

Data Science para Principiantes

- Qué estudio, y cuál fue mi camino en Data Science:
- Introducción a qué se verá en el curso y qué se logrará
- Qué es un Data Scientist?
- A qué fondos usamos Data Science?
- Varios ejemplos de los usos y logros del campo
- *Una semana típica*
- *Un año típico*
- Por qué emergió esta profesión en los últimos años y no antes?
- En qué campos trabajan los Data Scientists?
- Cuáles habilidades y herramientas usamos?
- Un workflow típico
- Formular la pregunta
- Recopilar datos
- Limpiar datos
- Dormir con tus datos
- Construir un modelo
- Validar su modelo
- Predecir el futuro
- Comunicar resultados

Qué es un Data Scientist?

"Una persona que construye sistemas que intentan encontrar patrones en datos"
—Yo

una cosa de que todos están hablando, para que todos están contratando, que ya tiene realmente una muy gran demanda.

- Por ello, puede ser una cosa que lleve mucha ilusión, y confusión, que se le pueda dejar a uno preguntando: "por qué Data Science? que alguien me explique todo esta alharaca."

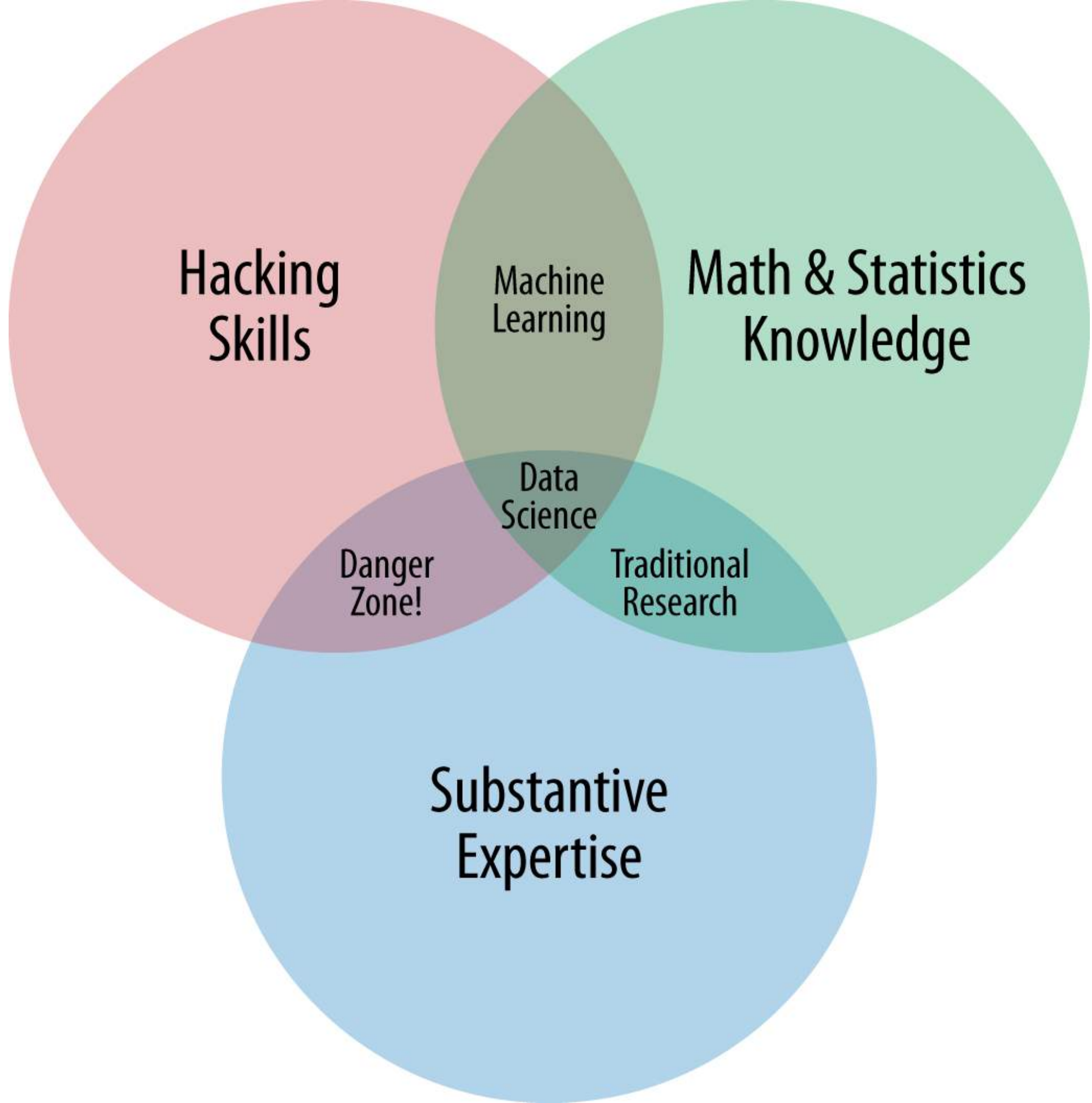
- Para este curso, voy a trabajar con algunas suposiciones:

- Ustedes no son Data Scientists

- Ustedes quieren aclarar el humo para saber lo que realmente hacemos

- La cosa es que: Data Science es un campo que abarca muchísimas cosas: temas de programación, matemáticas, y negocios. Lamentablemente, no podemos empezar desde el principio en cada uno de estos campos. En lugar, el ofrecimiento de este curso será la cosa más difícil de todas: cómo juntar estos temas para impactar a proyectos y organizaciones, que es, casi por definición, lo que es el campo de Data Science.

- Por lo mencionado, voy a empezar desde el inicio, explicando lo que es un Data Scientist, las cosas impresionantes que se pueden hacer



Una persona que programa mejor que estadísticos,
y que hace estadísticos mejor que programadores
^Los tres lados del triángulo: programación,
matemáticas, y experticia temática

Pues, eso me
parece difícil
lograr.

El "unicornio"

^Normalmente, encontramos personas que son fuertes en dos de los tres lados

^Vemos gente que viene de un solo lado del triángulo y les toca adquirir los demás

^Quizás sea bueno: así que Data Science es una disciplina que embarca muchísimas cosas, puede que le sirva a cada uno especializarse

A qué fondos usamos Data Science?

- Hasta ahora, hemos hablado en términos muy genéricos sobre lo que es data science. Ya les vamos a dar unos ejemplos de los logros, ambiciones y posibilidades de este campo tan vasto y poderoso.



