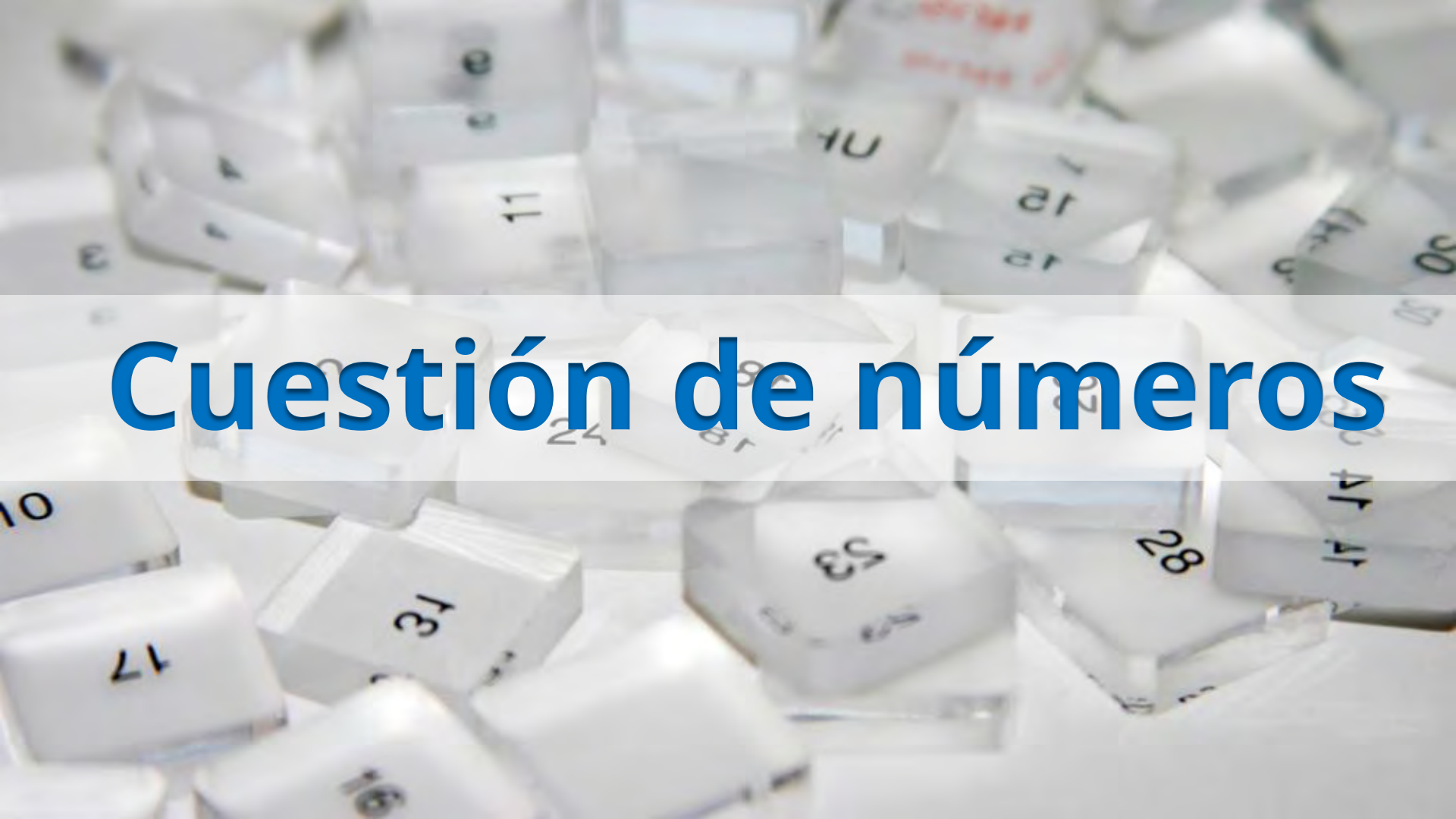


Cognición vs cálculo

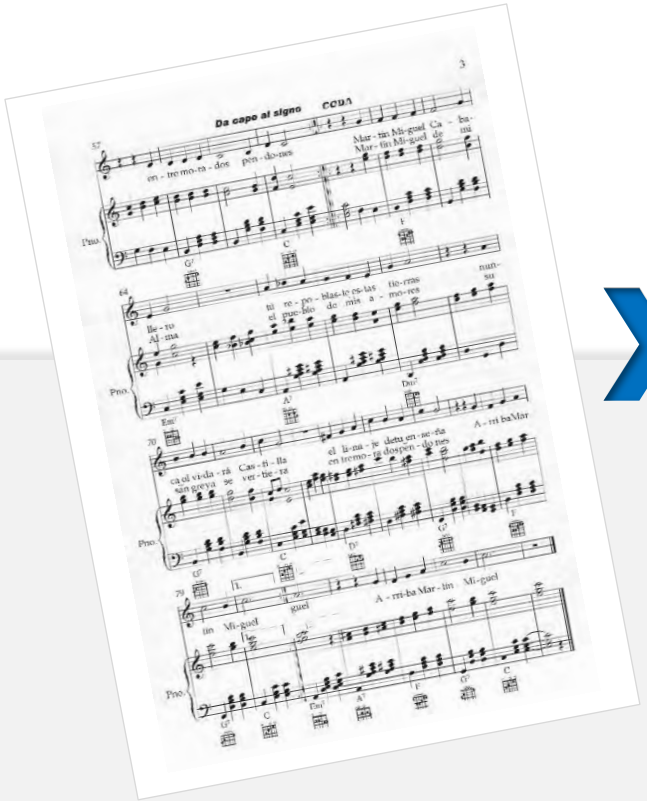


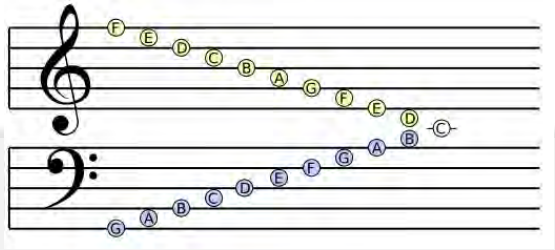






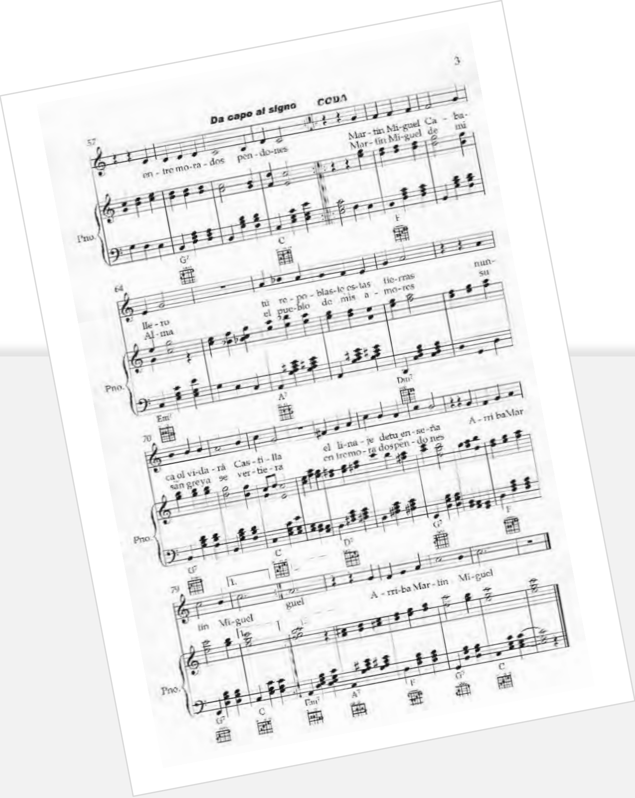
Cuestión de números



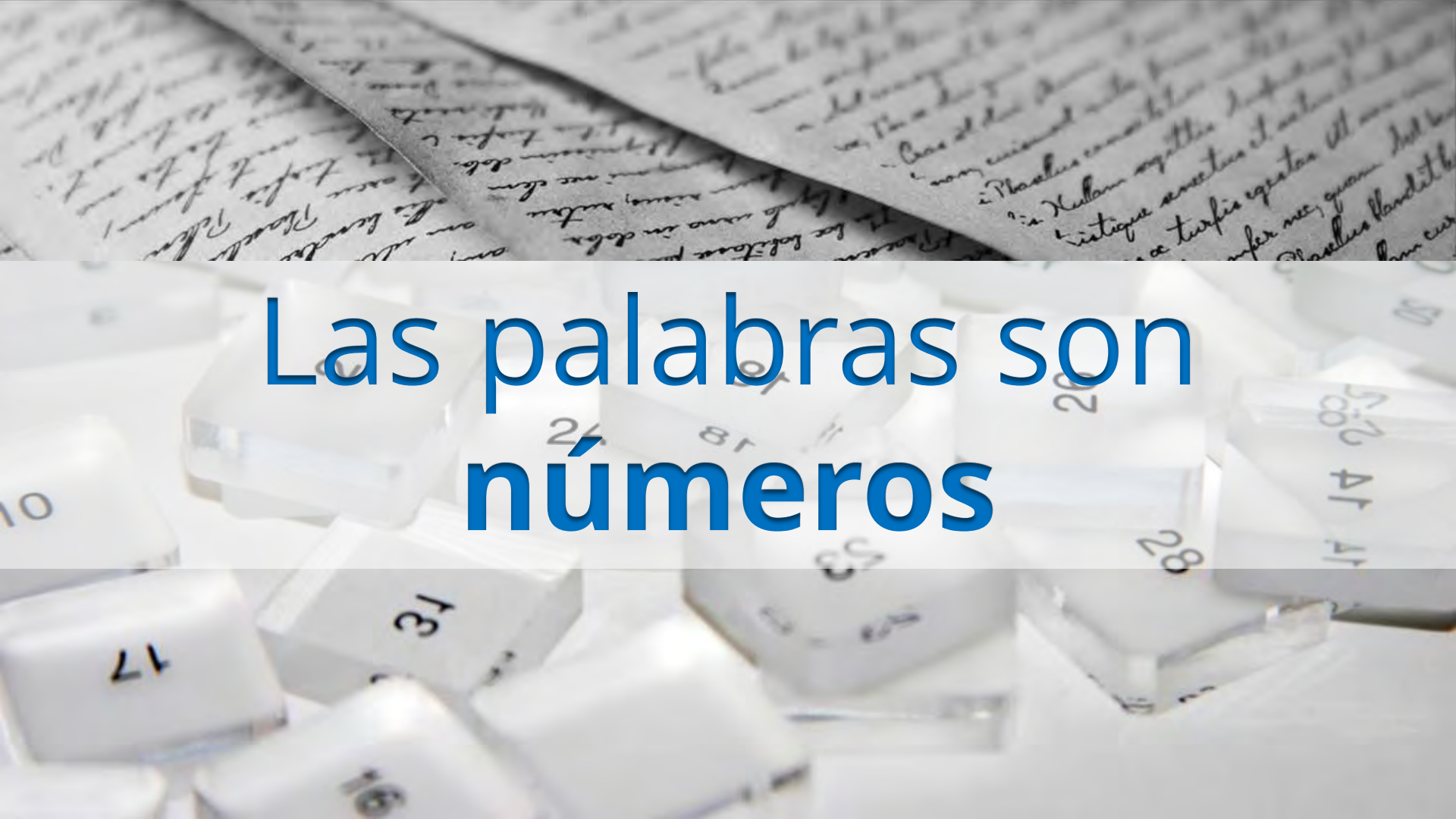


C
A
E
C
E
D
B

