

Laboratorio de datos, clase 1

Introducción a la materia

Prof. Enzo Tagliazucchi

tagliazucchi.enzo@googlemail.com

www.cocuco.org

Docentes



Enzo Tagliacruz
Prof. Adjunto (DF)
enzo@df.uba.ar



Sebastián Pinto
Ay. Primera (DF)
spinto@df.uba.ar



Tomás Cicchini
Ay. Primera (DF)
tomas.cicchini@gmail.com



Ariel Berardino
Ay. Segunda (DF)
ariberardino@gmail.com

Página web de la materia

<http://materias.df.uba.ar/lda2021c1/>



[Principal](#)

[Programa](#)

[Cronograma](#)

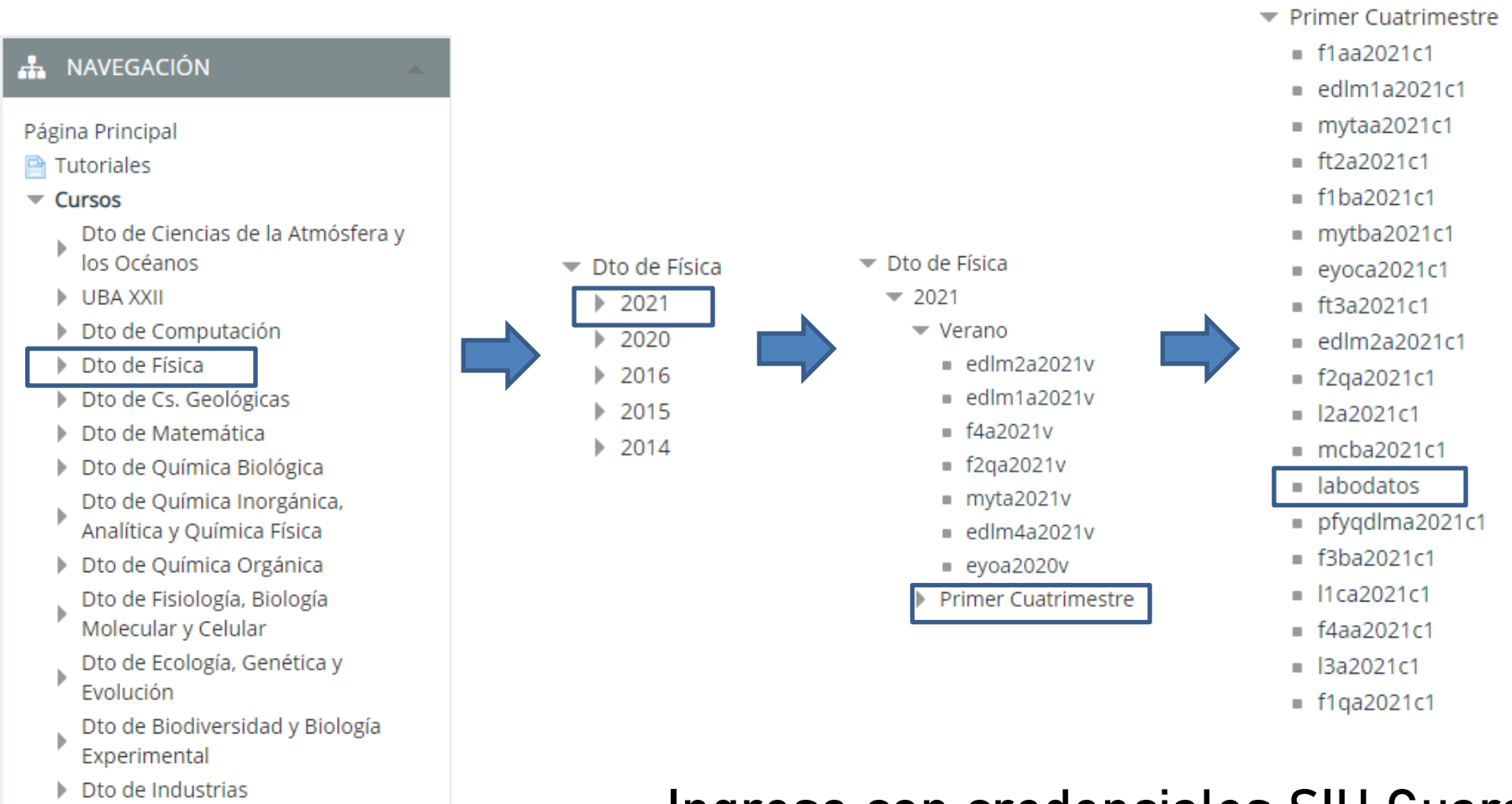
[Ejercicios a entregar](#)

[Régimen de aprobación](#)

[Bibliografía](#)

Campus virtual

<https://campus.exactas.uba.ar/>



Ingreso con credenciales SIU Guaraní

Formato de la materia

- Clases sincrónicas teórico-prácticas de 3 horas
- Mail unos días antes de cada clase para avisar temas
- Todas las clases grabadas y subidas al canal de YouTube del DF y al campus virtual
- Todos los apuntes (por ejemplo, este) y notebooks de Python subidos al campus virtual

¿Para quién es la materia?

- Materia obligatoria para la nueva Lic. en Cs. de Datos
- Materia optativa para las demás carreras de Exactas (espero 5 puntos)
- Materia optativa de doctorado (en trámite)

Criterio de aprobación

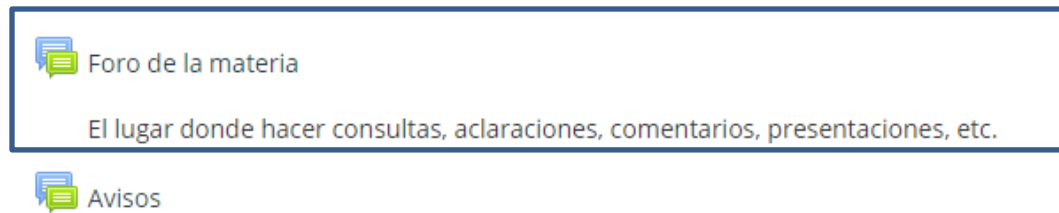
- No es necesario asistir de forma presencial (se puede cursar de forma asincrónica)
- Sí es necesario entregar las cuatro ejercitaciones individuales antes del deadline (aprox. 10 días)

9/4	Visualización de datos. Tipos de plots básicos y cuando usarlos: scatter, barras, boxplots, viol
	ENTREGA DE CONSIGNA EJERCICIOS INDIVIDUALES 1
13/4	Regresión lineal, cuadrados mínimos. Introducción a scikit-learn. Regresión lineal, cuadrados
16/4	Regresión lineal, polinomios. Concepto de overfiteado.
20/4	Regresión logística. Clasificación vs. regresión.
	FECHA LÍMITE ENTREGA EJERCICIOS INDIVIDUALES 1

- Trabajo final en grupos de 3 alumnos
- Nota final: 30% ejercicios, 70% trabajo final

Consejos para la materia

- Intenten formar grupos de estudio para ayudarse mutuamente (igual van a tener que armarlos para el TP final)
- Entregar la ejercitación es obligatorio: entreguen siempre algo, aunque no les guste.
- Usen el foro en el campus virtual para hacernos preguntas:



Notebooks de Python



Jupyter notebooks: combinación de código, texto e imágenes que se pueden correr desde un navegador

www.jupyter.org



Notebooks de Google Colab: combinación de código, texto e imágenes que se pueden correr desde un navegador en la nube de Google

<https://colab.research.google.com/>

Notebooks de Python

```
[ ] x0 = 1 # condicion inicial
    N = 1000
    dt = 2/1000
    x_tiempo = np.zeros(N) # crea un vector de longitud N donde vamos a ir poniendo las soluciones
    x_nuevo = x0
    x_tiempo[0] = x_nuevo
```

Ahora el bloque que se repite desde 1 hasta N lo corremos con una función llamada "for" que itera los números en una lista desde 1 hasta N, que creamos con el comando de numpy `np.arange(1,N)`

```
[ ] for i in np.arange(1,N):
    x_viejo = x_nuevo
    x_nuevo = x_viejo + x_viejo*dt # aca ponemos f(x_viejo)dt. En este caso, f(x_viejo) = x_viejo
    x_tiempo[i] = x_nuevo
```

