

Lecciones aprendidas haciendo productos de datos.

Data

on

the

rocks

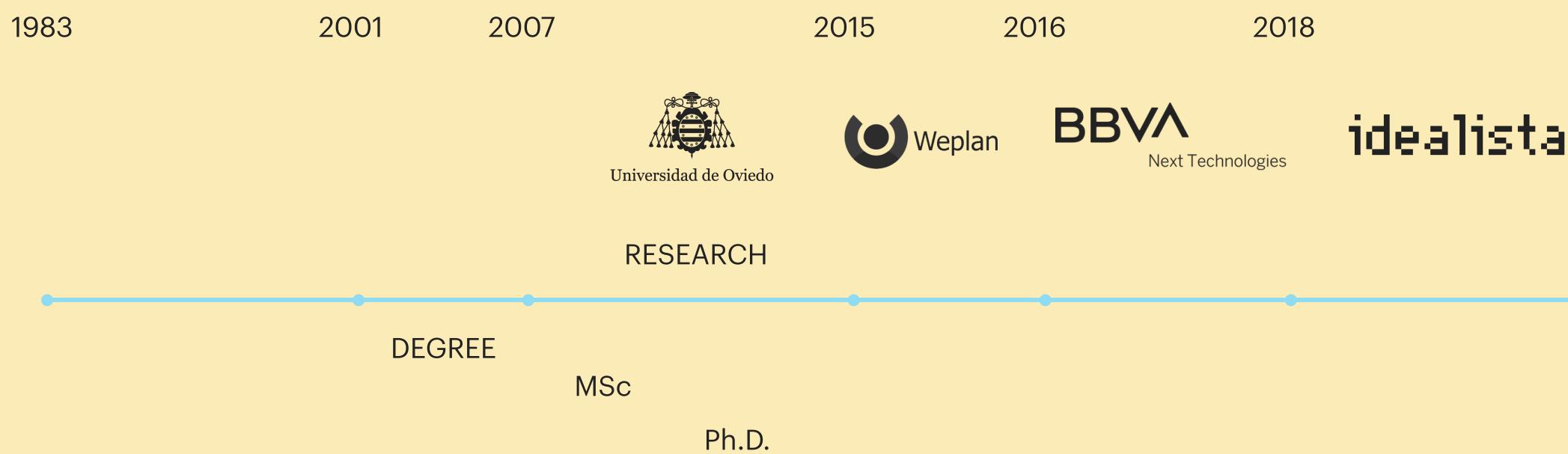
Data on the rocks es un evento presencial, bimensual, que organizamos desde Garaje de ideas y Gen/D en Madrid. En cada ocasión tratamos de un tema de actualidad en el mundo del dato, de la mano de los mejores expertos.

En esta oportunidad, en el mes de enero, hemos contado con la presencia de

Pelayo Arbués, Head of Data Science @ Idealista, quien nos ha contado sus lecciones aprendidas haciendo productos de datos.

Pelayo Arbués hoy día se desempeña como Head of data science en Idealista, pero previo a dicha posición ha recorrido un gran camino del que, sin dudas, ha aprendido muchas lecciones.

Si bien ha comenzado estudiando economía, luego de doctorarse y dedicarse varios años a la investigación, ha optado por cambiar de rumbo y aprender a programar. Sus inicios han sido en una startup llamada Weplan, luego continuó su carrera en la consultoría y en 2018 ha comenzado a trabajar en Idealista, particularmente en una central de la misma enfocada en Idealista Data, la vertical de negocios con la que venden datos a terceros y realizan valoraciones inmobiliarias.



¿Qué son los productos de datos?

Un producto de datos es una solución basada en datos que facilita un fin específico.

Dicho de otro modo, es un producto que está basado en datos que habitualmente se utiliza para tomar decisiones y que, por lo tanto, intenta cubrir una necesidad. Es decir, tiene una finalidad.

