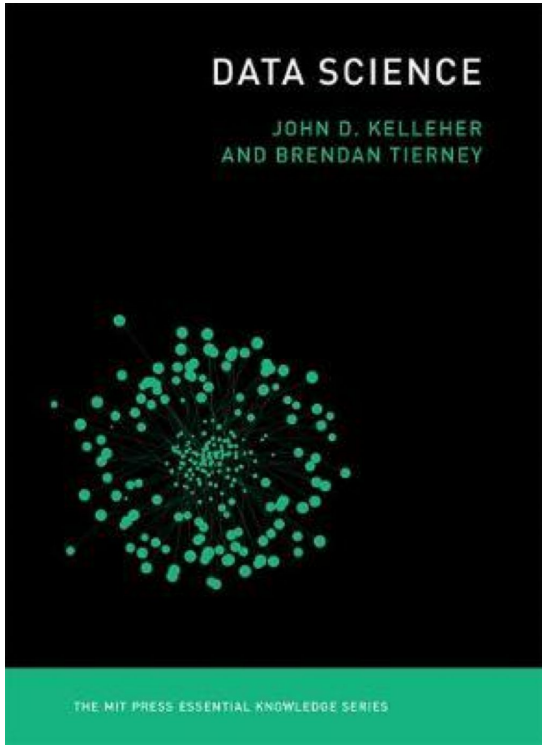
Argentina programa 4.0

Universidad Nacional de San Martín

Módulo 2 - Ciencia de Datos

Semana 2. Introducción a la Ciencia de Datos

¿Qué es la ciencia de datos?



Qué

"Un conjunto de principios, definiciones de problemas, algoritmos, y procesos para extraer **patrones no obvios** y **útiles** a partir de grandes conjuntos de datos."

Para qué

"Mejorar la toma de decisiones basándose en los conocimientos extraídos de (grandes) conjuntos de datos."

¿Qué es la ciencia de datos?

"Data Science is the art of turning data into actions."

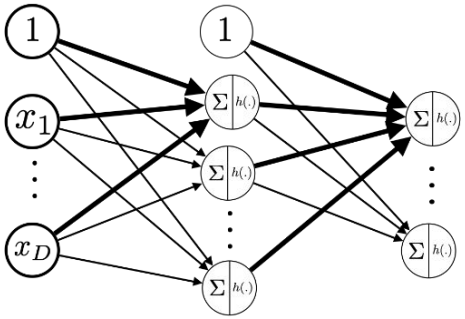
“La Ciencia de Datos es el arte de convertir a los datos en acciones”

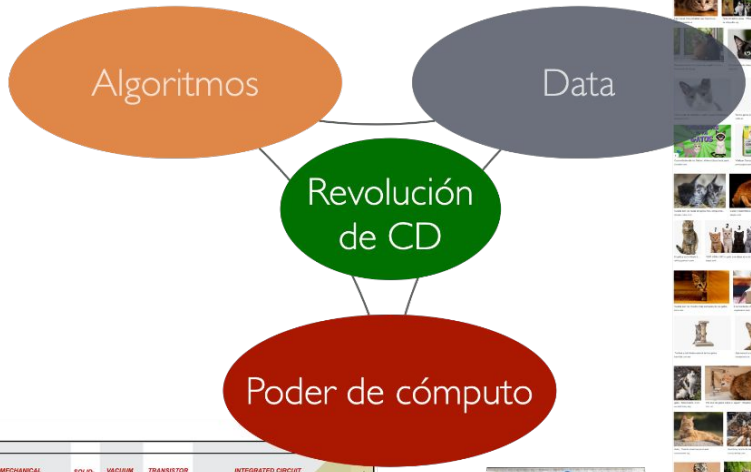
¿Qué es la ciencia de datos?

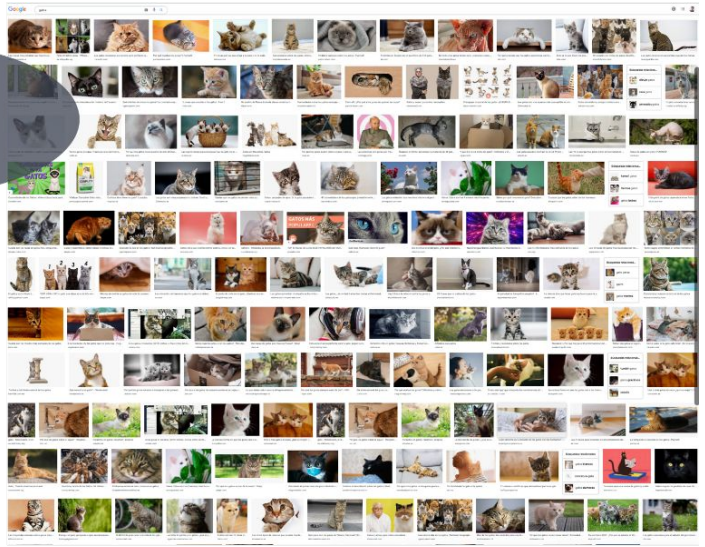
ES UNA

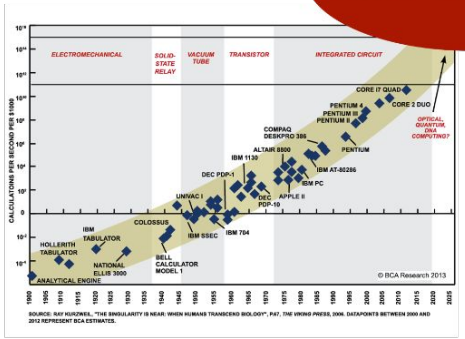
PRÁCTICA

¿Por qué ahora?