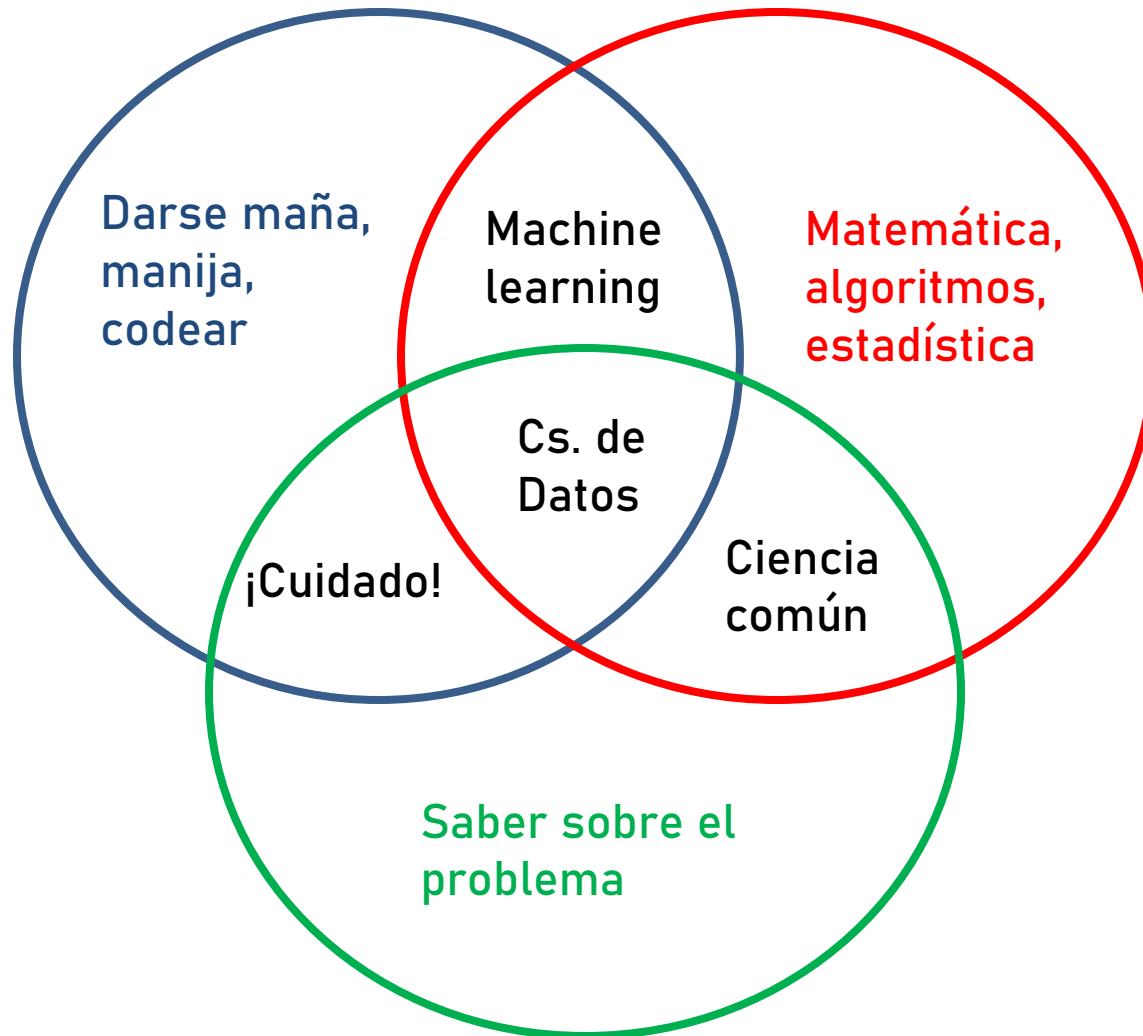
Entonces, en `x_tiempo` quedaron las soluciones de $x(t)$ en 1000 puntos separados por `dt` entre $t=0$ y $t=2$. Podemos graficar esto usando el comando `plt.plot`, donde antes tenemos que importar la librería `matplotlib.pyplot`

Cada clase va a estar acompañada por un Notebook

Los ejercicios se entregan como Notebook, con explicaciones (texto) y códigos que al ejecutarse cumplen la consigna (más sobre esto hoy)

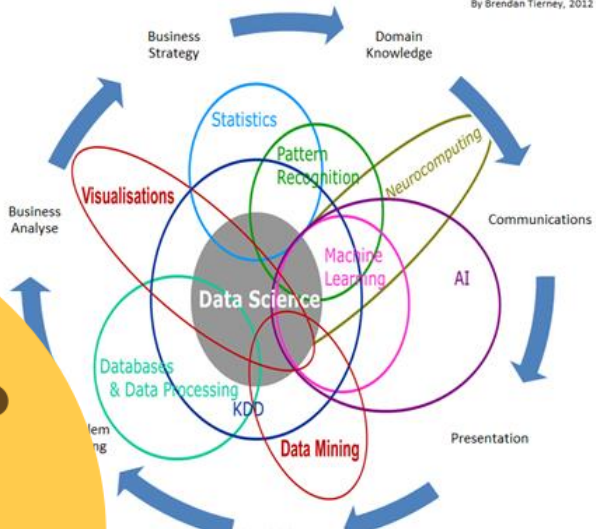
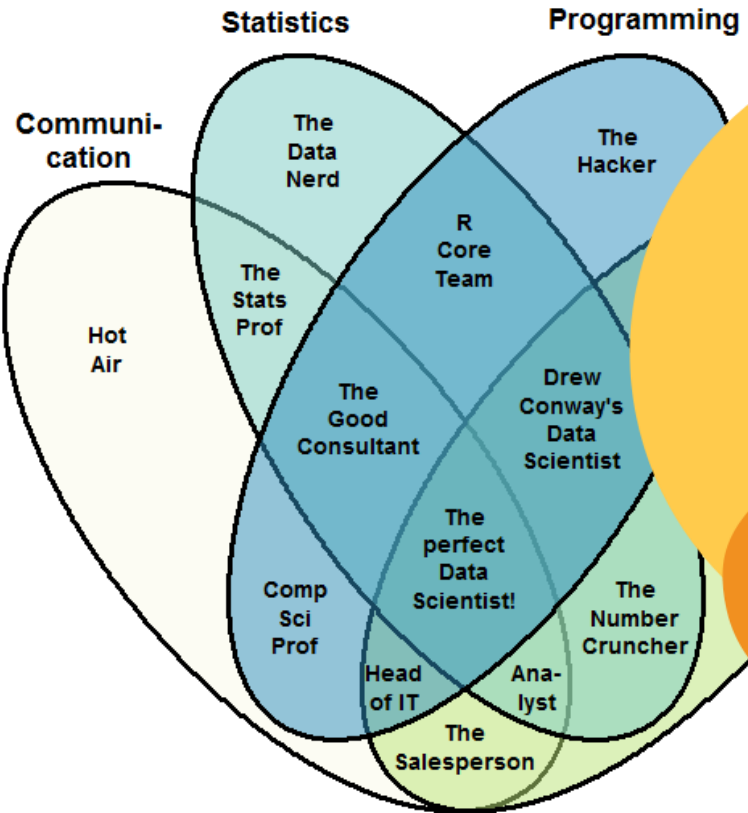
¿Preguntas?

Qué es la Cs. de Datos (primer intento)



Qué es la Cs. de Datos (segundo intento)

