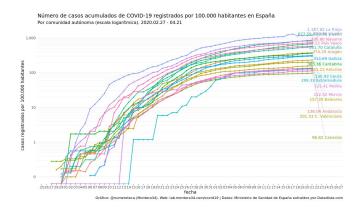
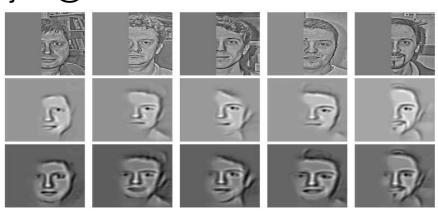


La ciencia del dato en la era de la COVID-19

Pedro Concejero pedro.concejero@u-tad.com



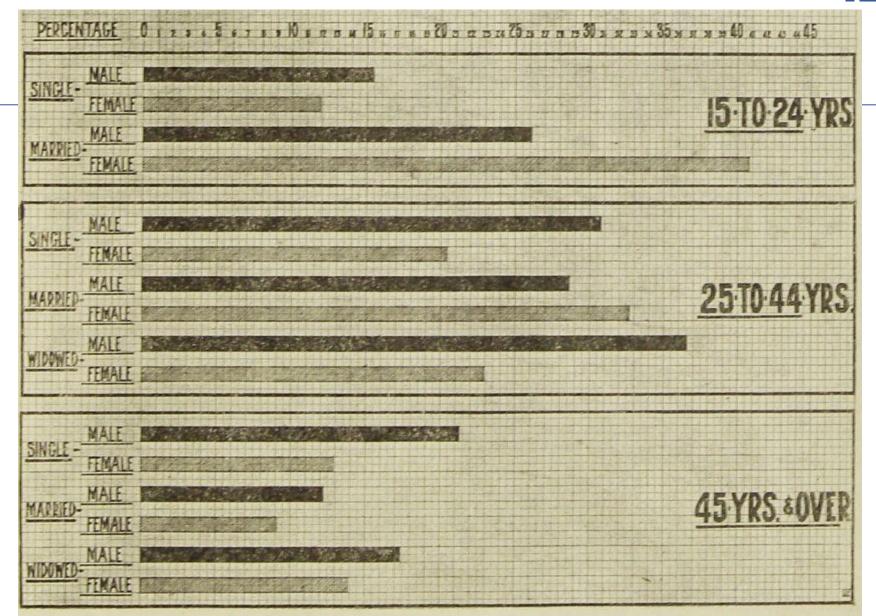




Introducción:

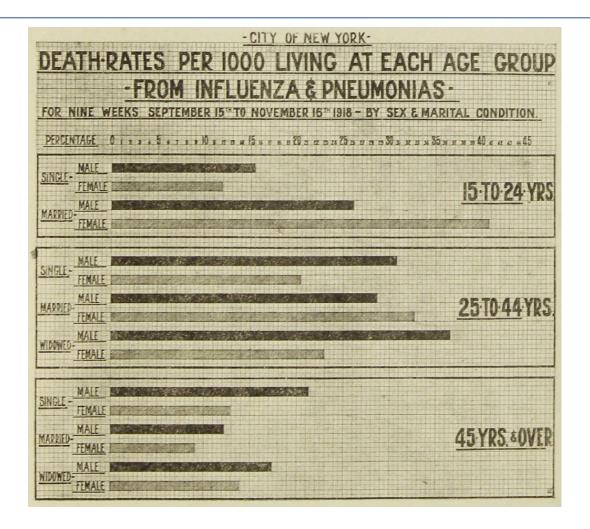
- ¿Estamos viviendo algo nuevo?
 - ¿De verdad no tenemos armas diferentes de las de nuestros antepasados?







https://es.wikipedia.org/wiki/Pandemia de gripe de 1918





Ahora: pendientes del (small?) data



Pendientes del dato

Justo y necesario alabar el magnífico trabajo de datadista (JJ Melero) para aglutinar datos COVID-19 en España: https://github.com/datadista/datasets/tree/master/COVID%2019



Extracción, limpieza y normalización de las tablas de la situación diaria acumulada de la enfermedad por el coronavirus SARS-CoV-2 (COVID-19) en España en un formato accesible y reutilizable. Datos actualizados diariamente según el calendario y ritmo de publicación del Ministerio de Sanidad.