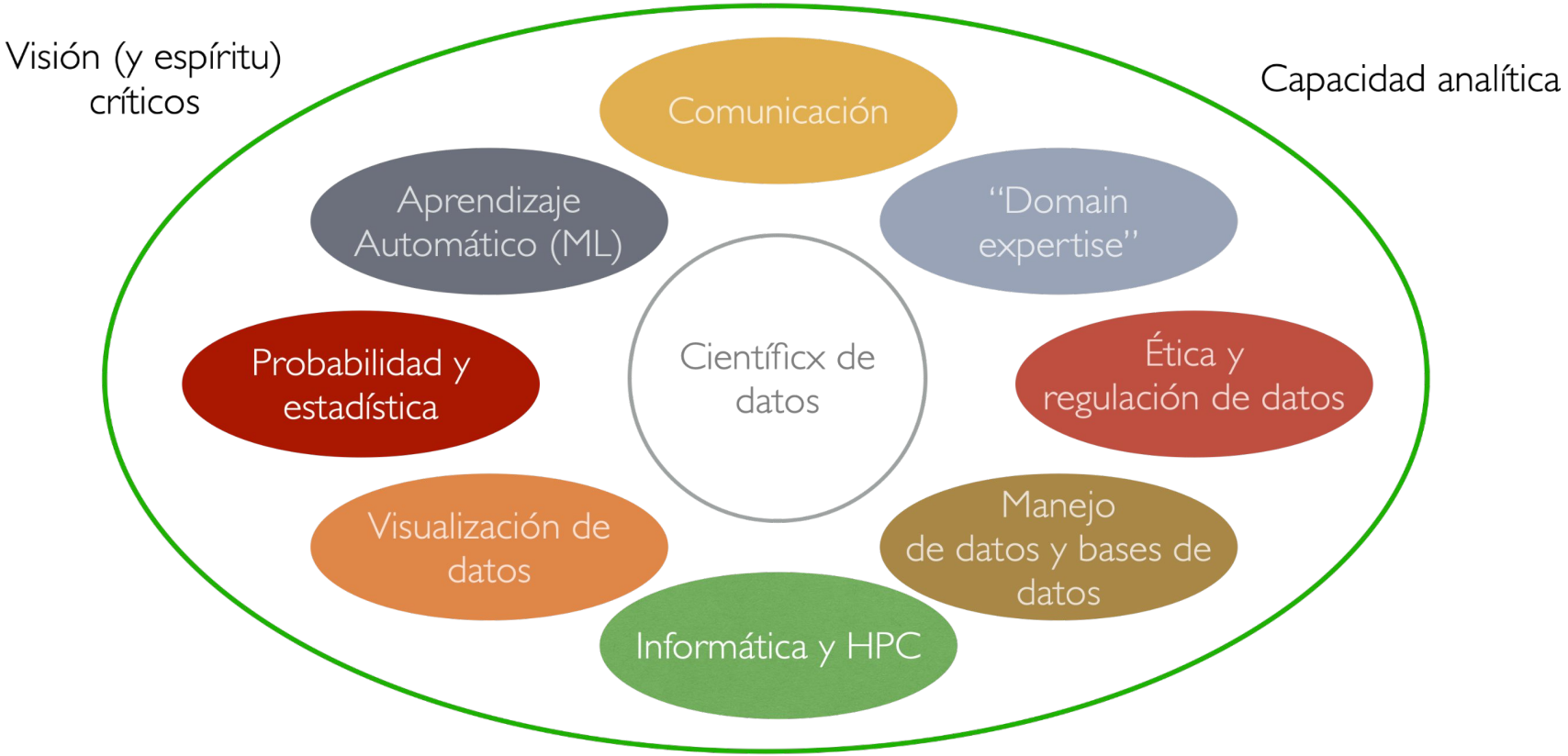






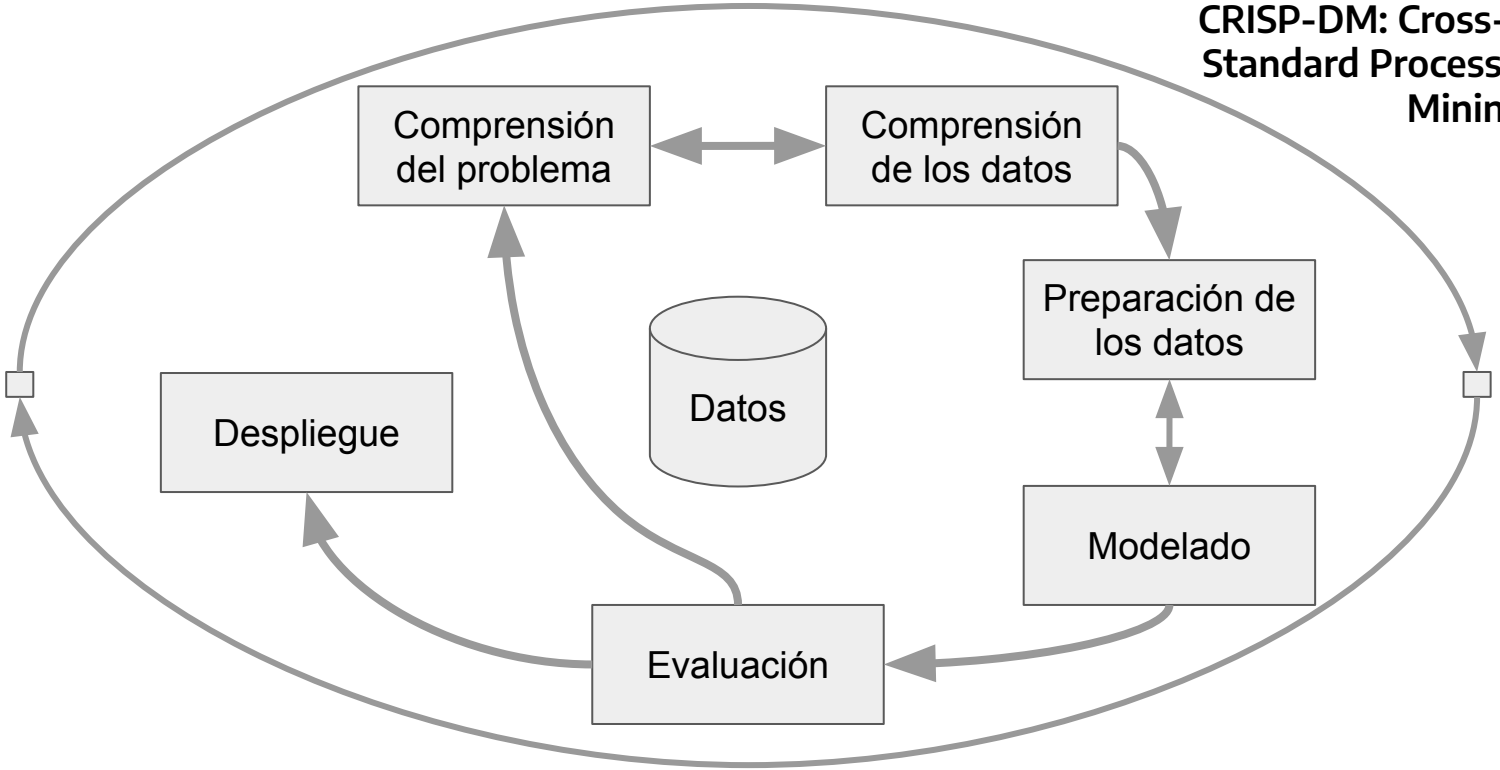


¿Qué necesita unx científicx de datos?



El proceso de ciencia de datos

CRISP-DM: Cross-Industry Standard Process for Data Mining (2006)



El proceso de ciencia de datos

- Precisar la pregunta de interés.
- Definir los objetivos del proyecto.
- Se toman en cuenta las necesidades del negocio.

- Identificar si los datos necesarios para el proyecto están disponibles.
- Se consideran asuntos de ética y privacidad

- Reuniones en grupos específicos (venta, marketing, ...)
- Interacciones con administradores de datos.

Preparación de los datos

- Crear un dataset coherente para el análisis posterior.
- Integrar datos de diversas fuentes (bases de datos).
- Controlar la calidad de los datos (faltantes, outliers, ...).

El proceso de ciencia de datos

- Extracción de patrones útiles de los datos.
- Involucra modelos sencillos y/o de machine learning.
- Experimentar con diversos algoritmos.
- Se identifican nuevos problemas con los datos, o con su preparación

Evaluación

- Evaluar el modelo en un contexto más amplio (p.ej., con datos que no se hayan visto antes).
- Evaluar si cumple los objetivos del proyecto.

Deployment

- Llevar la solución generada al ambiente del negocio.
- Planear la integración con soluciones existentes
- Planear la evaluación periódica del desempeño del modelo.

Los datos

- Estructura "tidy" de datos
- Tipo de datos
- Datos no estructurados.
- Consideraciones sobre el armado de datos.