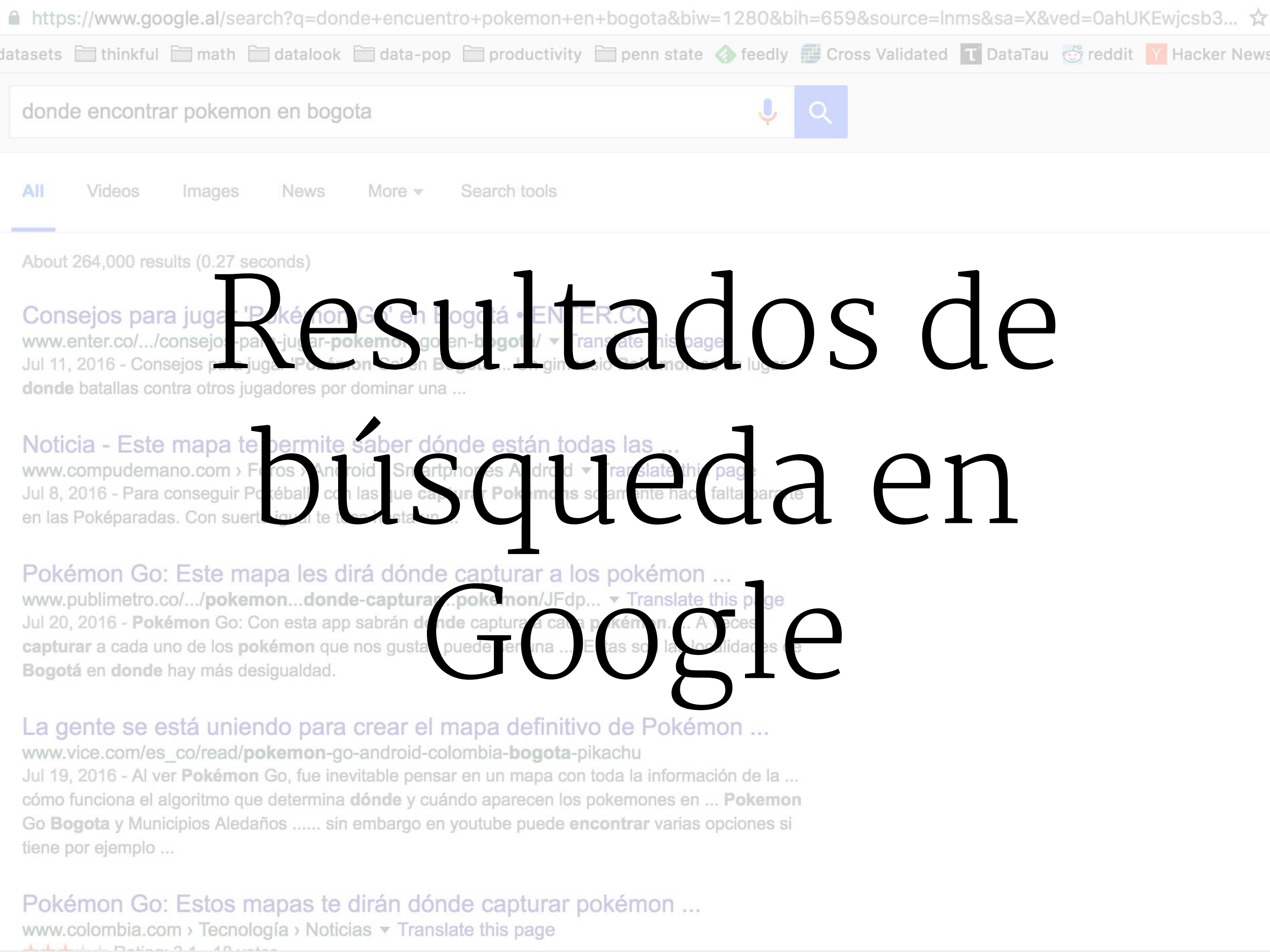
Coches de propulsión propia

de machine learning, un campo vasto abarcado por completo por lo que es data science.

- El algoritmo se llama una red neuronal convolucional, o en inglés el "convolutional neural network."
- Con este algoritmo, estamos básicamente imitando a lo que hace un ser humano detrás del volante.
- Entonces, deconstruyamos lo que hace el ser humano. En el caso más básico, mientras conduce, ve la curva de la ruta con sus ojos, y dado lo que vea gira el volante una cierta cantidad de grados para mantenerse en la ruta. Simple, no?
- Además, le toca a nuestro ser humano observar cosas como: obstrucciones en la ruta, señales de alto, etc., y tomar la buena acción en cuanto a esta información en seguida.
- Cómo podemos traducir esta tarea en la de una máquina? Deconstruyamos un poco más.
- Dijimos que un ser humano "observa" cosas en la calle. Qué significa observar? Les parece si cortamos el flujo de cosas que se ocurren en la calle en una serie de imágenes individuales, de la misma forma de la se puede decir que un video está compuesto de imágenes individuales?
- De ahí, vamos a decir que el ser humano hace nada más que ir recibiendo imágenes frente de él y toma una decisión - parar el coche, girar el volante, etc.
- al recibir cada uno.
- Y entonces, esto es lo que están haciendo los coches autónomos en Pittsburgh: ingiriendo imágenes de lo que esté pasando frente de ellos por

Predecir crimen en Rio de Janeiro

- Durante los juegos olímpicos, había una organización creando un mapa mostrando áreas de peligro en Rio de Janeiro. En realidad, lo que estaban haciendo era usar machine learning para predecir crimen en todos los locales de la ciudad, y visualizar los resultados en el mapa: áreas para que predijeron alta probabilidad de crimen viéndose como áreas de peligro.
- Obviamente, la pregunta es la siguiente: "cómo hicieron las predicciones?" La respuesta, machine learning.
- Para crear este model predictivo de machine learning, recopilaron muchos datos sobre crímenes antiguos: 14 millones de crímenes que tuvieron lugar en los últimos 5 años. Cada crimen hubiera traído varias cosas: el tipo crimen, o en otras palabras lo que pasó, junto con un conjunto de datos sobre el evento: la fecha, la hora del día, los coordinados geográficos, etc.
- Con estos modelos, digamos que entrenaron un modelo de machine learning. Así que cada crimen tiene un output - el tipo de crimen que se ocurrió - y un conjunto de inputs, se le pudo alimentar todo al model para que aprendiera los patrones entre los inputs y los outputs. Por ejemplo, puede que los datos muestren que en un cierto barrio, a medianoche, en el día de sabado en el verano, muchos crímenes se han ocurrido en el pasado. El modelo de machine learning aprendería este patrón, así que la próxima vez que surga esta situación - tal barrio, a medianoche, el sabado, el verno - predecirá que habrá mucho crimen igual.
- De ahí, los data scientists construyeron un sistema ingenieril para pasarle todas la predicciones al mapa, para que la gente pueda visualizar los resultados facilmente.