• También tienes disponibles otros datasets accesorios para cubrir la crisis por COVID-19 en España

Fuente de los datos:

- Ministerio de Sanidad, Consumo y Bienestar Social: https://www.mscbs.gob.es/profesionales/saludPublica/ccayes /alertasActual/nCov-China/situacionActual.htm y gabinete de prensa del Ministerio de Sanidad.
- Departamento de Seguridad Nacional: https://www.dsn.gob.es/gl/current-affairs/press-room
- Instituto de Salud Carlos III. Situación de COVID-19 en España https://covid19.isciii.es/
- BOE: Crisis Sanitaria COVID-19 https://www.boe.es/biblioteca_juridica/codigos/codigo.php?id=355&modo=2¬a=0&tab=2
- Ministerio de Transportes, Movilidad y Agenda Urbana https://www.mitma.gob.es/transporte-terrestre/punto-de-informacion-de-servicios-de-restauracion

Los datos para los días 29 de febrero, 1 de marzo, 7 de marzo y 8 de marzo provienen de las notas de prensa del Departamento de Seguridad Nacional y del gabinete de prensa del Ministerio de Sanidad.

PDFs originales de resumen de situación: Carpeta con los PDFs originales de los informes oficiales publicados por el Ministerio de Sanidad.



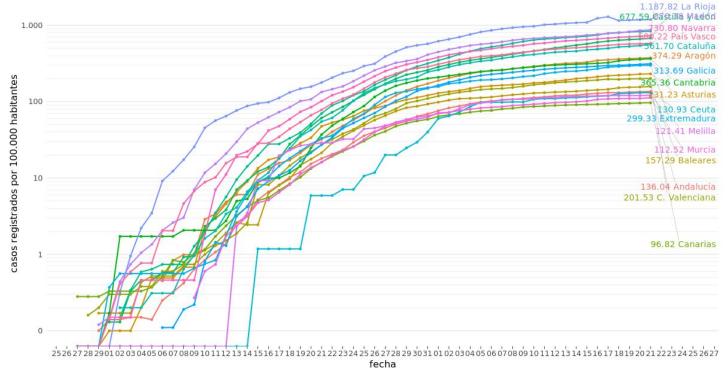
El papel clave de la comunicación y (estamos en s. XXI) de la visualización. ¿Lo estamos comunicando bien?



"Aplanar la curva"... ¿qué curva?

Infinitas gracias a montera34.com https://code.montera34.com/covid19/

Número de casos acumulados de COVID-19 registrados por 100.000 habitantes en España Por comunidad autónoma (escala logarítmica). 2020.02.27 - 04.21

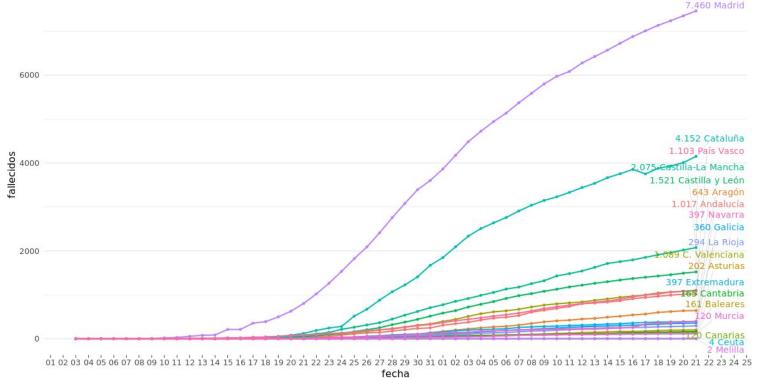




"Aplanar las curvas"... ¿qué curvas?

https://code.montera34.com:4443/numeroteca/covid19

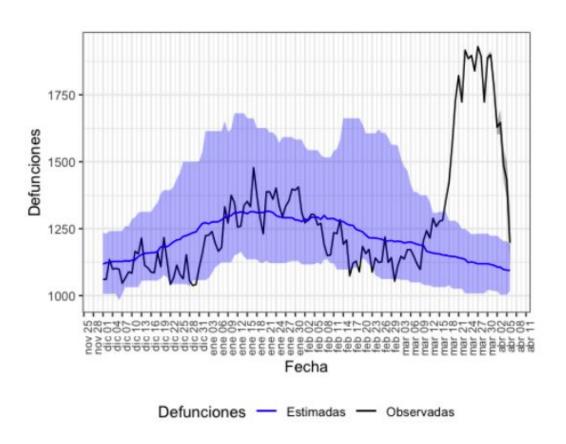
Número de fallecimientos acumulados por COVID-19 registrados en España Por comunidad autónoma (escala lineal). 2020.02.27 - 04.21





Los terribles datos: ¡necesario comunicar!