2013,9,30,2010,2012,-2,2145,2150,-7,EV,4106,N12521,EWR,GSO,64,445,20,12,2013-10-01T00:00:00Z
2013,9,30,2010,2016,-6,2119,2128,-9,EV,4312,N13969,EWR,DCA,35,199,20,16,2013-10-01T00:00:00Z
2013,9,30,2010,2015,-5,2100,2125,-25,B6,418,N249JB,JFK,BOS,35,187,20,15,2013-10-01T00:00:00Z
2013,9,30,2012,2015,-3,2117,2144,-27,B6,702,N339JB,JFK,BUF,51,301,20,15,2013-10-01T00:00:00Z
2013,9,30,2013,2000,13,2133,2120,13,US,2195,N747UW,LGA,DCA,41,214,20,0,2013-10-01T00:00:00Z
2013,9,30,2015,2030,-15,2221,2251,-30,EV,5209,N741EV,LGA,CHS,84,641,20,30,2013-10-01T00:00:00Z
2013,9,30,2015,2015,0,2244,2307,-23,UA,1545,N17730,EWR,IAH,174,1400,20,15,2013-10-01T00:00:00Z
2013,9,30,2016,2022,-6,2215,2246,-31,B6,135,N607JB,JFK,PHX,263,2153,20,22,2013-10-01T00:00:00Z
2013,9,30,2017,2022,-5,2132,2155,-23,B6,105,N298JB,JFK,ORD,112,740,20,22,2013-10-01T00:00:00Z
2013,9,30,2017,2017,0,2253,2320,-27,UA,771,N510UA,JFK,LAX,304,2475,20,17,2013-10-01T00:00:00Z
2013,9,30,2019,2023,-4,2225,2240,-15,EV,4085,N14188,EWR,OMA,149,1134,20,23,2013-10-01T00:00:00Z
2013,9,30,2021,2022,-1,2144,2200,-16,UA,1464,N76503,EWR,CLE,54,404,20,22,2013-10-01T00:00:00Z
2013,9,30,2022,2025,-3,2127,2140,-13,AA,2314,N3CUAA,JFK,BOS,37,187,20,25,2013-10-01T00:00:00Z
2013,9,30,2023,2025,-2,2315,2350,-35,B6,915,N712JB,JFK,SFO,327,2586,20,25,2013-10-01T00:00:00Z
2013,9,30,2026,2028,-2,2317,2340,-23,UA,1680,N31412,EWR,MIA,141,1085,20,28,2013-10-01T00:00:00Z
2013,9,30,2027,2034,-7,2129,2209,-40,9E,4127,N8790A,JFK,IAD,41,228,20,34,2013-10-01T00:00:00Z
2013,9,30,2028,1910,78,2255,2215,40,AA,21,N338AA,JFK,LAX,294,2475,19,10,2013-09-30T23:00:00Z
2013,9,30,2030,2045,-15,2141,2153,-12,B6,2680,N266JB,EWR,BOS,38,200,20,45,2013-10-01T00:00:00Z
2013,9,30,2030,2035,-5,2324,2340,-16,UA,374,N484UA,EWR,FLL,141,1065,20,35,2013-10-01T00:00:00Z
2013,9,30,2031,2040,-9,2228,2300,-32,9E,4033,N8924B,LGA,TYS,86,647,20,40,2013-10-01T00:00:00Z
2013,9,30,2032,1945,47,2115,2106,9,9E,3864,N602XJ,JFK,PHL,24,94,19,45,2013-09-30T23:00:00Z
2013,9,30,2032,2045,-13,2147,2225,-38,AA,371,N434AA,LGA,ORD,105,733,20,45,2013-10-01T00:00:00Z
2013,9,30,2033,2030,3,2143,2150,-7,WN,2520,N286WN,EWR,MDW,97,711,20,30,2013-10-01T00:00:00Z
2013,9,30,2034,2040,-6,2227,2248,-21,EV,4536,N25134,EWR,CVG,80,569,20,40,2013-10-01T00:00:00Z
2013,9,30,2034,2030,4,2204,2201,3,FL,354,N894AT,LGA,CAK,58,397,20,30,2013-10-01T00:00:00Z
2013,9,30,2035,2035,0,2142,2211,-29,9E,3395,N602LR,JFK,DCA,43,213,20,35,2013-10-01T00:00:00Z
2013,9,30,2035,2029,6,2318,2337,-19,UA,1039,N78524,EWR,AUS,191,1504,20,29,2013-10-01T00:00:00Z
2013,9,30,2041,2045,-4,2147,2208,-21,DL,985,N350NB,JFK,BOS,37,187,20,45,2013-10-01T00:00:00Z

Una tabla de datos

Encabezado (header)