Resultados de búsqueda en Google

- Un logro bien universal de data science es buscar resultados en Google.
- Al ingresar tu búsqueda, por ejemplo, "¿el cabello de Trump es bien real?", estás pasándolo un input a un modelo de machine learning, que predice en seguida un output, que es, simplemente, una lista de las páginas que más te complacerían.
- En este caso, cuáles son los inputs? Pues, es el texto. Y el modelo por su parte, al predecir, basa su decisión en cosas como: en cuáles páginas aparece tu texto? Que tan popular es cada página? Mejor surgirte una página con diez mil vistas mensualmente que una con 2 vistas, verdad?
- En este caso, quiero enfocarnos en el logro de ingeniería más que en el de matemáticas. Dijimos que data science es el campo de "construir sistemas, típicamente sistemas ingenieriles en su forma final, que intentan encontrar estructura o patrones en datos."
- Cuando buscas algo en Google, se te aparecen resultados buenos y relevantes dentro de unos milisegundos. Pensemos en esto un segundo. Ingresaste un pedazo de texto, y Google ha seleccionado inteligentemente entre 30 trillones de páginas de web un mucho, mucho menos de un segundo. Esto es un logro de ingeniería.
- Como Data Scientists, puede que sea fácil pensar que la parte matemática es siempre la demás: laboramos tanto en construir el algoritmo que nos recomienda las páginas, y al terminar, el hueco entre computer estos resultados y entregarles a nuestros usuarios es chico y trivial. En realidad, y muchas veces en las que sea el opuesto. Imagínate, computar y entregar estos resultados en un tiempo tan corto, y además, hacerlo para las millones de personas a la vez que están buscando cosas en Google.
- Un data scientist sería la persona trabajando en ambos lados del asunto: construyendo el modelo matemático de recomendación, y implementándolo para servir sus predicciones a mucha, mucha gente.



Playlists personalizados en Spotify

- Un cuarto logro muy importante de data science es el de personalización. Usando el ejemplo de Spotify, nos entregan cada semana el playlist de Descubrimiento Semanal: simplemente una lista música que Spotify piensa que te gustaría.
- La pregunta es: cómo se genera esta lista? Cómo se sabe cuál música te gustaría?
- Entra machine learning. Spotify usa algoritmos de machine learning para aprender tus preferencias de música, dado la música que has escuchado en el pasado. En cuanto a los ejemplos anteriores, tu comportamiento en el servicio en el pasado es el "input," y la lista de recomendaciones, el "output."
- La cosa tan chévere que está sucediendo acá es que: estos algoritmos son básicamente capaz de entender tus preferencias mejor que tú mismo. Si me preguntas: "Will, ¿qué de música te gusta escuchar?" Yo te puedo contestar: "pues, me gusta el reggaetón, el hip hop estadounidense y español, y últimamente una cantante de Kosovo."
- En realidad, otra persona puede dar la misma respuesta, pero pasa que en el fondo, le gusta escuchar artistas, ritmos y sonidos completamente diferentes.
- La verdad en este caso es muy, muy sutil, y puedo que simplemente no tengamos las palabras tan finas en nuestros idiomas para adecuadamente describir nuestras preferencias en solo unas frases. Puede que entregarle un USB con tu historia completa de escuchar sirva mucho mejor.
- En este sentido, los algoritmos data science de personalización te pueden entender mejor que tú mismo.

Una semana típica

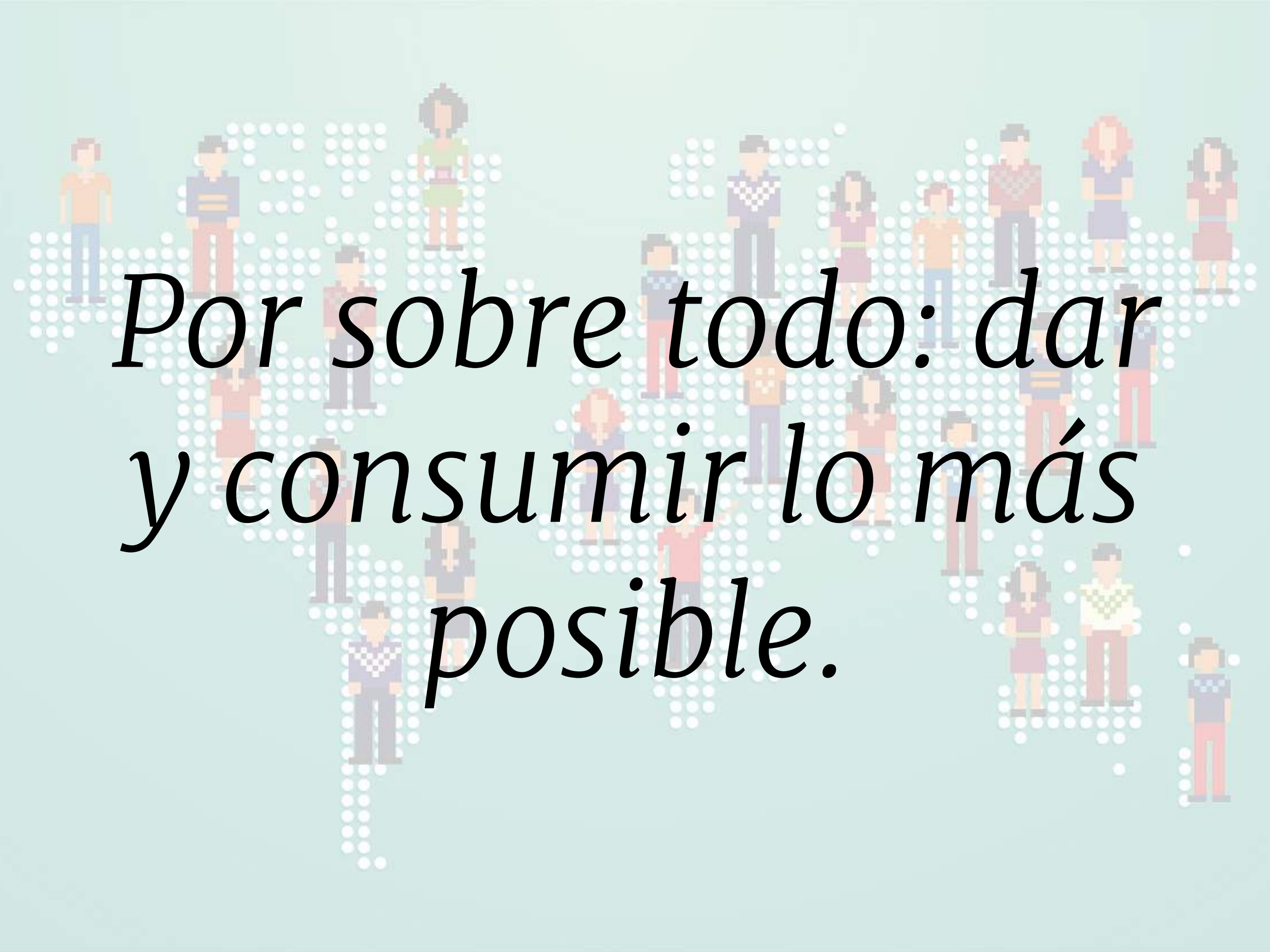
- >> Análisis exploratorio
- >> Construir un modelo matemático
- >> Construir un sistema ingenieril para automatizar dicho modelo
- >> Coordinar con el negocio
- >> Explicar resultados

- Análisis exploratorio: cuáles son los factores o comportamientos que contribuyen a que nuestros clientes nos dejan?
- Construir un modelo matemático: Escribe algún código que expone un framework para ingerir esos factores, codificados en números, y predice si un cliente nos está por dejar
- Construir un sistema ingenieril para automatizar el anterior: obvio que no lo queremos hacer a mano cada día: tenemos mejores cosas que hacer. Deja que la compu trabaje para nosotros
- Coordinar con el negocio: qué hicimos? Cómo se usan nuestros resultados? Convencerles de que el esfuerzo valió la pena, plata y tiempo, para que se queden contentos con nuestra presencia en el equipo, y listos para el próximo.
- Explicar resultados: "Que hiciste? Para que sirve?" te pregunta una persona que desafortunadamente no ha tomado una clase de matemáticas desde los 12.