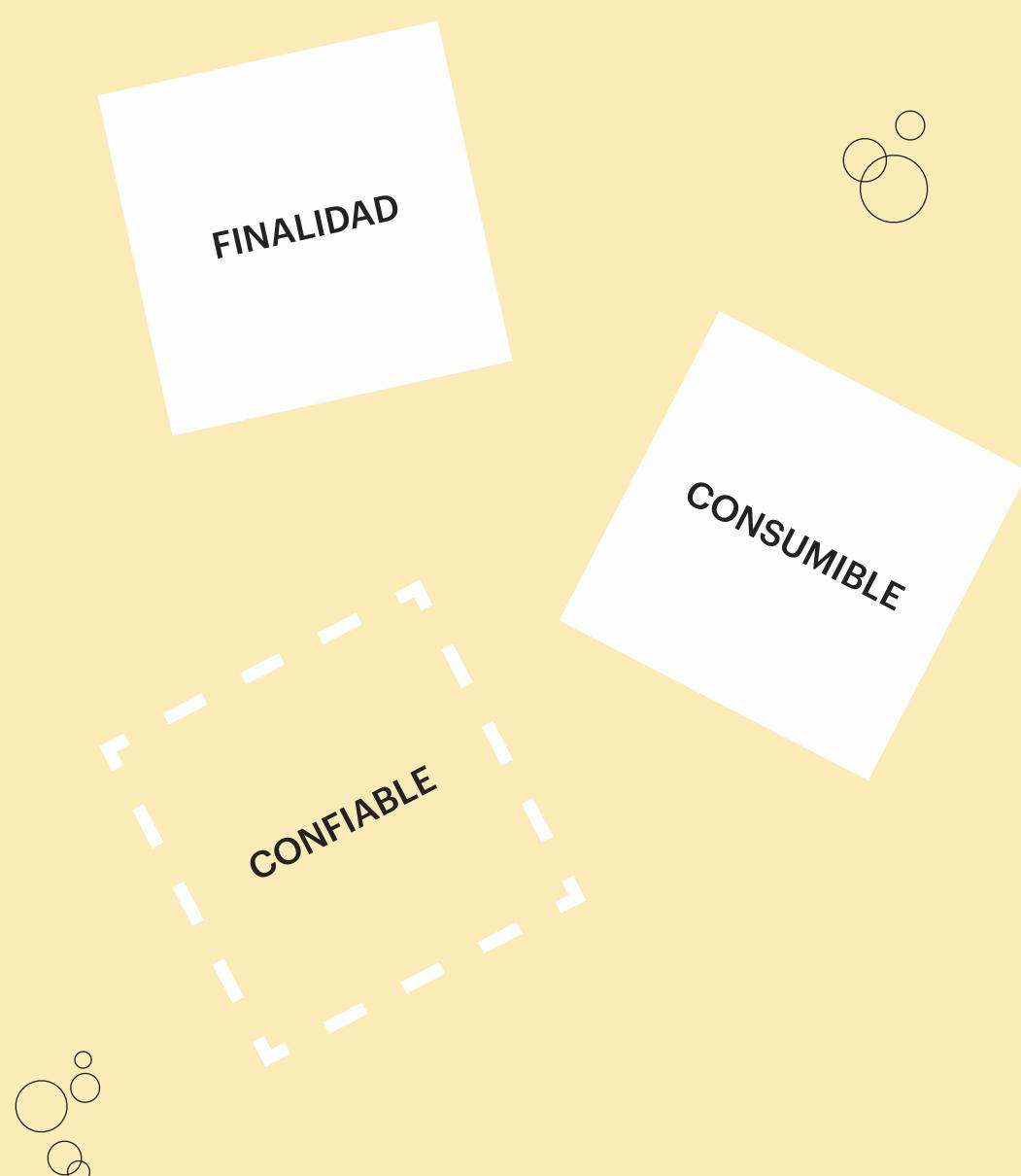
Pero... **¿Qué diferencia hay entre el dato como producto y el producto de datos?**

Muchas veces se considera el dato en bruto como producto de datos, pero este simplemente recoge información, no ha sido refinado y no tiene una finalidad específica que sirva para tomar una decisión.

En cambio, el producto de datos debe tener ciertas características con las que no cuenta el dato en bruto. Las mismas son: finalidad, confiabilidad y consumo.

Características que tradicionalmente se conocen como **Data Governance**.

