UNIVERSIDAD DE GRANADA

DECSAI

Sistemas Inteligentes para la Gestión de la Empresa

Tema 1. Introducción a la Ciencia de Datos

E.T.S. de Ingenierías Informática y de Telecomunicación

Universidad de Granada

Juan Gómez Romero
jgomez@decsai.ugr.es
Departamento de Ciencias de la Computación e Inteligencia Artificial
http://decsai.ugr.es

nature Vol 457 | 19 February 2009 | doi:10.1038/nature07634

LETTERS

Detecting influenza epidemics using search engine query data

Jeremy Ginsberg[1], Matthew H. Mohebbi[1], Rajan S. Patel[1], Lynnette Brammer[2], Mark S. Smolinski[1] & Larry Brilliant[1]

Seasonal influenza epidemics are a major public health concern, causing tens of millions of respiratory illnesses and 250,000 to 500,000 deaths worldwide each year[1]. In addition to seasonal influenza, a new strain of influenza virus against which no previous immunity exists and that demonstrates human-to-human transmission could result in a pandemic with millions of fatalities[2]. Early detection of disease activity, when followed by a rapid response, can reduce the impact of both seasonal and pandemic influenza[3,4]. One way to improve early detection is to monitor health-seeking behaviour in the form of queries to online search engines, which are submitted by millions of users around the world each day. Here we present a method of analysing large numbers of Google search queries to track influenza-like illness in a population. Because the relative frequency of certain queries is highly correlated with the percentage of physician visits in which a patient presents with influenza-like symptoms, we can accurately estimate the current level of weekly influenza activity in each region of the United States, with a reporting lag of about one day. This approach may make it possible to use search queries to detect influenza epidemics in areas with a large population of web search users.

Traditional surveillance systems, including those used by the US Centers for Disease Control and Prevention (CDC) and the European Influenza Surveillance Scheme (EISS), rely on both virological and clinical data, including influenza-like illness (ILI) physician visits. The CDC publishes national and regional data from these surveillance systems on a weekly basis, typically with a 1–2-week reporting lag.

In an attempt to provide faster detection, innovative surveillance systems have been created to monitor indirect signals of influenza activity, such as call volume to telephone triage advice lines[5] and over-the-counter drug sales[6]. About 90 million American adults are believed to search online for information about specific diseases or medical problems each year[7], making web search queries a uniquely valuable source of information about health trends. Previous attempts at using online activity for influenza surveillance have counted search queries submitted to a Swedish medical website (A. Hulth, G. Rydevik and A. Linde, manuscript in preparation), visitors to certain pages on a US health website[8], and user clicks on a search keyword advertisement in Canada[9]. A set of Yahoo search queries containing the words 'flu' or 'influenza' were found to correlate with virological and mortality surveillance data over multiple years[10].

Our proposed system builds on this earlier work by using an automated method of discovering influenza-related search queries. By processing hundreds of billions of individual searches from 5 years of Google web search logs, our system generates more comprehensive models for use in influenza surveillance, with regional and state-level estimates of ILI activity in the United States. Widespread global usage of online search engines may eventually enable models to be developed in international settings.

By aggregating historical logs of online web search queries submitted between 2003 and 2008, we computed a time series of weekly counts for 50 million of the most common search queries in the United States. Separate aggregate weekly counts were kept for every query in each state. No information about the identity of any user was retained. Each time series was normalized by dividing the count for each query in a particular week by the total number of online search queries submitted in that location during the week, resulting in a query fraction (Supplementary Fig. 1).

We sought to develop a simple model that estimates the probability that a random physician visit in a particular region is related to an ILI; this is equivalent to the percentage of ILI-related physician visits. A single explanatory variable was used: the probability that a random search query submitted from the same region is ILI-related, as determined by an automated method described below. We fit a linear model using the log-odds of an ILI physician visit and the log-odds of an ILI-related search query: $\text{logit}(I(t)) = \alpha \text{logit}(Q(t)) + \varepsilon$, where $I(t)$ is the percentage of ILI physician visits, $Q(t)$ is the ILI-related query fraction at time t, α is the multiplicative coefficient, and ε is the error term. $\text{logit}(p)$ is simply $\ln(p/(1-p))$.

Publicly available historical data from the CDC's US Influenza Sentinel Provider Surveillance Network (http://www.cdc.gov/flu/weekly) was used to help build our models. For each of the nine surveillance regions of the United States, the CDC reported the average percentage of all outpatient visits to sentinel providers that were ILI-related on a weekly basis. No data were provided for weeks outside of the annual influenza season, and we excluded such dates from model fitting, although our model was used to generate unvalidated ILI estimates for these weeks.

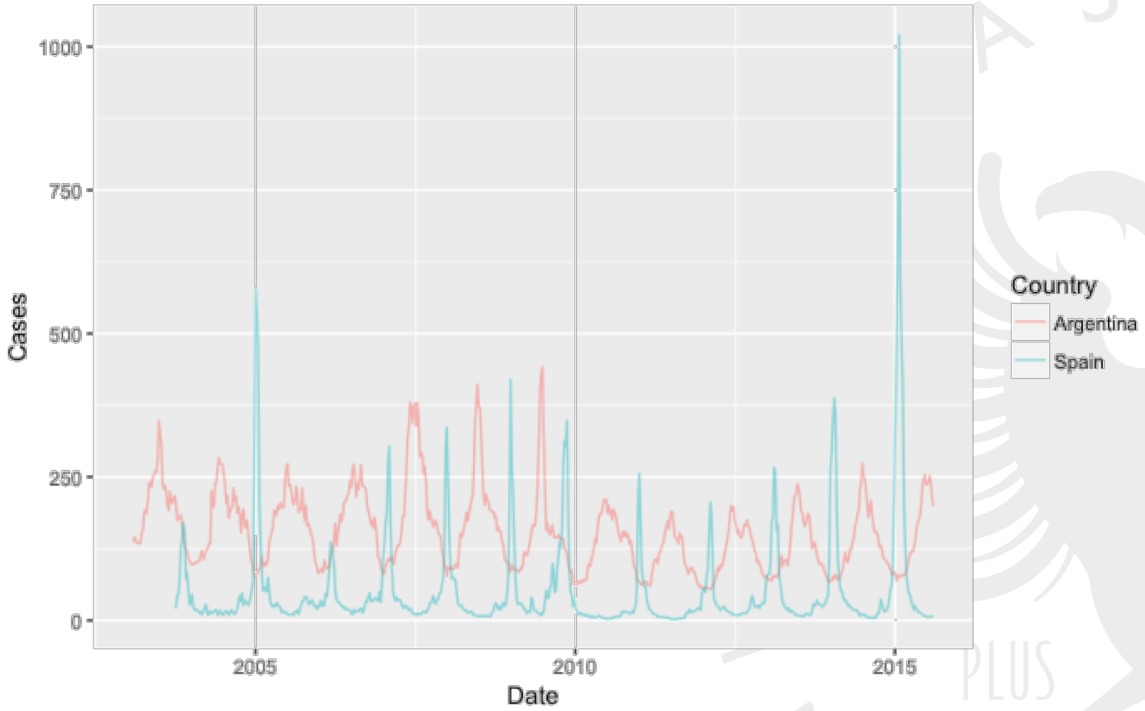
We designed an automated method of selecting ILI-related search queries, requiring no previous knowledge about influenza. We measured how effectively our model would fit the CDC ILI data in each region if we used only a single query as the explanatory variable, $Q(t)$. Each of the 50 million candidate queries in our database was separately tested in this manner, to identify the search queries which could most accurately model the CDC ILI visit percentage in each region. Our approach rewarded queries that showed regional variations similar to the regional variations in CDC ILI data: the chance that a random search query can fit the ILI percentage in all nine regions is considerably less than the chance that a random search query can fit a single location (Supplementary Fig. 2).

The automated query selection process produced a list of the highest scoring search queries, sorted by mean Z-transformed correlation across the nine regions. To decide which queries would be included in the ILI-related query fraction, $Q(t)$, we considered different sets of n top-scoring queries. We measured the performance of these models based on the sum of the queries in each set, and picked n such that we obtained the best fit against out-of-sample ILI data across the nine regions (Fig. 1).

[1]Google Inc., 1600 Amphitheatre Parkway, Mountain View, California 94043, USA. [2]Centers for Disease Control and Prevention, 1600 Clifton Road, NE, Atlanta, Georgia 30333, USA.

J. Ginsberg, M.H. Mohebbi, R.S. Patel, L. Brammer, M.S. Smolinski, L. Brilliant (2009) Detecting influenza epidemics using search engine query data. *Nature* 457, pp. 1012–1015.

```r
library(tidyr)
plotdata <-
  data %>%
  gather(key = Country, value = Cases, Argentina:Uruguay, na.rm = TRUE) %>%
  filter(Country == 'Spain' | Country == "Argentina")
library(ggplot2)
ggplot(plotdata, aes(x = Date, y = Cases, color = Country)) +
  geom_line(alpha = 0.5)
```