Un año típico

- >> Desarrollar relaciones
- >> Evangelizar una cultura de datos dentro de la empresa
- >> Construir herramientas para uso internal
- >> Construir sistemas y features para usuarios externos
- >> Construir y mantener sistemas ingenieriles de ETL

- Desarrollar relaciones: quién me ayuda a: entender un problema del negocio, conseguir recursos de tiempo y plata, promover mi solución a la gerencia?
- Evangelizar una cultura de datos: hay verdad en datos. Debemos siempre acompañar argumentos con números y hechos. Si no entiendes nada de datos, aprende (o pide que alguien de enseñe). Basicamente, súmate.
- Construir herramientas para uso internal: dashboards de KPI. Por ejemplo, tiempo útil de servidores, o disponibilidad de Customer Care.
- Construir sistemas y features para los usuarios externos: Si somos Spotify, un sistema de machine learning que aprende tus preferencias de música (realmente mejor que las entiendes tú), y te recomienda otras cosas que te puedan gustar
- Construir y mantener sistemas ingenieriles de ETL: Por lo general, nos toca a nosotros construir y mantener sistemas que ingieren datos de cualquier fuente (que sea Google Analytics, Marketo, Snowplow, etc.), los transforman (les aplican lógico de negocio), los guardan en un base de datos, y los exponen para consumo á través de una visualizacion, o acceso directo



*Por sobre todo: dar
y consumir lo más
posible.*

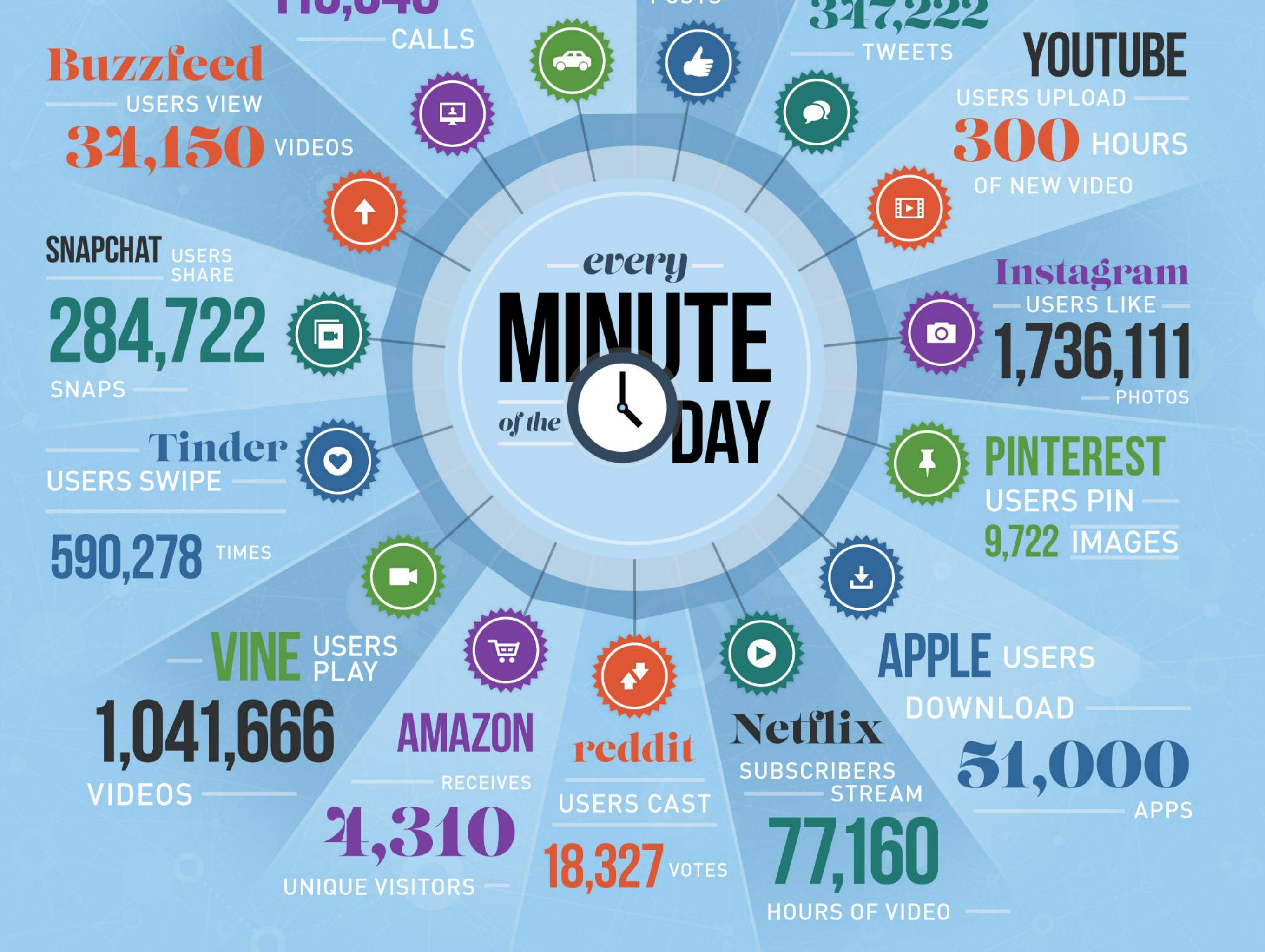
Dar y consumir lo más posible de la comunidad de matemáticas e ingeniería en el internet abierto

- Estar siempre leyendo, participando, dialogando, publicando y aprovechando. Esta comunidad existe online. En mi opinión, no puedes ser Data Scientist sin ser parte de dicha comunidad.

Por qué emergió
esta profesión en
los últimos años
y no antes?

Ya llueve datos.





- Las matemáticas no son nuevas. La programación no es nueva. Por qué ahora?
- Ya generamos un montón de datos. Tipo superficie de Everest. Ya tenemos el poder computacional para procesar estos datos, y a su vez construir representaciones del mundo verdaderamente fuertes.
 - Imagínate que eres radiólogo. Cada año, se te enseñan 500 rayos X - unos con tumores y unos sin. Aprendes hasta un cierto extremo los elementos que se caracterizan por tumores y los que no. Llegado el momento de analizar un nuevo rayo X y decirle al paciente si tiene tumor, usas lo que aprendiste y haces una predicción.
 - Un modelo de machine learning hace exactamente lo mismo: se le enseñan rayos X (codificados en números) y aprende los elementos que se caracterizan por tumores y los que no. En cambio, en vez de poder ingerir 500 rayos X al año, es capaz de ingerir 1,000,000 cada día. Llegado el momento de predecir. Cuál debería predecir mejor?

