La innovación empresarial a través de la Ciencia de Datos y la investigación matemática

Dae-Jin Lee

BCAM – Basque Center for Applied Mathematics

En esta charla

- BCAM – Basque Center for Applied Mathematics
- Unidad de transferencia de conocimiento: Data Science e Inteligencia Artificial
- Matemáticas e Innovación
- Algunos ejemplos
- Reflexiones sobre el futuro de la profesión

Sobre mí

- Licenciatura en Admón. y Dirección de Empresas (2002)
- Licenciatura en Ciencias y Técnicas Estadísticas (2004)
- Máster en Ingeniería Matemática (2006)
- Doctorado en Ingeniería Matemática, especialidad en Estadística (2010)

uc3m | Universidad Carlos III de Madrid

- Investigador postdoctoral en Commonwealth Scientific and Industrial Research Organization (CSIRO) en Melbourne, Australia, entre 2011 y 2014.
- División de Matemáticas, Informática y Estadística.

CSIRO

- Desde 2014, lidero el grupo de Estadística Aplicada, dentro del área de Ciencia de Datos e Inteligencia Artificial.

- Formado por investigadores senior, postdoc, estudiantes de doctorado, técnicos de apoyo a la investigación y estudiantes de grado o postgrado (internships).

(bcam) basque center for applied mathematics

- Investigación en el área de Estadística y labores de transferencia.

BCAM – Basque Center for Applied Mathematics

- El BCAM es un centro de investigación en el campo de las matemáticas aplicadas. Fue creado en 2008 por el Gobierno Vasco, la Universidad del País Vasco e Ikerbasque, la Fundación Vasca para la Ciencia. También cuenta con el apoyo de la Diputación Foral de Bizkaia e Innobasque.

EUSKO JAURLARITZA GOBIERNO VASCO

eman ta zabal zazu
Universidad del País Vasco
Euskal Herriko Unibertsitatea

ikerbasque
Basque Foundation for Science

Bizkaia
foru aldundia
diputación foral

innobasque
Berrikuntzaren Euskal Agentzia
Agencia Vasca de la Innovación

- Alineado con la Estrategia Vasca de Ciencia y Tecnología y colaborando con universidades, centros tecnológicos, empresas y el resto de agentes de I+D+i.

BCAM – Basque Center for Applied Mathematics

- El BCAM ha sido reconocido en dos ocasiones como Centro de Investigación de Excelencia Severo Ochoa por la Agencia Estatal de Investigación, una distinción que se otorga a las mejores instituciones de investigación del mundo en su campo.

EXCELENCIA
SEVERO
OCHOA

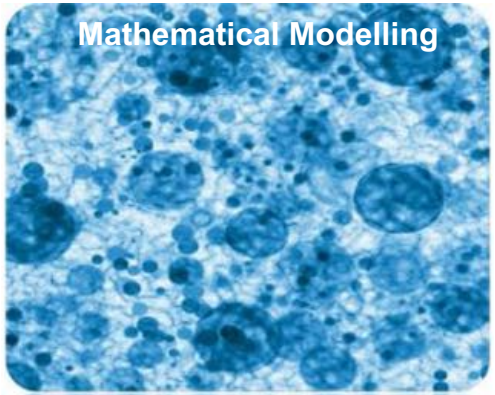
- Basado en un núcleo de investigadores altamente cualificados y una amplia red internacional de excelencia, el BCAM se ha convertido en un centro de referencia internacional en el campo de la Matemática Aplicada.

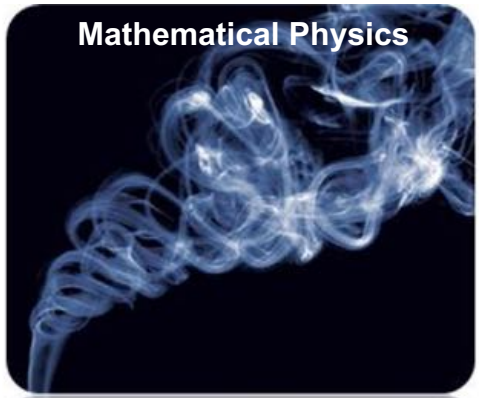
Nuestra misión

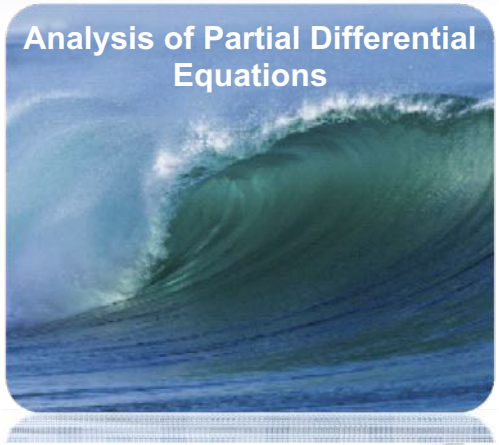
Las matemáticas al servicio de la Sociedad

- Realizar investigaciones en las fronteras de las matemáticas,
- Formar y atraer a científicos con talento,
- Transferir conocimientos a la industria y a los agentes de I+D+i,
- Difundir la importancia de las Matemáticas y sus aplicaciones en la sociedad.
- Actualmente formado por 14 líneas de investigación y más de 120 personas.

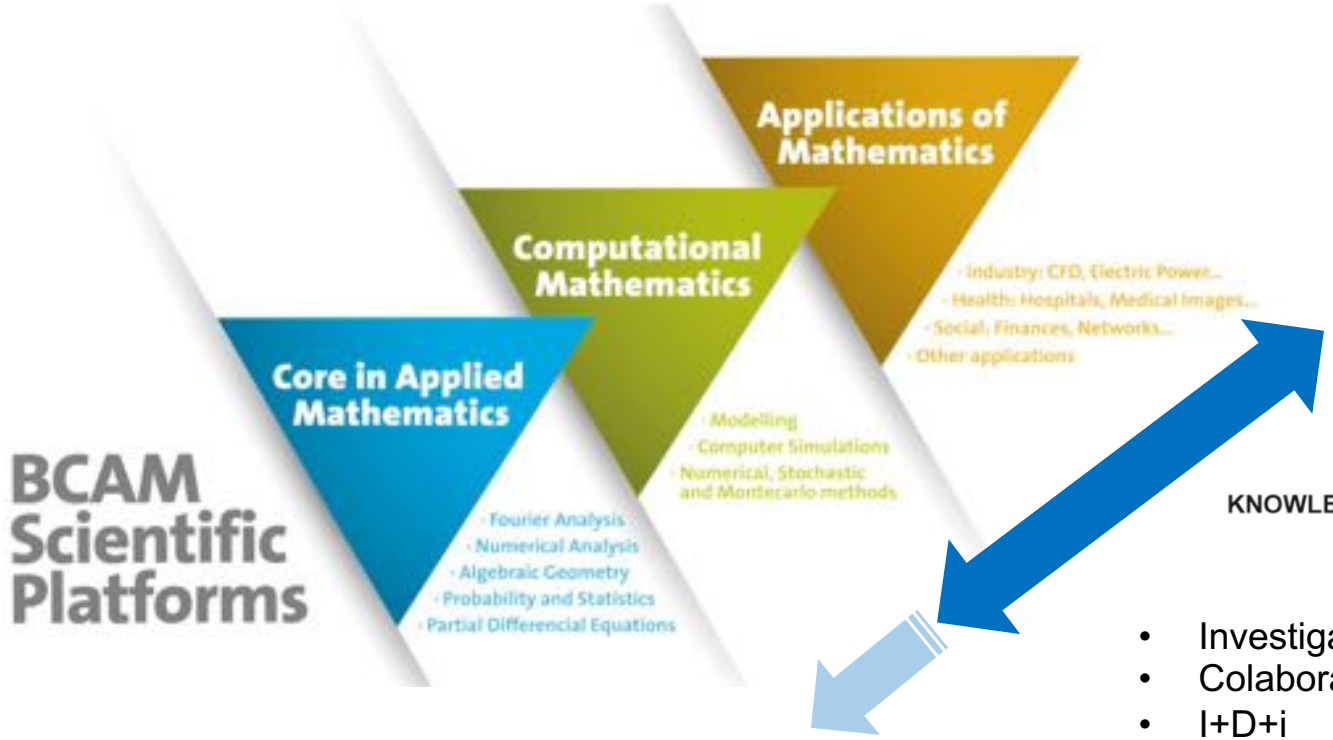
Áreas de investigación







Plataformas científicas



KNOWLEDGE TRANSFER UNIT

K T U

- Investigación aplicada
- Colaboración con la industria
- I+D+i
- Atracción de talento
- Formación avanzada

Objetivos de la KTU

Modelos de colaboración

Alianzas estratégicas

Posiciones conjuntas

Proyectos de Innovación y desarrollo

Supervisión de tésis de master en empresas

Doctorados industrials

Servicio de Asesoramiento a PYMES y Startups

Servicio de diagnóstico de Modelización Matemática

Desde 2019, la KTU cuenta también con el apoyo del "Departamento de Emprendimiento y Competitividad Empresarial" de la Diputación Foral de Bizkaia.