The Data Scientist Venn Diagram



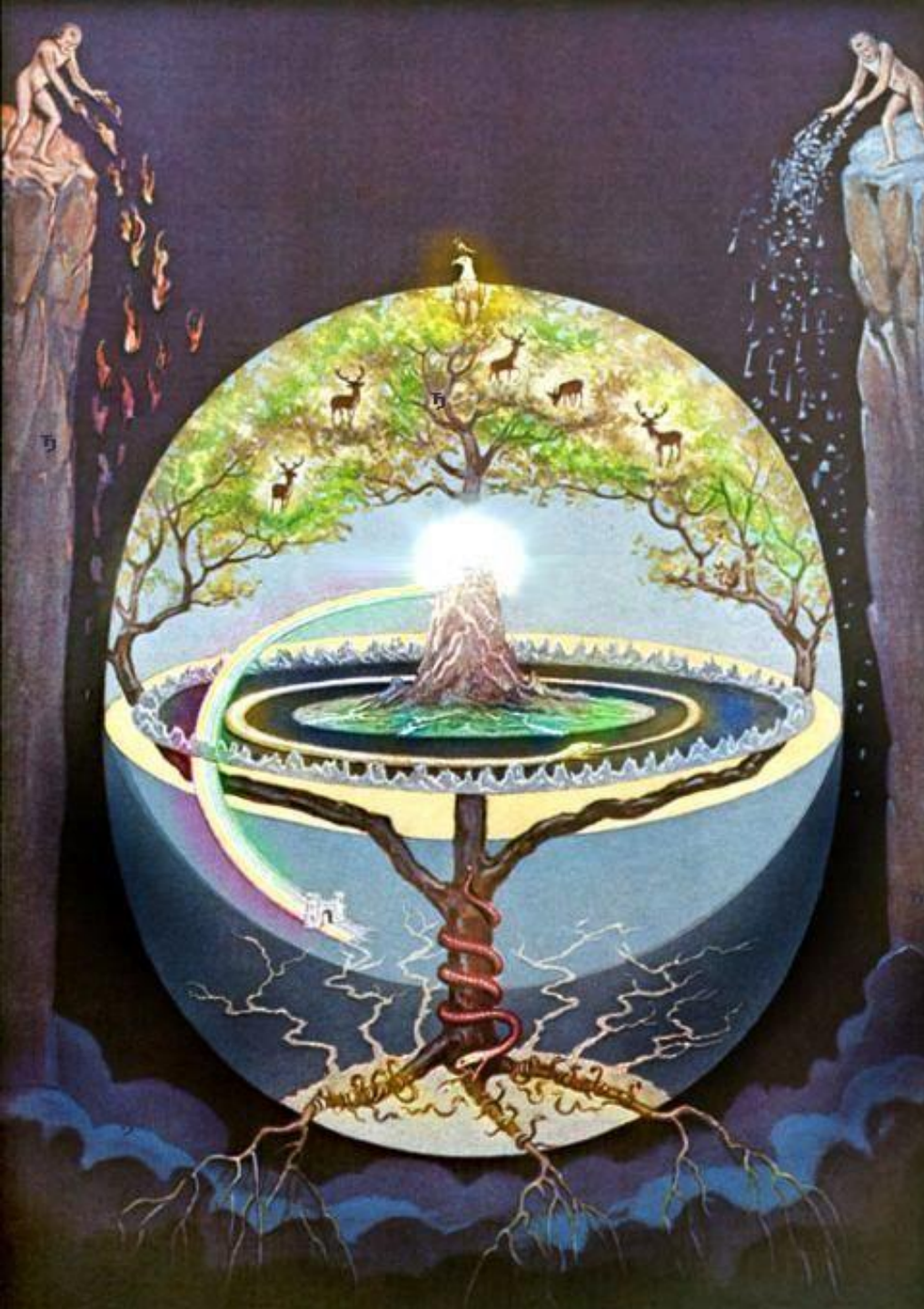
By Brendan Tierney, 2012

Qué es la Cs. de Datos (tercer intento)

Las ciencias naturales estudian las cosas que existen en el mundo: no tiene sentido hablar de una ciencia que estudia los unicornios

Además, las cosas que existen en el mundo cambian con el tiempo:

- Hace 500.000 años no había humanos, por lo que la psicología no habría tenido sentido.
- Hace 4.000.000.000 años no había vida, por lo que la biología no habría tenido sentido.
- Hace 100 años no había computadoras como las que tenemos hoy, por lo que buena parte de las ciencias de la computación no habrían tenido sentido.



¿Qué cosas existen?

Planetas

Estrellas

Sólidos

Líquidos

Gases

Seres vivos

Océanos

Humanos

Sociedades

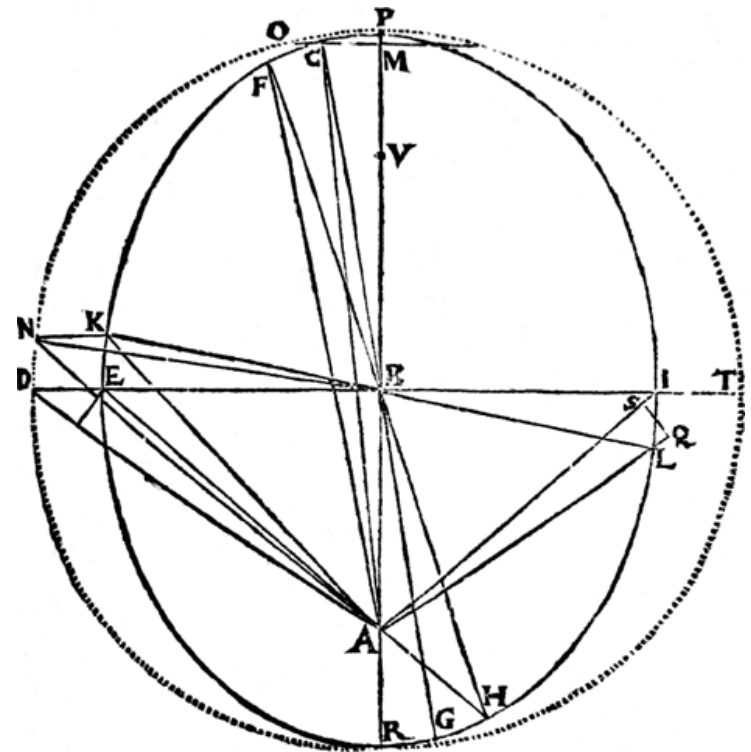
¿Datos?

Hace más de pocas décadas...

... había que trabajar muy duro para que los datos existan



Tycho Brahe (1546 -1601)

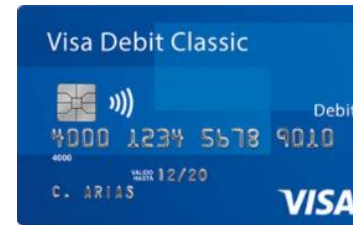


Johannes Kepler (1571 -1630)

Todo deja una secuela de datos

LN **Página12**

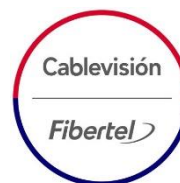
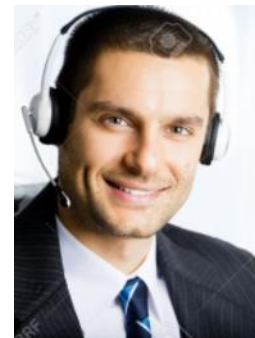
Clarín



Google



edenor



Pero... ¿no todos almacenan los datos, verdad?



Los datos existen

La Cs. de Datos busca responder preguntas prácticas y teóricas sobre los datos

(así como la Cs. de la computación tiene sub-disciplinas como ingeniería de software, teoría de lenguajes o teoría de la complejidad computacional)

Generalmente estamos en contacto con la parte más aplicada, y por un buen motivo...

Data Scientist: The Sexiest Job of the 21st Century

by **Thomas H. Davenport** and **D.J. Patil**

FROM THE OCTOBER 2012 ISSUE

Harvard
Business
Review

50 Best Jobs in America

★ Awards

This report ranks jobs according to each job's Glassdoor Job Score, determined by combining three factors:

Job Title	Median Base Salary	Job Satisfaction	Job Openings
#1 Front End Engineer	\$105,240	3.9/5	13,122
#2 Java Developer	\$83,589	3.9/5	16,136
#3 Data Scientist	\$107,801	4.0/5	6,542

Qué vemos en este curso

- Una introducción a los aspectos prácticos de la Cs. de Datos
- Desarrollar intuiciones sobre qué preguntas pueden resolverse con datos y dónde ir a buscarlos
- Hacer preguntas interesantes y responderlas usando datos
- Hacer ciencia de datos: plantear una pregunta, identificar una fuente de datos que sirva para responderla, obtener esos datos, evaluar la calidad de los datos, analizarlos, generar visualizaciones, generar un reporte y comunicar una historia que responda la pregunta planteada
- Aprender algunas técnicas y heurísticas, pero también actitudes: escepticismo, proactividad y el valor de probar y equivocarse (y darse cuenta)

Qué NO vemos en este curso

- Probabilidad y estadística desde un punto de vista formal o matemático
- Análisis de algoritmos desde un punto de vista formal o matemático
- Machine learning desde un punto de vista formal o matemático
- Deep learning
- Temas de ingeniería de datos y procesamiento de grandes volúmenes de datos
- Preguntas teóricas o fundacionales sobre Cs. de Datos
- Conocimiento de un campo científico específico