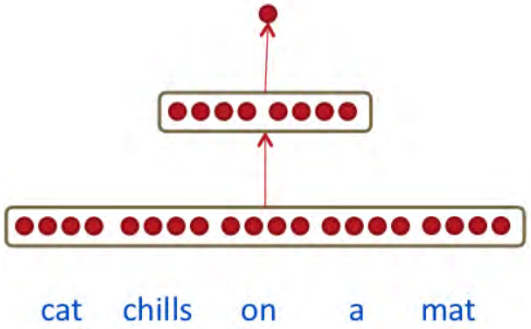
3
1
5
3
5
4
2



[1, 4, 5, 2, 3, 4, 2, …]

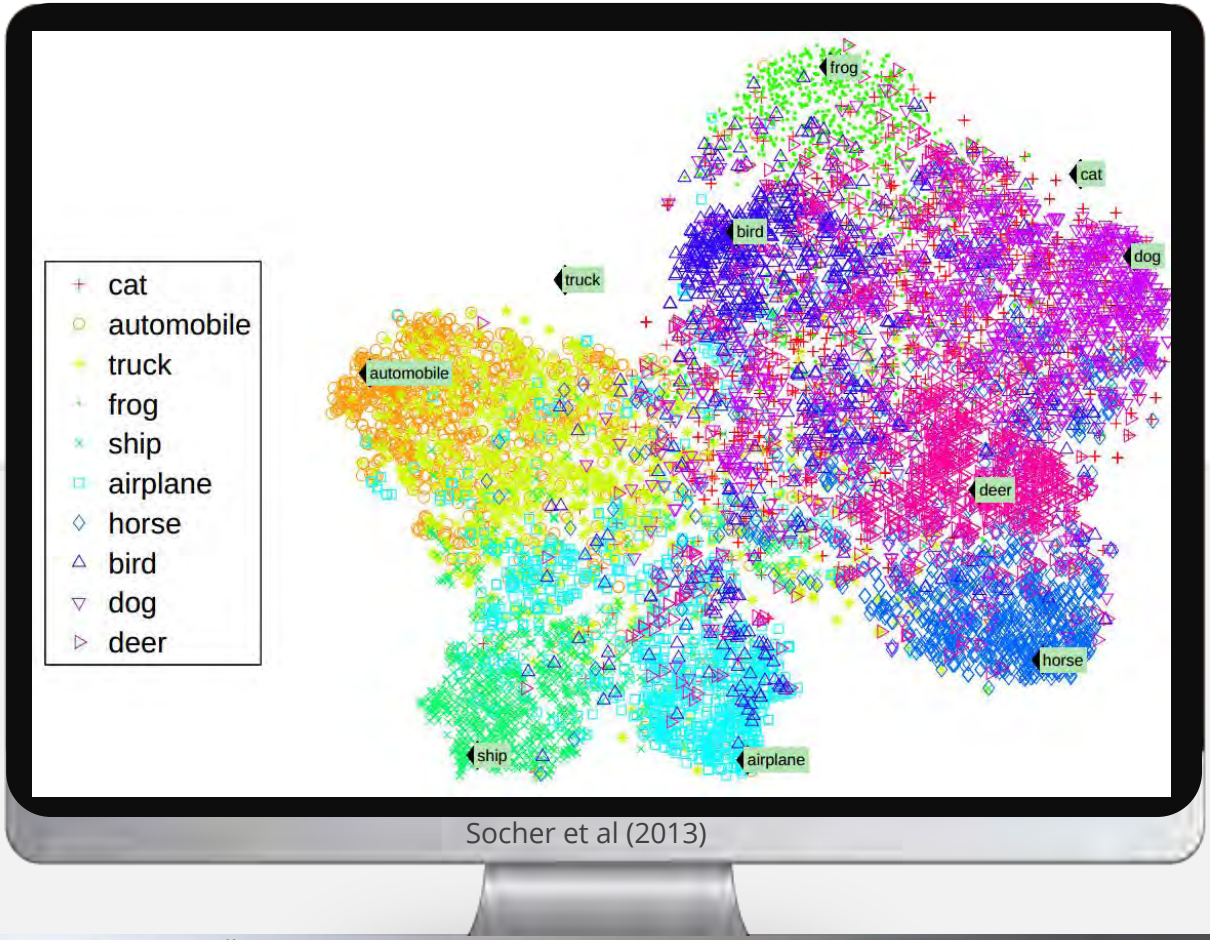

Las palabras son
números

Las palabras también pueden ser números



cat chills **on** a mat ❯ ⊕

cat chills **mushroom** a mat ❯ ⊖



Socher et al (2013)