https://www.google.org/flutrends/about/

BUSINESS INSIDER TECH FINANCE POLITICS STRATEGY L

The price of bitcoin has a 91% correlation with Google searches for bitcoin

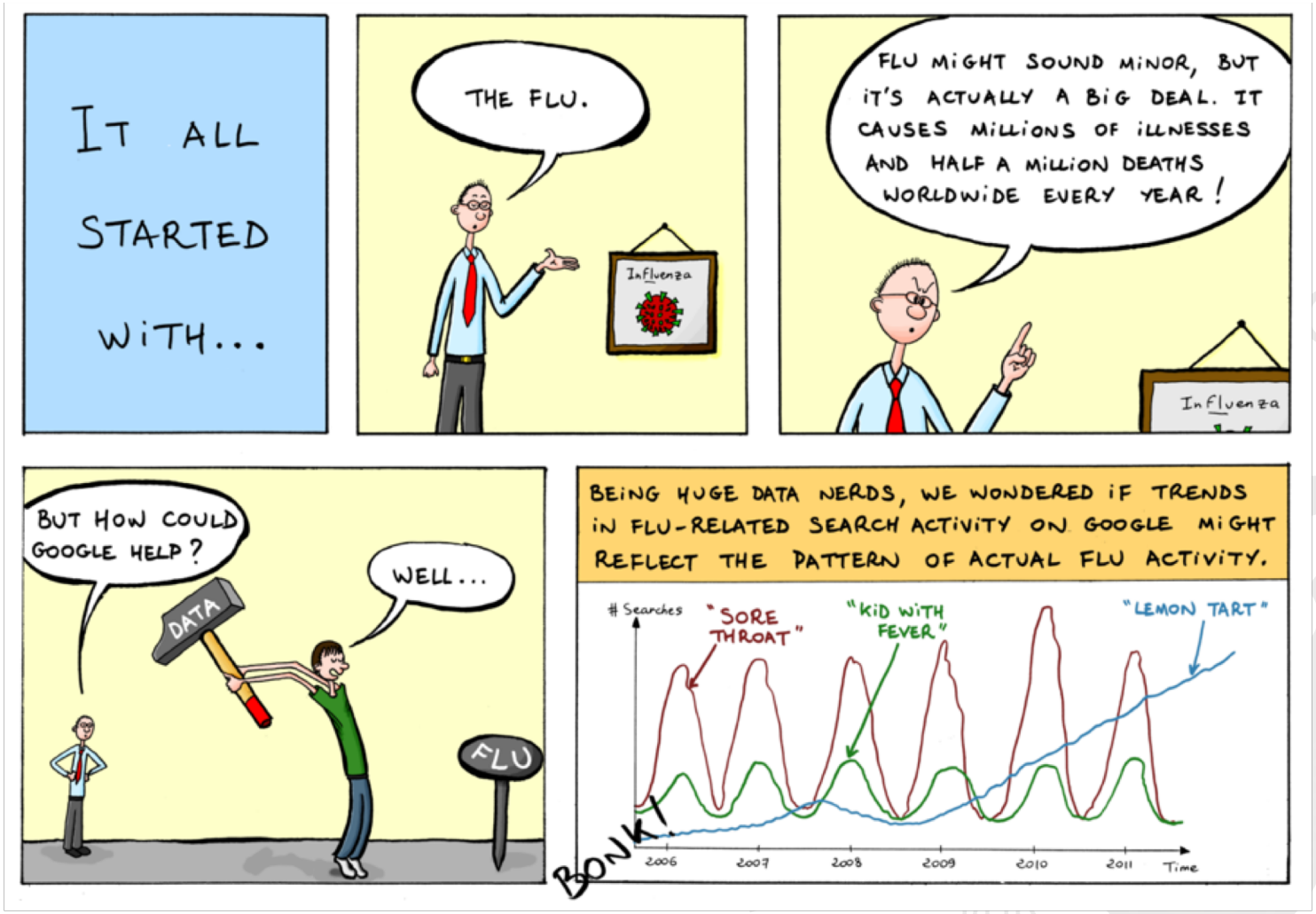
Jim Edwards
Sep. 19, 2017, 5:08 AM 42,207

FACEBOOK LINKEDIN TWITTER

The current price of bitcoin has a 91% correlation with the volume of Google search requests for bitcoin-related terms, according to a study by SEMrush, a search engine marketing agency.

The study drew from a database of 120 million US keyword searches linked to the cryptocurrency. The overall search volume of bitcoin-related keywords is estimated to be 51.4 million requests over a period of a year. It showed that the price of bitcoin in US dollars rose and fell largely in tandem with the number of search requests for terms like "bitcoin," "bitcoin price," and "bitcoin value."

At one level, the study merely confirms the obvious: As bitcoin becomes more expensive, and thus more exciting, more people search online to find out how it is doing. On the other hand, it is nice to see statistics confirm your hunches. While the study looked at the correlation between searches and prices, it did not — sadly — say whether searches *predicted* or *trailed* the bitcoin/dollar exchange rate.



https://www.google.com/trends/correlate/comic?p=1

=> Creación de modelos de predicción

Modelo formal que permite obtener la salida de un sistema ante unos estímulos previamente no conocidos

=> Soporte a la toma de decisiones

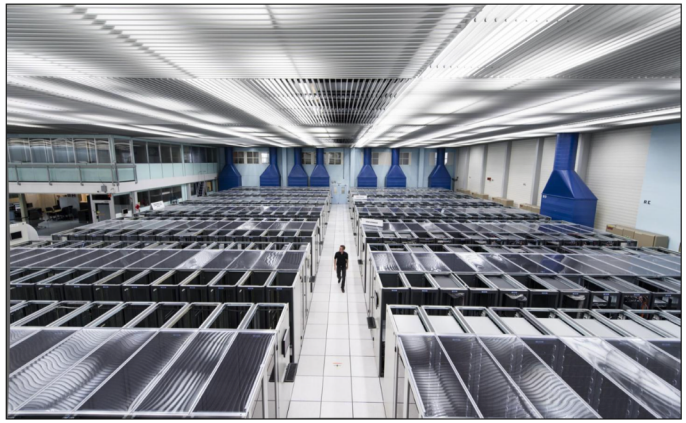
Plantear escenarios alternativos y valorar los resultados esperados

Big Data

Grandes conjuntos de datos permiten abordar problemas minimizando o incluso eliminando la necesidad de realizar muestreos aleatorios

Computing

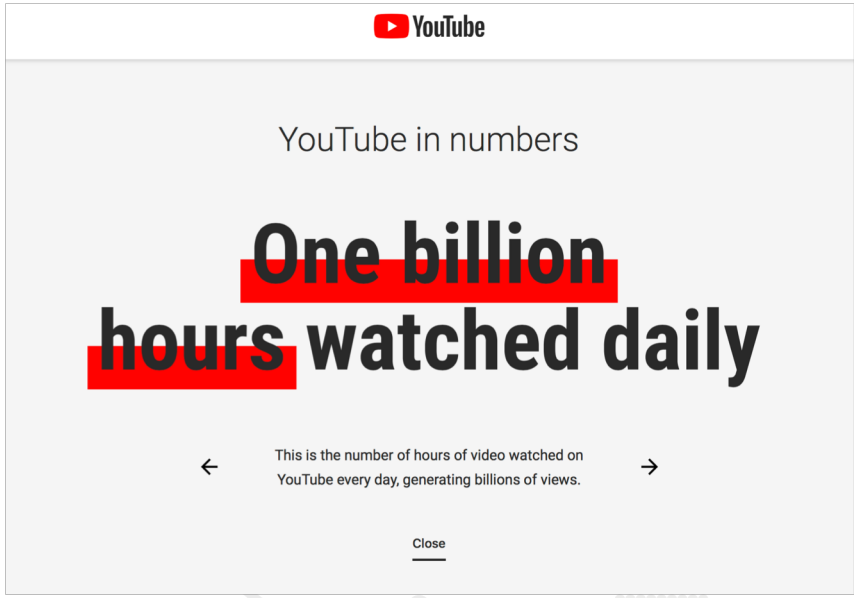
Experiments at CERN generate colossal amounts of data. The Data Centre stores it, and sends it around the world for analysis

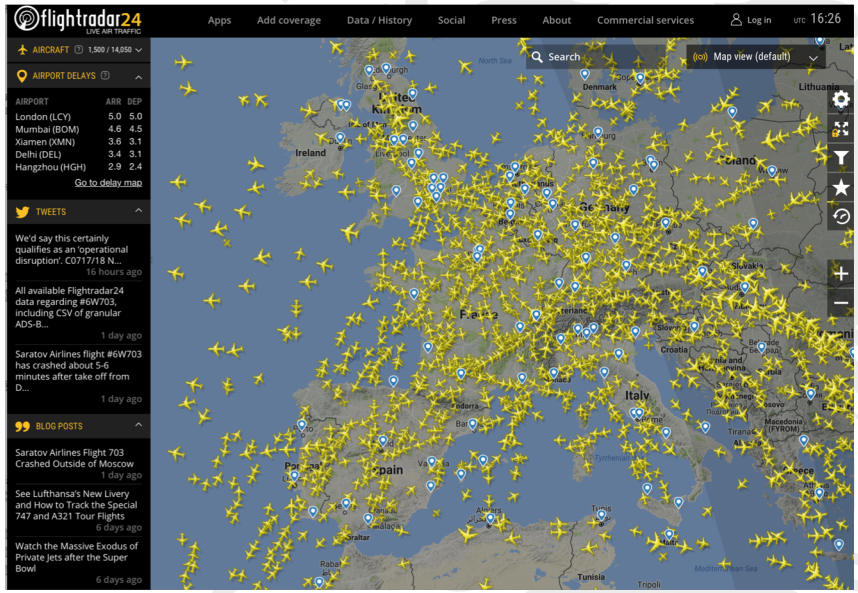


Approximately 600 million times per second, particles collide within the Large Hadron Collider (LHC). Each collision generates particles that often decay in complex ways into even more particles. Electronic circuits record the passage of each particle through a detector as a series of electronic signals, and send the data to the CERN Data Centre (DC) for digital reconstruction. The digitized summary is recorded as a "collision event". Physicists must sift through the 30 petabytes or so of data produced annually to determine if the collisions have thrown up any interesting physics.

CERN does not have the computing or financial resources to crunch all of the data on site, so in 2002 it turned to grid computing to share the burden with computer centres around the world. The Worldwide LHC Computing Grid (WLCG) – a distributed computing infrastructure arranged in tiers – gives a community of over 8000 physicists near real-time access to LHC data. The Grid builds on the technology of the World Wide Web, which was invented at CERN in 1989.

https://home.cern/about/computing





¿Quién genera datos?

Ciencia

Astronomía, genómica, física experimental, medioambiente

Ciencias Sociales y Humanidades

Libros escaneados, documentos históricos, datos sociales

Negocio y Comercio

Transacciones de mercados, logística, tráfico (aéreo, marítimo, carretera)

Entretenimiento y Ocio

Recursos multimedia en internet

Medicina

Bases de datos de pacientes, imágenes médicas

Industria

Sensores (redes eléctricas inteligentes, control de edificios, etc.)