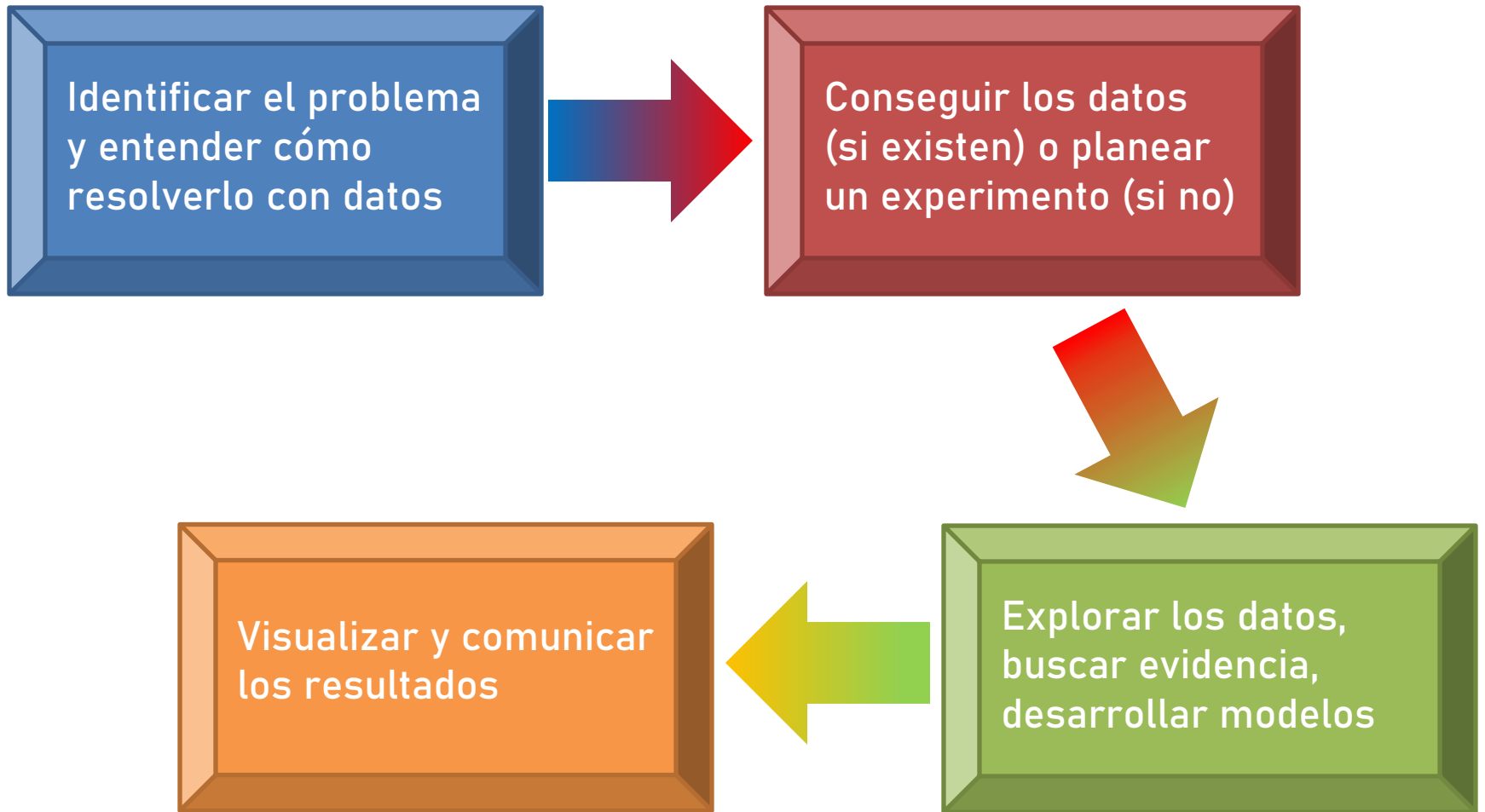
<http://lcd.exactas.uba.ar/>

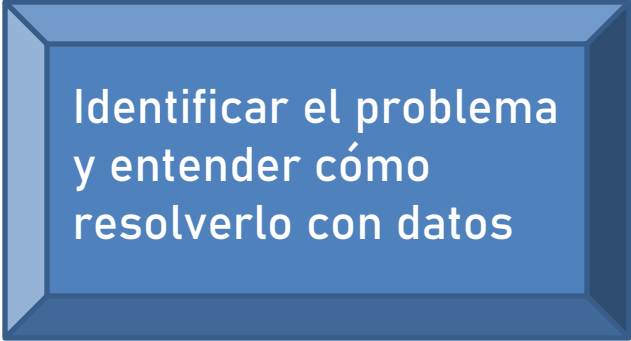
- + [Análisis I \(*\)](#)
- + [Álgebra I \(*\)](#)
- + [Algoritmos y Estructuras de Datos I \(*\)](#)
- + [Electiva de Introducción a las Ciencias Naturales](#)
- + [Análisis II \(*\)](#)
- + [Algoritmos y Estructuras de Datos II \(*\)](#)
- + [Laboratorio de Datos \(*\)](#)
- + [Análisis Avanzado](#)
- + [Álgebra Lineal Computacional \(*\)](#)
- + [Probabilidad](#)
- + [Algoritmos y Estructura de Datos III](#)
- + [Intr. a la Estadística y Ciencia de Datos](#)
- + [Intr. a la Investigación Operativa y Optimización](#)

Algunas preguntas teóricas

- ¿Qué problemas pueden resolverse únicamente analizando más y más datos?
- ¿Qué problemas pueden ser resueltos fácilmente usando modelos generales de machine learning o inteligencia artificial en vez de modelos específicos y por qué?
- ¿Para qué problemas tiene sentido generar instancias simuladas de datos de entrenamiento?
- ¿Cuántos datos son suficientes para resolver distintos tipos de problemas?
- ¿Cuál es la relación causal que existe entre nuestros modelos y los datos en los que se basan? ¿Es posible construir modelos sin cambiar el mundo? Y si no, ¿cómo tenemos en cuenta esa interacción para evitar círculos viciosos y profecías autocumplidas?

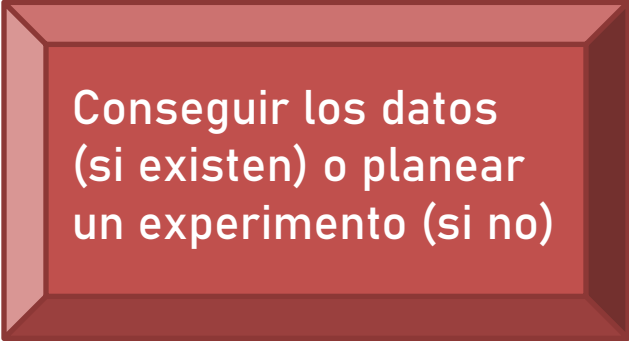
Qué vemos en este curso





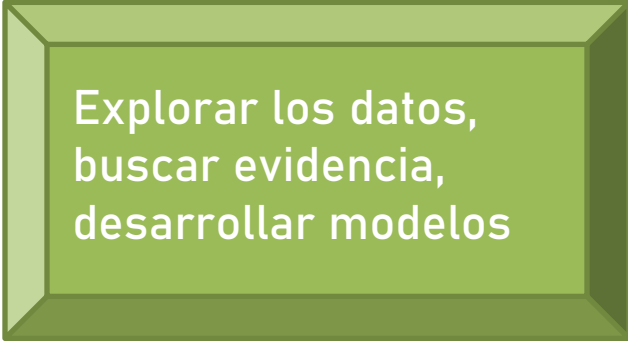
Identificar el problema
y entender cómo
resolverlo con datos

- Hablar con los especialistas y entender su relación con los datos
- Entender qué se sabe y qué no se sabe sobre el problema
- Si el problema todavía no fue abordado, ¿por qué?
- ¿Cuáles son los beneficios de trabajar en este problema y no en otro?
- ¿Quién se beneficia de que trabajemos en este problema y no en otro?
- ¿Existe realmente un problema?



Conseguir los datos
(si existen) o planear
un experimento (si no)

- ¿Qué fuentes de datos conocen los especialistas?
- ¿Hay datos en Internet que puedan servir para resolver el problema?
- Evaluar la calidad de los datos
- ¿Quién es el propietario de los datos? ¿Qué quiere a cambio?
- Si los datos no existen, ¿cómo podemos conseguirlos?
- ¿Cuántos datos necesitamos? ¿Cuáles son los posibles sesgos de las fuentes de datos? Los datos, ¿son independientes entre sí? ¿cuál es el espacio muestral?



Explorar los datos,
buscar evidencia,
desarrollar modelos

- ¿Tenemos una hipótesis? ¿O estamos explorando a ver qué encontramos?
- ¿Qué variables importantes podríamos no estar teniendo en cuenta?
- ¿Cuáles son nuestros propios sesgos a la hora de analizar los datos?
- Si el objetivo es elaborar un informe cuali- o cuantitativo, ¿están las conclusiones justificadas por nuestro análisis? ¿Estamos influenciados por las expectativas de quién va a leer el informe? ¿ocultamos resultados que contradicen las conclusiones?
- Si desarrollamos un modelo predictivo o de machine learning, ¿es el modelo trivial? ¿se basa el modelo únicamente en datos anteriores a lo que queremos predecir? ¿entendemos cómo funciona el modelo? ¿es ético?



Visualizar y comunicar
los resultados

- ¿Entendemos a quién le estamos comunicando, cuáles son sus conocimientos e intereses y cómo difieren de los nuestros?
- Implementar un mínimo sentido de la estética y la economía para presentar los resultados
- ¿Somos aburridos y cómo podemos dejar de serlo?
- ¿Estamos intentando convencer al otro de algo y por qué? ¿Estamos siendo honestos? ¿Estamos diciendo lo que el otro quiere escuchar?
- ¿Estamos distribuyendo correctamente el crédito por las contribuciones?
¿Entendemos el alcance de nuestro reporte?
- ¿Entendimos cómo respondió nuestra audiencia a lo que les dijimos?