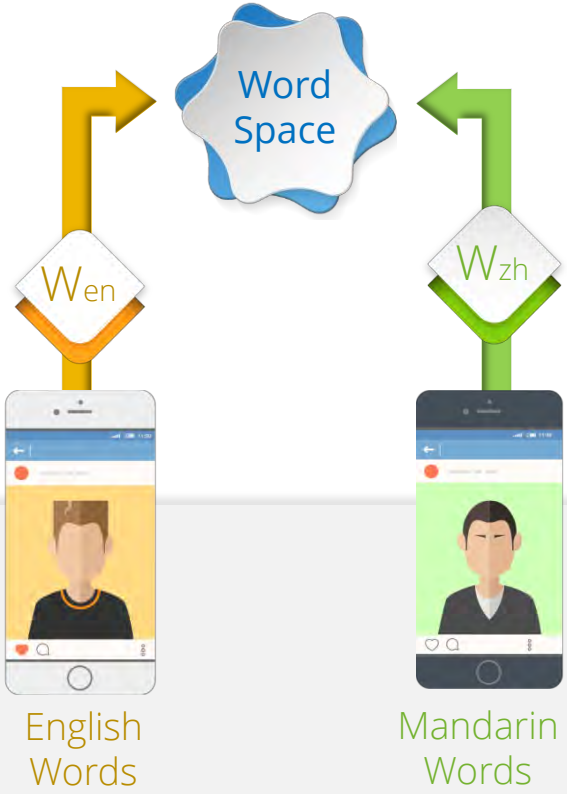
king − man + woman = queen

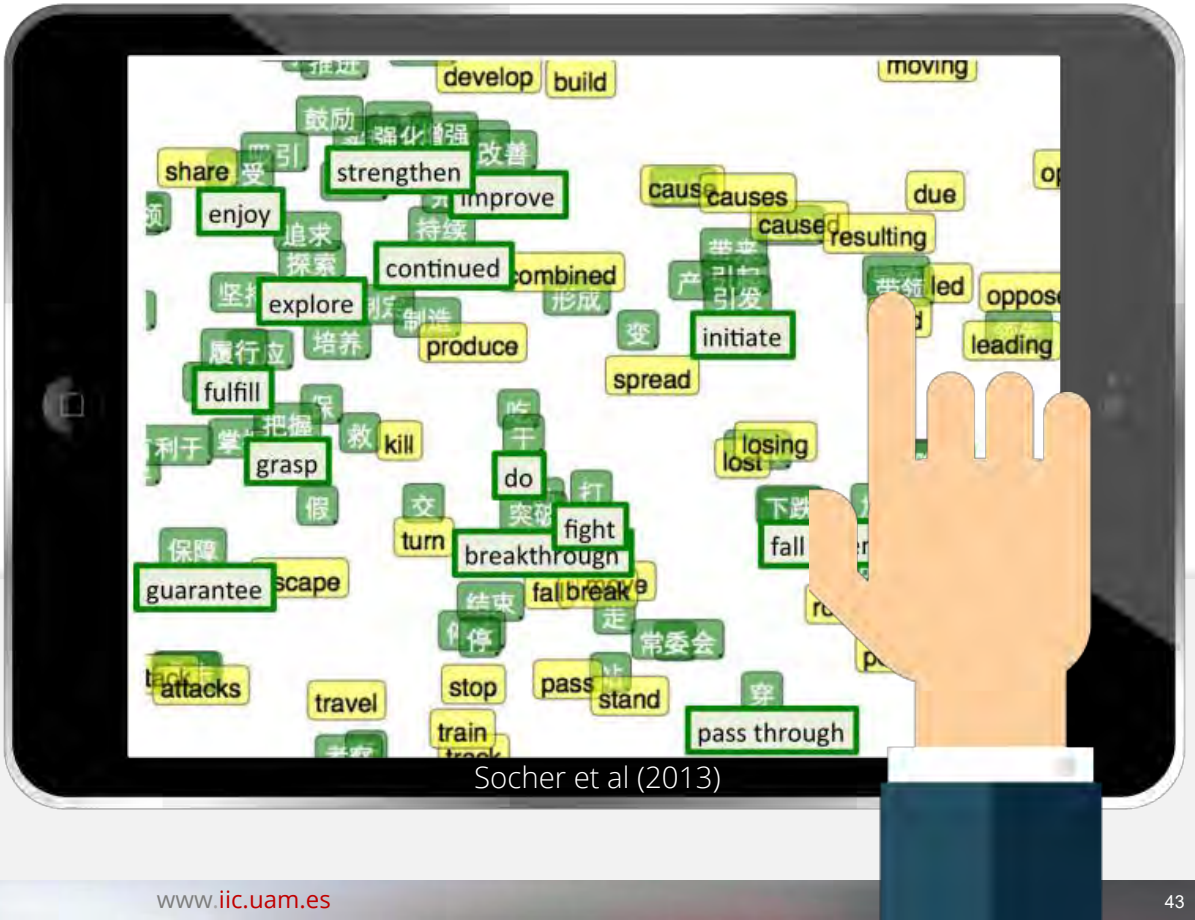
Obama − USA + Russia = Putin

human − animal = ethics

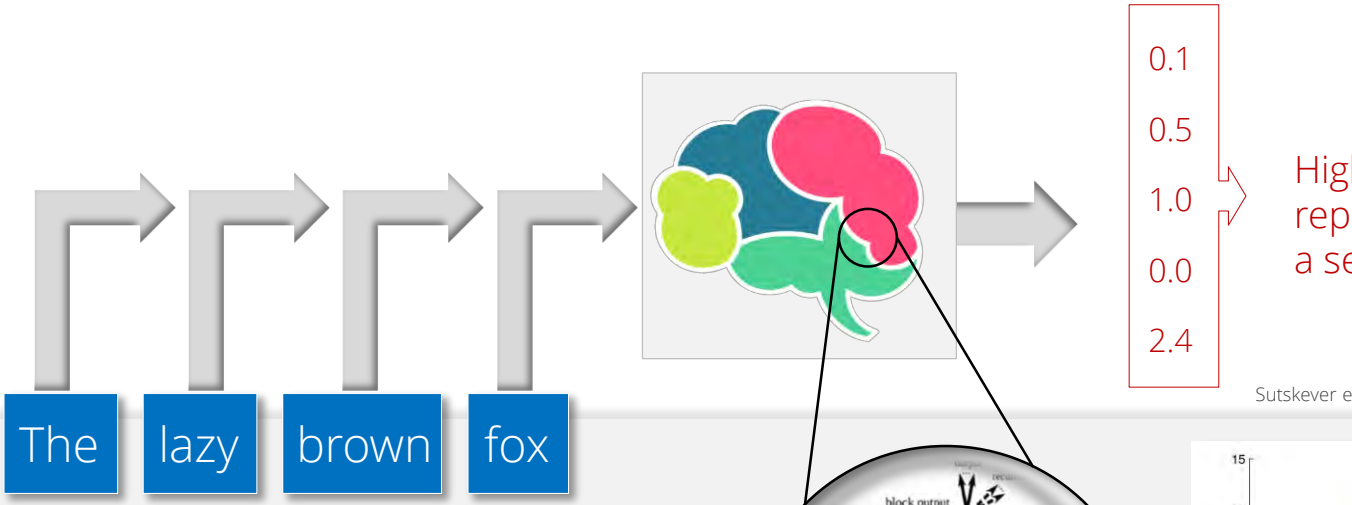
paella − Spain + Italy = risotto

Cristiano − Madrid + Barcelona = Messi



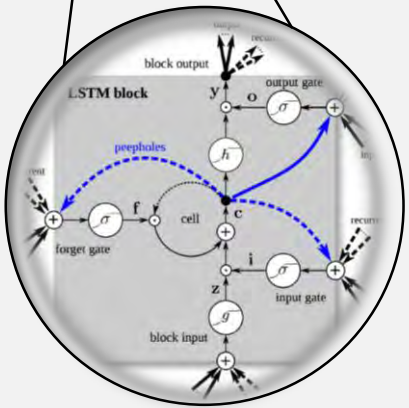


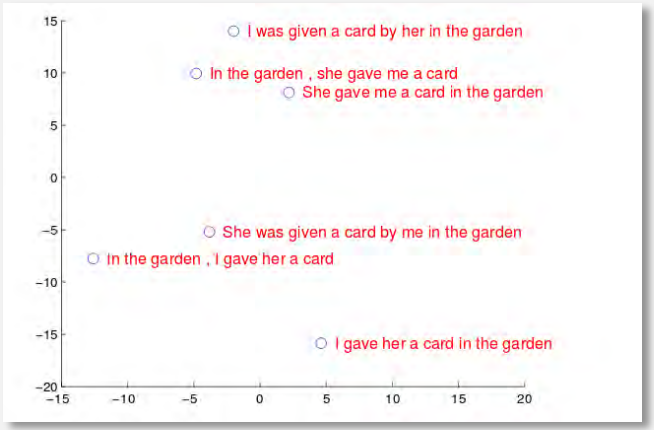
Socher et al (2013)



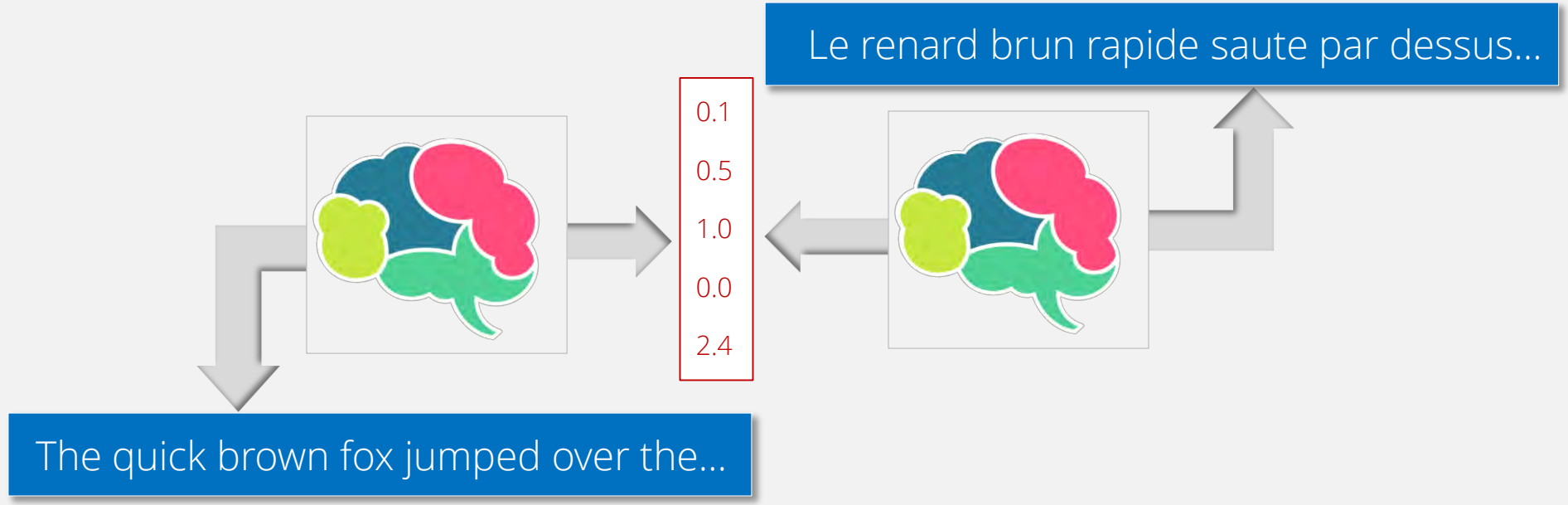
High dimensional representation of a sequence

Sutskever et al - Sequence to Sequence Learning with neural networks





Traducción automática



iic instituto de ingeniería del conocimiento

El Quijote escrito por NeuroCervanes

En un lugar de la Mancha está en su amada, y había de hacer regidor de tu amo que desde aquí está por el suelo.

-Por aquellos de las armas -respondió don Quijote-, que podrá poner y al presente y con la que debe de ser estimado en su parecer que tan desenvainado se le ha contado. Y así, no le hallará en el mundo, y así lo ha de decir que es la lanza a la princesa Sancho Panza, y se me estima y considere que mi tal, que yo haré de preguntar si a mí se ha de ser en el rey a la risa, que es usanza de la libertad que la compañía está en el mundo.

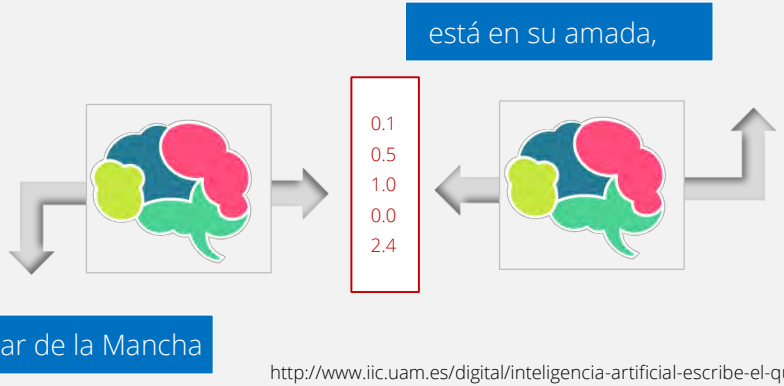
-Y ¿qué me ha de saber, Sancho -respondió don Quijote-, que todo vuestra merced se encamina su marido de mí bien o de lo que a mi señor está a la provecho de las muelas. Esta nueva amistad de su hermano me ha de hacer un caballero andante, y todo se encerrare en ella.

-¿Qué se lo ha de parecer -respondió don Quijote-, que yo pienso que el hermosuro hay señora a la carta que no lo piense la carta y en la salud, sino que lo haré yo, y es menester este tiempo de mi cuento, y no hay que procurar su padre de la puerta de los de los caballeros andantes; y así, acomodado a la mano del camino, dijo:

-Estos caballeros andantes me parecen que se alcanzan tantos de su historia, que es lo que ha de hacer en el estraño del gusto que vuestra merced debe de estar ciertos y menesterosos caballeros, que me está muy bien de oír estos ojos.