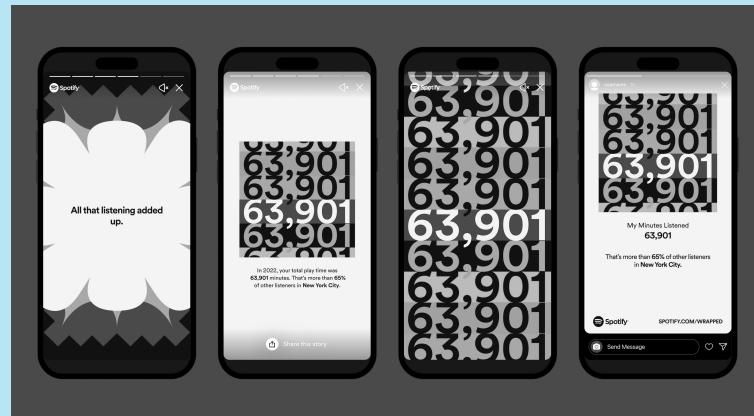
La gobernanza de lo que hacemos, pone orden y da sentido al trabajo en cuestión.



Tipos de productos de datos

1

Datos de conocimiento

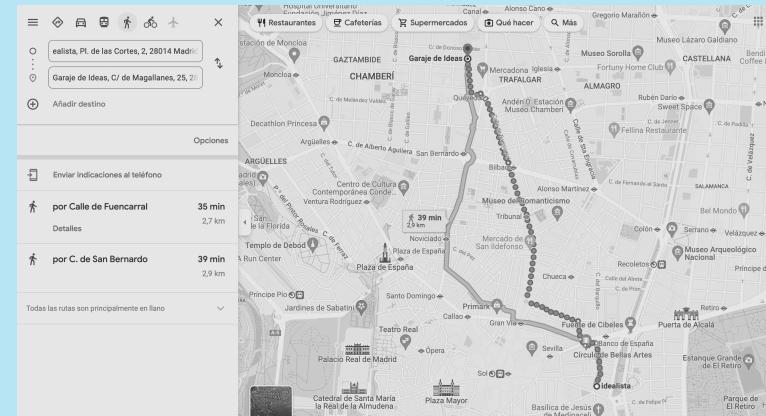


Son los clásicos insides que sirven para saber cómo van o cómo han ido las cosas. Ayudan en el proceso de la toma de decisiones.

Por ejemplo, el clásico BI, el reporting o cuando Spotify te hace saber a final de año cuántos minutos has escuchado algún artista y cuál es tu lista favorita, entre otras cosas.

2

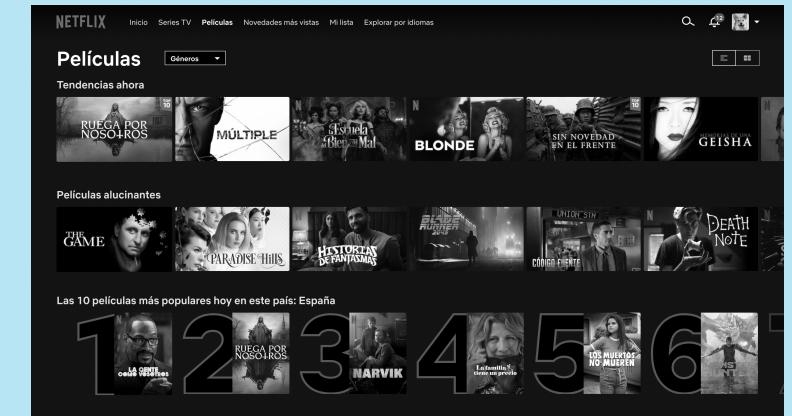
Optimización de procesos



Mejoran la eficiencia de las operaciones que tienes que llevar a cabo. Su finalidad es reducir los tiempos o los costes y extraer el máximo potencial del dato para optimizar la experiencia de productos existentes o crear nuevos. Por ejemplo, el routing de Google Maps.

3

Mejora de experiencia



Explotación o activación del dato para mejorar la experiencia del producto existente o la creación de nuevas líneas de ingreso. Por ejemplo, en Idealista la mejora de experiencia está directamente relacionada con la detección de mensajería spam y con el sistema de recomendación de inmuebles, entre otras cosas.