Explosión de información

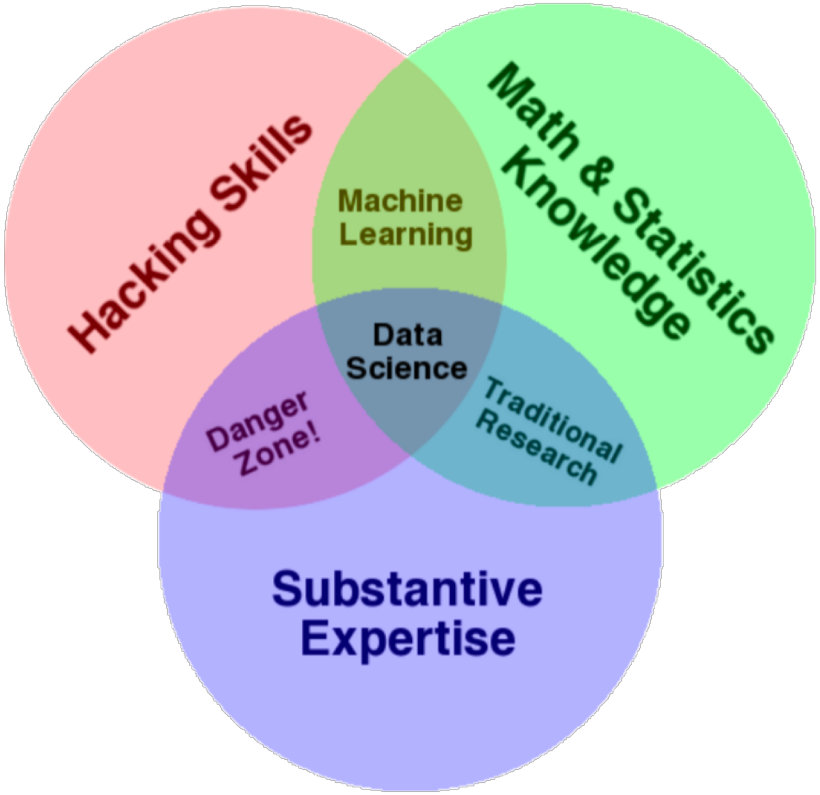
Nuevas herramientas para manejar datos

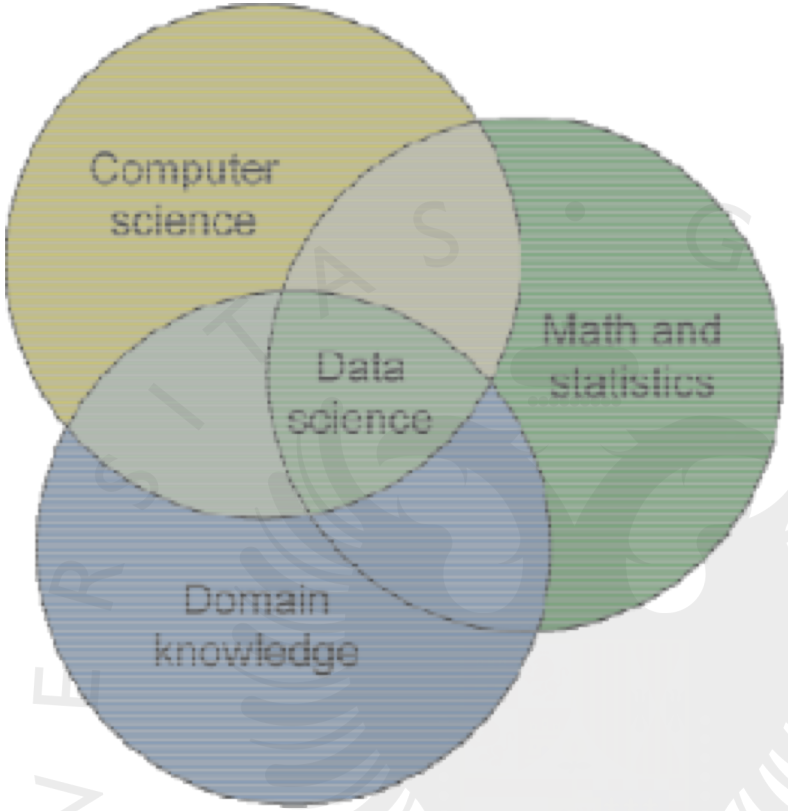
- Bajo precio de hardware
- Tecnologías de bases de datos
- Software para procesamiento distribuido

Podemos manejar grandes volúmenes de datos, pero resulta difícil extraer conocimiento

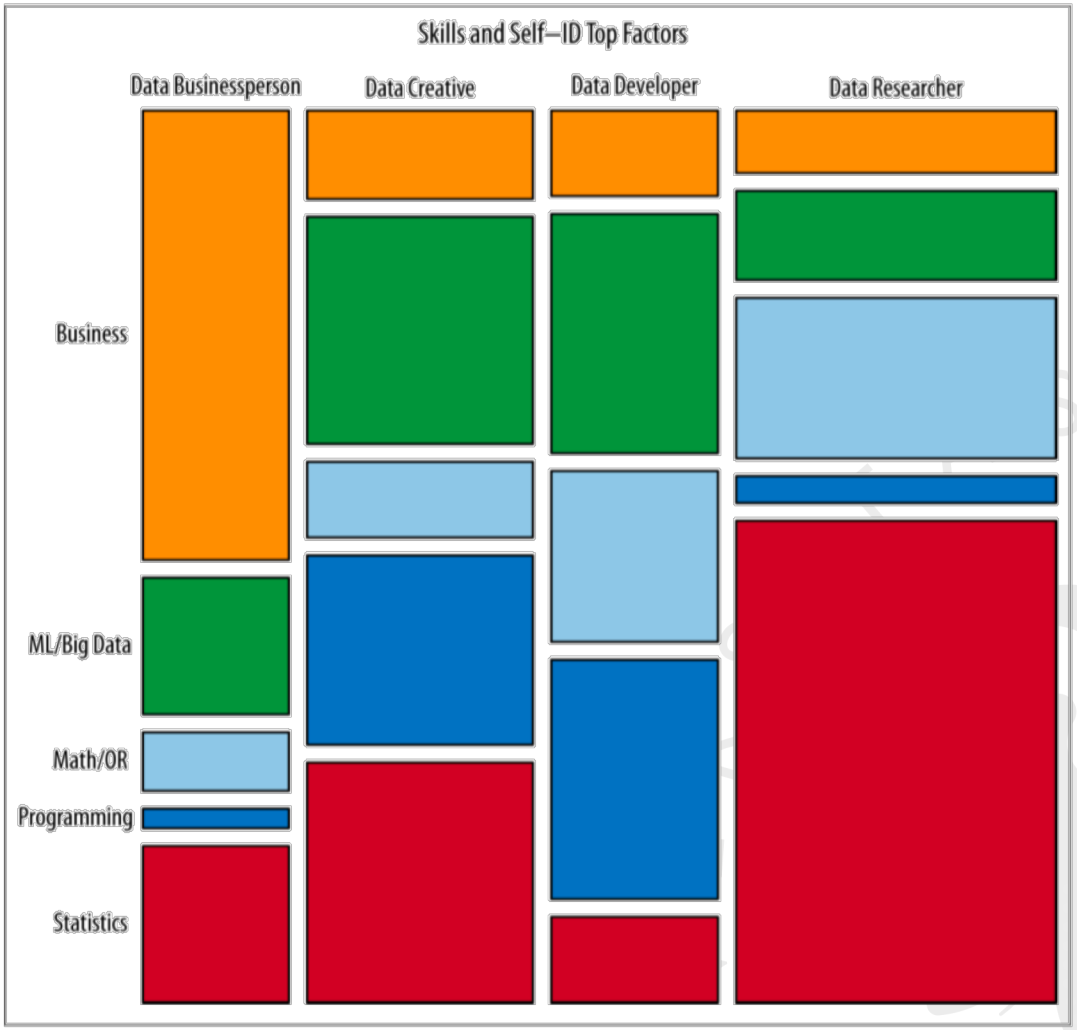
El progreso y la innovación ya no se ven obstaculizados por la capacidad de recopilar datos, sino por la capacidad de gestionar, analizar, sintetizar, visualizar, y descubrir el conocimiento de los datos recopilados de manera oportuna y en una forma escalable

Ciencia de Datos (*Data Science*) incorpora diferentes elementos y se basa en las técnicas y teorías de muchos campos, incluyendo las matemáticas, estadística, ingeniería de datos, reconocimiento de patrones y aprendizaje, computación avanzada, visualización, modelado de la incertidumbre, almacenamiento de datos y la informática de alto rendimiento **con el objetivo de extraer el significado de datos y la creación de productos de datos**