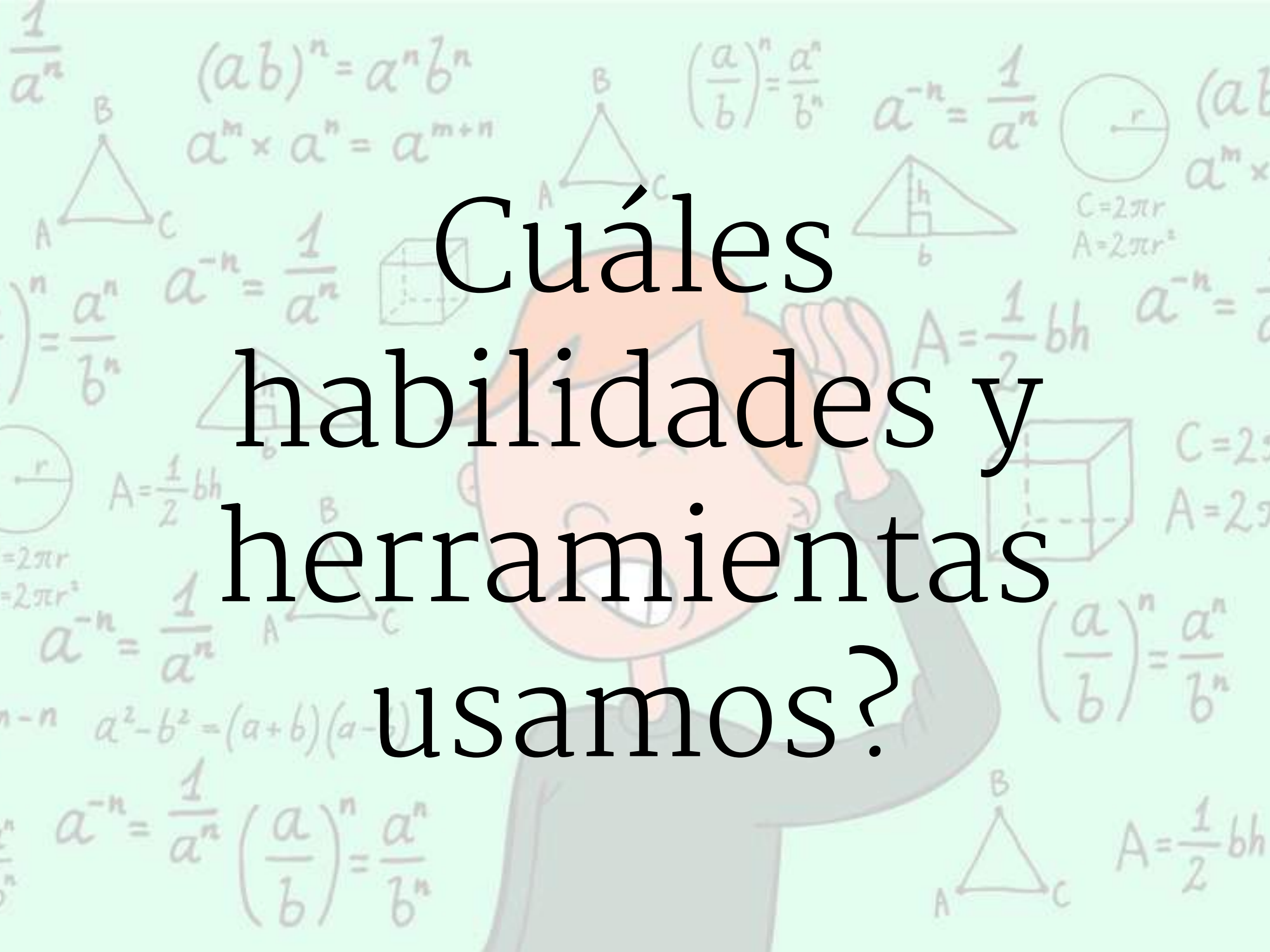


Poder de procesamiento

- Entre más experiencia tienes en la vida, mejor entiendes el mundo. En machine learning, entre más ejemplos le des al modelo, mejor predice. Para alimentar a la máquina con muchos ejemplos en un tiempo razonable, se necesitan computadoras rápidas. Ya tenemos esas computadoras.
- Según la ley de Moore, estas van volviéndose cada vez mas rápidas y menos caras.

En qué campos trabajan los Data Scientists?

- Casi todos. Medicina, agricultura, energia, gobierno, software.
- Empresas como Spotify, Etsy, Amazon, Facebook, Google, Twitter, Square.
- Digamos que hay Data Scientists que trabajan en academia y industria.
- Los que trabajan en academia se enfocan en avanzar el campo de un sentido teórico. Existen dentro del ecosistema académico y todo lo que exige: publicar recursos y papeles, pedir y conseguir concesiones, etc.
- Los que trabajan en industria están para, basicamente, enriquecer el negocio. Para mejorar un producto, ganar usuarios, construir sistemas ingenieriles, etc. Existen dentro de las realidades de trabajar como un ingeniero de software profesional: estar muy bien pagado, tratado y apoyado.
- Últimamente está apareciendo el Data Scientist académico trabajando en industria: avanzan el campo de Data Science de un sentido teórico, para que el negocio use sus conclusiones para enriquecerse. Se le exigen todas las mismas cosas a esta persona que el ecosistema académico, menos las concesiones: se le dan muchos fondos para hacer recursos, y se le paga un montón a su vez.
- Tu ejemplo de ingenieros hace 15 años y los data scientists de ahora

The background features a light green field filled with various mathematical concepts. It includes several algebraic identities such as $(ab)^n = a^n b^n$, $a^m \times a^n = a^{m+n}$, $\left(\frac{a}{b}\right)^n = \frac{a^n}{b^n}$, $a^{-n} = \frac{1}{a^n}$, and $a^{-n} = \frac{1}{a^n} \left(\frac{a}{b}\right)^n = \frac{a^n}{b^n}$. There are also geometric diagrams: a circle with radius r and formulas $C = 2\pi r$ and $A = 2\pi r^2$; a triangle with base b and height h and formula $A = \frac{1}{2}bh$; and a 3D cube. In the center, a cartoon character with orange hair and a grey shirt is shown from the chest up, with one hand on their chin in a thinking pose.

Cuáles habilidades y herramientas usamos?

Vamos de vuelta al diagrama de Venn de
antes.

Matemáticas

Estadísticos, machine learning, optimización

Estadísticos nos permite hacer cosas como: cuál tratamiento funciona mejor en mi sistema? Es Droga A, o Droga B, que sea mas eficaz en mis pacientes?

^Machine learning nos permite hacer cosas como: "Gracias por subir tu foto a Facebook. Hemos encontrado las caras de tu tía, tu madre, tu otra tía, tu hermano, y tu gatito Raúl. Quieres etiquetarlos?" Las matemáticas de machine learning son basadas en álgebra lineal, que es decir que no son realmente tan difíciles aprender.