Objetivo: potenciar la innovación de las pymes y startups a través de la transferencia de conocimiento matemático avanzado → Servicio de Asesoramiento Matemático para pymes y startups de Bizkaia

Áreas de especialización de la KTU

Área de ciencia de datos e inteligencia artificial

- Análisis de datos y modelos estadísticos
- Análisis y previsión de series temporales
- Aprendizaje automático
- Optimización a gran escala
- Inteligencia artificial

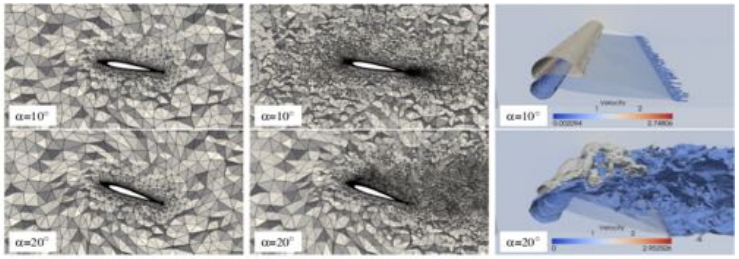
Aplicaciones:

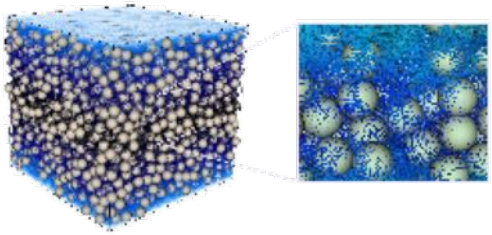
- Biomedicina, biología y agricultura
- Finanzas y seguros
- Energía
- Deportes e Industria 4.0.

Áreas de especialización de la KTU

Área de Modelización y Simulación Numérica

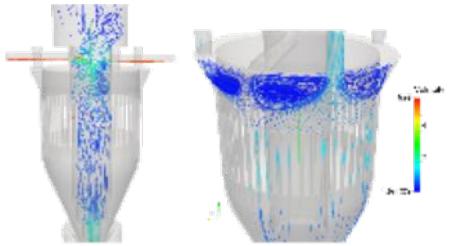
- Dinámica de fluídos computacional (CFD)
- Micro/Macro CFD
- Métodos de Elementos Finitos (adaptativos)





Aplicaciones:

- Manufactura avanzada
- Procesos industriales
- Energías renovables (wind and ocean)



Más de 10 años de transferencia de conocimiento en BCAM

¿Cuánto aportan las matemáticas al PIB?

RED ESTRATÉGICA EN MATEMÁTICAS

¿Cuánto aportan las matemáticas al PIB?

1) Diseño, modelaje, simulación y prototipado de productos.

2) Optimización de procesos productivos y de organización. Reducción de costes, y mejora de la eficiencia en la producción (producir al menor coste posible).

3) Análisis de datos. Big Data e Inteligencia Artificial. Análisis de grandes volúmenes de datos

Fundación Princesa de Asturias

PREMIO PRINCESA DE ASTURIAS
DE INVESTIGACIÓN CIENTÍFICA Y TÉCNICA 2020

YVES MEYER, INGRID DAUBECHIES, TERENCE TAO Y EMMANUEL CANDÈS

PREMIO PRINCESA DE ASTURIAS DE INVESTIGACIÓN CIENTÍFICA Y TÉCNICA 2020

Yves Meyer (francés), Ingrid Daubechies (belga y estadounidense), Terence Tao (australiano y estadounidense) y Emmanuel Candès (francés) han realizado contribuciones pioneras y trascendentales a las teorías y técnicas modernas del procesamiento matemático de datos y señales. Estas son base y soporte de la era digital –al permitir comprimir archivos gráficos sin apenas pérdida de resolución–, de la imagen y el diagnóstico médicos –al permitir reconstruir imágenes precisas a partir de un reducido número de datos– y de la ingeniería y la investigación científica –al eliminar interferencias y ruido de fondo–.

Matemáticas e innovación

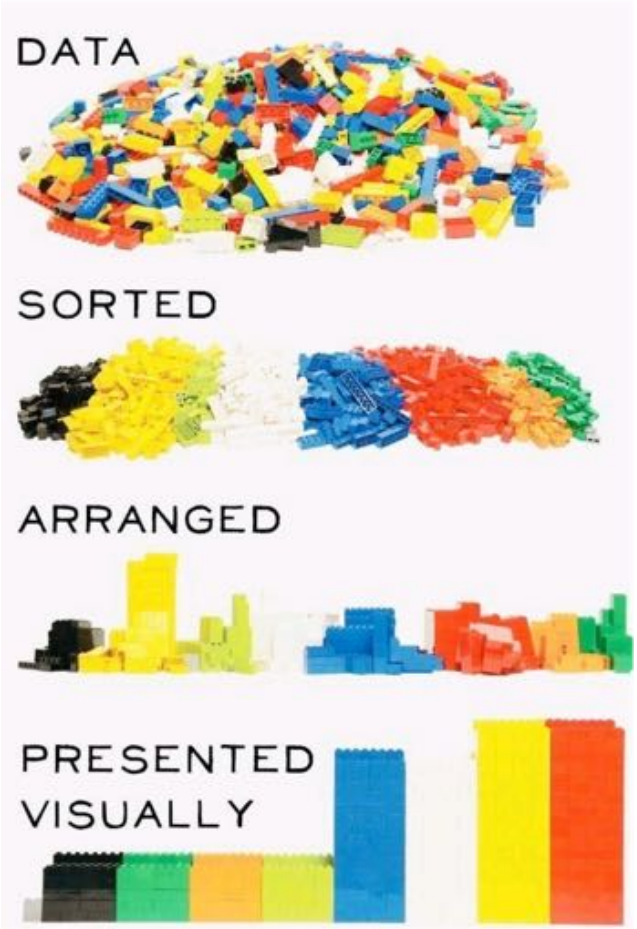
- Las Matemáticas contribuyen a cualquier proceso que involucre:
  - Optimizar
  - Tomar decisiones
  - Predecir

Ciencia de datos y la Industria

`dplyr` (grammar of data manipulation)

Pandas (**Pan**el **Da**ta **S**ystem)

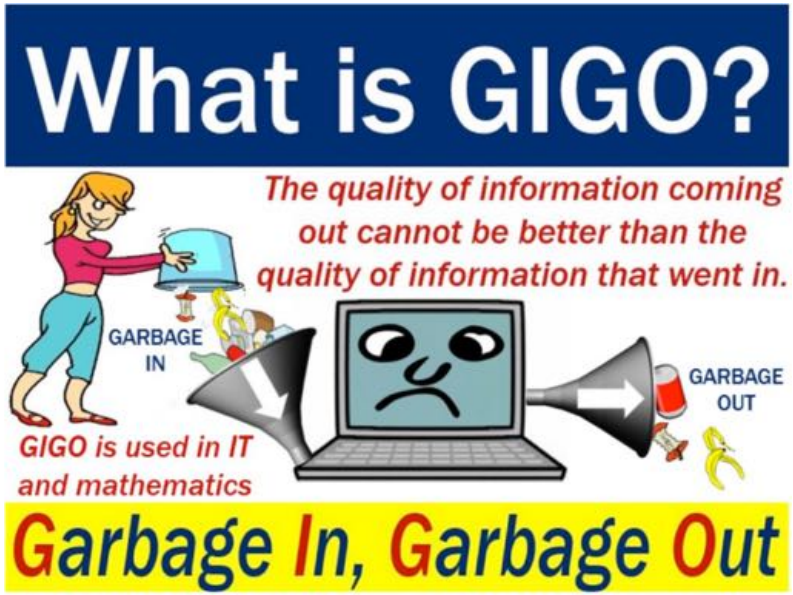
Etapas de un proyecto de datos



EXPLAINED
WITH A STORY

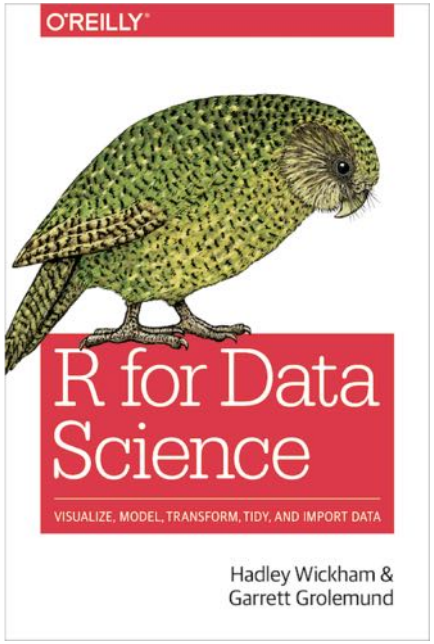
ACTIONABLE
(USEFUL)