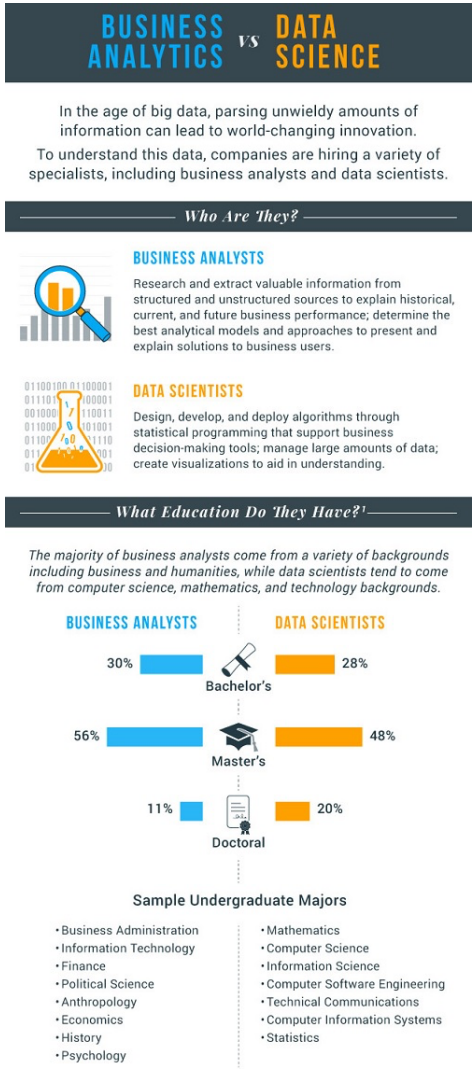




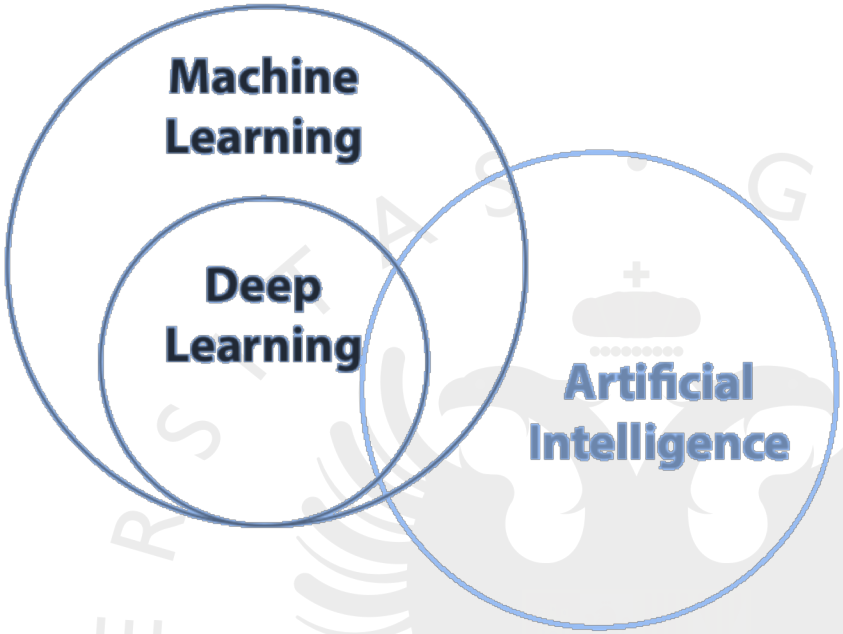
http://drewconway.com



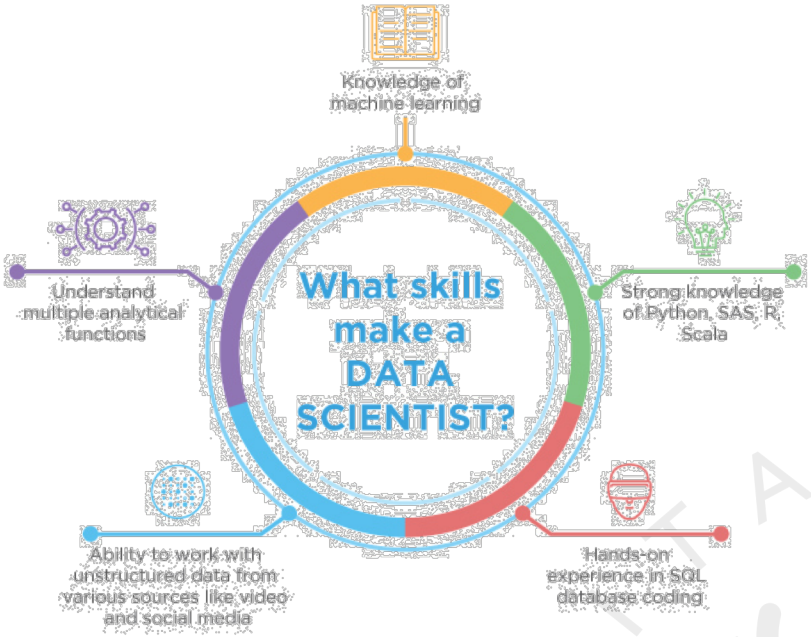
M. Vaisman, H. Harris, S. Murphy (2013) *Analyzing the analyzers: an introspective survey of Data Scientiest and their work*. O'Reilly

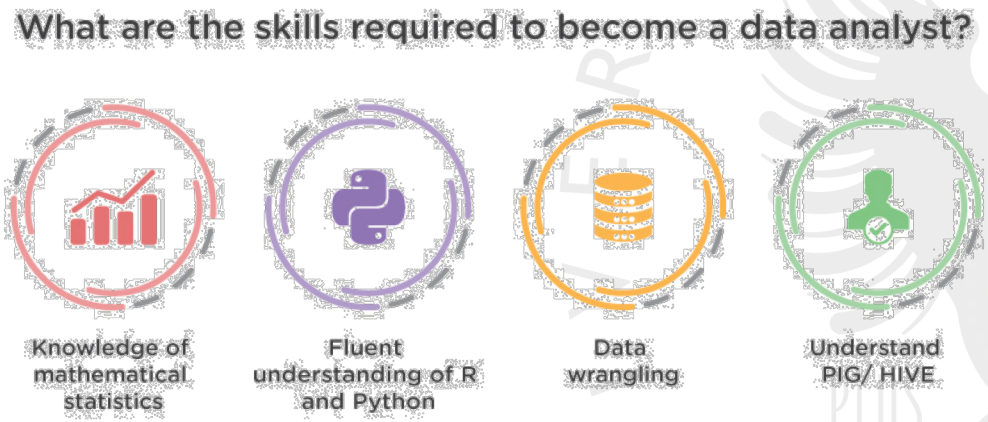


https://www.kdnuggets.com/2015/10/infographic-data-scientist-business-analyst-difference.html



Andrew W. Trask (2018) *Grokking Deep Learning*. Manning.





https://www.simplilearn.com/data-science-vs-data-analytics-vs-machine-learning-article

Cerveza => Pañales **60%**

Pañales => Cerveza **100%**

https://www.theregister.co.uk/2006/08/15/beer_diapers/

Tarjetas de crédito

"A division called MasterCard Advisors aggregates and analyzes 65 billion transactions from 1.5 billion cardholders in 210 countries in order to divine business and consumer trends. Then it sells that information to others. It discovered, among other things, that if people fill up their gas tanks at around four o'clock in the afternoon, they're quite likely to spend between $35 and $50 in the next hour at a grocery store or restaurant. ***A marketer might use that insight to print out coupons for a nearby supermarket on the back of gas-station receipts around that time of day.***"

Cesta de la compra

"The analytics team reviewed the shopping histories of women who signed up for its baby gift-registry. They noticed that these women bought lots of unscented lotion at around the third month of pregnancy, and that a few weeks later they tended to purchase supplements like magnesium, calcium, and zinc. ***The team ultimately uncovered around two dozen products that, used as proxies, enabled the company to calculate a "pregnancy prediction" score for every customer who paid with a credit card or used a loyalty card or mailed coupons.*** *The correlations even let the retailer estimate the due date within a narrow range, so it could send relevant coupons for each stage of the pregnancy."*

V. Mayer-Schonberger, K. Cukier (2013) *Big Data, A revolution that will transform how we live, work and think.* Harcourt Publishing.

ARTICLE

doi:10.1038/nature16961

Mastering the game of Go with deep neural networks and tree search

David Silver[1]*, Aja Huang[1]*, Chris J. Maddison[1], Arthur Guez[1], Laurent Sifre[1], George van den Driessche[1], Julian Schrittwieser[1], Ioannis Antonoglou[1], Veda Panneershelvam[1], Marc Lanctot[1], Sander Dieleman[1], Dominik Grewe[1], John Nham[2], Nal Kalchbrenner[1], Ilya Sutskever[2], Timothy Lillicrap[1], Madeleine Leach[1], Koray Kavukcuoglu[1], Thore Graepel[1] & Demis Hassabis[1]

The game of Go has long been viewed as the most challenging of classic games for artificial intelligence owing to its enormous search space and the difficulty of evaluating board positions and moves. Here we introduce a new approach to computer Go that uses 'value networks' to evaluate board positions and 'policy networks' to select moves. These deep neural networks are trained by a novel combination of supervised learning from human expert games, and reinforcement learning from games of self-play. Without any lookahead search, the neural networks play Go at the level of state-of-the-art Monte Carlo tree search programs that simulate thousands of random games of self-play. We also introduce a new search algorithm that combines Monte Carlo simulation with value and policy networks. Using this search algorithm, our program AlphaGo achieved a 99.8% winning rate against other Go programs, and defeated the human European Go champion by 5 games to 0. This is the first time that a computer program has defeated a human professional player in the full-sized game of Go, a feat previously thought to be at least a decade away.

All games of perfect information have an optimal value function, $v^*(s)$, which determines the outcome of the game, from every board position or state s, under perfect play by all players. These games may be solved by recursively computing the optimal value function in a search tree containing approximately b^d possible sequences of moves, where b is the game's breadth (number of legal moves per position) and d is its depth (game length). In large games, such as chess ($b \approx 35, d \approx 80$)[1] and especially Go ($b \approx 250, d \approx 150$)[1], exhaustive search is infeasible[2,3], but the effective search space can be reduced by two general principles. First, the depth of the search may be reduced by position evaluation: truncating the search tree at state s and replacing the subtree below s by an approximate value function $v(s) \approx v^*(s)$ that predicts the outcome from state s. This approach has led to superhuman performance in chess[4], checkers[5] and othello[6], but it was believed to be intractable in Go due to the complexity of the game[7]. Second, the breadth of the search may be reduced by sampling actions from a policy $p(a|s)$ that is a probability distribution over possible moves a in position s. For example, Monte Carlo rollouts[8] search to maximum depth without branching at all, by sampling long sequences of actions for both players from a policy p. Averaging over such rollouts can provide an effective position evaluation, achieving superhuman performance in backgammon[8] and Scrabble[9], and weak amateur level play in Go[10].