Un ejemplo (de mi trabajo)

¿a qué vendedor asigno cada *lead*?

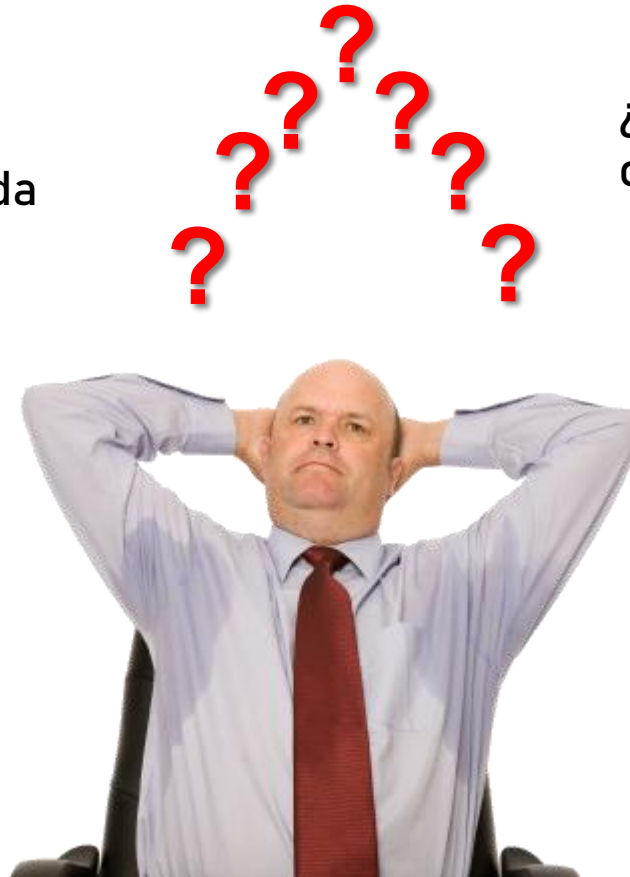
¿cuánto tengo que contactar cada *lead*?

¿a qué *lead* tengo que responder primero?

¿cuántas veces debo contactar cada *lead* hasta descartarla?

¿cuáles *leads* puedo abandonar?

¿tengo que cambiar de vendedor asignado a una *lead*?



La vida de un gerente de ventas



Velocity™

Velocity es un servicio basado en *cloud computing* que automatiza el proceso de asignar vendedores a *leads*, distribuyendo la carga entre vendedores de acuerdo a los criterios del gerente, asegurándose de que no queden vendedores sin asignar ni *leads* sin contactar.



Pero esto no resuelve los problemas que enfrenta el gerente de ventas, únicamente automatiza sus decisiones

Un ejemplo (de mi trabajo)

¿a qué vendedor asigno cada *lead*?

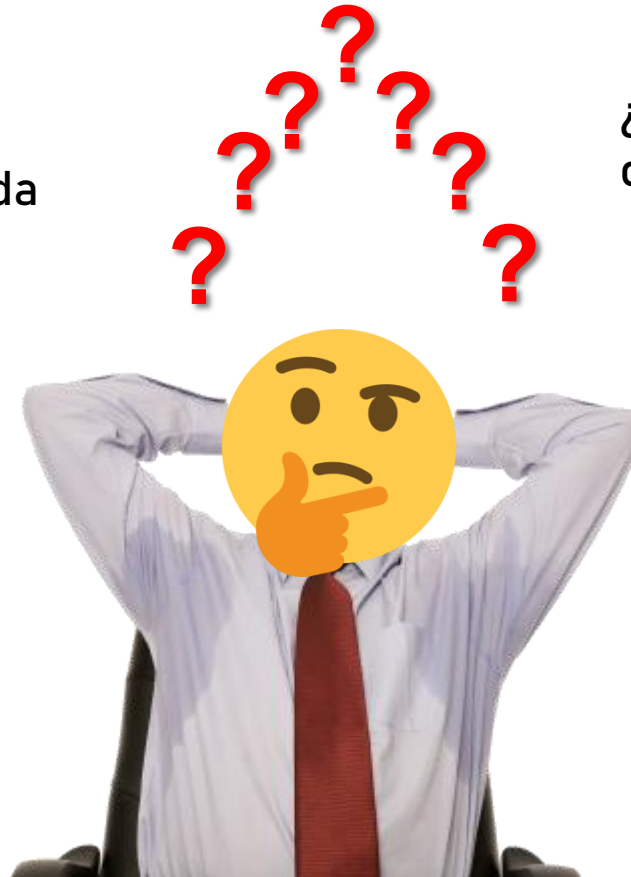
¿cuánto tengo que contactar cada *lead*?

¿a qué *lead* tengo que responder primero?

¿cuántas veces debo contactar cada *lead* hasta descartarla?

¿cuáles *leads* puedo abandonar?

¿tengo que cambiar de vendedor asignado a una *lead*?



La vida de un gerente de ventas

Pero Velocify guarda en sus bases de datos terabytes de información sobre transacciones pasadas, incluyendo:

- Información sobre cada *lead*. En muchos casos se trata de alguien que completa información en un formulario online a la espera de que lo contacte un vendedor. Por ejemplo: edad, sexo, nivel de educación, ubicación geográfica (IP), cómo llegó a la página web, fecha y hora y (en EEUU) raza (blanco, negro, hispano, nativo americano, islas del pacífico).
- Información sobre cada vendedor. Idem *lead*, junto con su récord de ventas pasadas.
- Información sobre la interacción entre vendedor y *lead*. ¿Cuánto tiempo se tardó hasta el primer contacto? ¿cuál es la duración media de cada llamada telefónica? ¿Qué dicen durante las llamadas (grabaciones)? ¿A qué hora responde las llamadas?
- Información sobre la venta. ¿Compró o no compró? Si no compró, ¿está interesado en un futuro o no quiere saber más nada?



Con nombre y apellido es posible comprar información a grandes compañías que a su vez compran información a otras compañías sobre cada *lead*

Datos Básicos y de Identificación

Apellido y Nombre	TAGLIAZUCCHI ENZO RODOLFO	Ver Informe Completo
Posible DNI	30.406.011	
CUIT/CUIL	20-30406011-6	
Edad Estimada	38 años	
Homónimos	En el Padrón de AFIP no se han identificado otras personas con mismo nombre/s y apellido.	
Personas con el mismo apellido	16 personas en Argentina tienen el apellido Tagliazucchi	

Participación en Sociedades (incluyendo si es o fue Socio, Autoridad o Representante de Sociedades inscriptas en la IGJ) [+](#)

Información Laboral [+](#)

Historial de Contratos con ART [+](#)

Informacion Comercial [+](#)

Inclusión en Padrón SIRCREB - Riesgo Fiscal [+](#)

Deudas de Impuestos Provinciales [+](#)

Causas en Corte Suprema de Justicia de la Nación [+](#)

Causas en Cámara Nacional de Apelaciones en lo Contencioso Administrativo Federal [+](#)

Causas en Cámara Nacional de Apelaciones en lo Civil y Comercial Federal [+](#)

Domicilios y Teléfonos Vinculados [+](#)

Constancia de Inscripcion en AFIP [+](#)

Actividades Registradas en AFIP [+](#)

Impuestos Registrados en AFIP [+](#)

Vencimiento de Impuestos Registrados en AFIP [+](#)

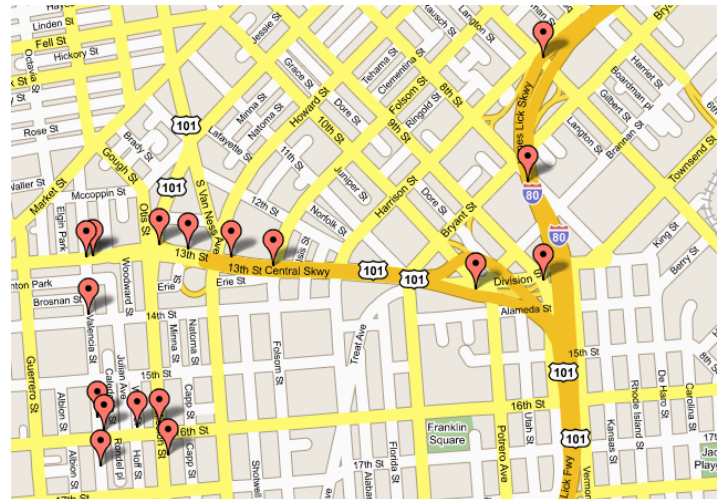
Obra Social, Empleador y Situacion Previsional [+](#)

Domicilio Actualizado y Fecha de Nacimiento [+](#)

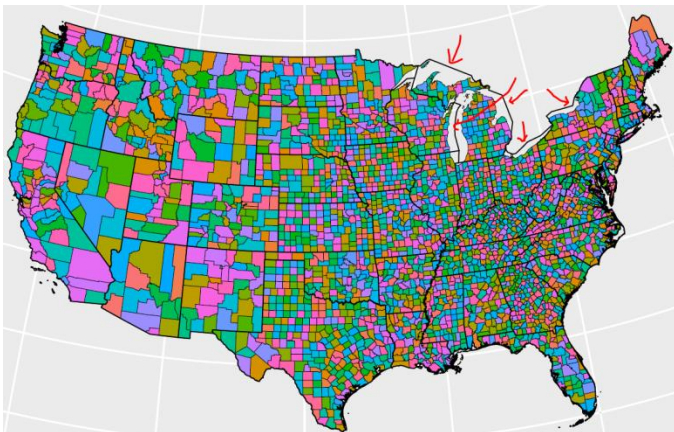
Cuánto gana, en qué gasta, composición núcleo familiar, impuestos, hipotecas, deudas, hobbies, suscripciones, apps de celular, etc.