-No hay más que decir -respondió don Quijote-, y aun lo será el caballero andante, pues tan buena ventura hay tan bien acometer con quien se me ha de acometer que no puede decir que aquí dio con el trabajo que se pierde el rey que mal suele arremeter de aquella caballería, que alguna vez se halla en el primer deseo de la otra manera, y yo soy alguna cosa que pecador de la señora Dulcinea del Toboso, que es mía que no podía decir de la carne de la vida en el corazón de Cardenio, y la duquesa estaba en los carros de su estraño espacio, en la cual con mucho

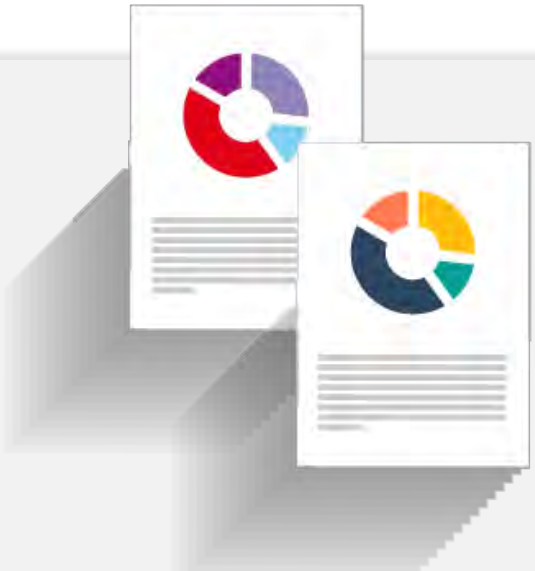
Generación de lenguaje
Neurocervantes



http://www.iic.uam.es/digital/inteligencia-artificial-escribe-el-quijote/

Clasificación automática de documentos

Recuperación de documentos similares



Perfilado de autores

GOBIERNO DE ESPAÑA | MINISTERIO DE INDUSTRIA, ENERGÍA Y TURISMO

Pliego del Ministerio de Industria, Secretaría de Estado de Telecomunicaciones y Sociedad de la Información.

Componente:

Investigadora: desarrollo de nuevos algoritmos de procesamiento del lenguaje natural con aplicación a la clasificación automática de solicitudes de patentes, en español e inglés.

Industrial: requisitos de funcionamiento en tiempo real, despliegue en entornos desconocidos a priori, sistema autocontenido, escalable, de fácil uso.

Resultados:

- ✓ **Nuevos algoritmos** desarrollados basados en tecnología de Deep Learning y cálculo en GPU.
- ✓ **Últimas tecnologías** de almacenamiento de datos: ElasticSearch, S3.
- ✓ **Despliegue modular, escalable y multiplataforma:** Docker, en local o en cloud de Amazon.
- ✓ **Solución completa** desde la fase de investigación hasta el despliegue.

  - Muy alta satisfacción por parte del cliente

amazon web services | ELASTIC SEARCH | riak | docker

Lynguo es una **herramienta de *monitorización*** en tiempo real que analiza el **contenido de las redes sociales** y proporciona una valoración sobre las **opiniones y emociones de los consumidores.**

- Escucha -

- ✓ Conversaciones
- ✓ Términos
- ✓ Marcas
- ✓ Cuentas
- ✓ Muros
- ✓ Vídeos
- ✓ Foros
- ✓ Blogs

- Analiza -

01. Análisis del contexto

02. Análisis léxico-semántico

03. Análisis gramatical

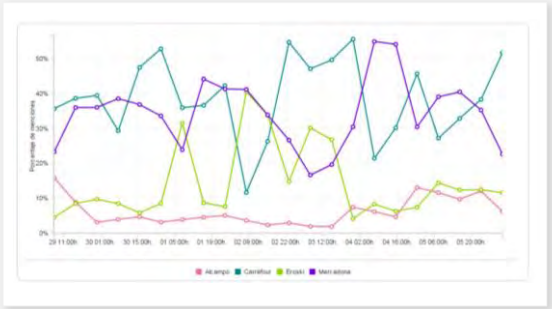
04. Algoritmos propios

- Responde-

- Responde -

KPIs

Compara **patrones y tendencias** a partir de información en tiempo real

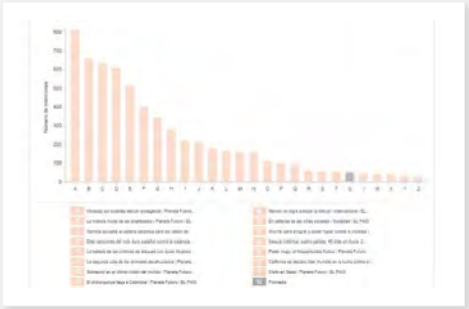


Opinión

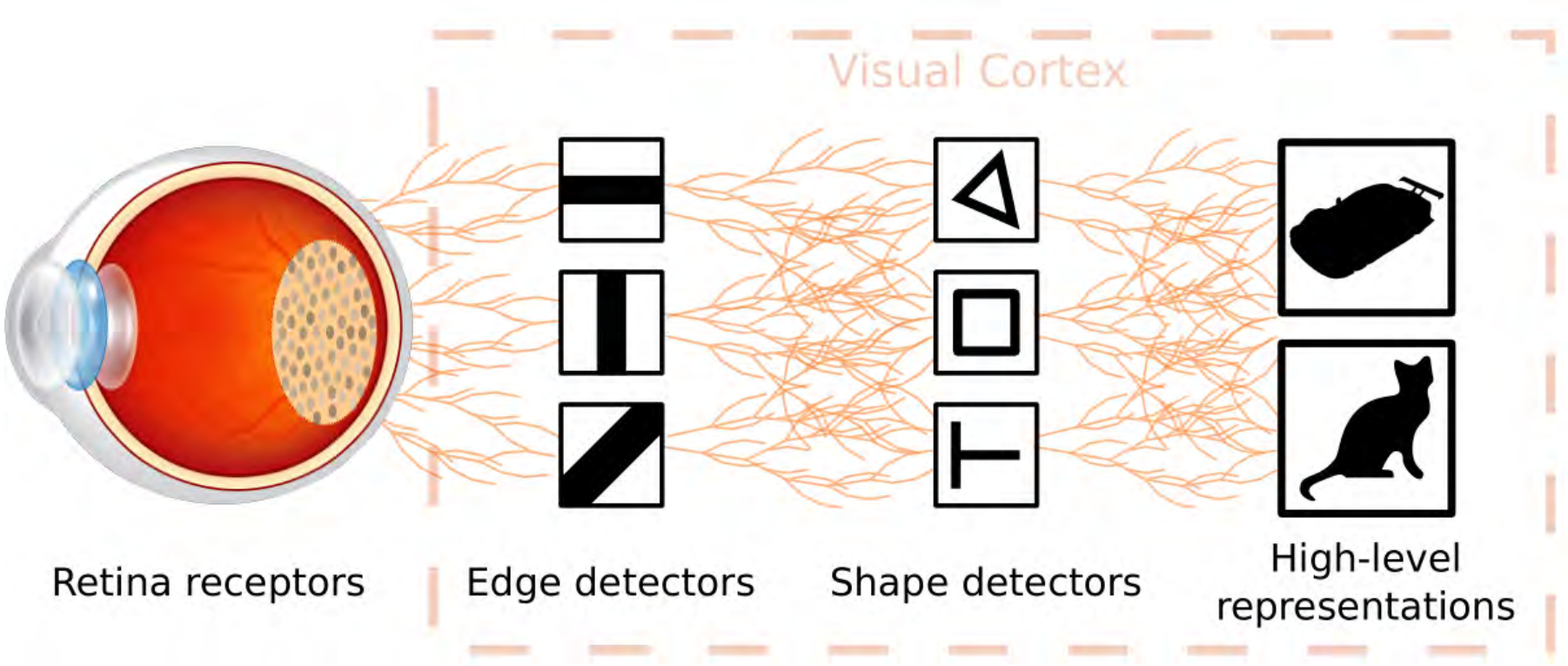
¿Y las **opiniones más presentes**? ¿Los tuits están altamente cargados emocionalmente? ¿De manera positiva, negativa?

Engagement

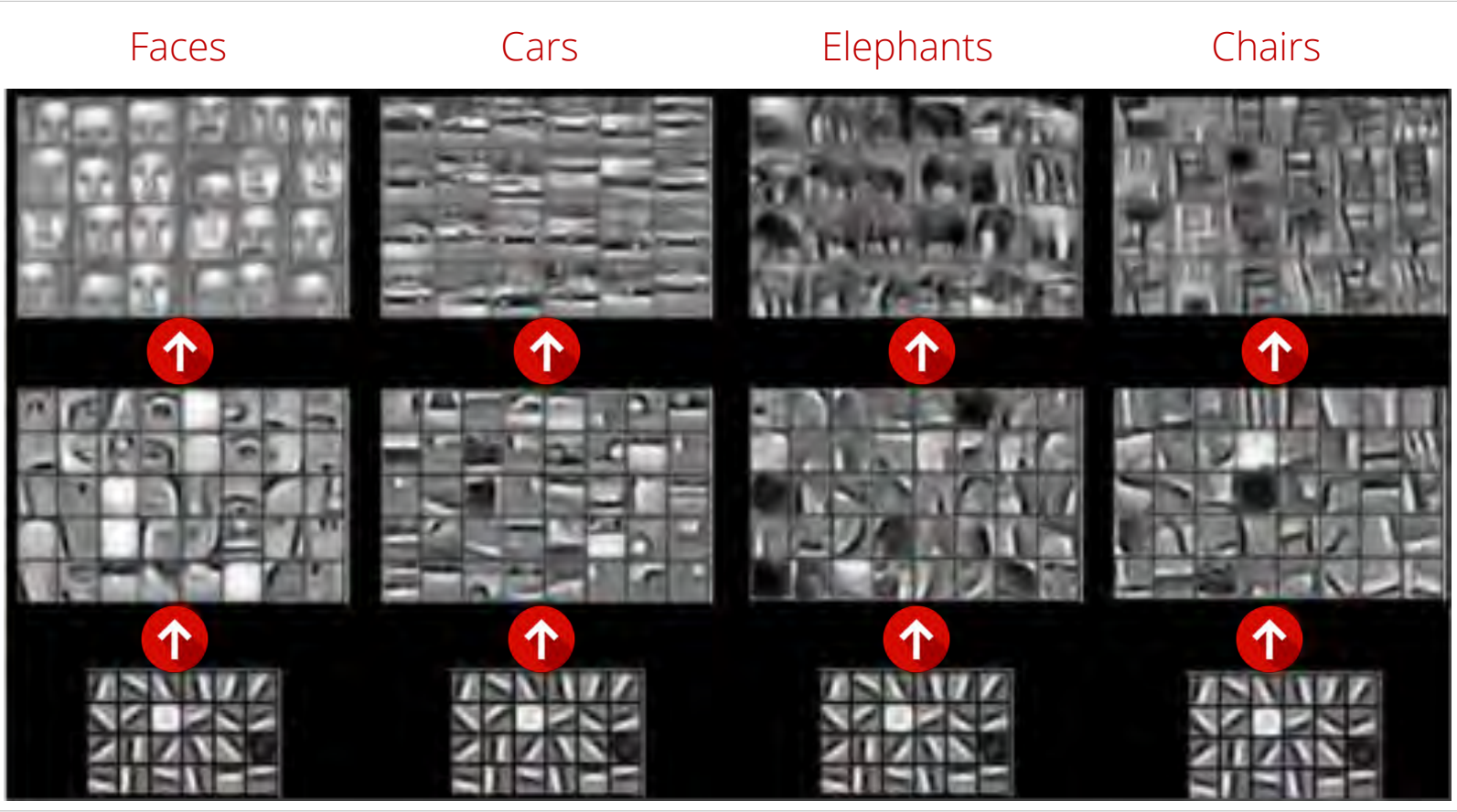
¿Cuáles son los **usuarios y temas estrella** sobre los que se está hablando en la red?