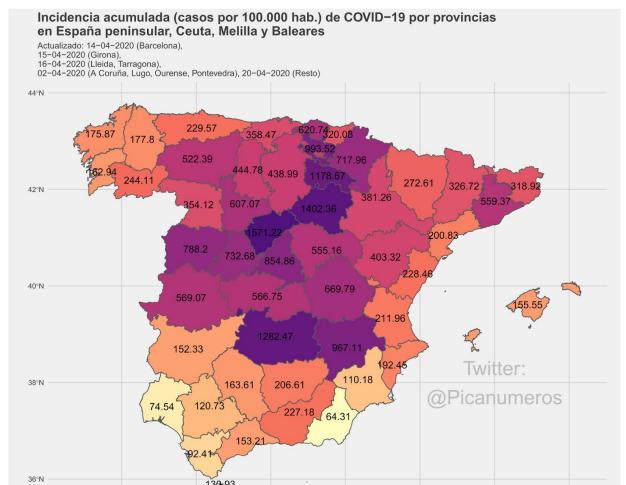
https://www.datanalytics.com/2020/04/08/momo-una-documentacion-oficiosa/





...y comunicar bien, y gestionar mejor!

https://picanumeros.wordpress.com/





...y comunicar bien, y gestionar mejor!

https://elpais.com/sociedad/2020-04-04/el-virus-se-ensana-con-la-espana-vacia.html

https://analisis.datosabiertos.jcyl.es/pages/coronavirus/situacin-actual#sit uacin-actual



Ahora: unos pocos conceptos esenciales necesarios (sí, de epidemiología)



Conceptos básicos a partir de ahora

(Tasa de) Prevalencia

https://es.wikipedia.org/wiki/Prevalencia

(Tasa de) Incidencia

https://es.wikipedia.org/wiki/Incidencia

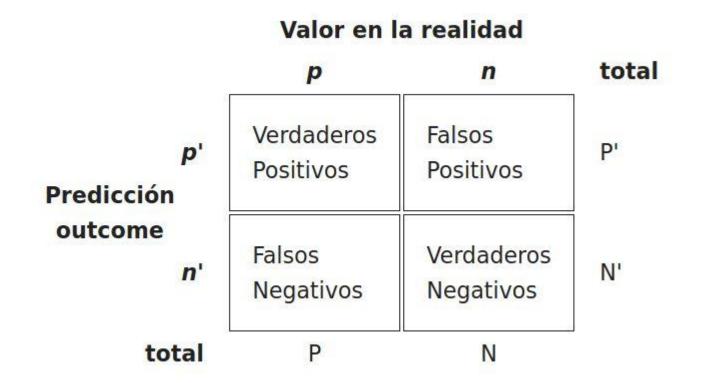
Detección rápida (screening)

- https://es.wikipedia.org/wiki/Cribado
 - https://en.wikipedia.org/wiki/Screening_(medicine)



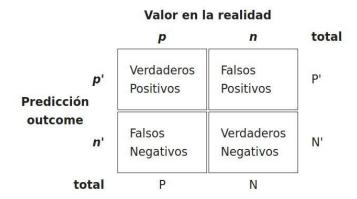
Conceptos básicos a partir de ahora

https://es.wikipedia.org/wiki/Curva ROC





Conceptos básicos a partir de ahora



Sensibilidad o Razón de Verdaderos Positivos (VPR)

o también razón de éxitos y, recuerdo en recuperación de información,

$$VPR = VP/P = VP/(VP + FN)$$

Especificidad (SPC) o Razón de Verdaderos Negativos

$$SPC = VN/N = VN/(FP + VN) = 1 - FPR$$



Pero... ¿de verdad comprendemos bien los datos?



Encuesta 1:

Cuál fue la primera causa de muerte *no natural* (que no sea por enfermedad, de cualquier tipo) en España en 2018?