README

Preguntar

	year	month	day	dep_time	sched_dep_time	dep_delay	arr_time	sched_arr_time	arr_delay	carrier	flight	tailnum	origin	dest	air_time	distance	hour	minute	time_hour
0	2013	1	1	517.0	515	2.0	830.0	819	11.0	UA	1545	N14228	EWR	IAH	227.0	1400	5	15	2013-01-01T10:00:00Z
1	2013	1	1	533.0	529	4.0	850.0	830	20.0	UA	1714	N24211	LGA	IAH	227.0	1416	5	29	2013-01-01T10:00:00Z
2	2013	1	1	542.0	540	2.0	923.0	850	33.0	AA	1141	N619AA	JFK	MIA	160.0	1089	5	40	2013-01-01T10:00:00Z
3	2013	1	1	544.0	545	-1.0	1004.0	1022	-18.0	B6	725	N804JB	JFK	BQN	183.0	1576	5	45	2013-01-01T10:00:00Z
4	2013	1	1	554.0	600	-6.0	812.0	837	-25.0	DL	461	N668DN	LGA	ATL	116.0	762	6	0	2013-01-01T11:00:00Z
5	2013	1	1	554.0	558	-4.0	740.0	728	12.0	UA	1696	N39463	EWR	ORD	150.0	719	5	58	2013-01-01T10:00:00Z
6	2013	1	1	555.0	600	-5.0	913.0	854	19.0	B6	507	N516JB	EWR	FLL	158.0	1065	6	0	2013-01-01T11:00:00Z
7	2013	1	1	557.0	600	-3.0	709.0	723	-14.0	EV	5708	N829AS	LGA	IAD	53.0	229	6	0	2013-01-01T11:00:00Z
8	2013	1	1	557.0	600	-3.0	838.0	846	-8.0	B6	79	N593JB	JFK	MCO	140.0	944	6	0	2013-01-01T11:00:00Z
9	2013	1	1	558.0	600	-2.0	753.0	745	8.0	AA	301	N3ALAA	LGA	ORD	138.0	733	6	0	2013-01-01T11:00:00Z
10	2013	1	1	558.0	600	-2.0	849.0	851	-2.0	B6	49	N793JB	JFK	PBI	149.0	1028	6	0	2013-01-01T11:00:00Z
11	2013	1	1	558.0	600	-2.0	853.0	856	-3.0	B6	71	N657JB	JFK	TPA	158.0	1005	6	0	2013-01-01T11:00:00Z
12	2013	1	1	558.0	600	-2.0	924.0	917	7.0	UA	194	N29129	JFK	LAX	345.0	2475	6	0	2013-01-01T11:00:00Z
13	2013	1	1	558.0	600	-2.0	923.0	937	-14.0	UA	1124	N53441	EWR	SFO	361.0	2565	6	0	2013-01-01T11:00:00Z
14	2013	1	1	559.0	600	-1.0	941.0	910	31.0	AA	707	N3DUAA	LGA	DFW	257.0	1389	6	0	2013-01-01T11:00:00Z
15	2013	1	1	559.0	559	0.0	702.0	706	-4.0	B6	1806	N708JB	JFK	BOS	44.0	187	5	59	2013-01-01T10:00:00Z
16	2013	1	1	559.0	600	-1.0	854.0	902	-8.0	UA	1187	N76515	EWR	LAS	337.0	2227	6	0	2013-01-01T11:00:00Z
17	2013	1	1	600.0	600	0.0	851.0	858	-7.0	B6	371	N595JB	LGA	FLL	152.0	1076	6	0	2013-01-01T11:00:00Z
18	2013	1	1	600.0	600	0.0	837.0	825	12.0	MQ	4650	N542MQ	LGA	ATL	134.0	762	6	0	2013-01-01T11:00:00Z
19	2013	1	1	601.0	600	1.0	844.0	850	-6.0	B6	343	N644JB	EWR	PBI	147.0	1023	6	0	2013-01-01T11:00:00Z
20	2013	1	1	602.0	610	-8.0	812.0	820	-8.0	DL	1919	N971DL	LGA	MSP	170.0	1020	6	10	2013-01-01T11:00:00Z
21	2013	1	1	602.0	605	-3.0	821.0	805	16.0	MQ	4401	N730MQ	LGA	DTW	105.0	502	6	5	2013-01-01T11:00:00Z
22	2013	1	1	606.0	610	-4.0	858.0	910	-12.0	AA	1895	N633AA	EWR	MIA	152.0	1085	6	10	2013-01-01T11:00:00Z
23	2013	1	1	606.0	610	-4.0	837.0	845	-8.0	DL	1743	N3739P	JFK	ATL	128.0	760	6	10	2013-01-01T11:00:00Z
24	2013	1	1	607.0	607	0.0	858.0	915	-17.0	UA	1077	N53442	EWR	MIA	157.0	1085	6	7	2013-01-01T11:00:00Z
25	2013	1	1	608.0	600	8.0	807.0	735	32.0	MQ	3768	N9EAMQ	EWR	ORD	139.0	719	6	0	2013-01-01T11:00:00Z
26	2013	1	1	611.0	600	11.0	945.0	931	14.0	UA	303	N532UA	JFK	SFO	366.0	2586	6	0	2013-01-01T11:00:00Z
27	2013	1	1	613.0	610	3.0	925.0	921	4.0	B6	135	N635JB	JFK	RSW	175.0	1074	6	10	2013-01-01T11:00:00Z
28	2013	1	1	615.0	615	0.0	1039.0	1100	-21.0	B6	709	N794JB	JFK	SJU	182.0	1598	6	15	2013-01-01T11:00:00Z
29	2013	1	1	615.0	615	0.0	833.0	842	-9.0	DL	575	N326NB	EWR	ATL	120.0	746	6	15	2013-01-01T11:00:00Z

Algunos nombres

- **Unidades**. Son los objetos principales de estudio. Por ejemplo, los vuelos en la tabla de recién. También llamadas instancias, ejemplos, entidades, casos, sujetos, registros, etc.

- **Variables**. Son propiedades, cualidades o cantidades de las unidades que se pueden medir.

- **Valores**. Son los estados de las variables una vez medida.

- **Observaciones**. Es el conjunto de valores medidos en condiciones similares. En general, hechas al mismo tiempo y para el mismo objeto. P.ej.: el alto y ancho del tronco de un árbol.