Conceptos básicos de cualquier proyecto con datos



- Gestión de datos
- Limpieza de datos
- Análisis de datos
- Modelización de los datos
- Organizar los datos en una forma adecuada para su representación gráfica o en tablas

Ciencia de datos y la Industria



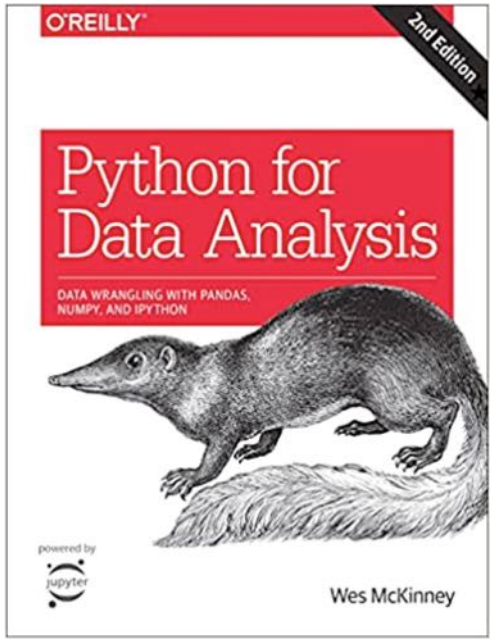
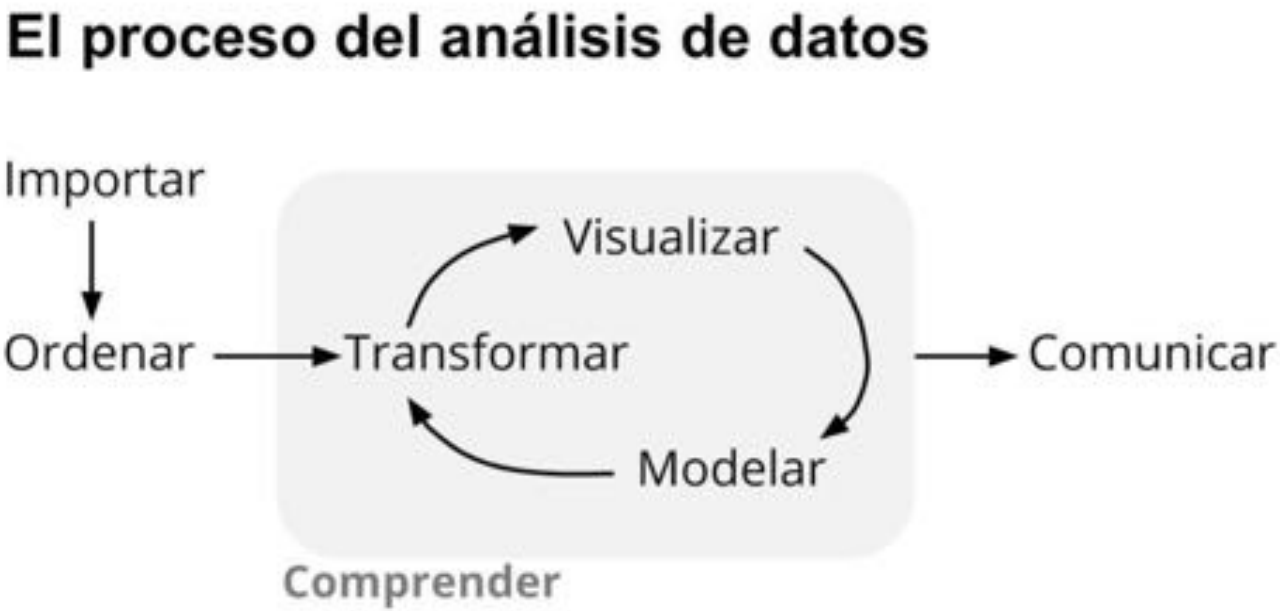


Diagrama tidyverse



Gestión de datos

Reino Unido "extravió" 16 mil casos de Covid-19 por un Excel

El error en los datos, que hizo que 15.841 pruebas positivas quedaran fuera de las cifras oficiales diarias, significa que más de 50.000 personas potencialmente infecciosas pueden haber sido pasadas por alto por los rastreadores de contactos y no se les dijo que se autoaislen.

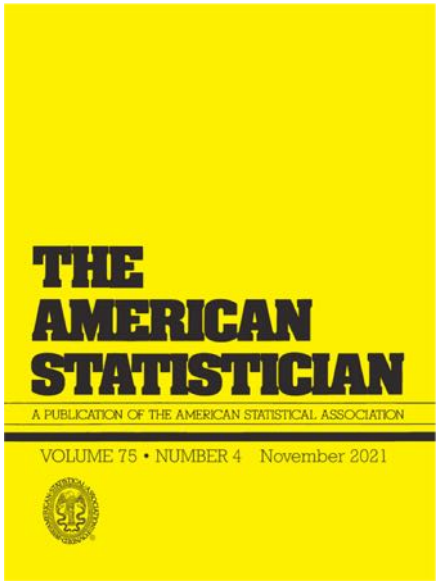
Gestión de datos

Data Organization in Spreadsheets

Karl W. Broman[a] and Kara H. Woo[b]

[a]Department of Biostatistics & Medical Informatics, University of Wisconsin-Madison, Madison, WI; [b]Information School, University of Washington, Seattle, WA

THE AMERICAN STATISTICIAN
2018, VOL. 72, NO. 1, 2–10
https://doi.org/10.1080/00031305.2017.1375989



Este artículo ofrece recomendaciones prácticas para organizar los datos de las hojas de cálculo para reducir los errores y facilitar los análisis posteriores.

Formato ancho a formato largo

	Province/State	Country/Region	Lat	Long	1/22/20	1/23/20	1/24/20	1/25/20	1/26/20	1/27/20	...	6/20/20	6/21/20	6/22/20	6/23/20	6/24/20	6/25/20
0	NaN	Afghanistan	33.0000	65.0000	0	0	0	0	0	0	...	28424	28833	29157	29481	29640	30175
1	NaN	Albania	41.1533	20.1683	0	0	0	0	0	0	...	1891	1962	1995	2047	2114	2192
2	NaN	Algeria	28.0339	1.6596	0	0	0	0	0	0	...	11631	11771	11920	12076	12248	12445
3	NaN	Andorra	42.5063	1.5218	0	0	0	0	0	0	...	855	855	855	855	855	855
4	NaN	Angola	-11.2027	17.8739	0	0	0	0	0	0	...	176	183	186	189	197	212

5 rows × 164 columns

	Id	Province_State	Country_Region	Date	ConfirmedCases
0	1	NaN	Afghanistan	2020-01-22	0.0
1	2	NaN	Afghanistan	2020-01-23	0.0
2	3	NaN	Afghanistan	2020-01-24	0.0
3	4	NaN	Afghanistan	2020-01-25	0.0
4	5	NaN	Afghanistan	2020-01-26	0.0
...	...	...	...	...	...
23557	32708	NaN	Zimbabwe	2020-04-03	9.0
23558	32709	NaN	Zimbabwe	2020-04-04	9.0
23559	32710	NaN	Zimbabwe	2020-04-05	9.0
23560	32711	NaN	Zimbabwe	2020-04-06	10.0
23561	32712	NaN	Zimbabwe	2020-04-07	11.0

23562 rows × 6 columns

Pandas `melt`

50 años de Data Science

JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2017, VOL. 26, NO. 4, 745–766
https://doi.org/10.1080/10618600.2017.1384734

50 Years of Data Science

David Donoho

Department of Statistics, Stanford University, Standford, CA

Tukey, John W. 1962. “The Future of Data Analysis.” *The Annals of Mathematical Statistics* 33 (1): 1–67.