- a. Suicidio
- b. Accidente de tráfico
- c. Ahogamiento



Los humanos no somos tan "óptimos" con los números

https://www.epdata.es/datos/mueren-espanoles-causas-muerte-datos-estadisticas/241/espana/106

https://en.wikipedia.org/wiki/Cognitive_bias

https://en.wikipedia.org/wiki/Hindsight_bias

https://en.wikipedia.org/wiki/Availability_heuristic

https://en.wikipedia.org/wiki/Base_rate_fallacy



Encuesta 2:

La policía ha comprado un alcoholímetro que detecta la embriaguez superior a los límites legales con un **porcentaje de FALSOS POSITIVOS del 5%**, aunque con la gran ventaja de que nunca falla en detectar una persona verdaderamente ebria.

Se sabe que **uno de cada mil conductores** está conduciendo de verdad ebrio.

Suponga que los policías detienen a un conductor al azar, y le hacen la prueba del alcoholímetro y da positiva.

Asumiendo que no sabemos nada más del conductor, ¿cuál es la probabilidad de que esté realmente ebrio?

- a. 0.95
- b. >0.8
- c. <0.05



Manejar los tests -de lo que sea- no es tan sencillo

- Requieren estudios que pueden ser muy complicados
- Especialmente si -como suele suceder- desconocemos la prevalencia (tasa base en la población).
- No es solo sensibilidad-especificidad:

Valor Predictivo Positivo (PPV)

o también "precisión" en recuperación de información

$$PPV = VP/(VP + FP)$$

$$PPV = P * VPR/P * VPR + (1 - P) * (1 - SPC)$$

Valor Predictivo Negativo (NPV)

$$NPV = VN/(VN + FN)$$

$$NPV = (1 - P)SPC/(1 - P)SPC + P(1 - VPR)$$



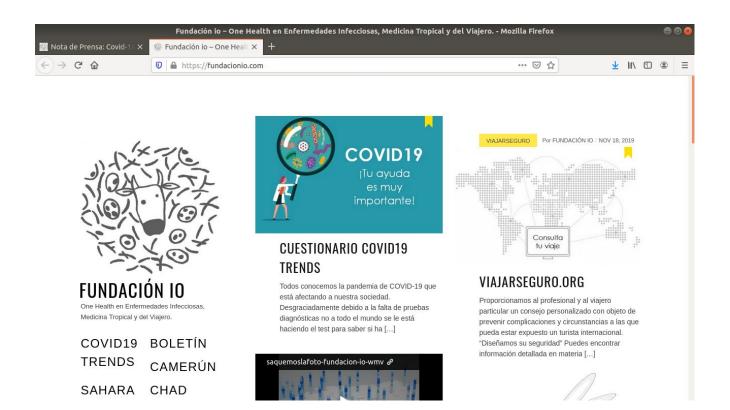
Detección rápida en la COVID-19



Fundación IO

https://fundacionio.com/

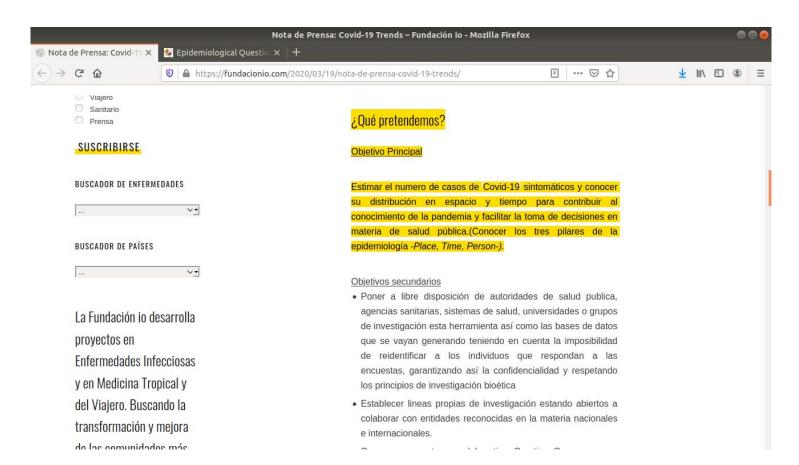
https://covid19.fundacionio.com/epidemiologicalquestionnairespain.aspx?Spanish