Una tabla de datos

Variable

Valor

Observación / DataPoint

	year	month	day	dep_time	sched_dep_time	dep_delay	arr_time	sched_arr_time	arr_delay	carrier	flight	tailnum	origin	dest	air_time	distance	hour	minute	time_hour
0	2013	1	1	517.0	515	2.0	830.0	819	11.0	UA	1545	N14228	EWR	IAH	227.0	1400	5	15	2013-01-01T10:00:00Z
1	2013	1	1	533.0	529	4.0	850.0	830	20.0	UA	1714	N24211	LGA	IAH	227.0	1416	5	29	2013-01-01T10:00:00Z
2	2013	1	1	542.0	540	2.0	923.0	850	33.0	AA	1141	N619AA	JFK	MIA	160.0	1089	5	40	2013-01-01T10:00:00Z
3	2013	1	1	544.0	545	-1.0	1004.0	1022	-18.0	B6	725	N804JB	JFK	BQN	183.0	1576	5	45	2013-01-01T10:00:00Z
4	2013	1	1	554.0	600	-6.0	812.0	837	-25.0	DL	461	N668DN	LGA	ATL	116.0	762	6	0	2013-01-01T11:00:00Z
5	2013	1	1	554.0	558	-4.0	740.0	728	12.0	UA	1696	N39463	EWR	ORD	150.0	719	5	58	2013-01-01T10:00:00Z
6	2013	1	1	555.0	600	-5.0	913.0	854	19.0	B6	507	N516JB	EWR	FLL	158.0	1065	6	0	2013-01-01T11:00:00Z
7	2013	1	1	557.0	600	-3.0	709.0	723	-14.0	EV	5708	N829AS	LGA	IAD	53.0	229	6	0	2013-01-01T11:00:00Z
8	2013	1	1	557.0	600	-3.0	838.0	846	-8.0	B6	79	N593JB	JFK	MCO	140.0	944	6	0	2013-01-01T11:00:00Z
9	2013	1	1	558.0	600	-2.0	753.0	745	8.0	AA	301	N3ALAA	LGA	ORD	138.0	733	6	0	2013-01-01T11:00:00Z
10	2013	1	1	558.0	600	-2.0	849.0	851	-2.0	B6	49	N793JB	JFK	PBI	149.0	1028	6	0	2013-01-01T11:00:00Z
11	2013	1	1	558.0	600	-2.0	853.0	856	-3.0	B6	71	N657JB	JFK	TPA	158.0	1005	6	0	2013-01-01T11:00:00Z
12	2013	1	1	558.0	600	-2.0	924.0	917	7.0	UA	194	N29129	JFK	LAX	345.0	2475	6	0	2013-01-01T11:00:00Z
13	2013	1	1	558.0	600	-2.0	923.0	937	-14.0	UA	1124	N53441	EWR	SFO	361.0	2565	6	0	2013-01-01T11:00:00Z
14	2013	1	1	559.0	600	-1.0	941.0	910	31.0	AA	707	N3DUAA	LGA	DFW	257.0	1389	6	0	2013-01-01T11:00:00Z
15	2013	1	1	559.0	559	0.0	702.0	706	-4.0	B6	1806	N708JB	JFK	BOS	44.0	187	5	59	2013-01-01T10:00:00Z
16	2013	1	1	559.0	600	-1.0	854.0	902	-8.0	UA	1187	N76515	EWR	LAS	337.0	2227	6	0	2013-01-01T11:00:00Z
17	2013	1	1	600.0	600	0.0	851.0	858	-7.0	B6	371	N595JB	LGA	FLL	152.0	1076	6	0	2013-01-01T11:00:00Z
18	2013	1	1	600.0	600	0.0	837.0	825	12.0	MQ	4650	N542MQ	LGA	ATL	134.0	762	6	0	2013-01-01T11:00:00Z
19	2013	1	1	601.0	600	1.0	844.0	850	-6.0	B6	343	N644JB	EWR	PBI	147.0	1023	6	0	2013-01-01T11:00:00Z
20	2013	1	1	602.0	610	-8.0	812.0	820	-8.0	DL	1919	N971DL	LGA	MSP	170.0	1020	6	10	2013-01-01T11:00:00Z
21	2013	1	1	602.0	605	-3.0	821.0	805	16.0	MQ	4401	N730MQ	LGA	DTW	105.0	502	6	5	2013-01-01T11:00:00Z
22	2013	1	1	606.0	610	-4.0	858.0	910	-12.0	AA	1895	N633AA	EWR	MIA	152.0	1085	6	10	2013-01-01T11:00:00Z
23	2013	1	1	606.0	610	-4.0	837.0	845	-8.0	DL	1743	N3739P	JFK	ATL	128.0	760	6	10	2013-01-01T11:00:00Z
24	2013	1	1	607.0	607	0.0	858.0	915	-17.0	UA	1077	N53442	EWR	MIA	157.0	1085	6	7	2013-01-01T11:00:00Z
25	2013	1	1	608.0	600	8.0	807.0	735	32.0	MQ	3768	N9EAMQ	EWR	ORD	139.0	719	6	0	2013-01-01T11:00:00Z
26	2013	1	1	611.0	600	11.0	945.0	931	14.0	UA	303	N532UA	JFK	SFO	366.0	2586	6	0	2013-01-01T11:00:00Z
27	2013	1	1	613.0	610	3.0	925.0	921	4.0	B6	135	N635JB	JFK	RSW	175.0	1074	6	10	2013-01-01T11:00:00Z
28	2013	1	1	615.0	615	0.0	1039.0	1100	-21.0	B6	709	N794JB	JFK	SJU	182.0	1598	6	15	2013-01-01T11:00:00Z
29	2013	1	1	615.0	615	0.0	833.0	842	-9.0	DL	575	N326NB	EWR	ATL	120.0	746	6	15	2013-01-01T11:00:00Z