John W. Tukey – The Future of Data Analysis (1962)

Identificó 4 puntos que impulsarían la Nueva Ciencia:

1. *Las teorías formales de la Estadística*
2. *El desarrollo acelerado de los ordenadores y los dispositivos de visualización*
3. *El reto, en muchos campos, de contar con más y conjuntos de datos*
4. *El énfasis en la cuantificación en una variedad cada vez mayor de disciplinas*

Las dos culturas

Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Cultures

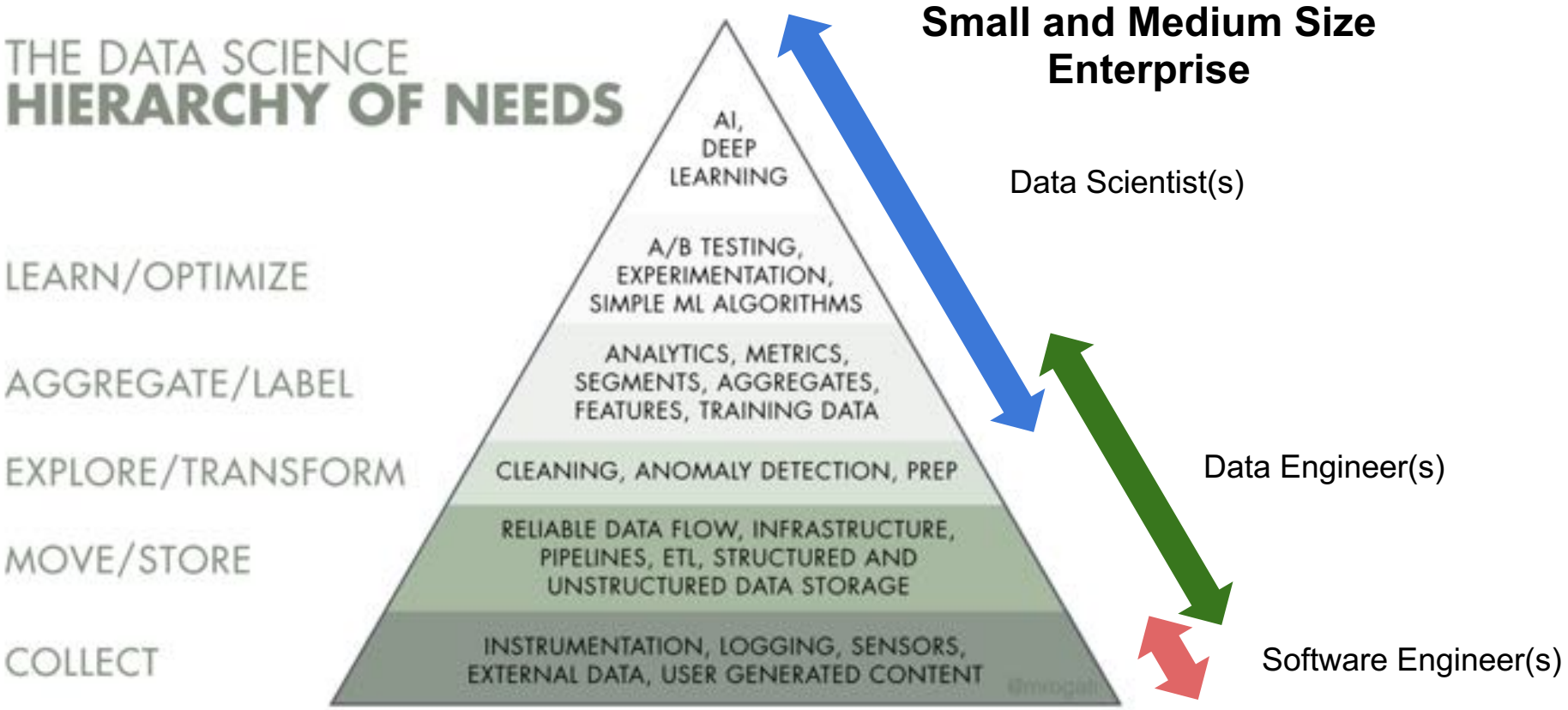
Leo Breiman

El análisis de los datos tiene dos objetivos:

- ***Predicción.*** *Ser capaz de predecir cuáles van a ser las respuestas a las futuras variables de entrada. Modelos Predictivos (que hacen hincapié en la precisión de la predicción realizada por diferentes algoritmos en varios conjuntos de datos). 2% de los Estadísticos.*

- ***Inferencia****. Para inferir cómo la naturaleza asocia las variable respuesta a las variables de entrada. "Cultura del modelado generativo". 98% de los Estadísticos.*

Jerarquía de necesidades en Data Science



Source: Monica Rogati

La ciencia de datos hoy

- Datos públicos (Open Data).
- Open Source Software (R, Python, Julia, etc.)
- Comunidad de usuarios online (stackoverflow)
- Competiciones tipo Kaggle, Numerai, Datathon, etc.
- Reproducibilidad (repositories Github), Notebooks, Rmarkdown.
- Dashboards (PowerBI, Shiny, Qlik, Tableau)
- Formación online (Datacamp, Coursera).

KTU Data Science Unit Services

1. **Contribute to your data strategy**
   - Which technologies and/or platforms do you use?
   - Identify training needs in Data Science
2. **Data exploration**
   - 80% of a Data Science project consist of data collection, cleaning and organization.
   - Data quality (“garbage in, garbage out”)
3. **Data science methods**
   - Which technique is the most appropriate for my company?
4. **Actions and recommendations**
   - Full diagnostics report.
   - Definition of a suitable Data Science project for your company.

Matemáticas e Innovación

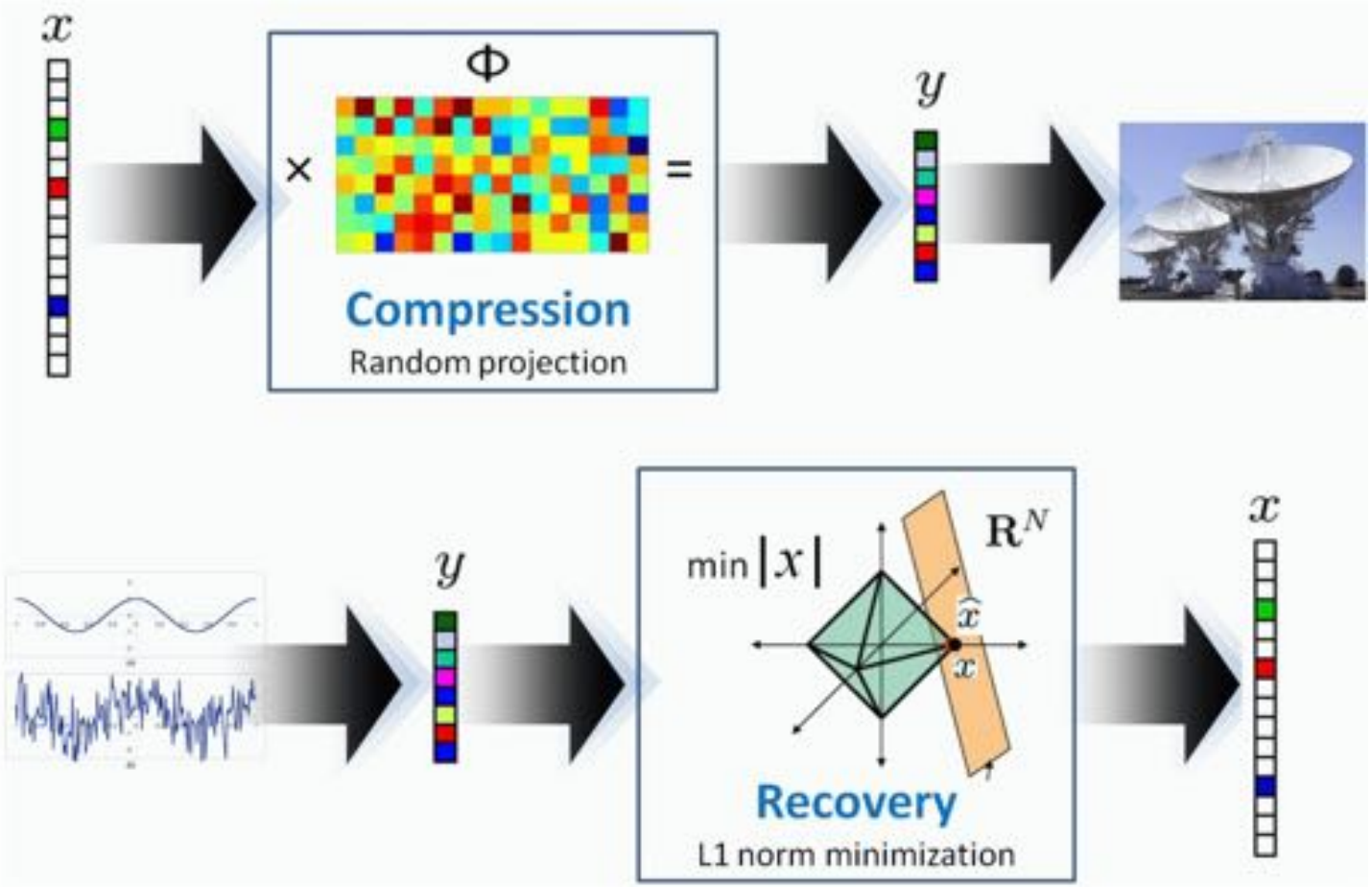
Fundación Princesa de Asturias

PREMIO PRINCESA DE ASTURIAS DE INVESTIGACIÓN CIENTÍFICA Y TÉCNICA 2020

YVES MEYER, INGRID DAUBECHIES, TERENCE TAO Y EMMANUEL CANDÈS

PREMIO PRINCESA DE ASTURIAS DE INVESTIGACIÓN CIENTÍFICA Y TÉCNICA 2020

Yves Meyer (francés), Ingrid Daubechies (belga y estadounidense), Terence Tao (australiano y estadounidense) y Emmanuel Candès (francés) han realizado contribuciones pioneras y trascendentales a las teorías y técnicas modernas del procesamiento matemático de datos y señales. Estas son base y soporte de la era digital –al permitir comprimir archivos gráficos sin apenas pérdida de resolución–, de la imagen y el diagnóstico médicos –al permitir reconstruir imágenes precisas a partir de un reducido número de datos– y de la ingeniería y la investigación científica –al eliminar interferencias y ruido de fondo–.