Fundación IO

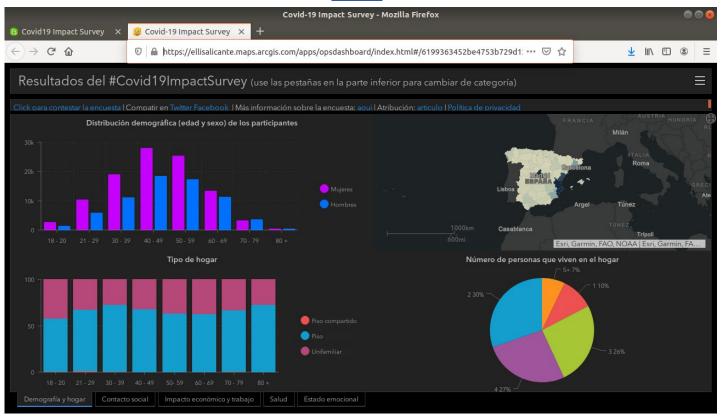
https://fundacionio.com/2020/03/19/nota-de-prensa-covid-19-trends/





Cuestionario impacto coronavirus

https://survey123.arcgis.com/share/d29378b51fe8496d8dd77f08ce73973f
https://ellisalicante.maps.arcgis.com/apps/opsdashboard/index.html#/6199363452be4753b729d122beb
a40bc





Pero ¿de qué hablábamos antes de la COVID-19?



Encuesta 3:

En qué año se considera que comienza la historia de redes reuronales artificiales?

- **a.** 1958
- **b.** 1943
- **C.** 1969



Some history

https://www.import.io/post/history-of-deep-learning/

1943 – The first mathematical model of a neural network

Walter Pitts and Warren McCulloch

Obviously, for machine and deep learning to work, we needed an established understanding of the neural networks of the human brain.

Walter Pitts, a logician, and Warren McCulloch, a neuroscientist, gave us that piece of the puzzle in 1943 when they created the first mathematical model of a neural network. Published in their seminal work "A Logical Calculus of Ideas Immanent in Nervous Activity", they proposed a combination of mathematics and algorithms that aimed to mimic human thought processes.

Their model – typically called McCulloch-Pitts neurons – is still the standard today (although it has evolved over the years).



Some history

https://www.import.io/post/history-of-deep-learning/

Alan Turing

Turing, a British mathematician, is perhaps most well-known for his involvement in code-breaking during World War II.

But his contributions to mathematics and science don't stop there. In 1947, he predicted the development of machine learning, even going so far as to describe the impact it could have on jobs.



In 1950, Turing proposed just such a machine, even hinting at genetic algorithms, in his paper "Computing Machinery and Intelligence." In it, he crafted what has been dubbed The Turing Test – although he himself called it The Imitation Game – to determine whether a computer can "think."

At its simplest, the test requires a machine to carry on a conversation via text with a human being. If after five minutes the human is convinced that they're talking to another human, the machine is said to have passed.

It would take 60 years for any machine to do so, although many still debate the validity of the results.



Encuesta 4:

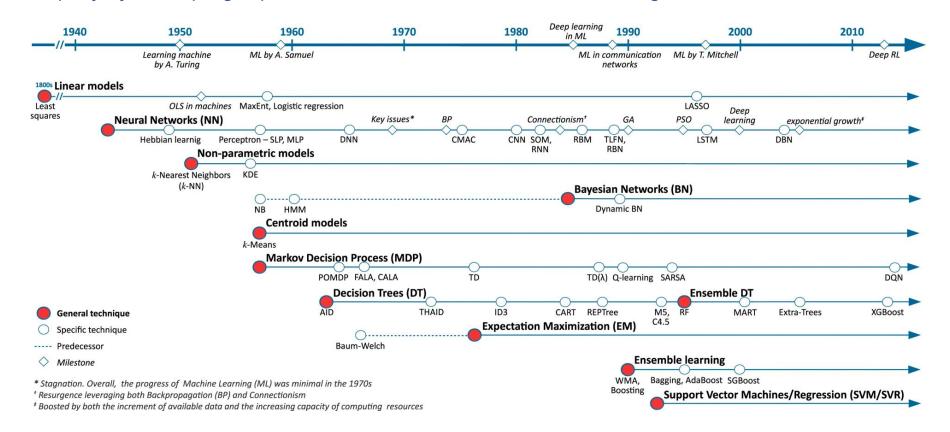
Cuál es la tasa de error del mejor modelo actual de reconocimiento de letras manuscritas (MNIST)?