^Optimización nos permite resolver el trayecto mas eficiente para visitar todas las capitales latinoamericanas en una sola gira en coche.



Ingeniería

Python, R, Scala

En teoría, podríamos implementar los modelos de arriba a mano, alimentándoles ejemplo por ejemplo con un lápiz y un papel de desecho. Lamentablemente, demoraríamos unas vidas en hacer que tal modelo pudiera distinguir entre una foto de un cocodrilo y la de un elefante. Por ello, usamos computadoras, que corren mucho más rápido que nuestros manos y mentes, y nunca se cansan.

^Primero, desarrollamos un solo modelo: intentando parámetros y representaciones diferentes, y validando los resultados. Para esto, usamos mayormente Python y R, que se nos hacen muy fácil hacer prototipos muy rápido, y que contienen muchísimas librerías de Data Science ya implementadas por la comunidad.

^Cuando tengamos un modelo que sirve, queremos construir un sistema ingenieril para hacer correr nuestro modelo cada cuando sin nuestra mano.

^R no nos sirve mucho para esto: es un lenguaje que no contiene casi sino librerías científicas; hacer un web app de nuestro modelo, por ejemplo, se nos hace difícil.

^Python es un lenguaje de uso general: librerías científicas constituyen una pequeña minoría de las cosas para las cuales se puede usar Python. Si queremos hacer ese web app de nuestro modelo, Python nos sirve mucho.

^Scala es otro lenguaje que va ganando popularidad en la comunidad.

Basicamente, si queremos construir un sistema súper rápida (porque el negocio o problema lo exige), nos sirve echar un vistazo a Scala.

El espacio del problema

Cuanto más lo entiendes,
mejor haces tu trabajo

Antes de siquiera escribir una línea de código, nos toca responder detenidamente a las siguientes preguntas:

^Cual es nuestro objetivo? Queremos investigar relaciones entre varios factores - por ejemplo, cuanto afecta la mala calidad del aire en Ciudad de Mexico, los costos de su sistema de salud? Por otro lado, nos interesa predecir algo - por ejemplo, dada la conversación actual en Twitter, quién va a ganar la presidencia en Estados Unidos?

^Cuál es la pregunta? Qué queremos modelar exactamente? Usando el ejemplo anterior, la conversación entre qué personas? Cuales geografías? Cuales fechas?

^Recuerda que nos toca enseñarle esta pregunta a una computadora. Una computadora no piensa por sí misma: hace exactamente lo que le digas que haga. Nunca más, nunca menos.

^Cómo definimos el éxito? Debe nuestra predicción sobrepasar un cierto nivel de confianza? A quiénes mostramos la predicción? Hay algunos a quienes no se la queremos mostrar (por ejemplo, para hacer un test que pruebe las adivinaciones de "expertos")?

^Es crucial entender que enterarte íntimamente del problema es la parte más importante. En mi opinión, es fácil creer que escribir código es lo único que equivale a productividad - que leer, deliberar y hablar tiene su lugar, pero entre más pronto que lleguemos a programar mejor. En realidad, te sirve muchísimo pasar la mayoría, aún la gran mayoría, del proceso en esta parte. Me tomó mucho tiempo para entender este balance y sentirme cómodo en él.

A person with their back to the camera stands in front of a wall covered in white chalk-like drawings. The drawings include various data science and business concepts: bar charts, line graphs, pie charts, a lightbulb, a network diagram, a city skyline, and the word 'FINANCE' written vertically. The person is wearing a light-colored shirt and dark pants.

Cómo llegar a ser Data Scientist?

Otra vez, es una cierta capacidad en matemáticas, programación y el problema en cuestión que queremos alcanzar.



Título universitario

Así que Data Science es una distinción relativamente nueva, los cursos al respecto son relativamente nuevos también. Como ejemplo, cuando yo entré en la universidad hace 9 años, básicamente no existían.

^Últimamente están apareciendo carreras en Data Science. Mientras más pasan los meses (o semanas, realmente), aparecerán muchas más.

^Este camino tiene sus pros y contras igual a cualquier carrera universitaria demás.

A group of people, mostly women, are sitting at long tables in a room, working on laptops. They are smiling and looking at their screens. There are many water bottles and some food containers on the tables. The word "Bootcamp" is written in large, black, serif font over the image.

Bootcamp

Además de carreras universitarias de Data Science, van apareciendo los bootcamps.

^El bootcamp es un nuevo paradigma de educación en el cual se aprenden las habilidades y herramientas prácticas que usamos en el trabajo.

^Este es un currículo intenso que se cumple en alrededor de 3 meses, asumiendo entre 40 y 60 horas de trabajo cada semana.

^Su ofrecimiento es bastante básico: las posiciones profesionales de Data Science exigen un cierto conjunto de habilidades y nosotros se las enseñamos en un tiempo corto. ¿De qué se puede quejar?



Tornado imparable

Esto es un término mío. Es mirarse en el espejo y decirse: "¿Qué me toca aprender? ¿Dónde se encuentra toda esta información? Sí, toda se encuentra en el internet abierto. Sí, Coursera me puede enseñar algoritmos de machine learning tan bien como cualquiera. Sí, competencias de predictive modeling me pueden dar mucha experiencia en resolver problemas del mundo real con herramientas de Data Science. Y sí, todo lo mencionado es efectivamente gratis acceder. Yo seré el tornado imparable. Yo voy a consumir todo lo que esté en mi camino. Yo no paré. Tu no me puedes detener. Yo me vuelvo el mejor en el mundo en esta cosa de Data Science. Yo soy el tornado imparable."



Escogiendo tu camino

Mi camino.

^Mi opinión, tocando temas de: costo, tiempo, burocracia, reconocimiento, eficacia de aprendizaje.

^Les sirve hacer un poco de todo.



Introducción a Python

Un workflow típico

Para el resto del curso, vamos a trabajar la siguiente pregunta: "¿Cual es la probabilidad que un tweet que viene de Colombia contiene la palabra 'yo?'" Se les enseña el workflow típico que se emprende para resolverla.



Formular la pregunta

Otra vez, una computadora no piensa por sí misma: hace exactamente lo que le digas que haga. No más, no menos. Entonces, por ejemplo, podríamos precisar nuestra pregunta de las siguientes maneras:

^"Que viene de Colombia" quiere decir que el tweet tiene coordenados geográficos dentro de los siguientes límites: "-78.31, 0.44, -70.71, 11.39."

^Asumimos que "yo" con acento, sin acento y de cualquier composición de letras minúsculas y mayúsculas constituyen la palabra original.

^"Que contiene la palabra 'yo'" quiere decir que una o más ocurrencias de la palabra "yo".

^En su conjunto, nuestra pregunta ya se lee: "Cual es la probabilidad que un tweet que origina dentro de los coordenados "-78.31, 0.44, -70.71, 11.39" contiene al menos una ocurrencia de la palabra 'yo' con cualquier composición de acentos y letras minúsculas y mayúsculas."