x
Φ
×
=
y
Compression
Random projection
y
min |x|
R^N
x̂
x
x
Recovery
L1 norm minimization

Microsoft

amazon

Uber

airbnb

SILICON VALLEY

pied piper

Algoritmo de compresión de archivos “middle-out”

SLASH GEAR TECH CARS GAMING ENTERTAINMENT SCIENCE FEATURES

HBO's 'Silicon Valley' inspires real compression algorithm Piper Pied

2

Lindsey Caldwell - May 4, 2015, 4:10am CDT

Dropbox's Lepton lossless image compression really uses a 'middle-out' algorithm

Devin Coldewey @techcrunch / 1:05 am CEST • July 15, 2016

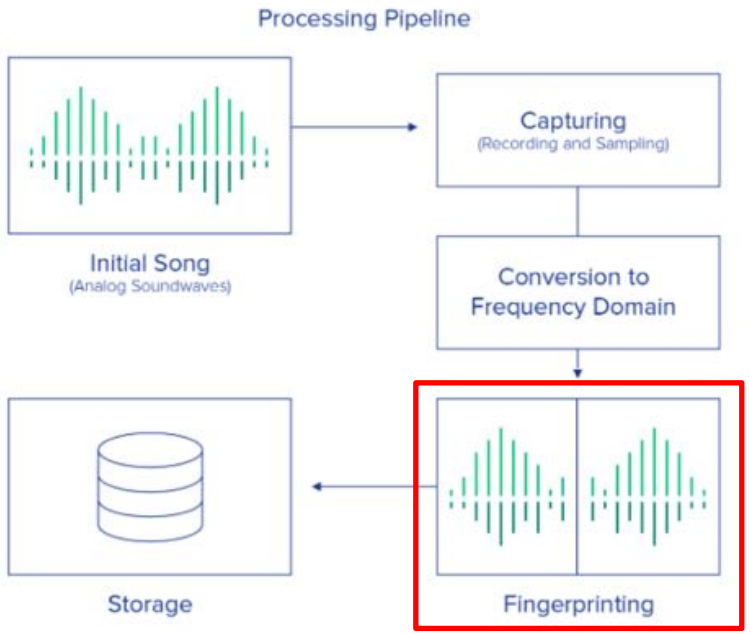
Comment

Shazam – Reconocimiento de canciones

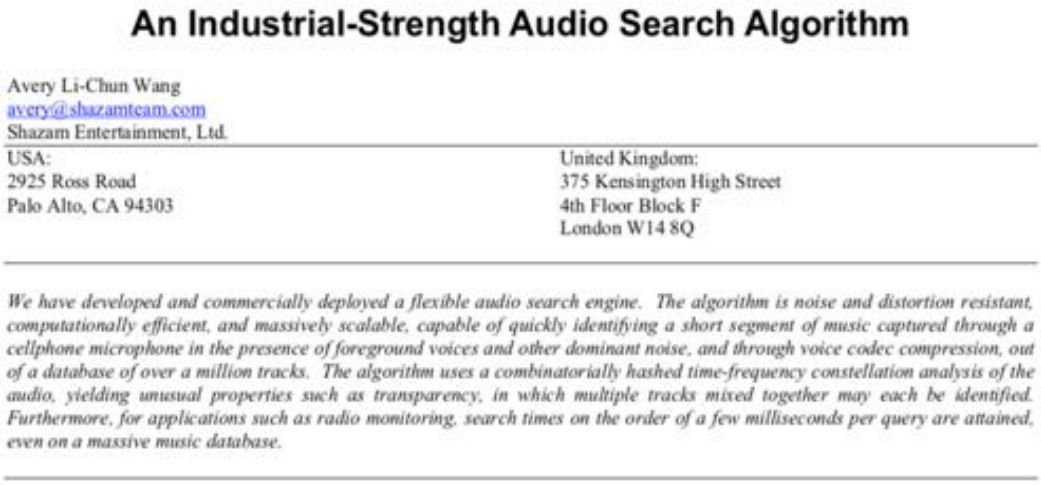
Shazam – Reconocimiento de canciones

Shazam – ¿Cómo funciona?

Hash Tag	Time in Seconds	Song
30 51 99 121 195	53.52	Song A by artist A
33 56 92 151 185	12.32	Song B by artist B
39 26 89 141 251	15.34	Song C by artist C
32 67 100 128 270	78.43	Song D by artist D
30 51 99 121 195	10.89	Song E by artist E
34 57 95 111 200	54.52	Song A by artist A
34 41 93 161 202	11.89	Song E by artist E



Shazam – ¿Cómo funciona?



An Industrial-Strength Audio Search Algorithm

2003

Avery Li-Chun Wang
avery@shazamteam.com
Shazam Entertainment, Ltd.

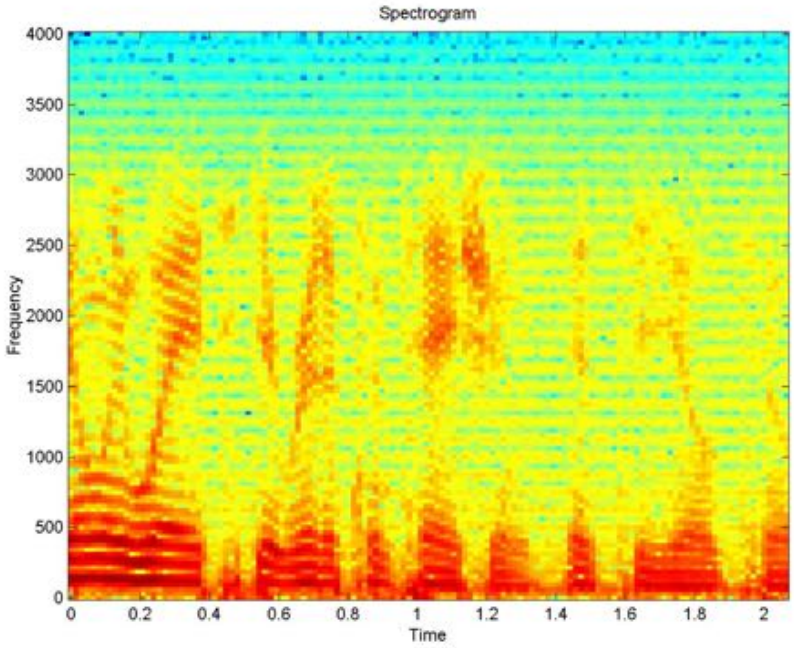
USA:
2925 Ross Road
Palo Alto, CA 94303

United Kingdom:
375 Kensington High Street
4th Floor Block F
London W14 8Q

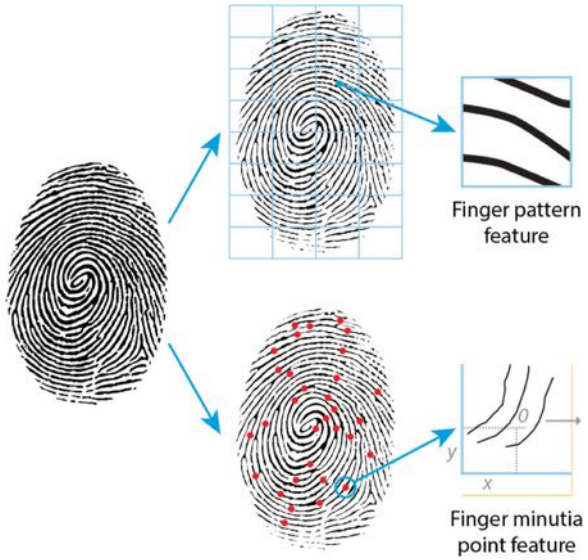
We have developed and commercially deployed a flexible audio search engine. The algorithm is noise and distortion resistant, computationally efficient, and massively scalable, capable of quickly identifying a short segment of music captured through a cellphone microphone in the presence of foreground voices and other dominant noise, and through voice codec compression, out of a database of over a million tracks. The algorithm uses a combinatorially hashed time-frequency constellation analysis of the audio, yielding unusual properties such as transparency, in which multiple tracks mixed together may each be identified. Furthermore, for applications such as radio monitoring, search times on the order of a few milliseconds per query are attained, even on a massive music database.

Shazam Entertaiment Ltd fue **fundada en 1999**
Su primer negocio fue catalogar **1,5 millones de canciones**
Shazam disponible en **Julio 2008** en Iphone
En 2010, la base de datos había aumentado a **8 millones de canciones**
En 2017, Shazam fue comprado por **Apple Inc.** (400 Millones de $)

Shazam – ¿Cómo funciona?