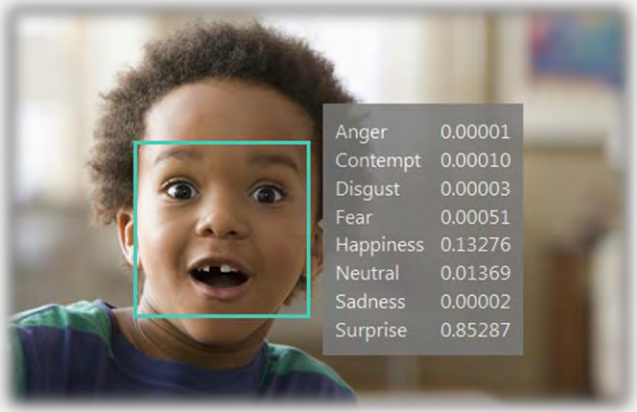
Las imágenes son **números**

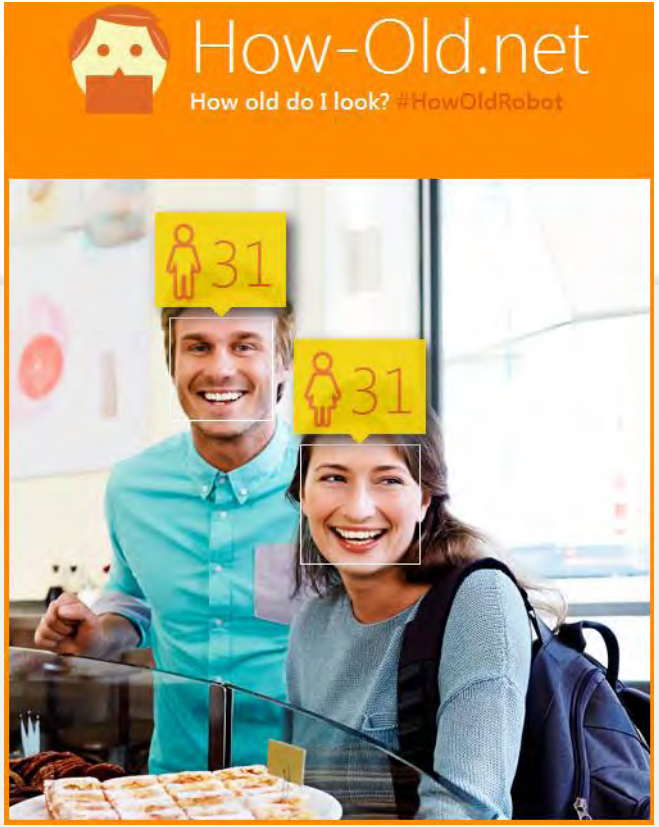

Visual Cortex
Retina receptors
Edge detectors
Shape detectors
High-level representations

Neuronas auto-aprendidas en una red neuronal artificial de varias capas



Ejemplos en reconocimiento de imágenes

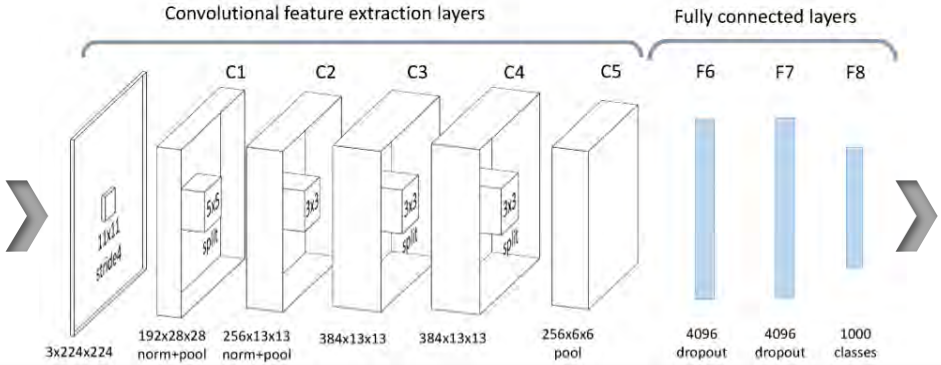




https://www.microsoft.com/cognitive-services/en-us/emotion-api

Extrayendo el estilo artístico de una imagen

Low level observations

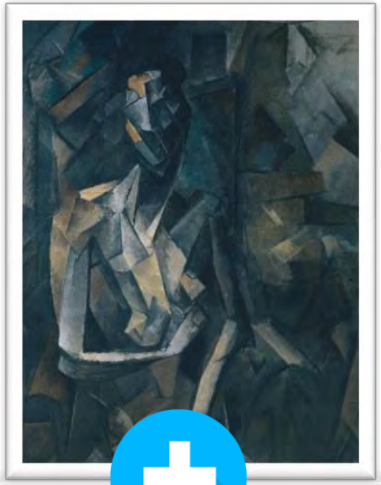


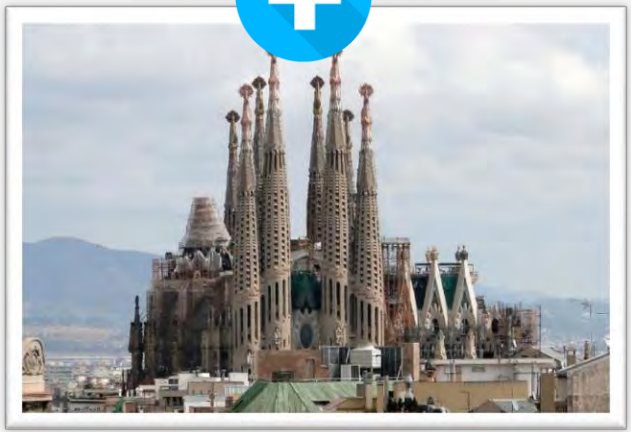
Style embedding

Gatys et al – A Neural Algorithm of Artistic Style
Bottou et al - Optimization Methods for Large-Scale Machine Learning

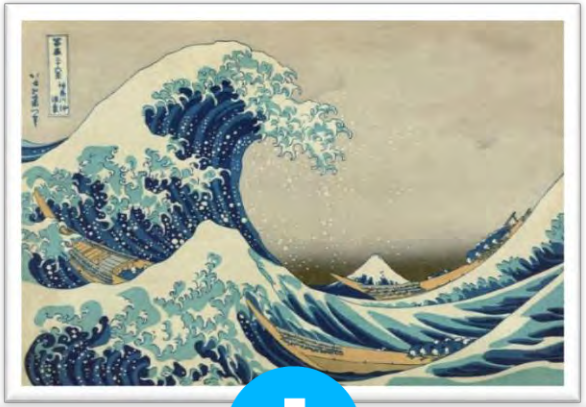
Generando imágenes que combinan contenido y estilo