Formato .csv

- Comma-separated-values
- Son archivos de texto, donde las columnas están separadas por una coma.
- Formato estándar para transmitir datos ordenados.

```
long,lat,nro_registro,comuna,calle_nombre,calle_altura,calle_chap
    nombre_cientifico,ancho_acera,estado_plantera,ubicacion_plant
    nivel_plantera,diametro_altura_pecho,altura_arbol
-58.52246854385779,-34.635242540536105,248212,9,Alvarez Jonte Av.
    Lagerstroemia,4.4,Ocupada,Regular,Bajo nivel,4,2
-58.522629550526105,-34.6354294300277,248235,9,Alvarez Jonte Av.,
    Melia,4.4,Ocupada,Regular,A nivel,70,10
-58.522684484956706,-34.635493195266,248260,9,Alvarez Jonte Av.,6
    4.4,Ocupada,Regular,A nivel,73,10
-58.5228272952872,-34.635658962044204,248278,9,Alvarez Jonte Av.,
    Melia,4.4,Ocupada,Regular,A nivel,57,11
-58.5221122386389,-34.634885150978604,248675,9,Alvarez Jonte Av.,
    Melia,4,Ocupada,Regular,A nivel,58,9
-58.522164769565705,-34.6349430221208,248701,9,Alvarez Jonte Av.,
    Melia,4,Ocupada,Regular,A nivel,57,9
-58.5218837652316,-34.634633450096196,248619,9,Alvarez Jonte Av.,
    Lagerstroemia,4,Ocupada,Regular,Elevada,30,8
-58.521941356652896,-34.634696896668004,248632,9,Alvarez Jonte Av
    Melia,4,Ocupada,Regular,A nivel,77,8
-58.522055938054294,-34.634823126844694,248655,9,Alvarez Jonte Av
    Ficus,4,Ocupada,Regular,A nivel,35,6
-58.51345241543199,-34.6258097317006,290862,10,Alvarez Jonte Av.,
    Tilia,4.2,Ocupada,Regular,A nivel,25,6
-58.5131743311085,-34.6255628519808,290866,10,Alvarez Jonte Av.,5
    Lagerstroemia,4.2,Sobreocupada,Regular,A nivel,10,4
-58.513109978134395,-34.6255057199335,290867,10,Alvarez Jonte Av.
    Lagerstroemia,4.2,Sobreocupada,Regular,A nivel,8,3
-58.5130459627116,-34.6254488874498,290868,10,Alvarez Jonte Av.,5
    Lagerstroemia,4.2,Sobreocupada,Regular,A nivel,8,3
-58.512806797262705,-34.6252349281503,290871,10,Alvarez Jonte Av.
    Ficus,4.2,Ocupada,Ochava,Elevada,131,8
-58.50746009448001,-34.6206822126709,290611,10,Alvarez Jonte Av.,
```

Tipos de datos

Enteros (int)

Números de punto flotante (float)

Tiempos (date / datetime)

Cadenas de caracteres (string)

	year	month	day	dep_time	sched_dep_time	dep_delay	arr_time	sched_arr_time	arr_delay	carrier	flight	tailnum	origin	dest	air_time	distance	hour	minute	time_hour
0	2013	1	1	517.0	515	2.0	830.0	819	11.0	UA	1545	N14228	EWR	IAH	227.0	1400	5	15	2013-01-01T10:00:00Z
1	2013	1	1	533.0	529	4.0	850.0	830	20.0	UA	1714	N24211	LGA	IAH	227.0	1416	5	29	2013-01-01T10:00:00Z
2	2013	1	1	542.0	540	2.0	923.0	850	33.0	AA	1141	N619AA	JFK	MIA	160.0	1089	5	40	2013-01-01T10:00:00Z
3	2013	1	1	544.0	545	-1.0	1004.0	1022	-18.0	B6	725	N804JB	JFK	BQN	183.0	1576	5	45	2013-01-01T10:00:00Z
4	2013	1	1	554.0	600	-6.0	812.0	837	-25.0	DL	461	N668DN	LGA	ATL	116.0	762	6	0	2013-01-01T11:00:00Z
5	2013	1	1	554.0	558	-4.0	740.0	728	12.0	UA	1696	N39463	EWR	ORD	150.0	719	5	58	2013-01-01T10:00:00Z
6	2013	1	1	555.0	600	-5.0	913.0	854	19.0	B6	507	N516JB	EWR	FLL	158.0	1065	6	0	2013-01-01T11:00:00Z
7	2013	1	1	557.0	600	-3.0	709.0	723	-14.0	EV	5708	N829AS	LGA	IAD	53.0	229	6	0	2013-01-01T11:00:00Z
8	2013	1	1	557.0	600	-3.0	838.0	846	-8.0	B6	79	N593JB	JFK	MCO	140.0	944	6	0	2013-01-01T11:00:00Z
9	2013	1	1	558.0	600	-2.0	753.0	745	8.0	AA	301	N3ALAA	LGA	ORD	138.0	733	6	0	2013-01-01T11:00:00Z
10	2013	1	1	558.0	600	-2.0	849.0	851	-2.0	B6	49	N793JB	JFK	PBI	149.0	1028	6	0	2013-01-01T11:00:00Z
11	2013	1	1	558.0	600	-2.0	853.0	856	-3.0	B6	71	N657JB	JFK	TPA	158.0	1005	6	0	2013-01-01T11:00:00Z
12	2013	1	1	558.0	600	-2.0	924.0	917	7.0	UA	194	N29129	JFK	LAX	345.0	2475	6	0	2013-01-01T11:00:00Z
13	2013	1	1	558.0	600	-2.0	923.0	937	-14.0	UA	1124	N53441	EWR	SFO	361.0	2565	6	0	2013-01-01T11:00:00Z
14	2013	1	1	559.0	600	-1.0	941.0	910	31.0	AA	707	N3DUAA	LGA	DFW	257.0	1389	6	0	2013-01-01T11:00:00Z
15	2013	1	1	559.0	559	0.0	702.0	706	-4.0	B6	1806	N708JB	JFK	BOS	44.0	187	5	59	2013-01-01T10:00:00Z
16	2013	1	1	559.0	600	-1.0	854.0	902	-8.0	UA	1187	N76515	EWR	LAS	337.0	2227	6	0	2013-01-01T11:00:00Z
17	2013	1	1	600.0	600	0.0	851.0	858	-7.0	B6	371	N595JB	LGA	FLL	152.0	1076	6	0	2013-01-01T11:00:00Z
18	2013	1	1	600.0	600	0.0	837.0	825	12.0	MQ	4650	N542MQ	LGA	ATL	134.0	762	6	0	2013-01-01T11:00:00Z
19	2013	1	1	601.0	600	1.0	844.0	850	-6.0	B6	343	N644JB	EWR	PBI	147.0	1023	6	0	2013-01-01T11:00:00Z
20	2013	1	1	602.0	610	-8.0	812.0	820	-8.0	DL	1919	N971DL	LGA	MSP	170.0	1020	6	10	2013-01-01T11:00:00Z
21	2013	1	1	602.0	605	-3.0	821.0	805	16.0	MQ	4401	N730MQ	LGA	DTW	105.0	502	6	5	2013-01-01T11:00:00Z
22	2013	1	1	606.0	610	-4.0	858.0	910	-12.0	AA	1895	N633AA	EWR	MIA	152.0	1085	6	10	2013-01-01T11:00:00Z
23	2013	1	1	606.0	610	-4.0	837.0	845	-8.0	DL	1743	N3739P	JFK	ATL	128.0	760	6	10	2013-01-01T11:00:00Z
24	2013	1	1	607.0	607	0.0	858.0	915	-17.0	UA	1077	N53442	EWR	MIA	157.0	1085	6	7	2013-01-01T11:00:00Z
25	2013	1	1	608.0	600	8.0	807.0	735	32.0	MQ	3768	N9EAMQ	EWR	ORD	139.0	719	6	0	2013-01-01T11:00:00Z
26	2013	1	1	611.0	600	11.0	945.0	931	14.0	UA	303	N532UA	JFK	SFO	366.0	2586	6	0	2013-01-01T11:00:00Z
27	2013	1	1	613.0	610	3.0	925.0	921	4.0	B6	135	N635JB	JFK	RSW	175.0	1074	6	10	2013-01-01T11:00:00Z
28	2013	1	1	615.0	615	0.0	1039.0	1100	-21.0	B6	709	N794JB	JFK	SJU	182.0	1598	6	15	2013-01-01T11:00:00Z
29	2013	1	1	615.0	615	0.0	833.0	842	-9.0	DL	575	N326NB	EWR	ATL	120.0	746	6	15	2013-01-01T11:00:00Z