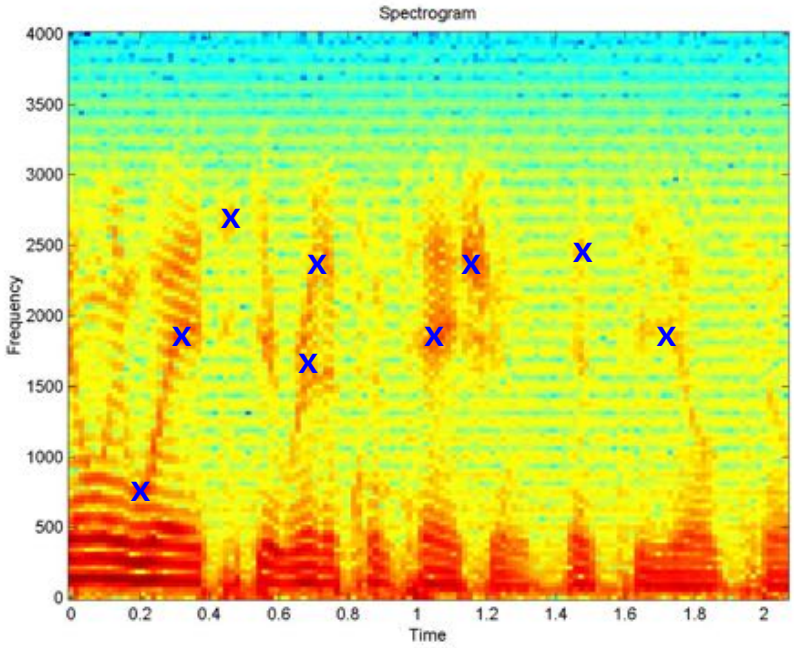


Espectograma

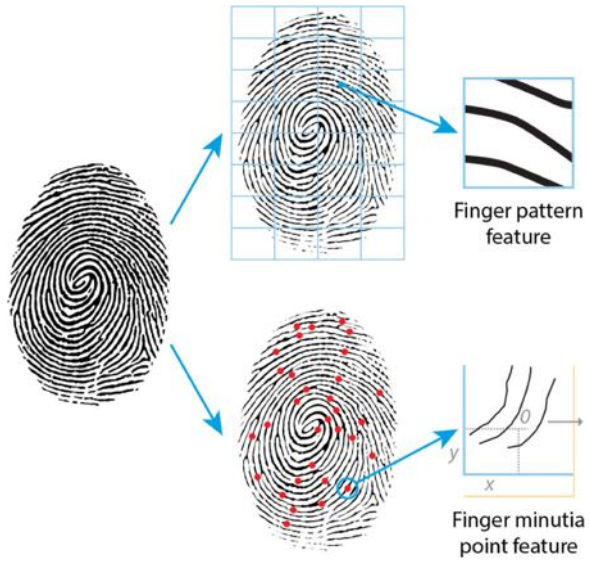


Audio fingerprinting algorithm

Shazam – ¿Cómo funciona?



Espectograma



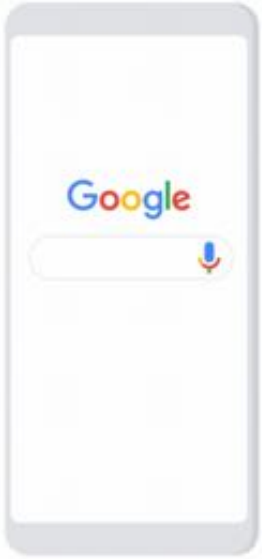
Audio fingerprinting algorithm

Google – Hum to search

15/10/2020

Google – Hum to search

15/10/2020



Otras aplicaciones similares

- Filtros de contenido (detección de copyright en archivos digitales)

Akinator: el genio adivinador

Creado en Francia en el 2005

Akinator adivina qué personaje está pensando el usuario, sea real o no, a través de preguntas sobre las características del mismo.

Para responder, el usuario tiene las opciones

- “Sí”
- “No”
- “No lo sé”
- “Probablemente”
- “Probablemente no”.

Akinator: el genio adivinador

elokence

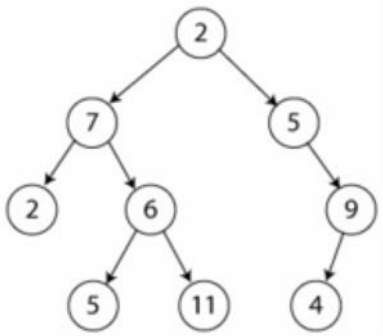
¿Cómo funciona Akinator?

- El algoritmo es “secreto”
- La idea del juego no es nueva

- Utiliza datos y “aprende” en base a las derrotas

¿Cómo funciona Akinator?

- **Sistema experto** (emplea conocimiento humano capturado en un ordenador para resolver problemas que normalmente resolverían humanos expertos).
- Árbol binario

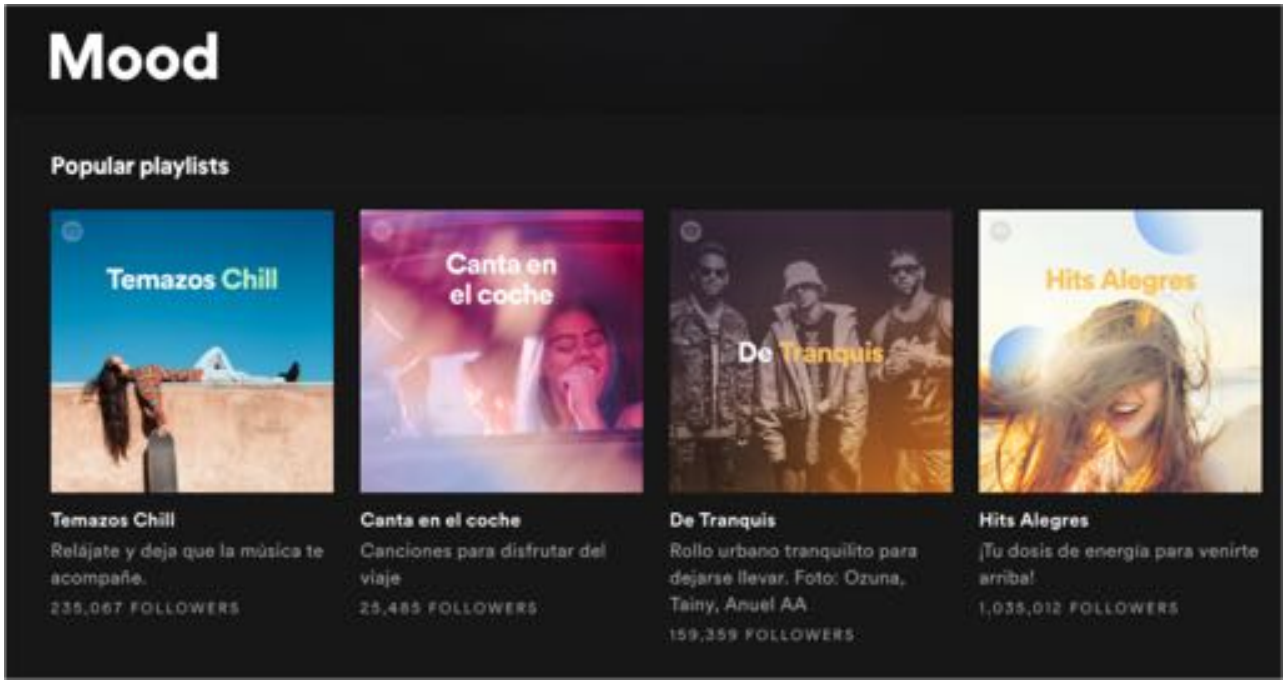


  - Reglas ó heurísticas
  - Probabilidades (Bayes)

- e.g: Asistentes personales, sistemas de recomendación (Amazon, Netflix, etc.)