Con la dirección es posible geolocalizar el domicilio (obtener latitud y longitud)



... y determinar en qué tracto reside la persona...



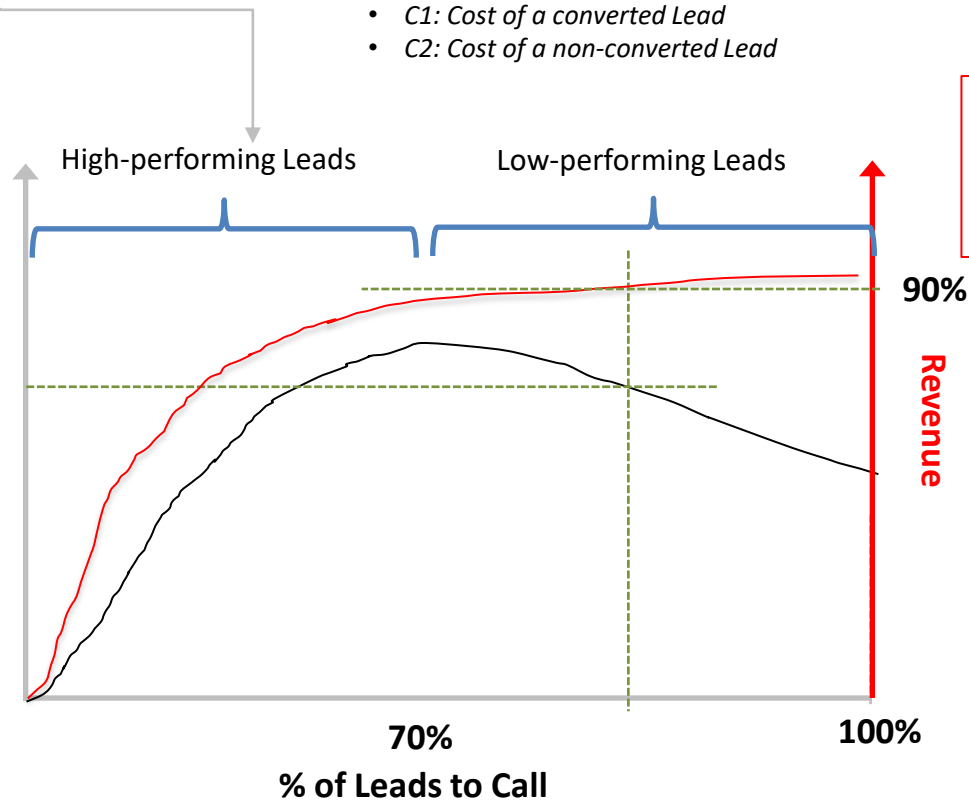
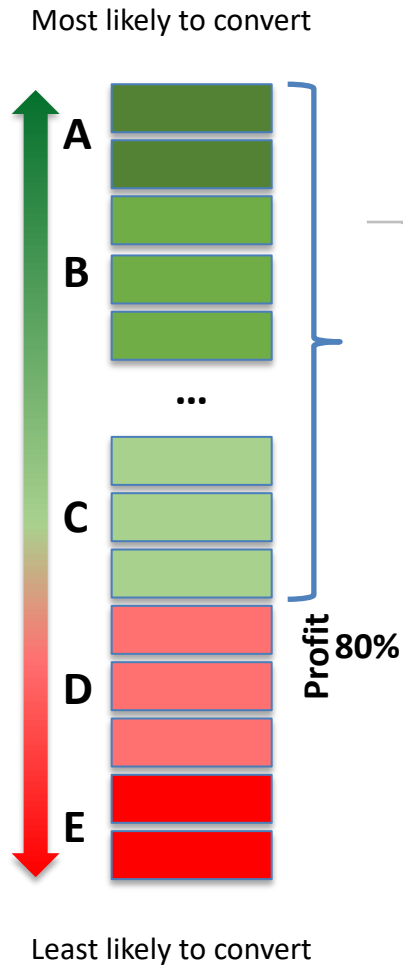
... para obtener información detallada socio-demográfica y económica del Censo de EEUU



Usar todos estos datos para asignar un score a cada lead que refleja la probabilidad que tiene de comprar o no, y estudiar el impacto en el revenue de únicamente contactar al top % de los leads (si hay más leads que vendedores)

$$Profit = \sum_n [1\{classifier\ predicted_+ \cap Actual_{converted}\} * (R_n - C1)] + \sum_n [1\{classifier\ predicted_+ \cap Actual_{not\ converted}\} * (-C2)]$$

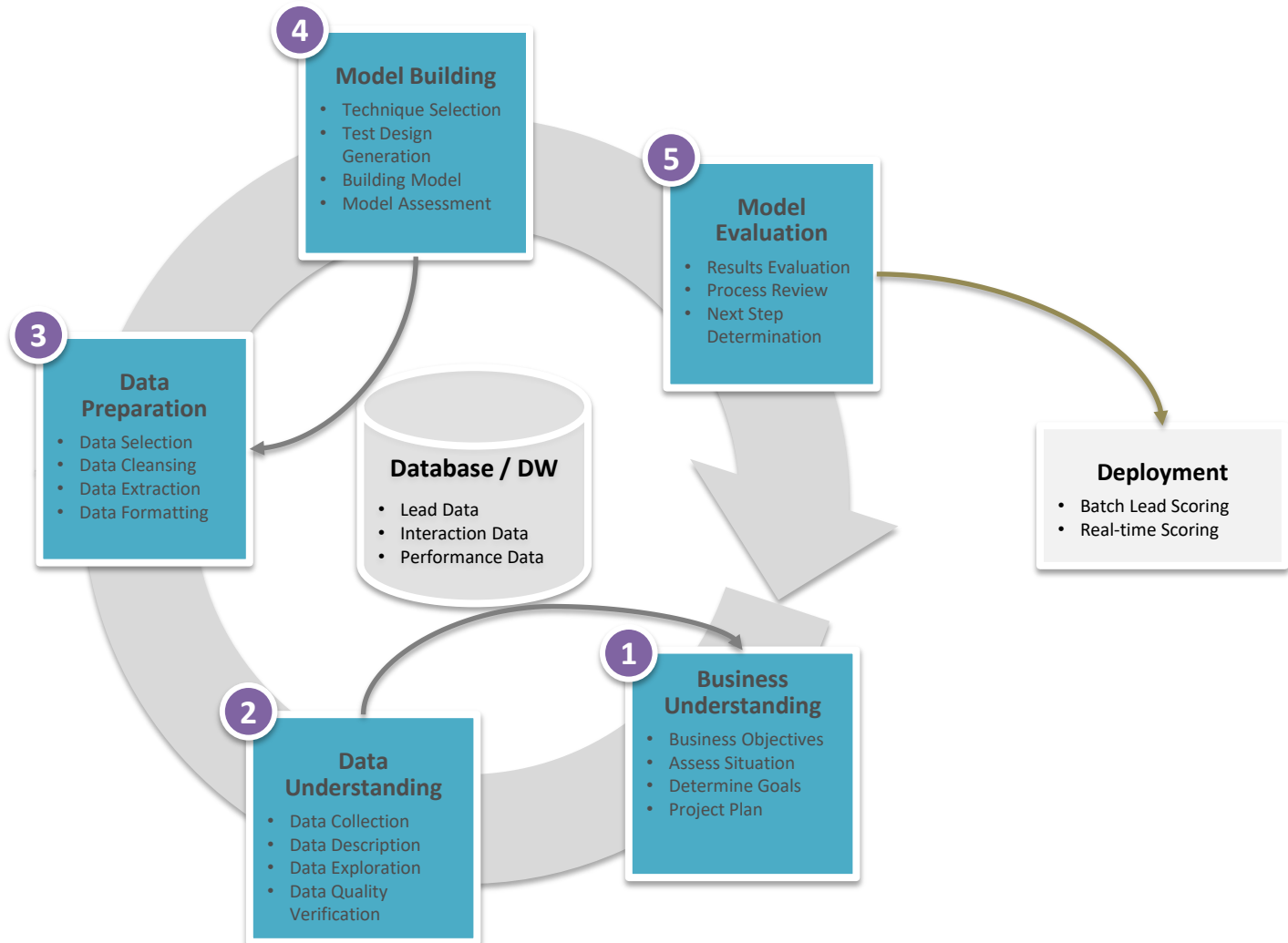
- n : Number of Leads
- R : Estimated Revenue of converted Lead
- $C1$: Cost of a converted Lead
- $C2$: Cost of a non-converted Lead