- a. 7.6%
- b. 0.7%
- c. 0.17%



"Historia" del Machine Learning

https://jisajournal.springeropen.com/articles/10.1186/s13174-018-0087-2/figures/5

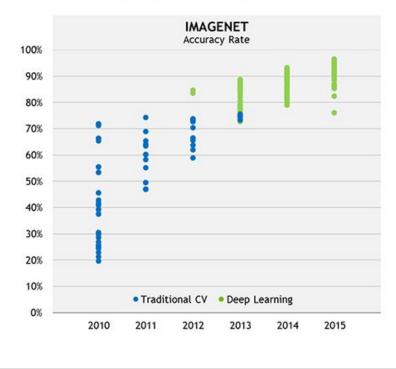




Deep Learning - salto cualitativo

https://blogs.nvidia.com/blog/2016/01/12/accelerating-ai-artificial-intelligence-gpus/

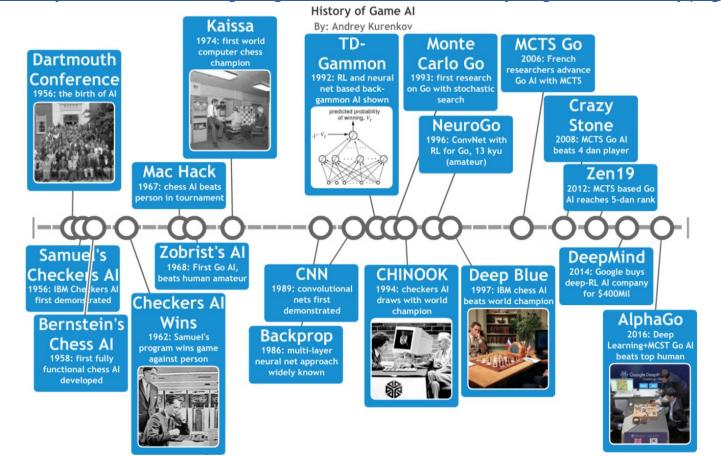






Deep Learning super-human performance in games

https://www.andreykurenkov.com/writing/images/2016-4-15-a-brief-history-of-game-ai/0-history.png





Deep Learning

Historia del deep learning 2012-2019:

https://towardsdatascience.com/milestones-of-deep-learning-1aaa9aef5b18

2019:

https://towardsdatascience.com/14-deep-learning-uses-that-blasted-me-away-2019-206a5271d98

¿Qué es lo que ha hecho de DL la tecnología más demandada en actualidad?

opensource software: libraries - TensorFlow, PyTorch...

widely available hardware (GPU's)

and most most importantly: shared training datasets and standardized evaluation methodologies:

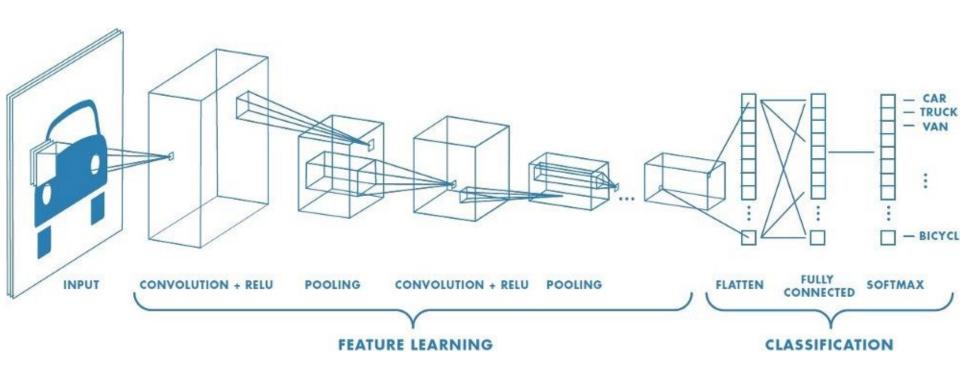
https://en.wikipedia.org/wiki/ImageNet

https://en.wikipedia.org/wiki/MNIST_database

https://github.com/Matuzas77/MNIST-0.17/blob/master/MNIST_final_solution.ipynb



Deep Learning





Deep Learning

http://web.eecs.umich.edu/~honglak/icml09-ConvolutionalDeepBeliefNetworks.pdf

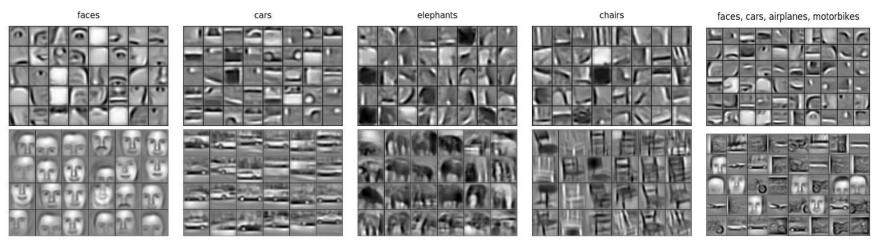
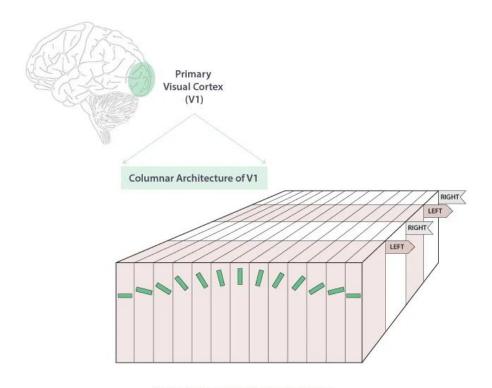


Figure 3. Columns 1-4: the second layer bases (top) and the third layer bases (bottom) learned from specific object categories. Column 5: the second layer bases (top) and the third layer bases (bottom) learned from a mixture of four object categories (faces, cars, airplanes, motorbikes).



Deep Learning y neurociencia: Hubel y Wiesel

https://knowingneurons.com/2014/10/29/hubel-and-wiesel-the-neural-basis-of-visual-perception/





Deep Learning

http://web.eecs.umich.edu/~honglak/icml09-ConvolutionalDeepBeliefNetworks.pdf



Figure 6. Hierarchical probabilistic inference. For each column: (top) input image. (middle) reconstruction from the second layer units after single bottom-up pass, by projecting the second layer activations into the image space. (bottom) reconstruction from the second layer units after 20 iterations of block Gibbs sampling.



Entonces, ¿y ahora qué?



DL everywhere - and privacy?

https://www.nytimes.com/interactive/2019/04/16/opinion/facial-recognition-new-york-city.html



To demonstrate how easy it is to track people without their knowledge, we collected public images of people who worked near Bryant Park (available on their employers' websites, for the most part) and ran one day of footage through Amazon's commercial facial recognition service. Our system detected 2,750 faces from a nine-hour period (not necessarily unique people, since a person could be captured in



Enriquecimiento, sinergias ciencia de datos y sectores

SNA, fake news, trolls...

The COVID-19 Social Media Infodemic: https://arxiv.org/abs/2003.05004
https://www.bloomberg.com/opinion/articles/2020-03-31/coronavirus-and-oth-er-falsehoods-are-believable-research-shows
https://behavioralscientist.org/loose-lips-the-psychology-of-rumor-during-crisi-s-coronavirus/

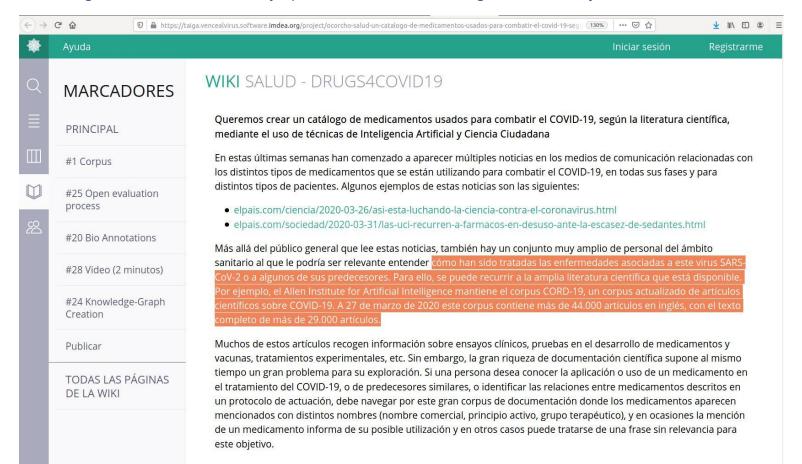
COVID -19 Social Science Research Tracker https://github.com/natematias/covid-19-social-science-research

Gestión de movimientos de población



ML para búsqueda eficaz de mejores tratamientos

https://taiga.vencealvirus.software.imdea.org/project/ocorcho-salud-un-catalogo-de-medicamentos-usados-para-combatir-el-covid-19-segun-la-literatura-cientifica-y-aplicando-tecnicas-de-inteligencia-artificial-y-ciencia-ciudadana/wiki/home





https://www.thelancet.com/pdfs/journals/landig/PIIS2589-7500(19)30124-4.pdf

Human versus machine in medicine: answer the question?