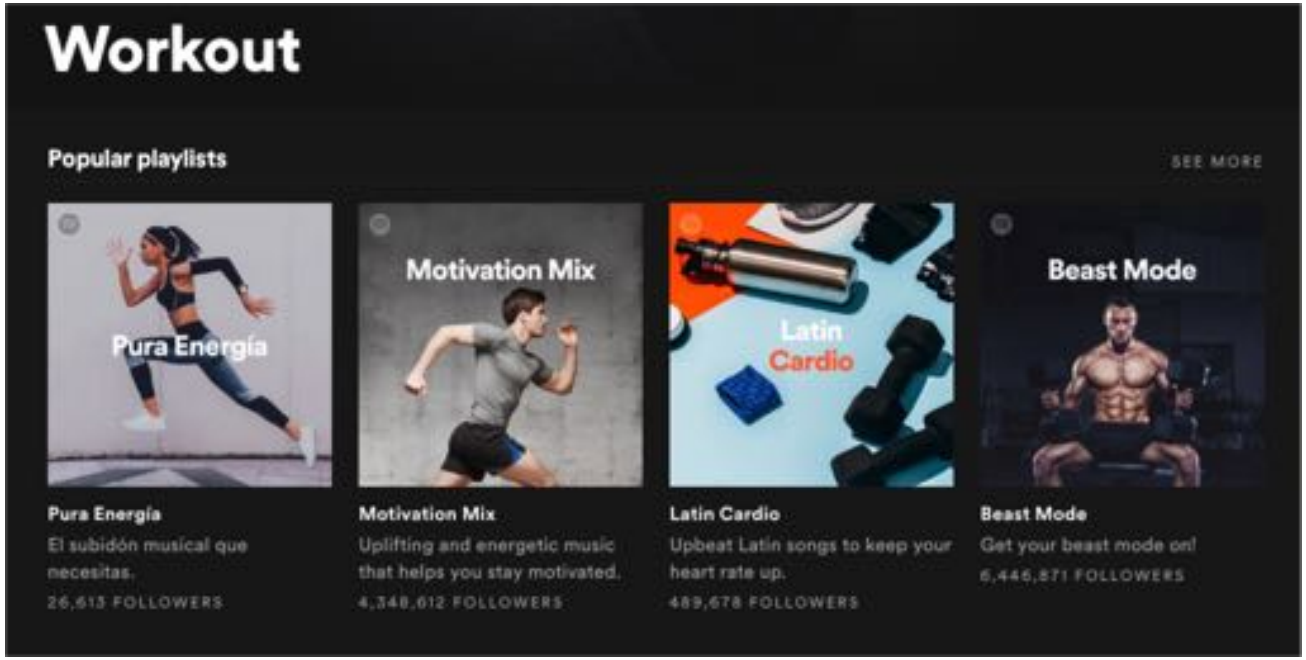
Otras aplicaciones “similares”

Spotify



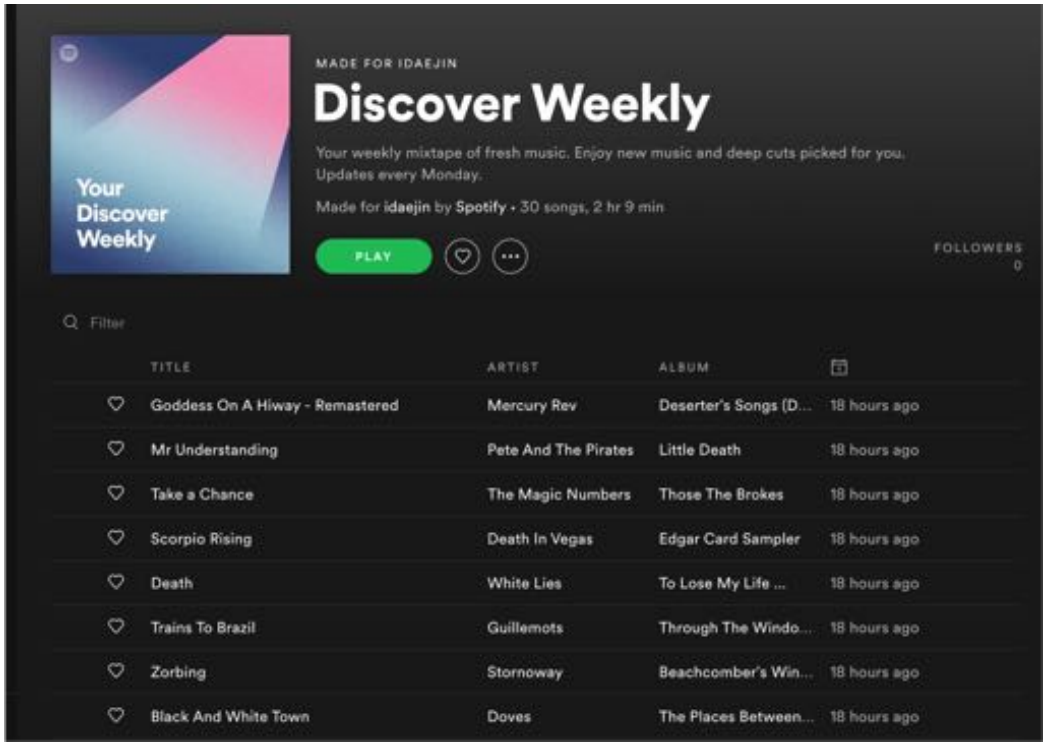
Otras aplicaciones “similares”

Spotify *“Similar artists …” “Discovery weekly” … “Playlists”*



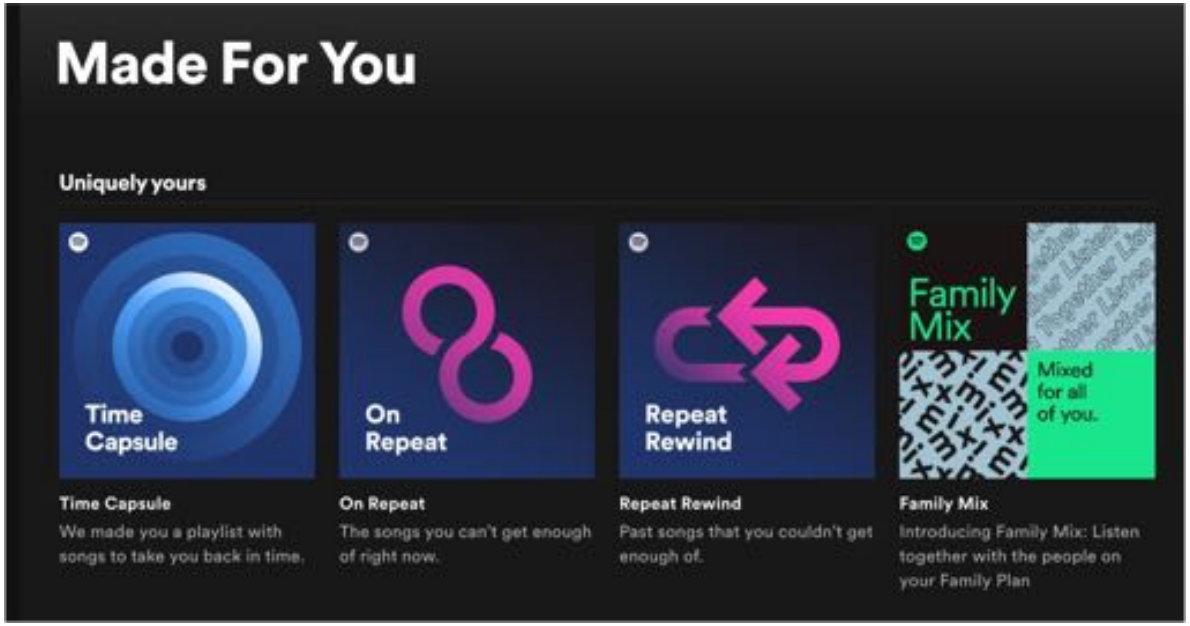
Otras aplicaciones “similares”

Spotify



Otras aplicaciones “similares”

Spotify

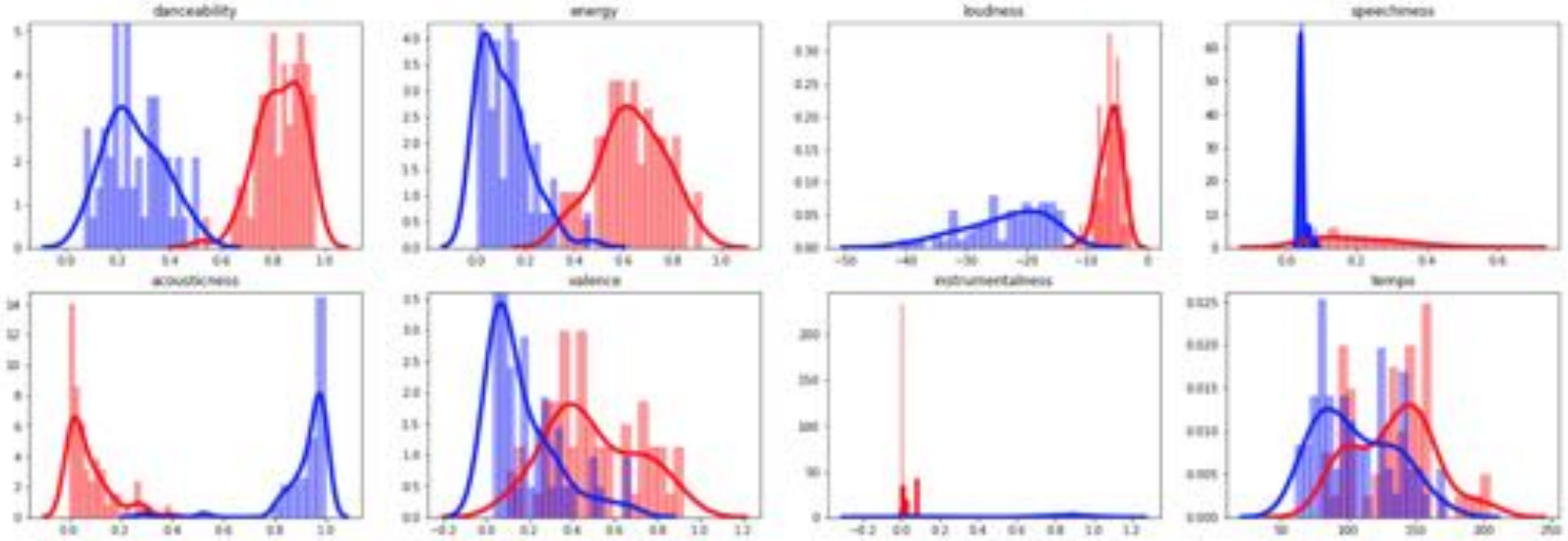


BART (*Bandits for Recommendations as Treatments*)