Monte Carlo tree search (MCTS)[11,12] uses Monte Carlo rollouts to estimate the value of each state in a search tree. As more simulations are executed, the search tree grows larger and the relevant values become more accurate. The policy used to select actions during search is also improved over time, by selecting children with higher values. Asymptotically, this policy converges to optimal play, and the evaluations converge to the optimal value function[12]. The strongest current Go programs are based on MCTS, enhanced by policies that are trained to predict human expert moves[13]. These policies are used to narrow the search to a beam of high-probability actions, and to sample actions during rollouts. This approach has achieved strong amateur play[13–15]. However, prior work has been limited to shallow policies[13–15] or value functions[16] based on a linear combination of input features.

Recently, deep convolutional neural networks have achieved unprecedented performance in visual domains: for example, image classification[17], face recognition[18], and playing Atari games[19]. They use many layers of neurons, each arranged in overlapping tiles, to construct increasingly abstract, localized representations of an image[20]. We employ a similar architecture for the game of Go. We pass in the board position as a 19×19 image and use convolutional layers to construct a representation of the position. We use these neural networks to reduce the effective depth and breadth of the search tree: evaluating positions using a value network, and sampling actions using a policy network.

We train the neural networks using a pipeline consisting of several stages of machine learning (Fig. 1). We begin by training a supervised learning (SL) policy network p_σ directly from expert human moves. This provides fast, efficient learning updates with immediate feedback and high-quality gradients. Similar to prior work[13,15], we also train a fast policy p_π that can rapidly sample actions during rollouts. Next, we train a reinforcement learning (RL) policy network p_ρ that improves the SL policy network by optimizing the final outcome of games of self-play. This adjusts the policy towards the correct goal of winning games, rather than maximizing predictive accuracy. Finally, we train a value network v_θ that predicts the winner of games played by the RL policy network against itself. Our program AlphaGo efficiently combines the policy and value networks with MCTS.

Supervised learning of policy networks

For the first stage of the training pipeline, we build on prior work on predicting expert moves in the game of Go using supervised learning[13,21–24]. The SL policy network $p_\sigma(a|s)$ alternates between convolutional layers with weights σ, and rectifier nonlinearities. A final softmax layer outputs a probability distribution over all legal moves a. The input s to the policy network is a simple representation of the board state (see Extended Data Table 2). The policy network is trained on randomly

[1]Google DeepMind, 5 New Street Square, London EC4A 3TW, UK. [2]Google, 1600 Amphitheatre Parkway, Mountain View, California 94043, USA.
*These authors contributed equally to this work.

D. Silver et al. (2016) Mastering the game of Go with deep neural networks and tree search. *Nature* 529, pp. 484-489.

http://www.alphagomovie.com/

RESEARCH

COMPUTER SCIENCE

A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play

David Silver[1*†], Thomas Hubert[1*], Julian Schrittwieser[1*], Ioannis Antonoglou[1], Matthew Lai[1], Arthur Guez[1], Marc Lanctot[1], Laurent Sifre[2], Dharshan Kumaran[1], Thore Graepel[1], Timothy Lillicrap[1], Karen Simonyan[1], Demis Hassabis[1†]

The game of chess is the longest-studied domain in the history of artificial intelligence. The strongest programs are based on a combination of sophisticated search techniques, domain-specific adaptations, and handcrafted evaluation functions that have been refined by human experts over several decades. By contrast, the AlphaGo Zero program recently achieved superhuman performance in the game of Go by reinforcement learning from self-play. In this paper, we generalize this approach into a single AlphaZero algorithm that can achieve superhuman performance in many challenging games. Starting from random play and given no domain knowledge except the game rules, AlphaZero convincingly defeated a world champion program in the games of chess and shogi (Japanese chess), as well as Go.

The study of computer chess is as old as computer science itself. Charles Babbage, Alan Turing, Claude Shannon, and John von Neumann devised hardware, algorithms, and theory to analyze and play the game of chess. Chess subsequently became a grand challenge task for a generation of artificial intelligence researchers, culminating in high-performance computer chess programs that perform at a superhuman level (1, 2). However, these systems are highly tuned to their domain and cannot be generalized to other games without substantial human effort, whereas general game-playing systems (3, 4) remain comparatively weak.

A long-standing ambition of artificial intelligence has been to create programs that can instead learn for themselves from first principles (5, 6). Recently, the AlphaGo Zero algorithm achieved superhuman performance in the game of Go by representing Go knowledge with the use of deep convolutional neural networks (7, 8), trained solely by reinforcement learning from games of self-play (9). In this paper, we introduce AlphaZero, a more generic version of the AlphaGo Zero algorithm that accommodates, without special casing, a broader class of game rules. We apply AlphaZero to the games of chess and shogi, as well as Go, by using the same algorithm and network architecture for all three games. Our results demonstrate that a general-purpose reinforcement learning algorithm can learn, tabula rasa—without domain-specific human knowledge or data, as evidenced by the same algorithm succeeding in multiple domains—superhuman performance across multiple challenging games.

A landmark for artificial intelligence was achieved in 1997 when Deep Blue defeated the human world chess champion (1). Computer chess programs continued to progress steadily beyond human level in the following two decades. These programs evaluate positions by using handcrafted features and carefully tuned weights, constructed by strong human players and programmers, combined with a high-performance alpha-beta search that expands a vast search tree by using a large number of clever heuristics and domain-specific adaptations. In (10) we describe these augmentations, focusing on the 2016 Top Chess Engine Championship (TCEC) season 9 world champion Stockfish (11); other strong chess programs, including Deep Blue, use very similar architectures (1, 12).

In terms of game tree complexity, shogi is a substantially harder game than chess (13, 14): It is played on a larger board with a wider variety of pieces; any captured opponent piece switches sides and may subsequently be dropped anywhere on the board. The strongest shogi programs, such as the 2017 Computer Shogi Association (CSA) world champion Elmo, have only recently defeated human champions (15). These programs use an algorithm similar to those used by computer chess programs, again based on a highly optimized alpha-beta search engine with many domain-specific adaptations.

AlphaZero replaces the handcrafted knowledge and domain-specific augmentations used in traditional game-playing programs with deep neural networks, a general-purpose reinforcement learning algorithm, and a general-purpose tree search algorithm.

Instead of a handcrafted evaluation function and move-ordering heuristics, AlphaZero uses a deep neural network $(\mathbf{p}, v) = f_\theta(s)$ with parameters θ. This neural network $f_\theta(s)$ takes the board position s as an input and outputs a vector of move probabilities $\mathbf{p}$ with components $p_a = \Pr(a|s)$ for each action a and a scalar value v estimating the expected outcome z of the game from position s, $v \approx \mathbb{E}[z|s]$. AlphaZero learns these move probabilities and value estimates entirely from self-play; these are then used to guide its search in future games.

Instead of an alpha-beta search with domain-specific enhancements, AlphaZero uses a general-purpose Monte Carlo tree search (MCTS) algorithm. Each search consists of a series of simulated games of self-play that traverse a tree from root state s_{root} until a leaf state is reached. Each simulation proceeds by selecting in each state s a move a with low visit count (not previously frequently explored), high move probability, and high value (averaged over the leaf states of

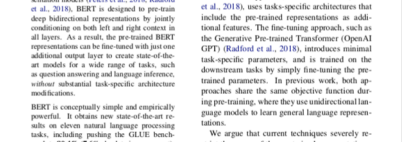


Fig. 1. Training AlphaZero for 700,000 steps. Elo ratings were computed from games between different players where each player was given 1 s per move. (**A**) Performance of AlphaZero in chess compared with the 2016 TCEC world champion program Stockfish. (**B**) Performance of AlphaZero in shogi compared with the 2017 CSA world champion program Elmo. (**C**) Performance of AlphaZero in Go compared with AlphaGo Lee and AlphaGo Zero (20 blocks over 3 days).

Silver et al., Science 362, 1140–1144 (2018) 7 December 2018

npj | Digital Medicine

www.nature.com/npjdigitalmed

ARTICLE OPEN

Scalable and accurate deep learning with electronic health records

Alvin Rajkomar[1,2], Eyal Oren[1], Kai Chen[1], Andrew M. Dai[1], Nissan Hajaj[1], Michaela Hardt[1], Peter J. Liu[1], Xiaobing Liu[1], Jake Marcus[1], Mimi Sun[1], Patrik Sundberg[1], Hector Yee[1], Kun Zhang[1], Yi Zhang[1], Gerardo Flores[1], Gavin E. Duggan[1], Jamie Irvine[1], Quoc Le[1], Kurt Litsch[1], Alexander Mossin[1], Justin Tansuwan[1], De Wang[1], James Wexler[1], Jimbo Wilson[1], Dana Ludwig[2], Samuel L. Volchenboum[3], Katherine Chou[1], Michael Pearson[1], Srinivasan Madabushi[1], Nigam H. Shah[4], Atul J. Butte[4], Michael D. Howell[1], Claire Cui[1], Greg S. Corrado[1] and Jeffrey Dean[1]

Predictive modeling with electronic health record (EHR) data is anticipated to drive personalized medicine and improve healthcare quality. Constructing predictive statistical models typically requires extraction of curated predictor variables from normalized EHR data, a labor-intensive process that discards the vast majority of information in each patient's record. We propose a representation of patients' entire raw EHR records based on the Fast Healthcare Interoperability Resources (FHIR) format. We demonstrate that deep learning methods using this representation are capable of accurately predicting multiple medical events from multiple centers without site-specific data harmonization. We validated our approach using de-identified EHR data from two US academic medical centers with 216,221 adult patients hospitalized for at least 24 h. In the sequential format we propose, this volume of EHR data unrolled into a total of 46,864,534,945 data points, including clinical notes. Deep learning models achieved high accuracy for tasks such as predicting in-hospital mortality (area under the receiver operator curve [AUROC] across sites 0.93–0.94), 30-day unplanned readmission (AUROC 0.75–0.76), prolonged length of stay (AUROC 0.85–0.86), and all of a patient's final discharge diagnoses (frequency-weighted AUROC 0.90). These models outperformed traditional, clinically-used predictive models in all cases. We believe that this approach can be used to create accurate and scalable predictions for a variety of clinical scenarios. In a case study of a particular prediction, we demonstrate that neural networks can be used to identify relevant information from the patient's chart.

npj Digital Medicine (2018)1:18; doi:10.1038/s41746-018-0029-1

INTRODUCTION

The promise of digital medicine stems in part from the hope that, by digitizing health data, we might more easily leverage computer information systems to understand and improve care. In fact, routinely collected patient healthcare data are now approaching the genomic scale in volume and complexity.[1] Unfortunately, most of this information is not yet used in the sorts of predictive statistical models clinicians might use to improve care delivery. It is widely suspected that use of such efforts, if successful, could provide major benefits not only for patient safety and quality but also in reducing healthcare cost.[2–5]

In spite of the richness and potential of available data, scaling the development of predictive models is difficult because, for traditional predictive modeling techniques, each outcome to be predicted requires the creation of a custom dataset with specific variables.[6] It is widely held that 80% of the effort in an analytic model is preprocessing, merging, customizing, and cleaning datasets,[6,7] not analyzing them for insights. This profoundly limits the scalability of predictive models.