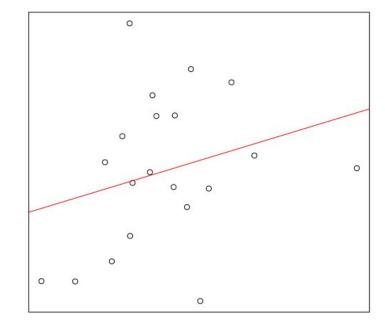
models. From more than 20 000 unique abstracts, fewer than 100 studies met their eligibility criteria for the systematic review and only 25 met their inclusion criteria for the meta-analysis. These 25 studies compared the performance of deep learning solutions to health-care professionals for 13 different specialty areas, only two of which-breast cancer and dermatological cancerswere represented by more than three studies. The meta-analysis suggests equivalent performance of deep learning algorithms and health-care professionals in the 14 studies that used the same out-of-sample validation dataset to compare their performances, showing a pooled sensitivity of 87.0% (95% CI 83.0-90.2) for deep learning models and 86.4% (79.9-91.0) for health-care professionals, and a pooled specificity of 92.5% (85.1–96.4) for deep learning models and 90.5%(80.6–95.7) for health-care professionals. This work nicely illustrates the challenge of attempting to compare Al with humans for medical applications, and the authors rightly qualify their conclusion with a detailed list of potential confounders and limitations. The eventual sample size representing a broad swath of the domain of medicine underlines the need for a deeper dive into the literature.2



Pero también ... ay

https://www.datanalytics.com/2020/04/15/cosas-que-ocurriran-sin-lugar-a-dudas-tras-el-coronavirus/

Mi 'cherry-picking' por países favorito



algo sobre el coronavirus



Unas palabras finales...

En este periodo debemos aprender

¿No son aplicables los mismos conceptos de epidemiología a entornos comerciales?

Algunos de ellos ya lo son (medidas de rendimiento predictivo)

¿Y las técnicas de screening?



Unas palabras finales...

¿Te gusta la ciencia de datos? ¿Conoces R? Únete: http://r-es.org/





Unas palabras finales...

grupo R madRid:

https://www.meetup.com/es-ES/Grupo-de-Usuarios-de-R-de-Madrid/





¡Gracias!

Pedro Concejero

pedro.concejero@u-tad.com

pedro.concejerocerezo@gmail.com

twitter: https://twitter.com/concejeropedro



Referencias adicionales

Curso de epidemiología:

https://online.stat.psu.edu/stat507/01/intro/

Screening studies:

https://online.stat.psu.edu/stat507/node/17/

US NIH (National Institudes of Health) on COVID-19:

https://www.nih.gov/health-information/coronavirus

Covid19 evolution simulator using SIR model (España, J. M. Benlloch):

http://covid19.jmbenlloch.net/sir

El coronavirus y la leyenda del tablero de ajedrez:

https://www.jotdown.es/2020/03/el-coronavirus-y-la-leyenda-del-tablero-de-ajedrez/

The Origin and Virulence of the 1918 "Spanish" Influenza Virus1:

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2720273/

El gran problema de las pruebas rápidas COVID-19

https://www.eldiario.es/sociedad/internacional-rapidos-COVID-19-absoluta-mayoria 0 101769923

2.html

Estudio seroprevalencia COVID-19

https://www.isciii.es/Noticias/Noticias/Paginas/Noticias/EstudioSeroprevalencia.aspx

Actividades del mundo R COVID-19

https://rviews.rstudio.com/2020/04/07/some-select-covid-19-modeling-resources/



Referencias adicionales

Epidemic calculator:

http://gabgoh.github.io/COVID/index.html

Why outbreaks like coronavirus spread exponentially, and how to "flatten the curve":

https://www.washingtonpost.com/graphics/2020/world/corona-simulator/

Modeling pandemics (Arthur Charpentier):

https://freakonometrics.hypotheses.org/60482

Convolutional neural networks

https://ml4a.github.io/ml4a/convnets/

Progress in Artificial Intelligence (inc. performance human-machine)

https://en.wikipedia.org/wiki/Progress in artificial intelligence

MNIST classifier with average 0.17% error (with TensorFlow code):

https://github.com/Matuzas77/MNIST-0.17

On-the-fly recognition of handwritten characters - 3D visualization of CNN

https://www.cs.ryerson.ca/~aharley/vis/conv/