Lecciones aprendidas

La cara mundana y poco glamorosa de los datos

1

Gobierna los datos

Es muy habitual entre las personas que trabajan con datos diferir en la medición de los mismos y que ello repercuta en un conflicto. Mucho tiene que ver Excel y esto se debe a que funciona como un silo de información que, por más que se comparta, no se disponibiliza a toda la compañía y, como no está gobernado, resulta muy difícil de trazar.

La solución: el **Data Governance**. Este evita los datos en conflicto estandarizando los conflictos, las redundancias e identificando cuáles son los problemas que tienes en los datos, descubrirlos y reutilizarlos.

Un gran aliado son los **metadatos** que indican quien utiliza esa fuente de datos, cuanto se consume, en qué otros sitios se está consumiendo y permite seguir la traza de una manera más o menos automática.

¡Pero atención! Cuidado con la **reutilización de datos**. Muchas veces la gente acaba tomando decisiones reutilizando un producto que existe sin saber ni para qué se generó, si está actualizado o si está funcionando bien, básicamente porque no hay ningún proceso de calidad del dato.

Y para tomar decisiones puedes basarte en tu experiencia, hacer research cualitativo o puedes buscar en los datos. Porque la finalidad, no olvidéis, es que tu empresa tome mejores decisiones en ambientes de incertidumbre. Y si se tiene todo organizado y documentado, los metadatos indican quién lo creó, cuándo se actualizó y si se sigue actualizando o no, puedes crear una traza, un seguimiento, un linaje.

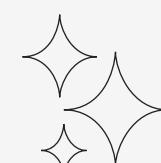
2

No supongas que el dato es de calidad

Lo primero que tienes que hacer cuando haces un producto de datos es mirar su calidad, ya que muchas veces no es evidente que está roto.

Es una falacia considerar que los datos son objetivos y perfectos. No pueden serlo porque son recogidos, tratados, transformados y, a partir de ello, se toman decisiones desde la propia sensorización; tú estás eligiendo qué sensorizas. Y no pueden ser perfectos porque son simplificaciones de la realidad y cuando modelas encima de los datos simplificas aún más la realidad y la comprimes.

Es importante tenerlo en cuenta a la hora de trabajar con datos porque lo que quieras es que la cosa apunte en la dirección correcta y que no se desvíe mucho.