Another challenge is that the number of potential predictor variables in the electronic health record (EHR) may easily number in the thousands, particularly if free-text notes from doctors, nurses, and other providers are included. Traditional modeling approaches have dealt with this complexity simply by choosing a very limited number of commonly collected variables to consider.[8] This is problematic because the resulting models may produce imprecise predictions: false-positive predictions can overwhelm physicians, nurses, and other providers with false alarms and concomitant alert fatigue,[9] which the Joint Commission identified as a national patient safety priority in 2014.[10] False-negative predictions can miss significant numbers of clinically important events, leading to poor clinical outcomes.[11,12] Incorporating the entire EHR, including clinicians' free-text notes, offers some hope of overcoming these shortcomings but is unwieldy for most predictive modeling techniques.

Recent developments in deep learning and artificial neural networks may allow us to address many of these challenges and unlock the information in the EHR. Deep learning emerged as the preferred machine learning approach in machine perception problems ranging from computer vision to speech recognition, but has more recently proven useful in natural language processing, sequence prediction, and mixed modality data settings.[13–17] These systems are known for their ability to handle large volumes of relatively messy data, including errors in labels

[1]Google Inc, Mountain View, CA, USA; [2]University of California, San Francisco, CA, USA; [3]University of Chicago Medicine, Chicago, IL, USA and [4]Stanford University, Stanford, CA, USA
Correspondence: Alvin Rajkomar (alvinrajkomar@google.com)
These authors contributed equally: Alvin Rajkomar, Eyal Oren

Received: 26 January 2018 Revised: 14 March 2018 Accepted: 26 March 2018
Published online: 08 May 2018

Published in partnership with the Scripps Translational Science Institute

npj nature partner journals

LETTER

doi:10.1038/nature21056

Dermatologist-level classification of skin cancer with deep neural networks

Andre Esteva[1*], Brett Kuprel[1*], Roberto A. Novoa[2,3], Justin Ko[2], Susan M. Swetter[2,4], Helen M. Blau[5] & Sebastian Thrun[6]

Skin cancer, the most common human malignancy[1–3], is primarily diagnosed visually, beginning with an initial clinical screening and followed potentially by dermoscopic analysis, a biopsy and histopathological examination. Automated classification of skin lesions using images is a challenging task owing to the fine-grained variability in the appearance of skin lesions. Deep convolutional neural networks (CNNs)[4,5] show potential for general and highly variable tasks across many fine-grained object categories[6–11]. Here we demonstrate classification of skin lesions using a single CNN, trained end-to-end from images directly, using only pixels and disease labels as inputs. We train a CNN using a dataset of 129,450 clinical images—two orders of magnitude larger than previous datasets[12]—consisting of 2,032 different diseases. We test its performance against 21 board-certified dermatologists on biopsy-proven clinical images with two critical binary classification use cases: keratinocyte carcinomas versus benign seborrheic keratoses; and malignant melanomas versus benign nevi. The first case represents the identification of the most common cancers, the second represents the identification of the deadliest skin cancer. The CNN achieves performance on par with all tested experts across both tasks, demonstrating an artificial intelligence capable of classifying skin cancer with a level of competence comparable to dermatologists. Outfitted with deep neural networks, mobile devices can potentially extend the reach of dermatologists outside of the clinic. It is projected that 6.3 billion smartphone subscriptions will exist by the year 2021 (ref. 13) and can therefore potentially provide low-cost universal access to vital diagnostic care.

There are 5.4 million new cases of skin cancer in the United States[1] every year. One in five Americans will be diagnosed with a cutaneous malignancy in their lifetime. Although melanomas represent fewer than 5% of all skin cancers in the United States, they account for approximately 75% of all skin-cancer-related deaths, and are responsible for over 10,000 deaths annually in the United States alone. Early detection is critical, as the estimated 5-year survival rate for melanoma drops from over 99% if detected in its earliest stages to about 14% if detected in its latest stages. We developed a computational method which may allow medical practitioners and patients to proactively track skin lesions and detect cancer earlier. By creating a novel disease taxonomy, and a disease-partitioning algorithm that maps individual diseases into training classes, we are able to build a deep learning system for automated dermatology.

Previous work in dermatological computer-aided classification[12,14,15] has lacked the generalization capability of medical practitioners owing to insufficient data and a focus on standardized tasks such as dermoscopy[16–18] and histological image classification[19–21]. Dermoscopy images are acquired via a specialized instrument and histological images are acquired via invasive biopsy and microscopy; whereby both modalities yield highly standardized images. Photographic images (for example, smartphone images) exhibit variability in factors such as zoom, angle and lighting, making classification substantially more challenging[22–24]. We overcome this challenge by using a data-driven approach—1.41 million pre-training and training images make classification robust to photographic variability. Many previous techniques require extensive preprocessing, lesion segmentation and extraction of domain-specific visual features before classification. By contrast, our system requires no hand-crafted features; it is trained end-to-end directly from image labels and raw pixels, with a single network for both photographic and dermoscopic images. The existing body of work uses small datasets of typically less than a thousand images of skin lesions[16,18,19], which, as a result, do not generalize well to new images. We demonstrate generalizable classification with a new dermatologist-labelled dataset of 129,450 clinical images, including 3,374 dermoscopy images.

Deep learning algorithms, powered by advances in computation and very large datasets[25], have recently been shown to exceed human performance in visual tasks such as playing Atari games[26], strategic board games like Go[27] and object recognition[28]. In this paper we outline the development of a CNN that matches the performance of dermatologists at three key diagnostic tasks: melanoma classification, melanoma classification using dermoscopy and carcinoma classification. We restrict the comparisons to image-based classification.

We utilize a GoogleNet Inception v3 CNN architecture[29] that was pre-trained on approximately 1.28 million images (1,000 object categories) from the 2014 ImageNet Large Scale Visual Recognition Challenge[6], and train it on our dataset using transfer learning[30]. Figure 1 shows the working system. The CNN is trained using 757 disease classes. Our dataset is composed of dermatologist-labelled images organized in a tree-structured taxonomy of 2,032 diseases, in which the individual diseases form the leaf nodes. The images come from 18 different clinician-curated, open-access online repositories, as well as from clinical data from Stanford University Medical Center. Figure 2a shows a subset of the full taxonomy, which has been organized clinically and visually by medical experts. We split our dataset into 127,463 training and validation images and 1,942 biopsy-labelled test images.

To take advantage of fine-grained information contained within the taxonomy structure, we develop an algorithm (Extended Data Table 1) to partition diseases into fine-grained training classes (for example, amelanotic melanoma and acrolentiginous melanoma). During inference, the CNN outputs a probability distribution over these fine classes. To recover the probabilities for coarser-level classes of interest (for example, melanoma) we sum the probabilities of their descendants (see Methods and Extended Data Fig. 1 for more details).

We validate the effectiveness of the algorithm in two ways, using nine-fold cross-validation. First, we validate the algorithm using a three-class disease partition—the first-level nodes of the taxonomy, which represent benign lesions, malignant lesions and non-neoplastic

[1]Department of Electrical Engineering, Stanford University, Stanford, California, USA. [2]Department of Dermatology, Stanford University, Stanford, California, USA. [3]Department of Pathology, Stanford University, Stanford, California, USA. [4]Dermatology Service, Veterans Affairs Palo Alto Health Care System, Palo Alto, California, USA. [5]Baxter Laboratory for Stem Cell Biology, Department of Microbiology and Immunology, Institute for Stem Cell Biology and Regenerative Medicine, Stanford University, Stanford, California, USA. [6]Department of Computer Science, Stanford University, Stanford, California, USA.
*These authors contributed equally to this work.

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova
Google AI Language
{jacobdevlin,mingweichang,kentonl,kristout}@google.com

Abstract

We introduce a new language representation model called **BERT**, which stands for **B**idirectional **E**ncoder **R**epresentations from **T**ransformers. Unlike recent language representation models (Peters et al., 2018; Radford et al., 2018), BERT is designed to pre-train deep bidirectional representations by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT representations can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, *without* substantial task-specific architecture modifications.

BERT is conceptually simple and empirically powerful. It obtains new state-of-the-art results on eleven natural language processing tasks, including pushing the GLUE benchmark to **80.4%** (**7.6%** absolute improvement), MultiNLI accuracy to **86.7%** (**5.6%** absolute improvement) and the SQuAD v1.1 question answering Test F1 to **93.2** (**1.5** absolute improvement), outperforming human performance by **2.0**.

1 Introduction

Language model pre-training has shown to be effective for improving many natural language processing tasks (Dai and Le, 2015; Peters et al., 2017, 2018; Radford et al., 2018; Howard and Ruder, 2018). These tasks include sentence-level tasks such as natural language inference (Bowman et al., 2015; Williams et al., 2018) and paraphrasing (Dolan and Brockett, 2005), which aim to predict the relationships between sentences by analyzing them holistically, as well as token-level tasks such as named entity recognition (Tjong Kim Sang and De Meulder, 2003) and SQuAD question answering (Rajpurkar et al., 2016), where models are required to produce fine-grained output at the token-level.