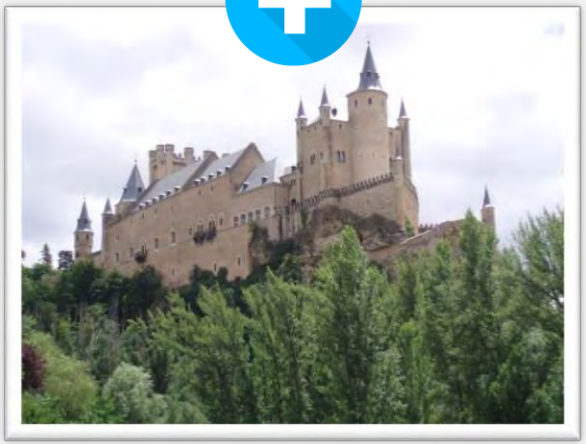
Combinación de estilos artísticos



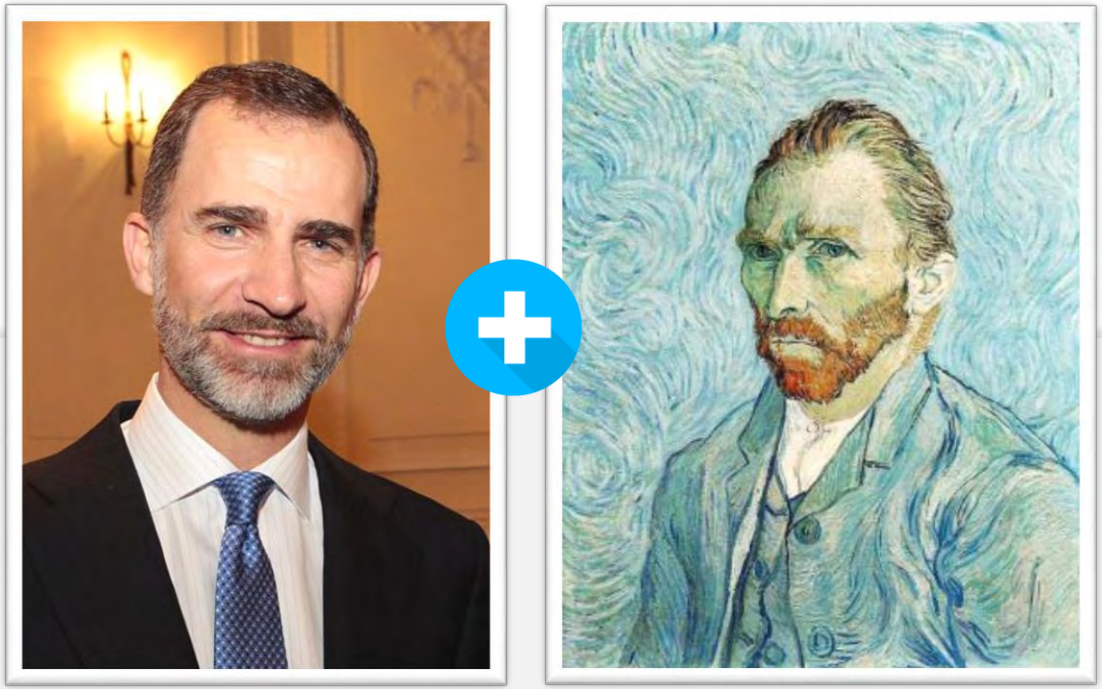


Combinación de estilos artísticos

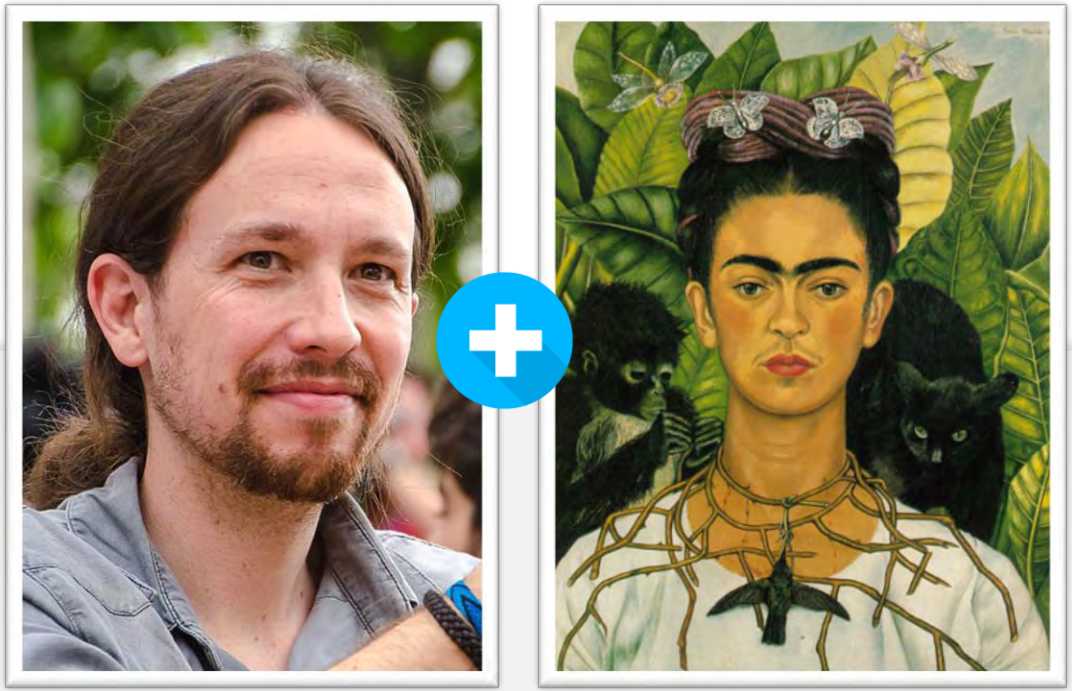


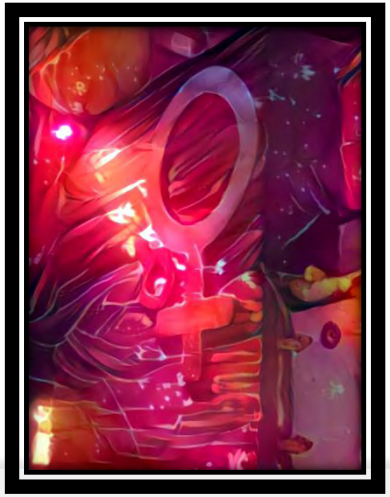


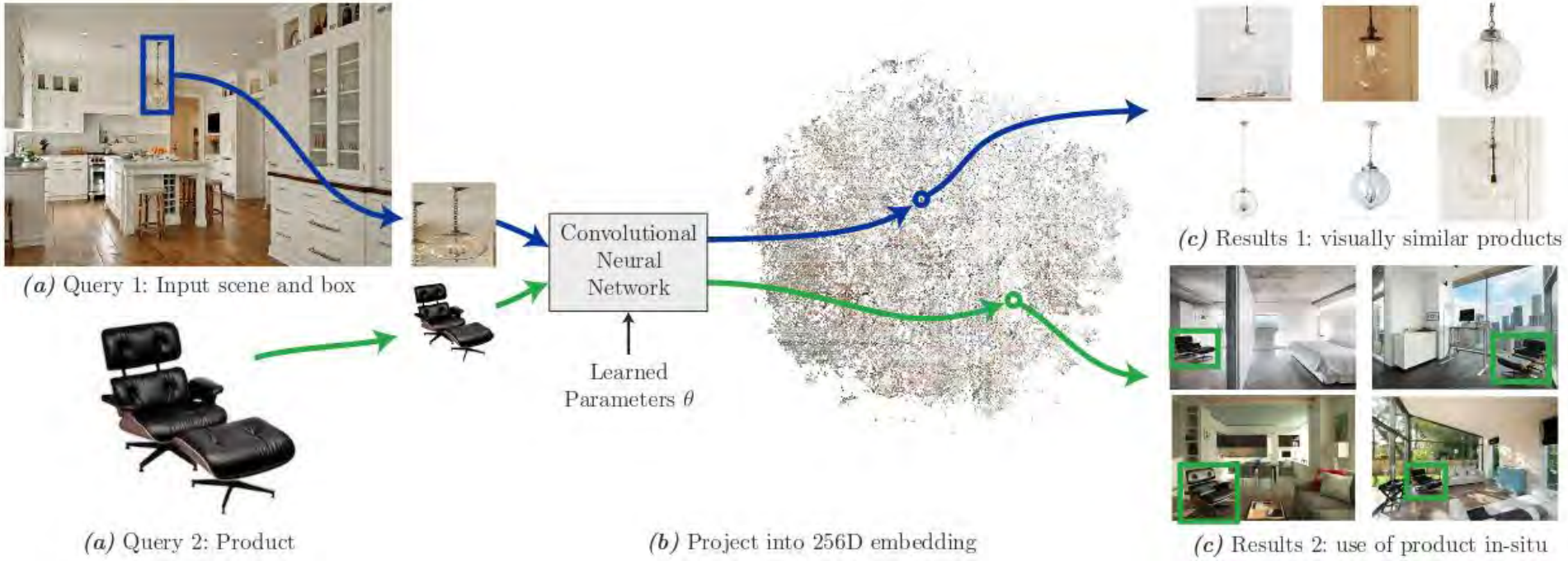
Combinación de estilos artísticos



Combinación de estilos artísticos



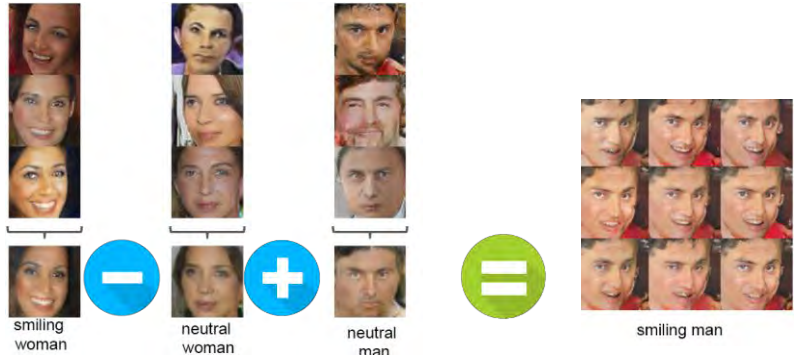




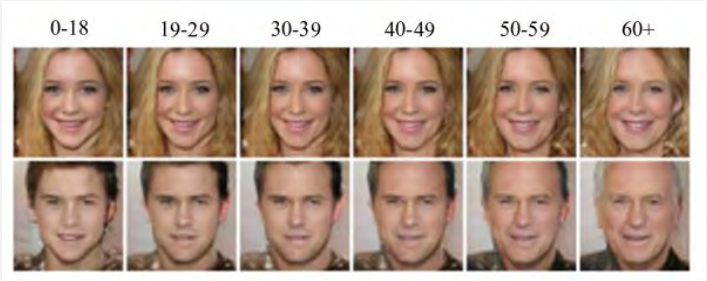
Bell and Bala - Learning visual similarity for product design with convolutional neural networks

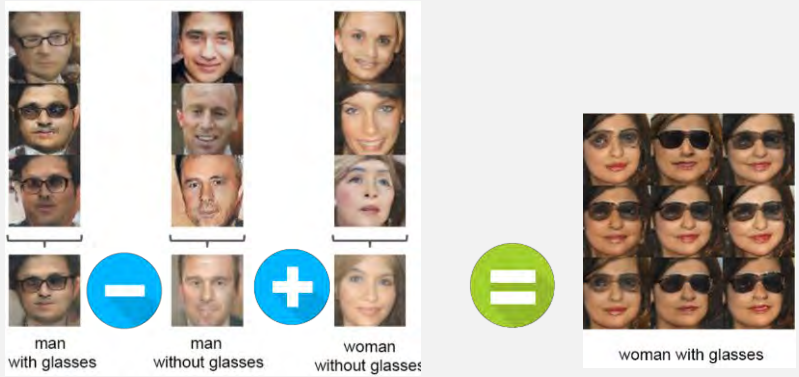
Restando ventanas

Modificando caras



Sumando años





Alec Radford, Luke Metz, Soumith Chintala - Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks
Antipov et al – Face Aging with Conditional Generative Adversarial Networks