3

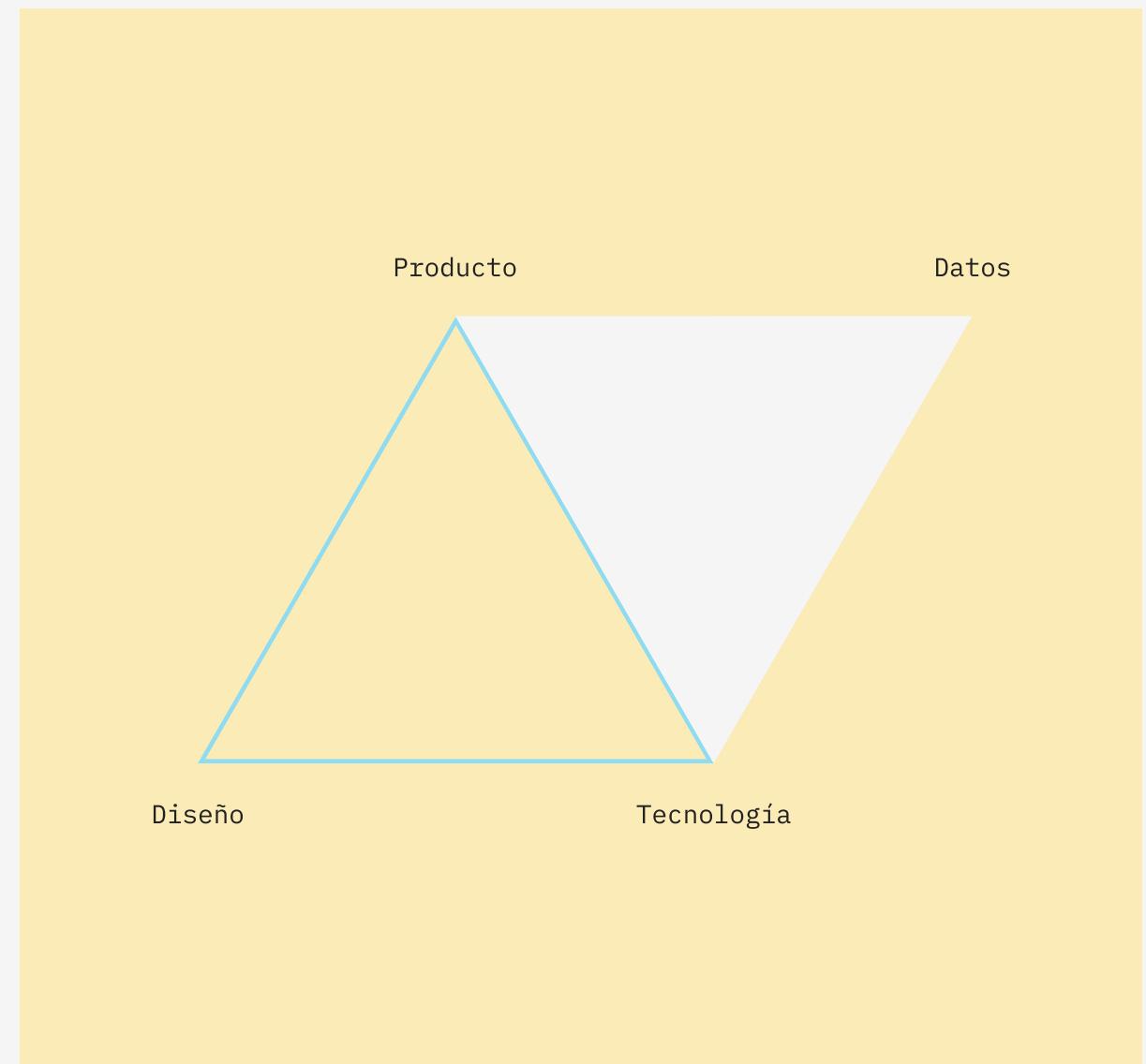
Si no tienes datos, créalos

Una de las primeras cosas a tener en cuenta a la hora de llevar a cabo un proyecto es si el dato con el que contamos es el mínimo que necesitarías (**MVD-Minimum Valuable Data**).

Muchas veces los datos con los que contamos no valen para tomar una decisión super fina, pero de todos modos se puede evidenciar una tendencia general de una población que sea un poquito más amplia a esa exacta a la que estás mirando.

Es importante definir cuál es el mínimo de datos que se precisa para tomar una decisión, si sirven o no y, en caso de que no sean suficientes, crearlos.

Generar datos es caro y hacerlo de manera inteligente es una operativa muy fuerte, pero de todos modos hay herramientas muy útiles como la generación de datos sintéticos o el active learning que es muy buena la hora de anotar datos.



4

Trabajar en equipo

Hoy día es muy importante hablar de **product trio**, producto-diseño-tecnología.

Decisiones que antes las tomaba un solo departamento, ahora las toman en conjunto.

Pero como el equipo de datos no forma parte del product trio muchas veces llega tarde al producto. Y ello genera que no se pueda avanzar desde el principio. Por ejemplo, en el caso de que haya un sistema de personalización de datos, se podría avanzar recogiendo datos o etiquetando, entre otras cosas.

Al fin y al cabo, la finalidad es resolver un problema de negocio y más allá del equipo, cuantos más cerebros trabajen en la solución, mejor.

5

A veces el dato despierta ideas

La combinación entre entender el negocio y entender el dato permite descubrir muchas oportunidades. Tener tiempo para pensar e investigar permite al equipo de datos entender el negocio y descubrir los puntos de dolor. Lo que ayuda a crear productos y soluciones que harán que aquellos que no pertenecen al equipo puedan entender el potencial que tienen estos datos.

Estandarizar procesos y tener herramientas que generen eficiencia ayudan a mantener la calidad del dato y a controlar mejor las cosas. Montar un sistema que permita mantener muchos proyectos con muy pocas manos es una gran solución para generar tiempo, poder pensar y tener nuevas ideas. **Al fin y al cabo, entender el negocio y ayudar a que el negocio entienda el dato, generará que este lo conozca y, por lo tanto, pueda consumirlo de la manera más eficiente.**



6

Abraza la incertidumbre

La gestión de proyectos de datos es aún peor que en proyectos de software porque los productos de datos son altamente inciertos; no es hacer un desarrollo que has hecho en reiteradas ocasiones y que puedes definir cuánto tiempo llevaría. Cuando incorporas el dato en la ecuación, el mismo ingresa con un componente de variabilidad propio del mismo, por lo que generalmente el proyecto tiene mucha incertidumbre.