There are two existing strategies for applying pre-trained language representations to downstream tasks: *feature-based* and *fine-tuning*. The feature-based approach, such as ELMo (Peters et al., 2018), uses task-specific architectures that include the pre-trained representations as additional features. The fine-tuning approach, such as the Generative Pre-trained Transformer (OpenAI GPT) (Radford et al., 2018), introduces minimal task-specific parameters, and is trained on the downstream tasks by simply fine-tuning the pre-trained parameters. In previous work, both approaches share the same objective function during pre-training, where they use unidirectional language models to learn general language representations.

We argue that current techniques severely restrict the power of the pre-trained representations, especially for the fine-tuning approaches. The major limitation is that standard language models are unidirectional, and this limits the choice of architectures that can be used during pre-training. For example, in OpenAI GPT, the authors use a left-to-right architecture, where every token can only attended to previous tokens in the self-attention layers of the Transformer (Vaswani et al., 2017). Such restrictions are sub-optimal for sentence-level tasks, and could be devastating when applying fine-tuning based approaches to token-level tasks such as SQuAD question answering (Rajpurkar et al., 2016), where it is crucial to incorporate context from both directions.

In this paper, we improve the fine-tuning based approaches by proposing BERT: **B**idirectional **E**ncoder **R**epresentations from **T**ransformers. BERT addresses the previously mentioned unidirectional constraints by proposing a new pre-training objective: the "masked language

Automatic Handgun Detection Alarm in Videos Using Deep Learning

Roberto Olmos[1], Siham Tabik[1], and Francisco Herrera[1,2]

[1]Soft Computing and Intelligent Information Systems research group
[2]Department of Computer Science and Artificial Intelligence, University of Granada, 18071 Granada, Spain. emails: siham@ugr.es, herrera@decsai.ugr.es

February 20, 2017

Abstract

Current surveillance and control systems still require human supervision and intervention. This work presents a novel automatic handgun detection system in videos appropriate for both, surveillance and control purposes. We reformulate this detection problem into the problem of minimizing false positives and solve it by building the key training data-set guided by the results of a deep Convolutional Neural Networks (CNN) classifier, then assessing the best classification model under two approaches, the sliding window approach and region proposal approach. The most promising results are obtained by Faster R-CNN based model trained on our new database. The best detector show a high potential even in low quality youtube videos and provides satisfactory results as automatic alarm system. Among 30 scenes, it successfully activates the alarm after five successive true positives in less than 0.2 seconds, in 27 scenes. We also define a new metric, Alarm Activation per Interval (AApI), to assess the performance of a detection model as an automatic detection system in videos.

Index terms— Classification, Detection, Deep learning, Convolutional Neural Networks (CNNs), Faster R-CNN, VGG-16, Alarm Activation per Interval

1 Introduction

The crime rates caused by guns are very concerning in many places in the world, especially in countries where the possession of guns is legal or was legal for a period of time. The last statistics reported by the United Nations Office on Drugs and Crime (UNODC) reveals that the number of crimes involving guns

PROFICIENT

Deep Learning for Energy-Efficient Building Control

Buildings account for more than one-third of the worldwide energy consumption, and they are an equally important source of CO2 emissions. A critical task to improve energy efficiency is to adapt HVAC (heat, ventilation, and air-conditioning) equipment operation to energy demand, minimizing consumption while keeping occupants comfortable. Model-predictive control (MPC) approaches have proved successful to generate HVAC operational control plans, but they still present several drawbacks: they require constructing a detailed physical model of each building, which implies a considerable effort, and have important performance limitations, which strongly limit the opportunities for implementing new intelligent optimization procedures.

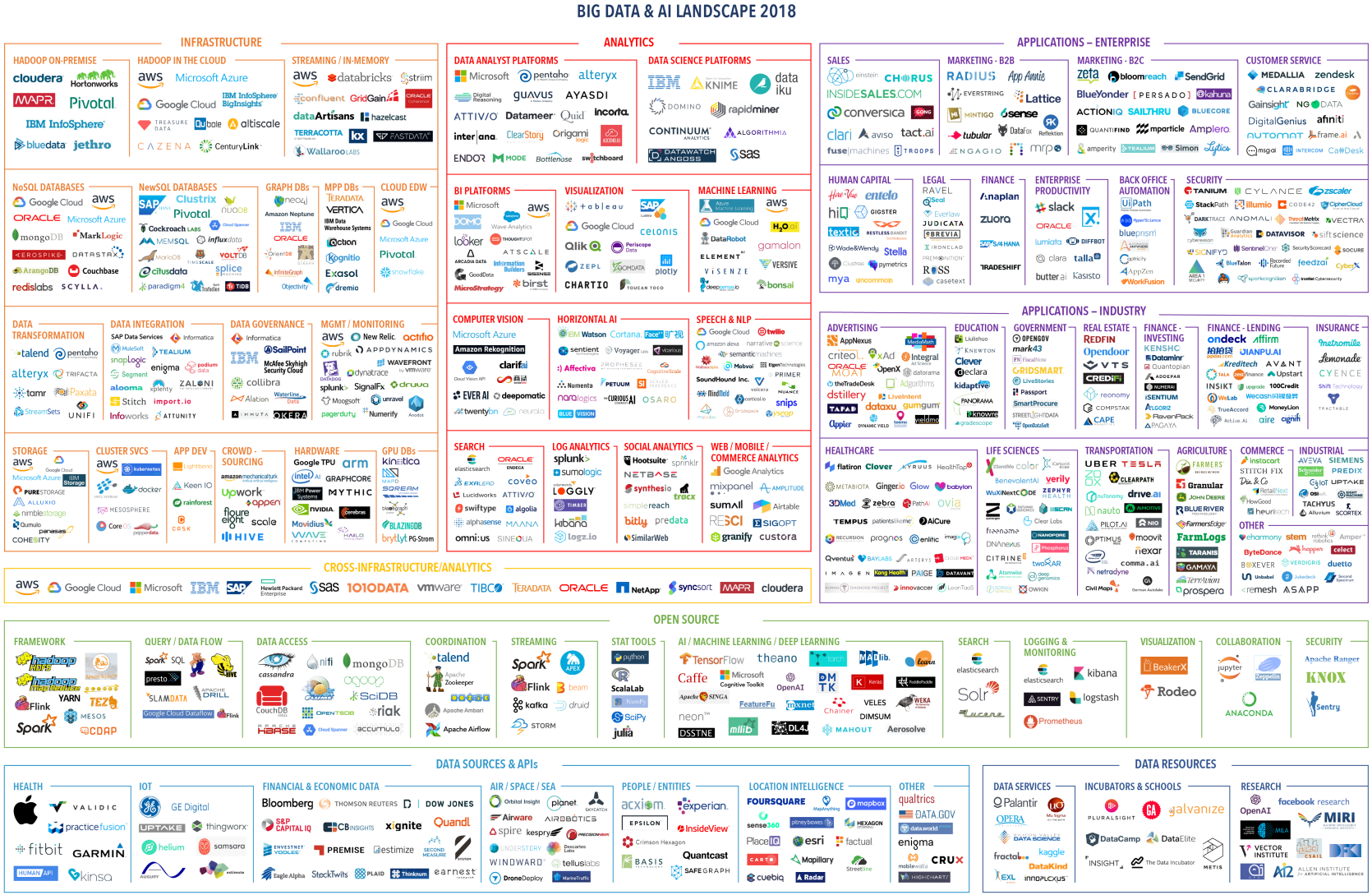
PROFICIENT aims at solving these issues by developing novel deep reinforcement learning techniques capable of: (1) learning a more efficient predictive model of the building from sensor data; and (2) optimizing the computation of operational plans without using heuristic knowledge.

Stay tuned for updates!

Support and contact

Principal Investigator: Dr. Juan Gómez-Romero

***PROFICIENT** is a 2-year project funded by the **EXPLORA** programme of the Spanish Ministry of Science, Innovation and Universities in 2018-2020 (TIN2017-91223-EXP).*



Final 2018 version, updated 07/15/2018

mattturck.com/bigdata2018

FIRSTMARK
EARLY STAGE VENTURE CAPITAL

Herramientas

- KNIME
- WEKA
- KEEL
- **Tableau**

Lenguajes

- Java, Scala
- **R**
- Python

Bibliotecas

- scikit-learn
- **tensorflow**
- **caret**
- PyTorch
- h2o

Frameworks

- Spark

Recursos

- MOOC (Coursera, Udacity, edX, etc.)
- **Kaggle**

Cloud

- Microsoft Azure
- Google Cloud
- Amazon Web Services

SCIENTIFIC REPORTS

OPEN

Unique in the Crowd: The privacy bounds of human mobility

Yves-Alexandre de Montjoye[1,2], César A. Hidalgo[1,3,4], Michel Verleysen[2] & Vincent D. Blondel[2,5]

[1]Massachusetts Institute of Technology, Media Lab, 20 Ames Street, Cambridge, MA 02139 USA, [2]Université catholique de Louvain, Institute for Information and Communication Technologies, Electronics and Applied Mathematics, Avenue Georges Lemaître 4, B-1348 Louvain-la-Neuve, Belgium, [3]Harvard University, Center for International Development, 79 JFK Street, Cambridge, MA 02138, USA, [4]Instituto de Sistemas Complejos de Valparaíso, Paseo 21 de Mayo, Valparaíso, Chile, [5]Massachusetts Institute of Technology, Laboratory for Information and Decision Systems, 77 Massachusetts Avenue, Cambridge, MA 02139, USA.