Coloreando fotos

Nguyen et al - Plug & Play Generative Networks: Conditional Iterative Generation of Images in Latent Space, http://demos.algorithmia.com/colorize-photos/

Creando imágenes sintéticas

Volcano

Nguyen et al - Plug & Play Generative Networks: Conditional Iterative Generation of Images in Latent Space, http://demos.algorithmia.com/colorize-photos/

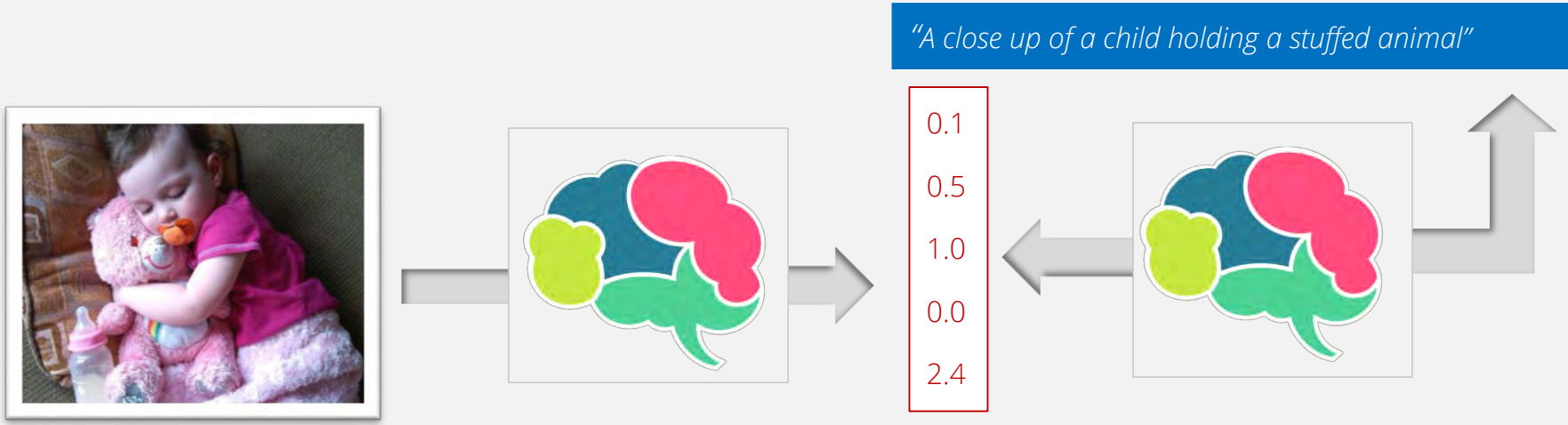
Creando imágenes sintéticas

Redshank | *Ant* | *Monastery*

Nguyen et al - Plug & Play Generative Networks: Conditional Iterative Generation of Images in Latent Space, http://demos.algorithmia.com/colorize-photos/

Mezclando números

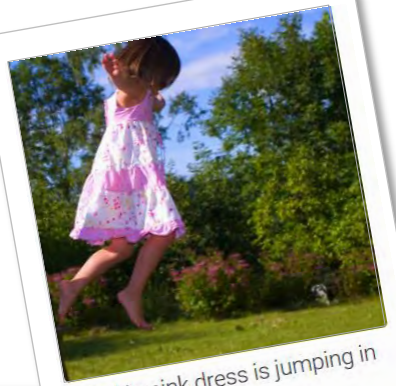
Descripción de imágenes



"man in black shirt is playing guitar."

"construction worker in orange safety vest is working on road."

"two young girls are playing with lego toy."

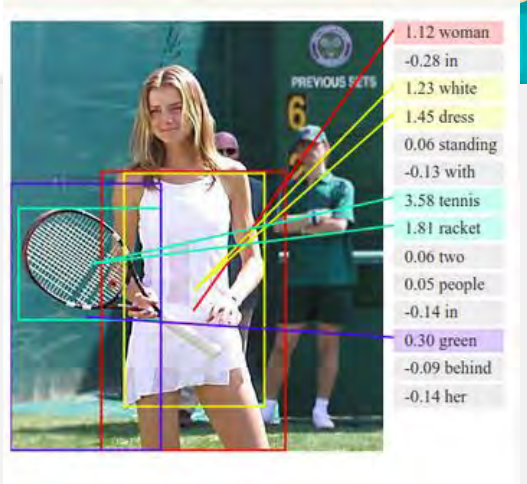


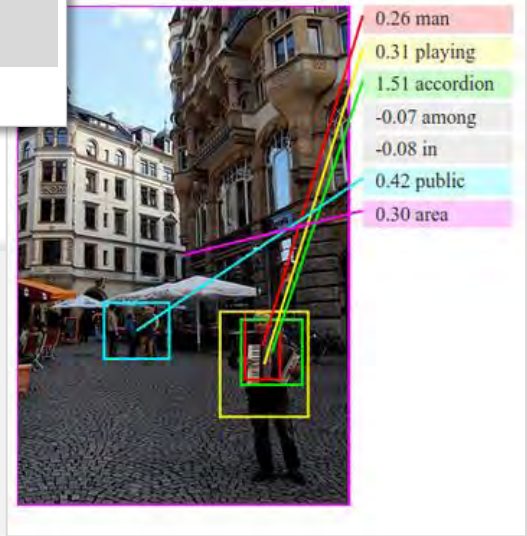
"girl in pink dress is jumping in air."

"black and white dog jumps over bar."

"young girl in pink shirt is swinging on swing."

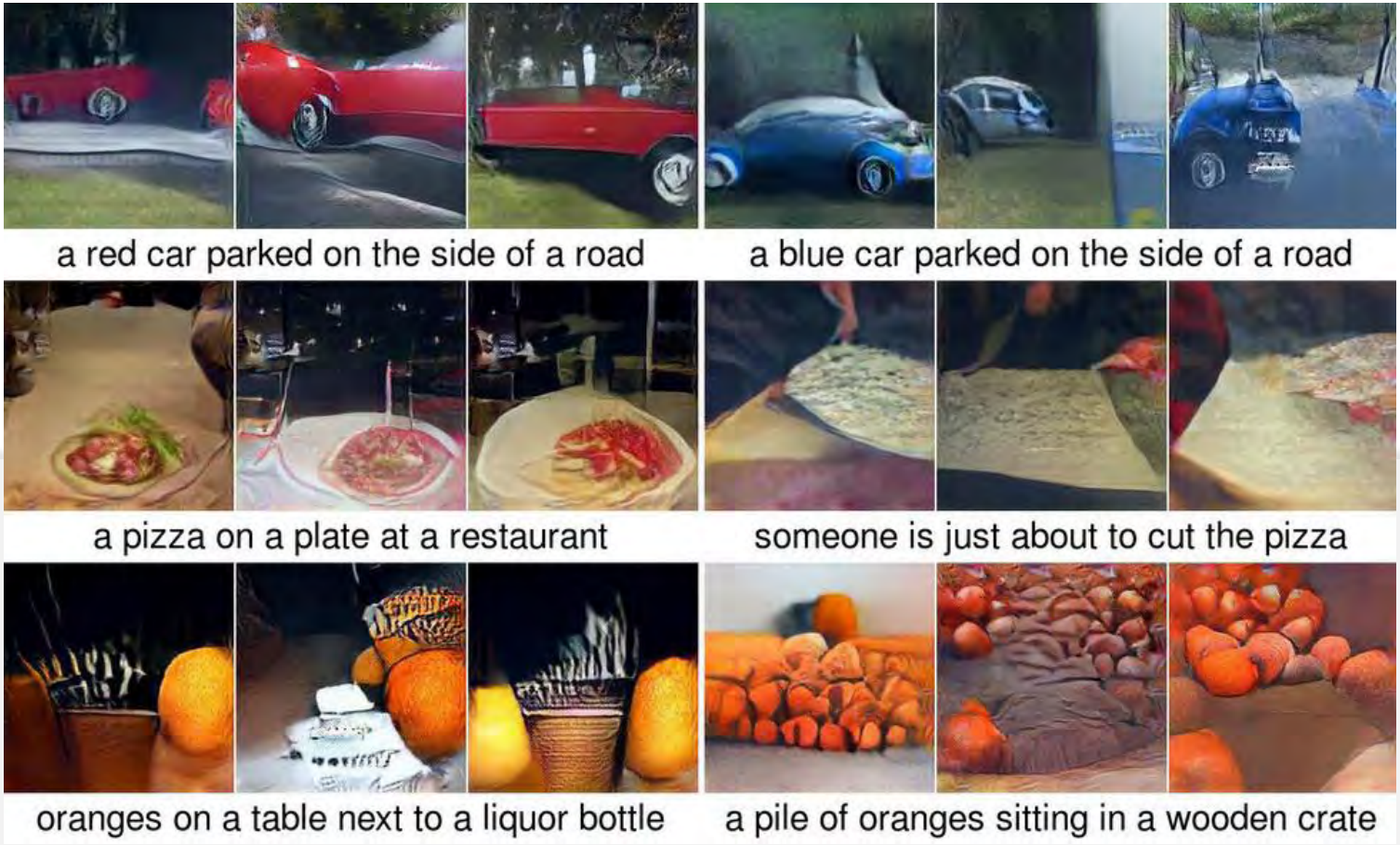
Deep Visual-Semantic Alignments for Generating Image Descriptions
http://cs.stanford.edu/people/karpathy/deepimagesent/





Deep Visual-Semantic Alignments for Generating Image Descriptions
http://cs.stanford.edu/people/karpathy/deepimagesent/

Generación de imágenes en base a textos



Nguyen et al - Plug & Play Generative Networks: Conditional Iterative Generation of Images in Latent Space