Otras aplicaciones “similares”

Spotify

Detecta “estilos”



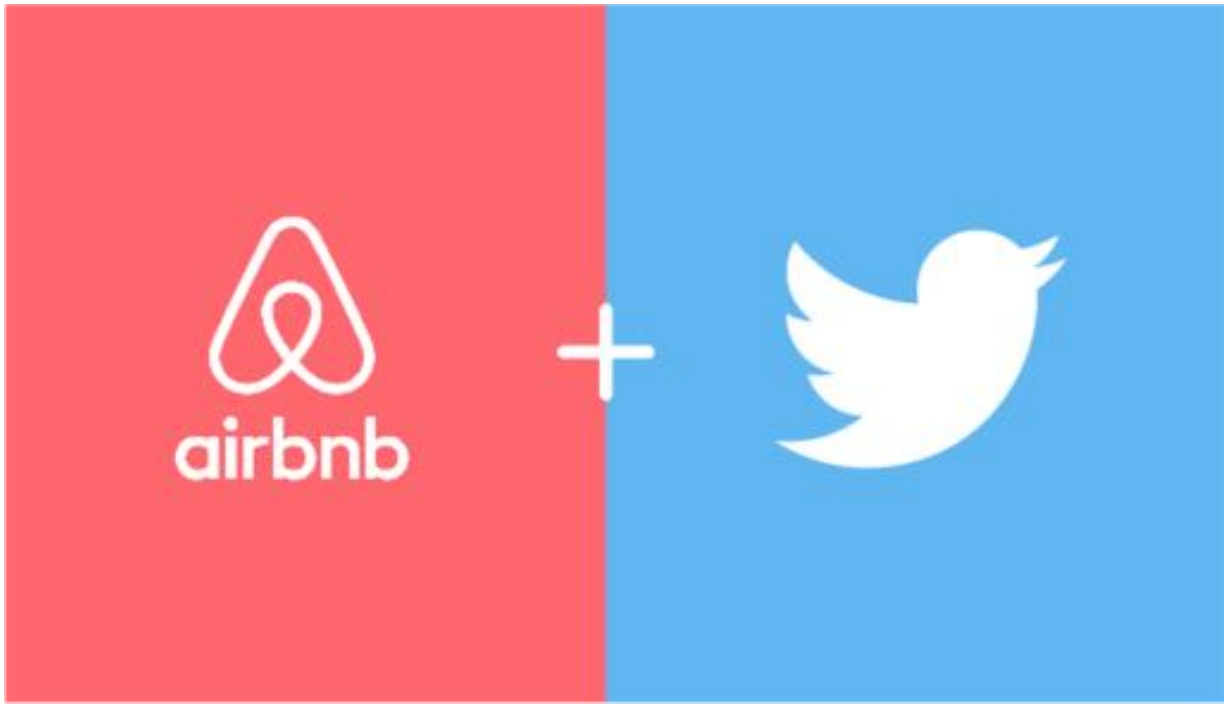
MÉTODO CIENTÍFICO

¿Y en la empresa?

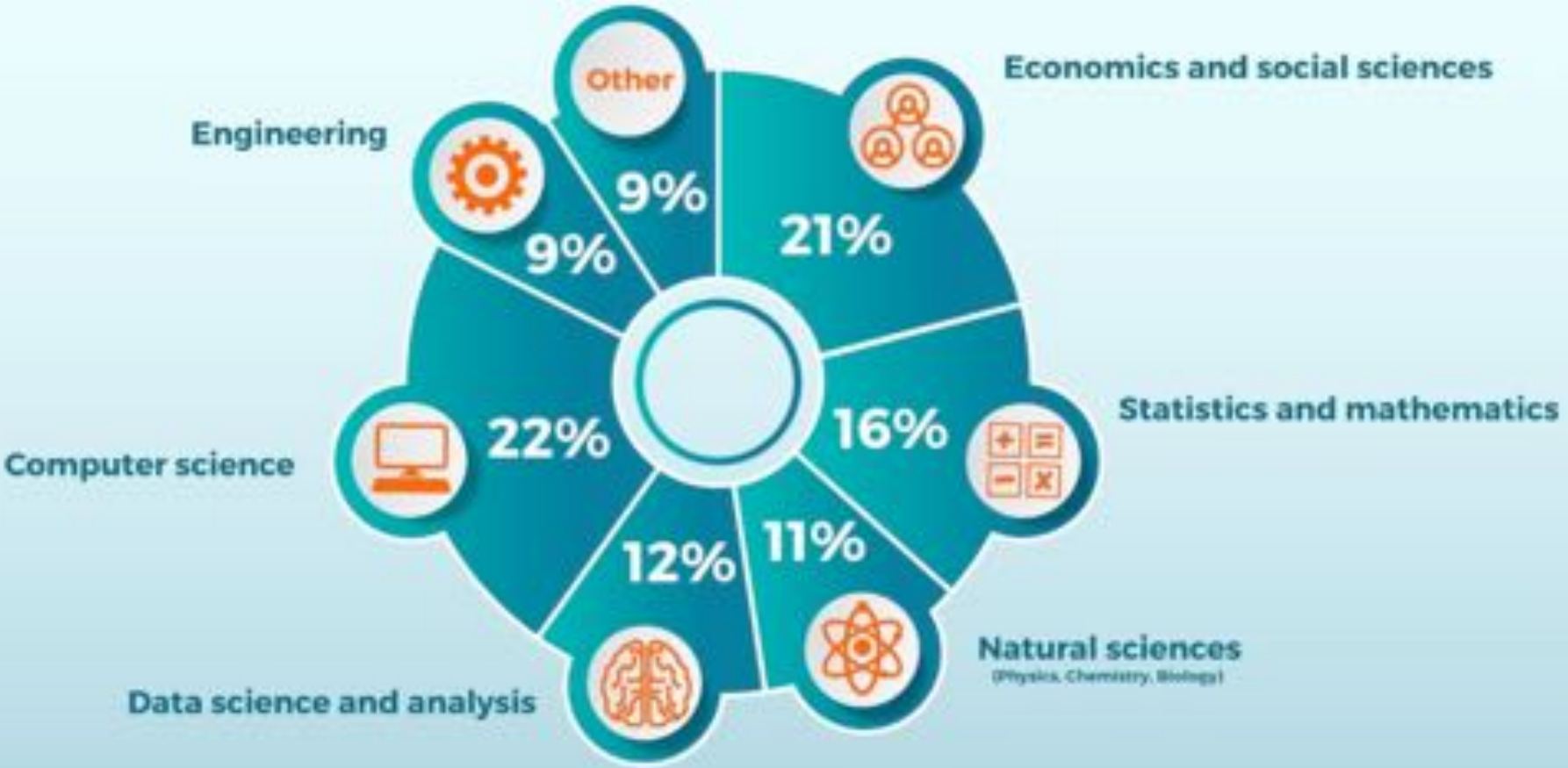
Lean Startup: innovación basada en el método científico

1. Plantear una **hipótesis** de partida
2. Escoger unos **indicadores adecuados** para comprobar la hipótesis
3. **Lanzar un producto mínimo viable** para validar la hipótesis
4. **Analizar los resultados** en función del feedback de los usuarios y los indicadores
5. **Replantear** nuevas hipótesis y volver a empezar

¿Qué será lo próximo?



Area of academic studies



Data Science en la Industria

1. **Identificar el problema**
   - Experiencia en el sector
2. **Datos**
   - Diseño experimental, recolección de los datos.
   - El 80% de un proyecto de Ciencia de Datos consiste en la recogida, limpieza y organización de los datos.
   - Calidad de los datos (“garbage in, garbage out”).
3. **Métodos**
   - Técnicas (clasificación supervisada/no-supervisada).
   - Predicción (series temporales)
   - Interpretables.
4. **Sugerencias**
   - Informe de diagnóstico complete.

Eskerrik asko

www.twitter.com/BCAMBilbao

www.linkedin.com/company/bcam-basque-center-for-applied-mathematics

www.youtube.com/user/BCAMchannel