SUBJECT AREAS:
APPLIED PHYSICS
APPLIED MATHEMATICS
STATISTICS
COMPUTATIONAL SCIENCE

Received
1 October 2012

Accepted
4 February 2013

Published
25 March 2013

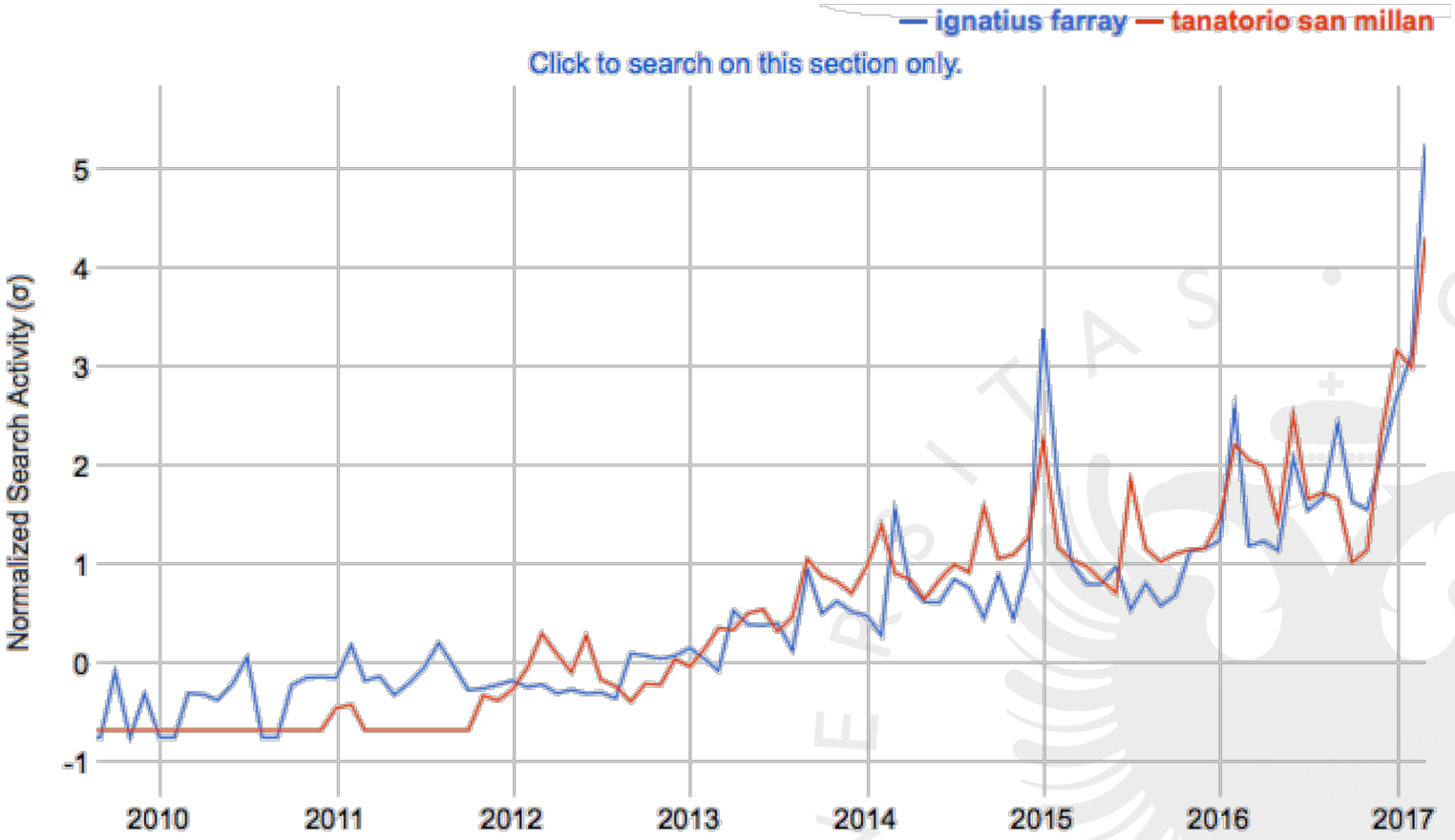
Correspondence and requests for materials should be addressed to Y.-A. de M. (yva@mit.edu)

We study fifteen months of human mobility data for one and a half million individuals and find that human mobility traces are highly unique. In fact, in a dataset where the location of an individual is specified hourly, and with a spatial resolution equal to that given by the carrier's antennas, four spatio-temporal points are enough to uniquely identify 95% of the individuals. We coarsen the data spatially and temporally to find a formula for the uniqueness of human mobility traces given their resolution and the available outside information. This formula shows that the uniqueness of mobility traces decays approximately as the 1/10 power of their resolution. Hence, even coarse datasets provide little anonymity. These findings represent fundamental constraints to an individual's privacy and have important implications for the design of frameworks and institutions dedicated to protect the privacy of individuals.

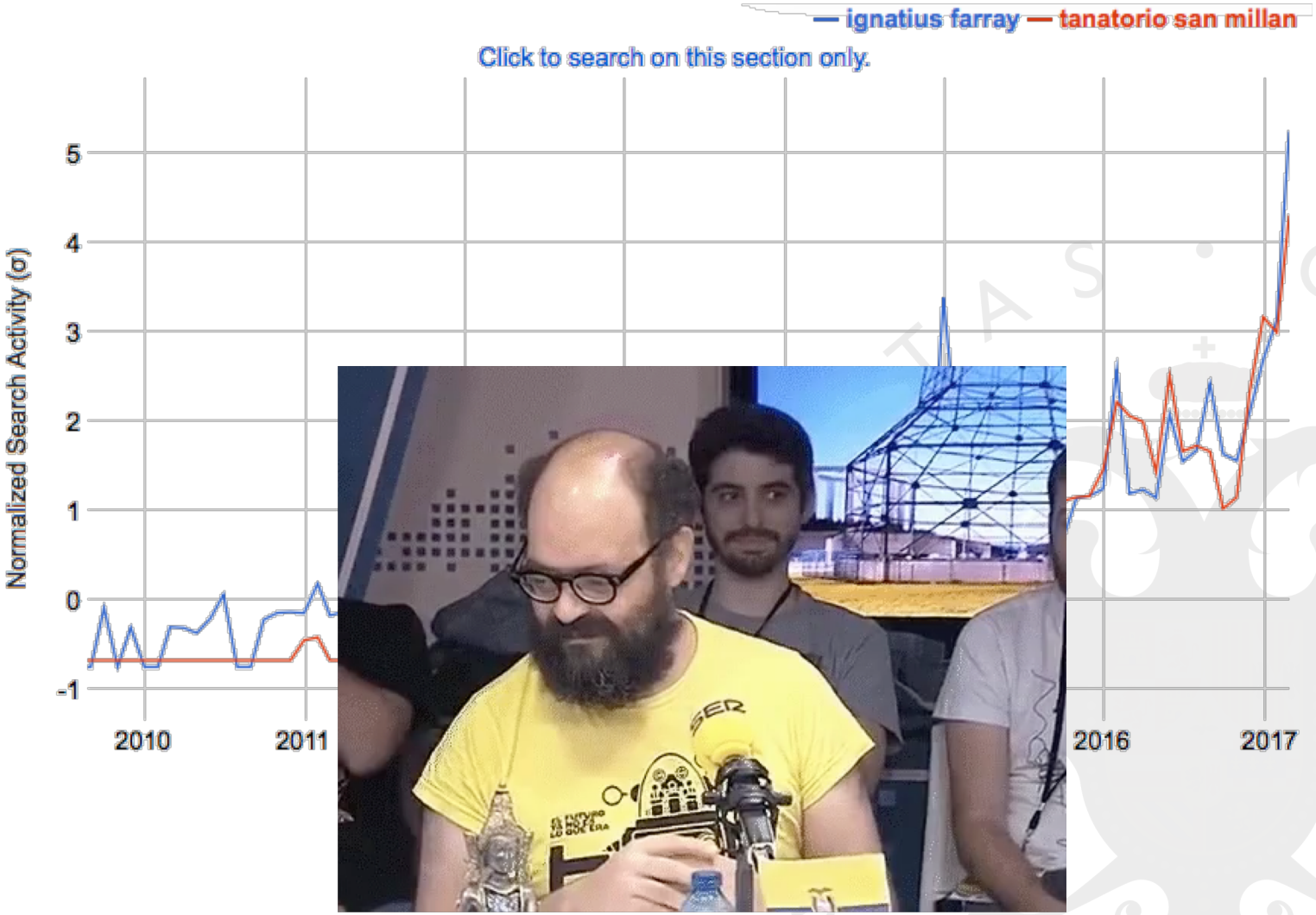
Derived from the Latin Privatus, meaning "withdraw from public life," the notion of privacy has been foundational to the development of our diverse societies, forming the basis for individuals' rights such as free speech and religious freedom[1]. Despite its importance, privacy has mainly relied on informal protection mechanisms. For instance, tracking individuals' movements has been historically difficult, making them de-facto private. For centuries, information technologies have challenged these informal protection mechanisms. In 1086, William I of England commissioned the creation of the Doomsday book, a written record of major property holdings in England containing individual information collected for tax and draft purposes[2]. In the late 19th century, de-facto privacy was similarly threatened by photographs and yellow journalism. This resulted in one of the first publications advocating privacy in the U.S. in which Samuel Warren and Louis Brandeis argued that privacy law must evolve in response to technological changes[3].

Y.A. de Montjoye, C.A. Hidalgo, M. Verleysen, V.D. Blondel (2013) Unique in the crowd: the privacy bounds of human mobility. *Nature Scientific Reports* 3, no. 1376.

High-Level Expert Group on Artificial Intelligence (2018)
Draft – Ethics Guidelines for Trustworthy AI.



https://www.google.com/trends/correlate/search?e=Ignatius+Farray&e=tanatorio+san+millan&t=monthly&p=es&shift=2#default,20


ignatius farray
tanatorio san millan
Click to search on this section only.
Normalized Search Activity (σ)
5
4
3
2
1
0
-1
2010
2011
2016
2017

V. Mayer-Schonberger, K. Cukier (2013) *Big Data, A revolution that will transform how we live, work and think.* Harcourt Publishing.

C. Rudder (2014) *Dataclysm*. 4th State.

C. Hidalgo (2015) *Why information grows.* Penguin Books.