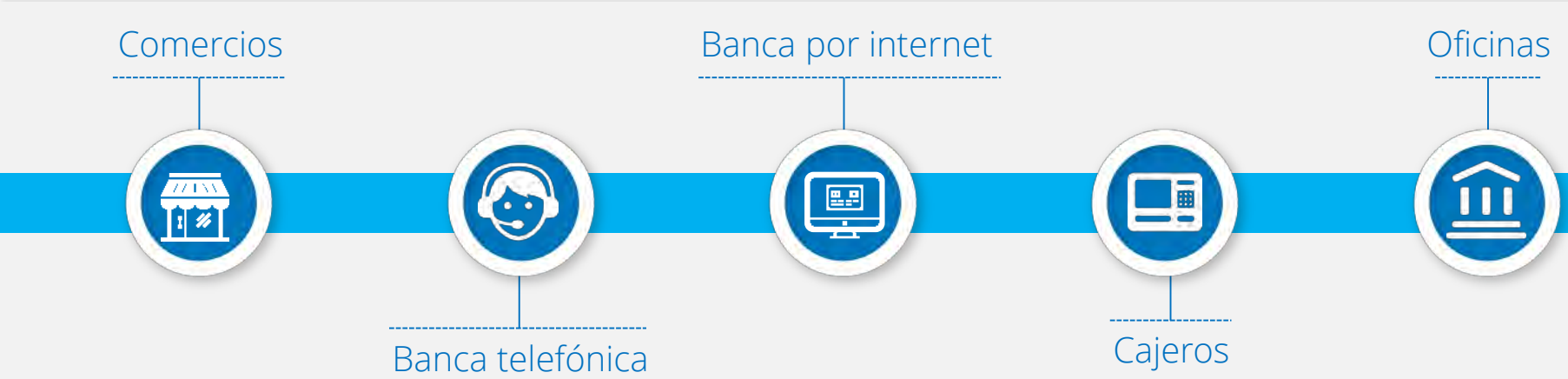
Las personas son **números**

Una persona **no** es un número...

¡Son **muchos** números!

Lynx

Objetivo: perfilado completo del cliente y detección de comportamientos fraudulentos



Operaciones financieras:

- Compras
- Retirada de efectivo
- Transferencias
- Recibos
- Recarga de móviles
- Préstamos, etc.

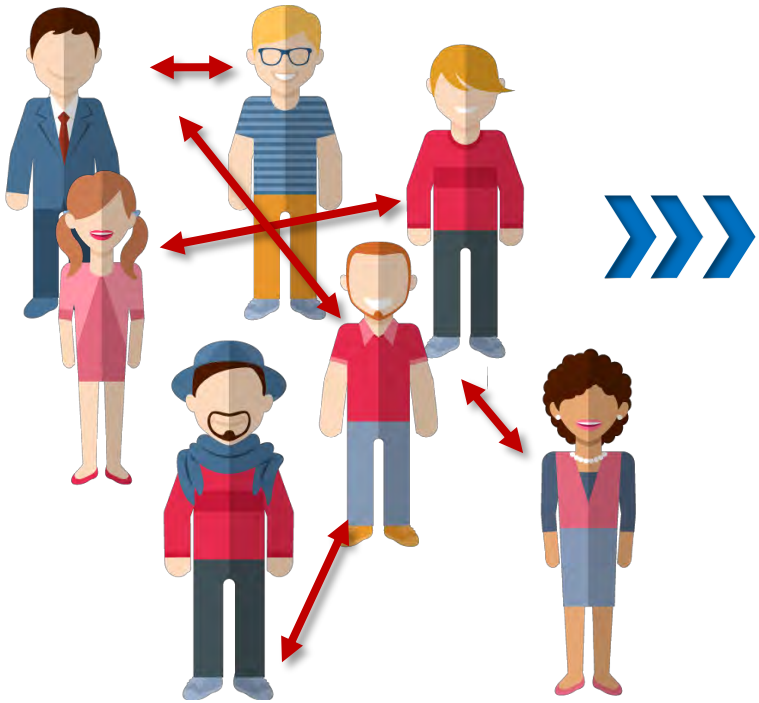
Operaciones no financieras:

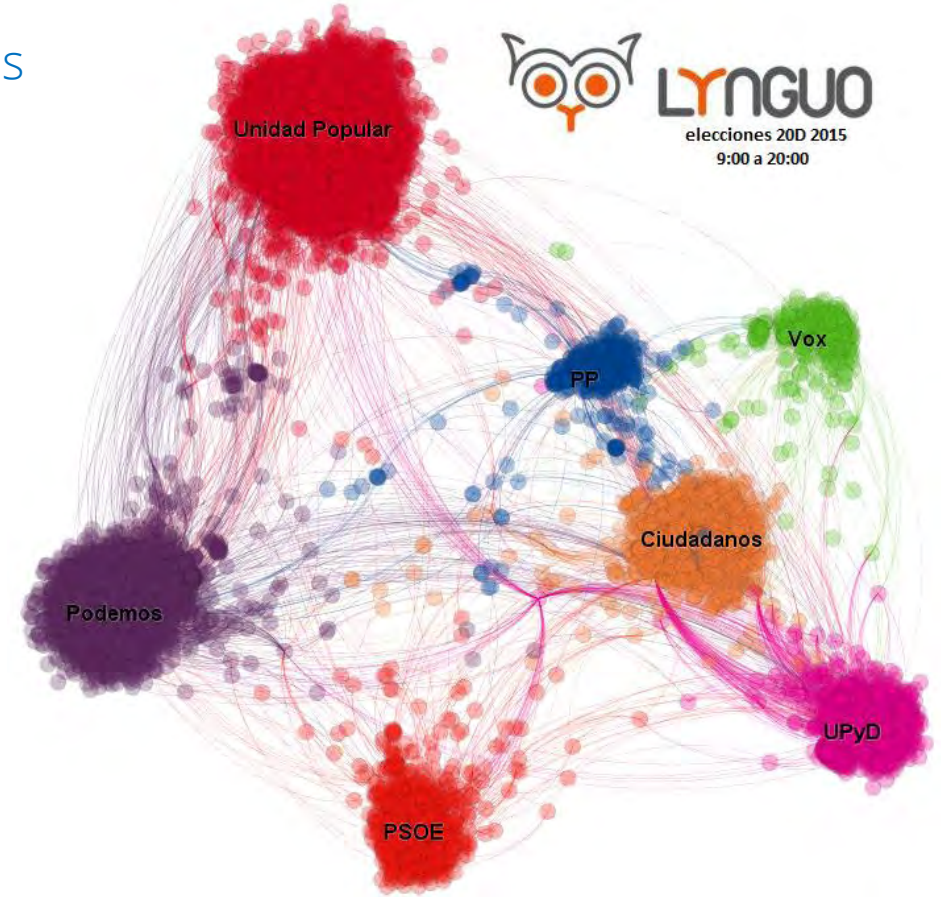
- Consulta de saldo
- Cambio de PIN
- Consulta de movimientos, etc.

- Criba cv
- Llamada Tfonica.
- Pruebas Psicométricas
- Breve entrevista
- Dinámica

Foco	Modelo Azar	Modelo Propuesto
Sólo puedo revisar 160 CVs	*5 Aptos*	*10 Aptos*
Seleccionar 10 Aptos	Reviso 320 CVs	Reviso 160 CVs

Grupos de personas y relaciones entre ellas





AROS y eAROS es un servicio de consultoría para una gestión eficaz de la red social corporativa de las empresas a través de su monitorización y análisis.

Proporcionar los indicadores clave de desempeño social (S-KPIs)

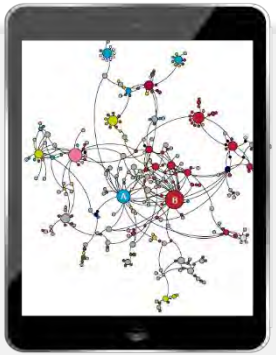
Localizar y caracterizar las comunidades no formales

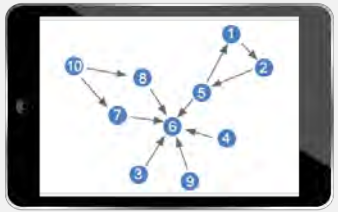
Ubicar el talento y la experiencia.

La instalación de plataformas sociales corporativas permite **canalizar e impulsar los procesos colaborativos** clave para el éxito de una organización.

Pero como medir y gestionar el **impacto social** de estas plataformas colaborativas.

Detectar a los innovadores y influencers.





Identificar los cuellos de botella en los procesos.

Reasignar responsabilidades para mejorar la comunicación entre equipos.

En resumen

Integración de toda clase de fuentes de datos, incluyendo imágenes, textos, video, datos de personas…

Soluciones de datos de mayor nivel de análisis:

predictivas y prescriptivas

Nuevas tecnologías de análisis multimedia:

practicas y efectivas

Solamente necesitas los expertos adecuados…

¡Cuenta con nosotros!

Álvaro Barbero Jiménez

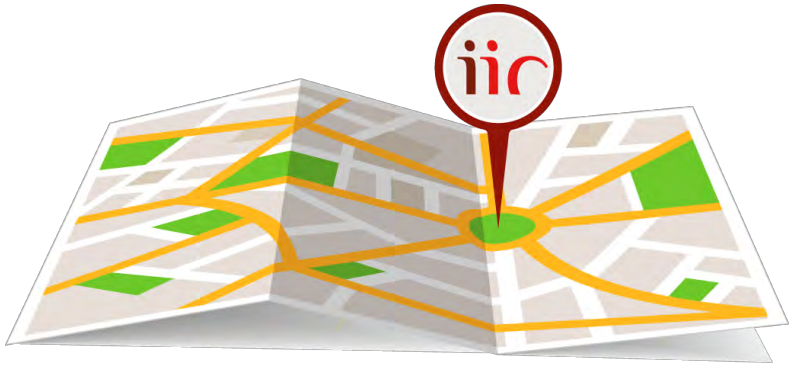
PhD, PMP, Chief Data Scientist at
Instituto de Ingeniería del Conocimiento (IIC)

Alvaro Barbero

@albarjip

albarji.deviantart.com

www.iic.uam.es



Elementos gráficos de apoyo obtenidos en:

designed by freepik.com

pixabay