¿Qué metodología de trabajo utilizáis? ¿Sois ágiles?

Muchas veces, cuando se trabaja el dato se encuentran situaciones a resolver que, si no se trabaja con un deadline, desvirtúan el proyecto y generan que uno intente resolver un problema que igual ya no es el conflicto original.

Por un lado, trabajar con la incertidumbre es muy difícil porque debes contemplar la posibilidad de que surja un inconveniente y que este lleve tiempo, pero por otro también dificulta la tarea de trabajar con temas que no apetecen.

Para evitar el desgaste, una solución puede ser compaginar las labores arduas con otros proyectos que tengan un recorrido más largo, pero que sean más interesantes. De este modo, se procrastina menos porque a la gente le apetece más hacerlo y provoca que quieran terminar rápido el trabajo operativo, puesto que tienen ideas para el otro proyecto.

La ciencia de datos es un proceso parecido a la investigación

Se parte de una hipótesis, se recogen datos, se analizan y muchas veces puedes descubrir un problema de los datos que genere que se deba volver a reformular la hipótesis.

Es importante brindarle el contexto a la gente de datos, aclarar cuál es el problema en cuestión y no solo demandar un dato; porque en muchas ocasiones hay una forma mejor de resolverlo que no sea simplemente sacar un dato.

7

Comunica creando

Muchas veces por la falta de experiencia de la gente en producto, analítica y datos, cuesta mucho que digan qué necesitan. A veces tampoco tienen definidos los requerimientos o el problema.

En estos casos, **es muy útil prototipar en pocas semanas y colocarlo frente al cliente**, así le brindamos la posibilidad de comunicar si es o no lo que esperaban del producto.

Si bien al principio puede ser frustrante que digan que no porque piensas que has perdido tiempo, en realidad has ganado mucho tiempo en el que estás conociendo las necesidades de tu cliente.

8

La experimentación no es fácil

Parte del valor del producto de datos proviene de ahorrar tiempos, pero ¿cuántos experimentos hacéis al mes? Lo ideal es que lo hagáis todo el rato, pero para ello debes tener una plataforma bien montada, y tiempo. Obviamente, para una startup la situación no es la misma, ya que con muy poco volumen de usuarios es muy difícil la experimentación.

Hay veces que no tienen el volumen suficiente y deben tener los experimentos corriendo mucho tiempo y eso lo invalida. O miden efectos que son demasiado pequeños como para medirlos con confianza.

Pero experimentar no es hacer 3 o 4 experimentos, experimentar es perder tiempo y liarla, por ejemplo, porque se ha planteado mal la recogida del dato por cuestiones de calidad del dato que no se han considerado o porque se han desplegado cosas por poner foco en otro tema y resulta que no se ha mirado otro punto importante en concreto mientras se rompía otro muchísimo más importante en otro lado.