^Como bien ves, el ambigüedad no les queda bien a las computadoras.

Recopilar datos

En data science, conseguimos datos de varios tipos de fuentes: datos estadísticos, por ejemplo un sencillo CSV desde un repositorio público; datos que hemos recopilado a través de nuestro producto, por ejemplo los clics que han hecho nuestros usuarios; o un API.

^Un API es una interfaz que nos permite interactuar con un servicio y acceder sus datos a su vez. A mí me gusta pensarlo como un grifo de la cocina: le conectamos un tubo nuestro y nos llegan sus datos. Además, normalmente tenemos la opción de pasarle al grifo algunos parámetros que precisan los datos que queremos y los que no queremos.

^Vamos a acceder al API de Twitter, que se nos hace fácil recopilar los tweets que necesitamos.

^Accederemos a un corriente consecutivo de tweets, para que todos sean de la misma época corta. Tweets de hace mucho más tiempo - por ejemplo, aquellos de la época de la Copa América - se van a comportar de una manera fundamentalmente diferente.



Limpiar datos

Datos del mundo real son sucios. Para hacernos la vida mucho más fácil, emprendemos el proceso de "limpiarlos."

^Un tweet puede contener muchas cosas. Esperar que todos no contengan sino letras minúsculas bonitas sencillas sería bastante optimístico.

^Convertir todo en letras minúsculas.

^Quitar diacríticos.

^Quitar números.

^Quitar puntuación.

Dormir con tus datos

Dormir con tus datos es decir que al recibirlos, queremos familiarizarnos de una manera íntima: de inspeccionarlos y visualizarlos de tantas formas como es práctico. Entre mejor que hagamos esta etapa, con más eficacia y menos iteración podemos construir modelos en adelante.

^Concretamente, en esta etapa, puede que:

^Calculemos estadísticos sumarios. En nuestro ejemplo, usando los tiempos de interarribo, computamos cosas como el mean, median, maximum, 25th percentile, 75 percentile, variance, etc. Estas cosas nos dan una idea de como se comporta este valor.

^Creemos un histogram - un gráfico muestra, de una lista de números, cuales ocurren más.

^Visualizemos un variable en contra de otro. Supón que tenemos datos sobre cuantos tweets contenían nuestra palabra clave a diario el los últimos 6 meses y el calendario de la selección colombiana. Se surge más la palabra en tweets en los días en los que juegan la selección? Un gráfico se nos hace fácil entender esta relación.

^Intentemos explicar como son los datos a nuestros colegas. Obvio que no se entiende algo nada bien si no se puede articularlo.

Construir un modelo

- En el primer caso, no queremos enterar íntimamente de las relaciones que existan entre un input en un sistema y un output que nos importa.
- Por ejemplo, como afecta la hora a que se escribe un tweet la probabilidad de que contiene la palabra "yo?"
- Usualmente, usamos este conocimiento para alterar un proceso del negocio. Por ejemplo, si estamos creando un bot que va a retwittear todos los tweets que contienen nuestra palabra, pero no tenemos los recursos para hacerlo correr todo el día, usaríamos el conocimiento de que los tweets matutinos tienen más probabilidad de contener la palabra, y por eso lo desplegaríamos por la mañana en vez de la noche.
- En el segundo caso, no interesa más la capacidad de predecir.
- Sobra decir que cuanto más entendemos las relaciones entre los inputs y outputs, mejor construimos o diseñamos el modelo.
- Dicho eso, en este caso, la tarea es más, dados varios inputs, predecir el futuro.
- Por ejemplo, dado el género del autor del tweet, la hora a la que se escribió, y los coordenados geográficos exactos, predecir la probabilidad de que el tweet contiene la palabra.
- Antes de platicar más sobre por qué queríamos construir un modelo, quiero hablar porque no lo queríamos en el primer lugar.
- Un modelo nos permite preguntar todas las cosas y recibir todas las respuestas. Imagínate que tengamos un dado sencillo. Si lo rodáramos 10,000 veces, aproximadamente cuántos "unos" podemos esperar? Ya sabemos. Cual es el valor máximo que veremos? Ya sabemos. Se realizan más "doses" o "cuatros?" Ya sabemos. Un dado es un modelo de un proceso estadístico, que nos enteró todo de él.
- En realidad, es mucho más común observar datos, e ir intentando reconstruir el proceso que los generó en el primer lugar. Así lo hacemos en nuestro ejemplo.

^Aquí, suponemos que nuestro proceso se puede modelar como un "beta distribution."

Validar su modelo

Predecir si un tweet en el futuro contendrá la palabra "yo" no es nada grave. Por otro lado, predecir si alguien tiene cáncer es algo diferente. Entonces, por cuestiones del costo (en los varios significados de la palabra) de la predicción, es crucial que validemos nuestros modelos antes de desplegarlos.

^En el caso de predecir cáncer, quisieramos tener una idea de la precisión del modelo antes de usarlo en pacientes reales. A este fondo, nos encargamos de evaluarlo en datos que, crucialmente, no fueron usados para entrenarlo, tal como los de los pacientes nuevos. Dicho de otra manera, nos encargamos de evaluar la precisión del modelo en datos que jamás había visto antes.

^Este proceso se llama cross validation. En el caso básico, dividimos nuestros datos en dos partes: entrenamos el modelo con la primera, y lo evaluamos con la segunda. Lamentablemente, este tema es un poco fuera del alcance del curso. Lo cubriremos en el curso avanzado.

^En nuestro ejemplo de los tweets, por efectos prácticos, lo vamos a omitir.



Predecir el futuro

Con un modelo entrenado, somos capaz de predecir eventos futuros. En nuestro ejemplo, podemos hablar muy específicamente sobre las probabilidades de que tweets futuros contengan nuestra palabra.

^En casos más de machine learning, ingerimos inputs para predecir la clase de una observación. Por ejemplo, puede que ingeramos una foto recién subida a Facebook para predecir cuál cara de amigos tuyos esté representada.

^Es importante notar que un modelo sirve mayormente como herramienta de automatización. En el ejemplo anterior, existe efectivamente un conjunto de reglas que engarza los colores de pixeles en una foto con el nombre de la persona representada. Por ejemplo, que tal color en tal y tal y tal pixel y otro color en tal y tal y tal pixel significa que estamos mirando una foto de nuestra amiga Laura. Así que cada foto contiene un montón de pixeles, este conjunto de reglas será demasiado complejo para nuestra comprensión. Y aún si lo pudieramos comprender, demoraríamos horas en clasificar cada foto manualmente. Por eso, utilizar modelos que automatizan el proceso.