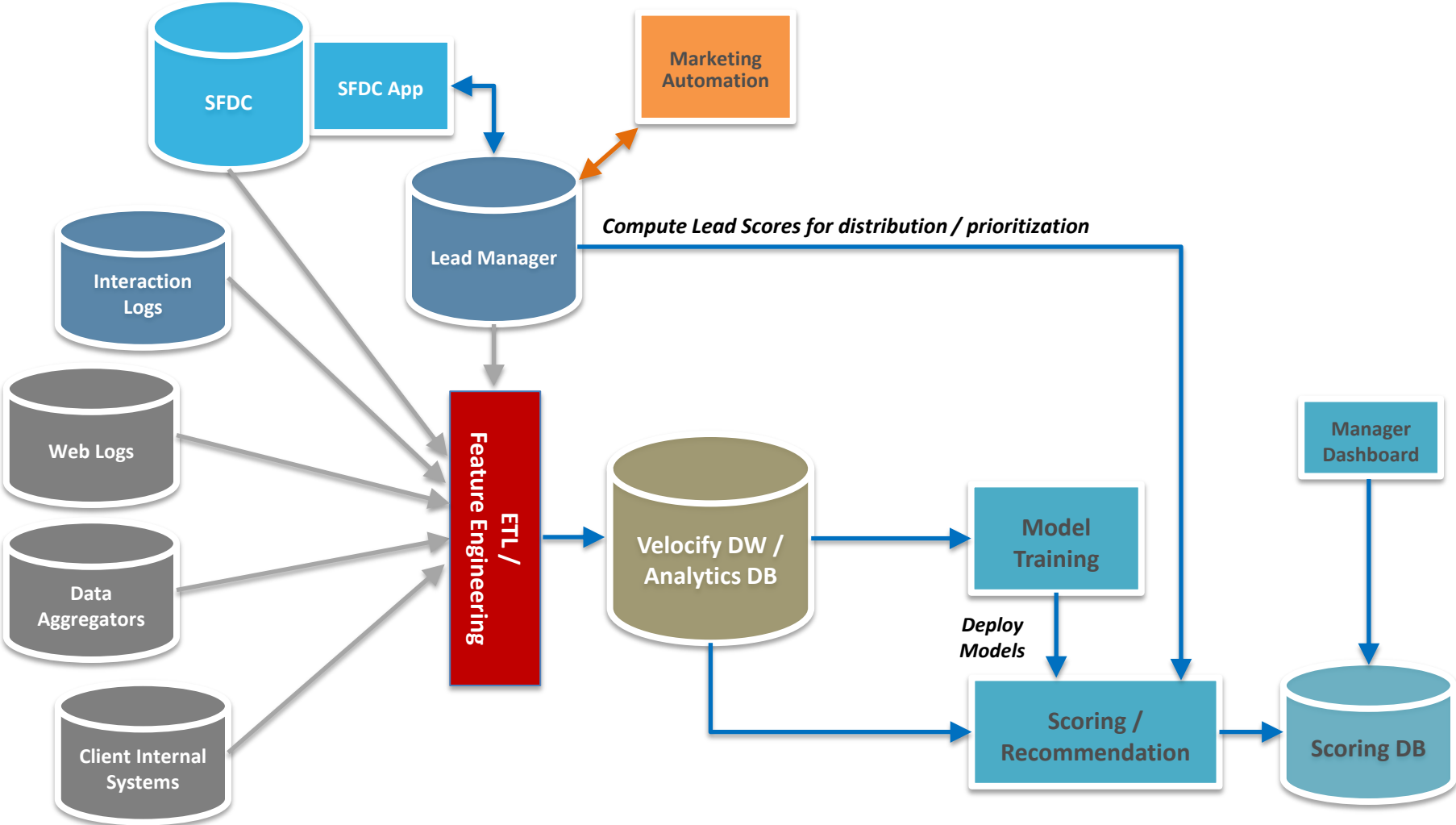
Example:

- Call 70% of Leads
- Generate 90% of Revenue
- Earn 80% of max Profit

Exploración de datos y desarrollo de modelos



Integración de la solución con el sistema



Ejemplo: una institución educativa

- **Sample Data (previous 6 months)**

- Contacted Leads:
- Qualified Leads:
- Converted Leads:

- **Available Features**

- Speed to Call , Speed-to-Contact,
- Campus Id
- Program of Interest
- Financial Aid
- Gender
- Previous Education
- State
- VA Benefits
- Lead Vendor

- **Calculated Features**

- » Degree of Interest
- » Distance to Campus (zipcode)
- » When the lead was added: hour, day, month, weekday
- » When the lead was first contacted: hour, day, month, weekday

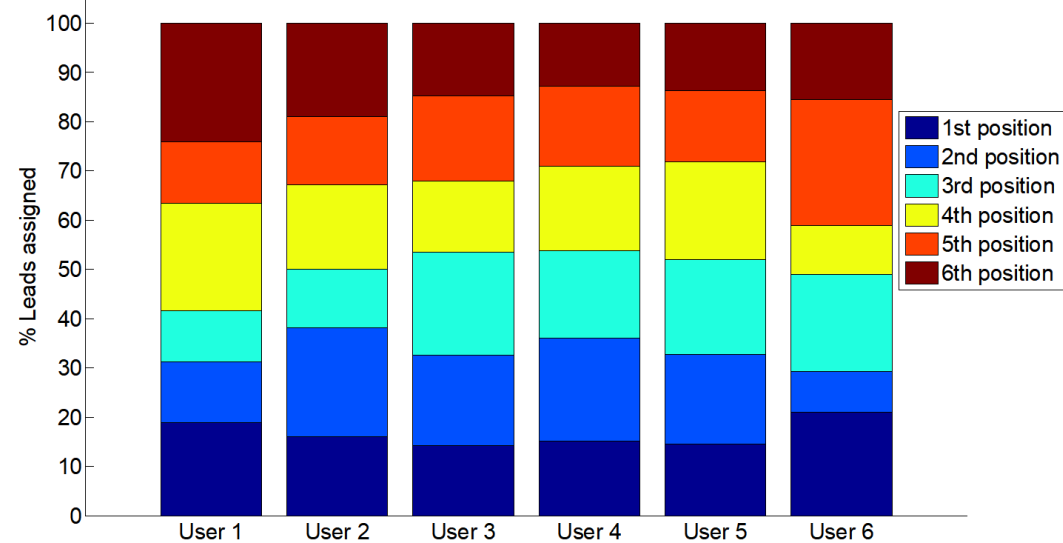
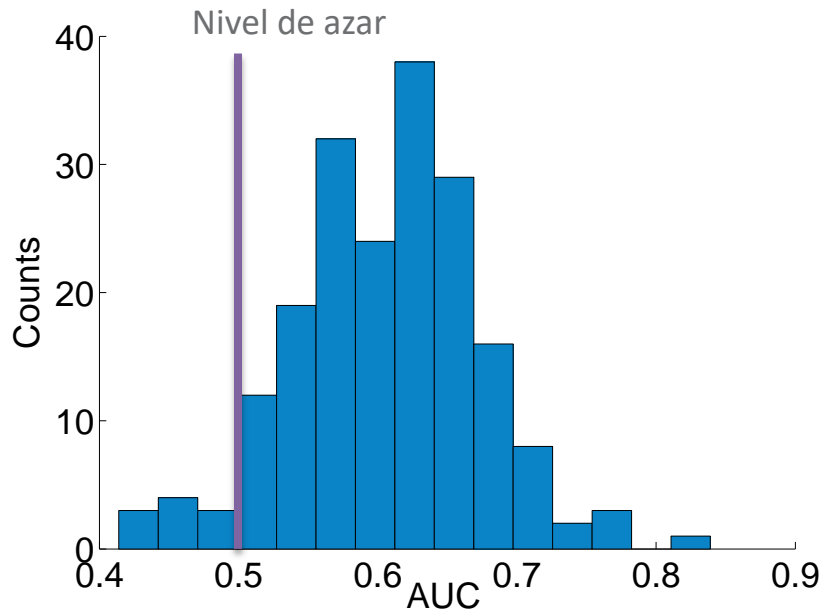
- **External Features**

- » Ethnicity

Datos comprados, o incluso inferidos a partir del nombre de la persona (funciona bastante bien)

El modelo es capaz de estimar la probabilidad de que un lead se convierta. Eso permite asignar un score de calidad, y distribuir los leads entre vendedores de forma homogénea de acuerdo a su score.