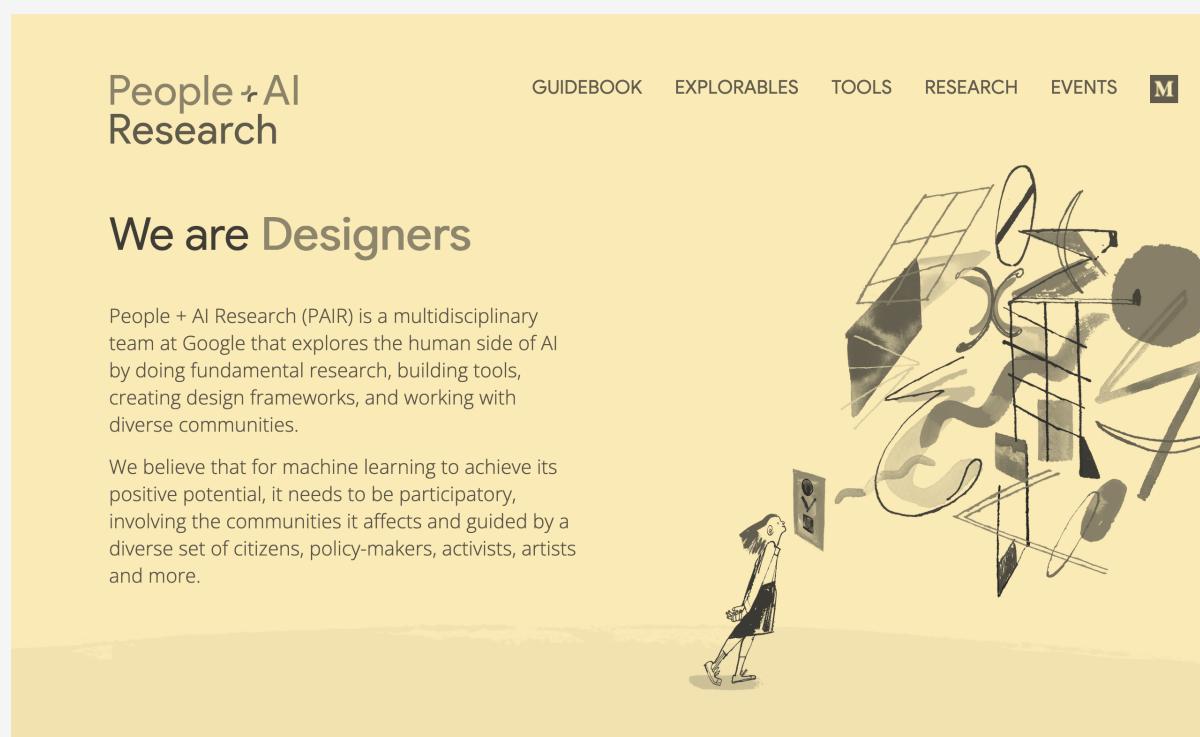
9

Da explicaciones a tus usuarios

No escondas la incertidumbre de tus estimaciones, sé transparente sobre lo que tu algoritmo puede y no puede hacer. **Es importante brindar al cliente una explicación, un análisis y ayudar a gestionar las expectativas.**

La gente quiere un número, no quiere un intervalo de confianza ni un rango, quieren un número concreto, aunque no sea representativo de lo que estamos haciendo. Y la realidad es que, al entregar un número, lo que se está haciendo es enterrar la incertidumbre y se vuelve peligroso porque, al fin y al cabo, no se sabe con qué se está tomando la decisión. Es valioso siempre explicar los beneficios, no la tecnología o lo que está por debajo.

Un recurso muy interesante es: **Google People + AI guidebook**



Une el algoritmo y el humano.
Muy útil para explicar mejor a
los clientes lo que puede llegar
a lograr tu solución de
productos de datos.

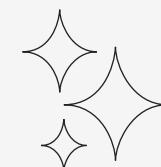
10

Usa IA si no te queda más remedio

Utiliza **inteligencia artificial** o **machine learning** si no te queda más remedio, porque el grado de complejidad que añade cuando lo tienes en producción es brutal.