Comunicar resultados

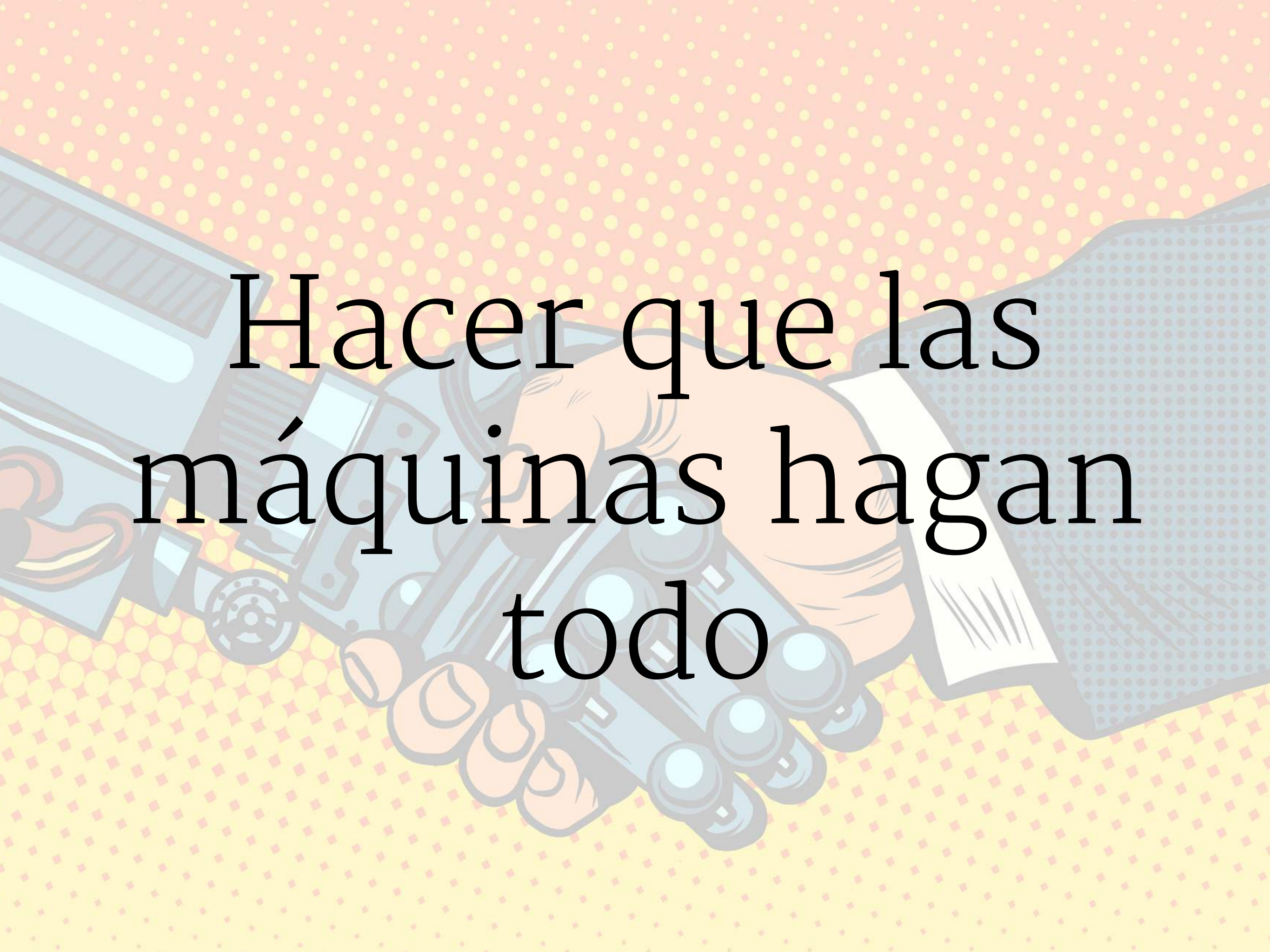
Como Data Scientists, nos pagan para entender el mundo de una manera más matemáticamente rigurosa que los demás. Por ello, nos toca poder comunicar nuestros resultados no solo con otros Data Scientists, sino con el jefe de ventas, el pasante en marketing, el CEO y nuestra madre. ^Comunicarse bien es algo santo. Es otra inteligencia por su propio derecho. Es muy difícil y vale la pena enfocarte en hacerlo muy bien.

^Otra vez, comunicar bien los resultados hace que:

^Otros miembros del negocio entiendan lo que Data Science puede hacer para la organización

^La gerencia entienda que la inversión que hizo valió el tiempo y la plata

^Entendamos nuestro propio trabajo aún mejor. Fue Einstein que dijo: "si no puedes explicar una cosa de una manera muy simple, no la entiendes nada bien."



Hacer que las máquinas hagan todo

Hasta ahora hemos cubierto dos lados del triángulo: el de las matemáticas y el del negocio. Ya se nos encarga el último: el de ingeniería.

^Después de construir un modelo, no queremos estar presionando botones cada noche nosotros para hacerlo correr. Por ello, construimos un sistema ingenieril alrededor de él.

^Así, dejamos a las computadoras a que hagan nuestro trabajo mientras dormimos nosotros.

^En realidad, estamos simplemente agregando otro estrato de automatización. Así nos enfocamos en el próximo problema del negocio.

Una casa en el espacio vectorial

- Como ya hemos dicho, modelos matemáticos no entienden sino números. No pueden interpretar fotos en su forma original, texto crudo, etc.
- Por ello, en el fondo, muchas tareas de data science nos piden codificar cosas en términos numéricos. Esto nos permite utilizar e interactuar con modelos y herramientas fundamentales en data science. Además, representar entidades numericamente nos permite descubrir, y luego visualizar, similitudes entre ellas.
- Hay varias estrategias canónicas para codificar texto, fotos, futbolistas, etc. Dicho eso, esta es la tipa de cosa que no tiene "límite:" cuanto más creatividad le llevas al proceso, más formas encuentras.
- Por ejemplo, se puede codificar un producto de e-commerce con valores de su peso, su valor, sus dimensiones físicas, etc. Dado esto, qué tal representarlo por todas las personas que lo han comprado en los últimos 30 días? O el número de veces que se ha comprado en cada de las últimas 25 semanas? Tener "vistas" diferentes de tus datos siempre se nos hace bien.



Métricos de distancia

- Con datos codificados en el "espacio de vector," podemos ejecutar varios métricos de distancia para encontrar la similaridad (que es el opuesto de la distancia, en términos cualitativos y técnicos) entre ellos.
- Métricos canónicos se nombran: distancia euclidean, distancia cosine, distancia jaccard y distancia manhattan.
- Cada uno se les presta a situaciones diferentes.

Un sencillo sistema de recomendación



Cierre del curso

- ^Lo que es un Data Scientist
- ^Como llegar a ser Data Scientist
- ^Las herramientas y habilidades que necesitamos
- ^El workflow típico para un problema sencillo
- ^Afortunadamente para aquellos que quedan inmensamente emocionado por aprender más, cada sección y subsección de este curso merece un curso entero en su propio derecho.
- ^Por otro lado, desafortunadamente para aquellos que quieren un camino cortito para volverse Data Scientist: nos falta un montón.
- ^En ambos casos, Data Science es un campo extramadamente divertido por sobre todo. Los invito a que sigan el camino con nosotros.
- ^Invitación a resolver los desafíos y tomar el examen
- ^Invitación a dejar dudas en la plataforma o en