Si los vendedores llaman a los leads en el orden dictado por este score en vez de hacerlo al azar, es posible estimar un (importante) incremento en el revenue de la empresa.



Pero, ¿por qué funciona el modelo?

Remover datos y evaluar la precisión del modelo

- **Sample Data (previous 6 months)**

- Contacted Leads:
- Qualified Leads:
- Converted Leads:

- **Available Features**

- Speed to Call, Speed-to-Contact,
- Campus Id
- Program of Interest
- Financial Aid
- Gender
- Previous Education
- State
- VA Benefits
- Lead Vendor

- **Calculated Features**

- » Degree of Interest
- » Distance to Campus (zipcode)
- » When the lead was added: hour, day, month, weekday
- » When the lead was first contacted: hour, day, month, weekday

- **External Features**

- » Ethnicity

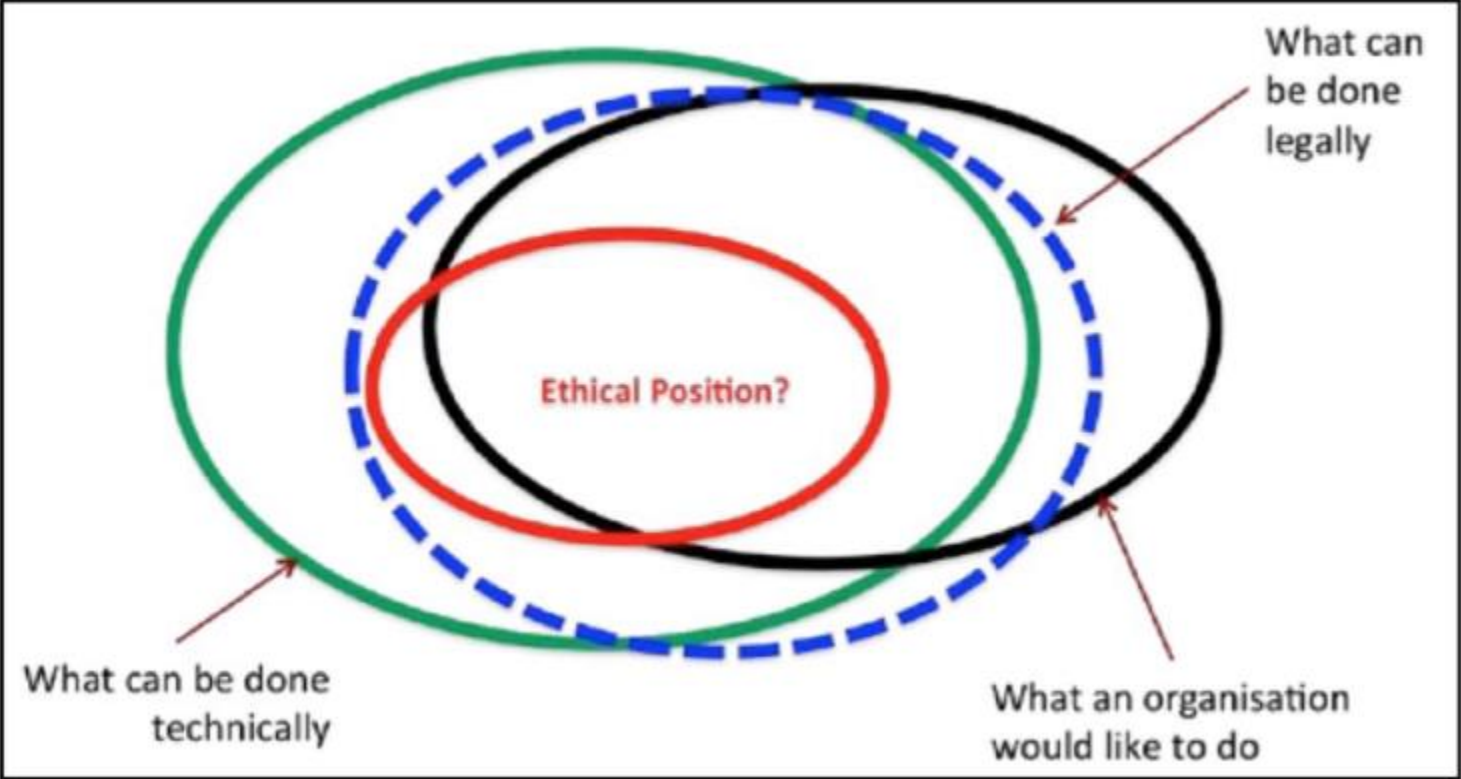


Hacer ciencia de datos tiene implicaciones éticas

- ¿Entendemos en qué se basan nuestras predicciones?
- ¿Es justo reforzar sesgos que aparecen en los datos? ¿Es ético?
- ¿Actúan nuestros modelos como profecías auto-cumplidas?
- ¿Qué significa ÉXITO en el contexto en el cual desarrollamos el modelo?
- ¿Dieron los sujetos consentimiento para usar sus datos?
¿Respetamos su privacidad y su anonimato?
- Si el modelo se equivoca o es injusto, ¿hay una instancia de supervisión humana?
- ¿Tienen las personas afectadas por el modelo una oportunidad de dialogar y discutir los resultados con un ser humano?
- Y sobre todo...

¿Cui bono?

¿quién se beneficia?



**YouTube vows to recommend fewer
conspiracy theory videos**

Site's move comes amid continuing pressure over i
platform for misinformation and extremism

**The Reason This "Racist Soap
Dispenser" Doesn't Work on
Black Skin**

Amazon Prime and the racist algorithms

**MACHINES TAUGHT BY PHOTOS
LEARN A SEXIST VIEW OF
WOMEN**

**Facial recognition software
is biased towards white
men, researcher finds**

Biases are seeping into software

**YouTube's Restricted Mode Is Hiding
Some LGBT Content [Update]**

**Google Translate's Gender
Problem (And Bing Translate's,
And Systran's...)**

Detección de criminales mediante expresiones faciales

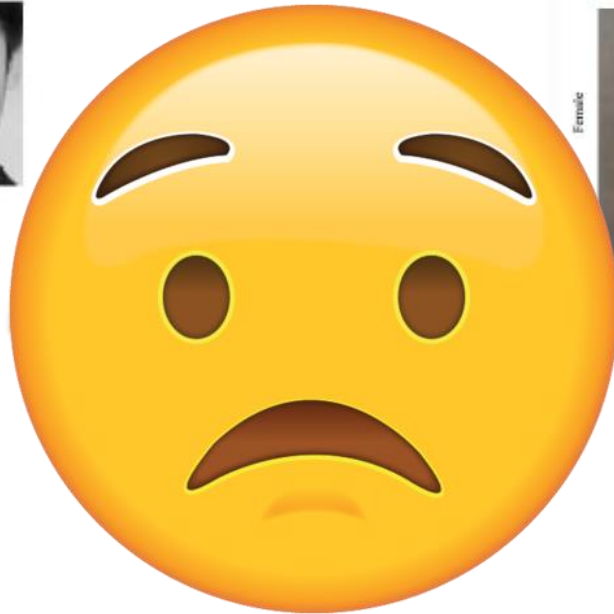
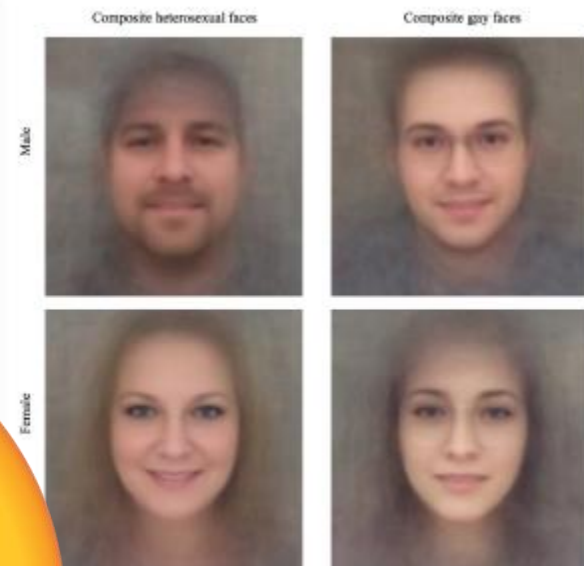


(a) Three samples in criminal ID photo set S_c .



(b) Three samples in non-criminal ID photo set S_n
Figure 1. Sample ID photos in our data set.

Detección de sexo mediante expresiones faciales



Próxima clase:

Introducción a distintos tipos de datos en Python, principalmente arrays (numpy) y dataframes (pandas)

Presentación del dataset en el que basamos ejemplos y los primeros ejercicios (actitudes respecto a vacunación contra COVID-19)