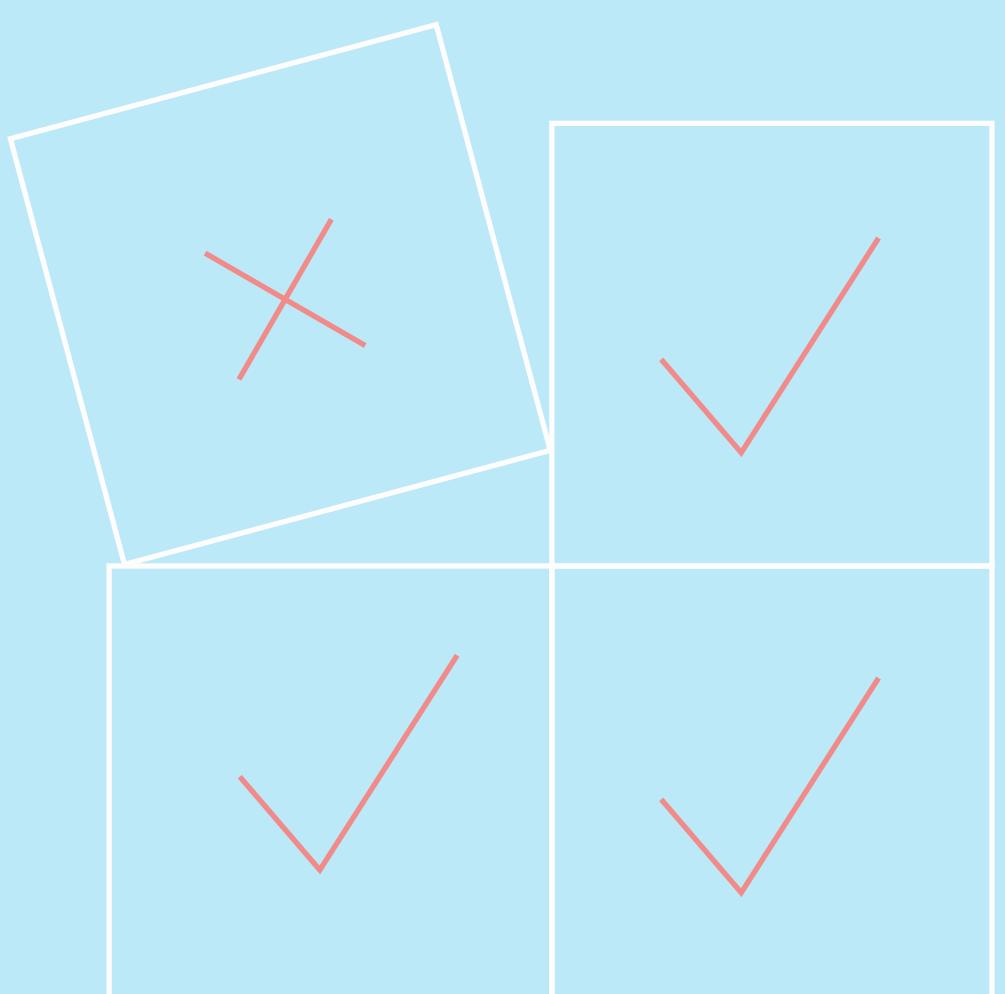
Si bien muchas veces un algoritmo resolvería la cuestión, hay que plantearse cuál es la ganancia de tener ese algoritmo y el coste de tenerlo en los sistemas corriendo.

Muchas veces las reglas heurísticas no son de gran fiabilidad o precisión, pero son mucho más fáciles de mantener y, en el punto que las reglas heurísticas son tan complicadas que no las puedes mantener, pueden ser compensadas con el algoritmo más complejo.



En conclusión

- La gobernanza de datos suena aburrida, pero es muy necesaria, poner orden suele ser muy rentable.
- Asegúrate de que los datos pasan por unos mínimos controles de calidad.
- Si tienes una idea y no tienes datos, trata de crearlos.
Cuando no existen, se pueden generar datos, es importante usar la imaginación, ya que hay datos que pueden servir como proxy o como primera interacción o datos públicos que se pueden traer de línea o de donde sea.
- Juega en equipo con otros perfiles y deja que los datos también descubran oportunidades. Ve los proyectos como un problema de equipo, no veamos las empresas como cajas, hay que verlo como gente intentando resolver problemas. Cuantos más cerebros diversos hay, mejor.
- Sé transparente sobre para qué sirve tu producto y cómo se gestionan los datos. Cuando se utilicen algoritmos, explícale a la gente que se está usando.
- Preguntaros: ¿realmente necesitas usar IA o es suficiente con un heurístico o con algo mucho más sencillo?



Contactos

Pelayo Arbués, Head of Data Sciene @ Idealista.

Correo: gonzalezpelayo@gmail.com

Linkedin: <https://www.linkedin.com/in/pelayoarbues/>

Óscar Casado, Head of Data @ Garaje de ideas.

Correo: oscar.casado@garajedeideas.com

Linkedin: <https://www.linkedin.com/in/oscar-casado-610513279/>

<https://garajedeideas.com/>

Arancha García, Directora y CoFundadora @ Gen/D.

Correo: arancha.garcia@gend.es

Linkedin: <https://www.linkedin.com/in/arancha-garcia-garcia/>

<https://gend.es/>



Garaje de ideas



Gen/D

Data, Design & Digital