Variables categóricas

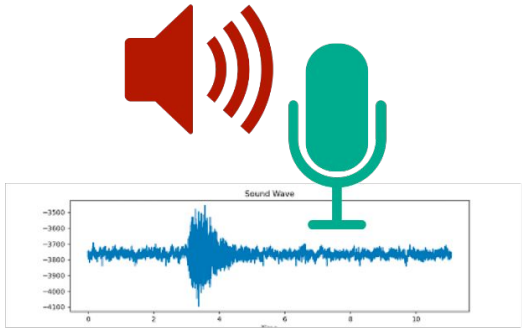
Todos los tipos pueden funcionar como **variables categóricas**

Dividen al dataset en grupos, con un número significativo de miembros en cada grupo.

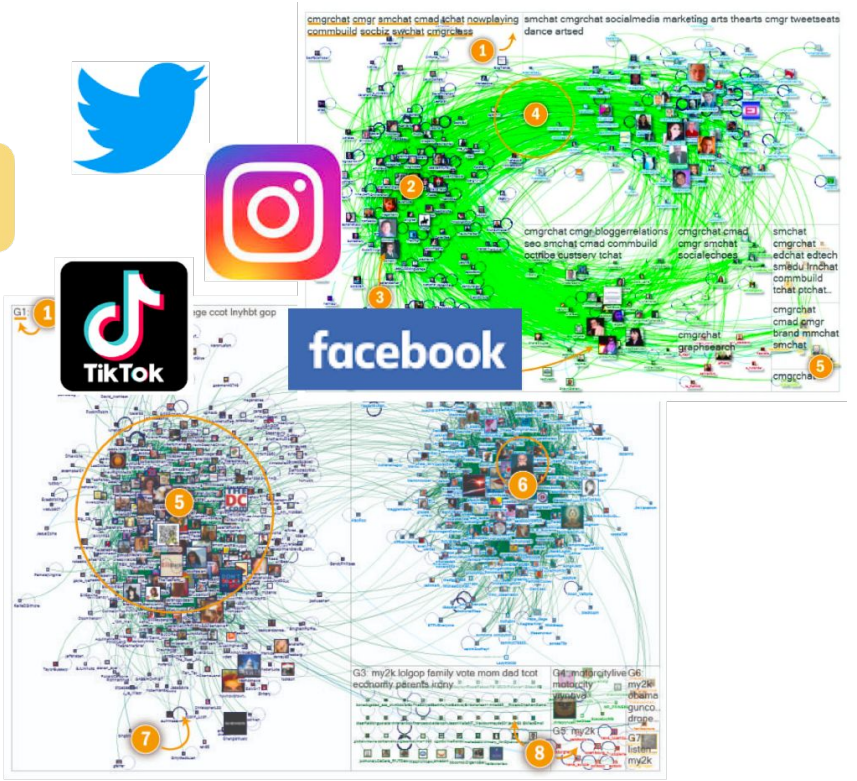
¿Cómo venimos con esto?

https://www.menti.com/alzt591v9o6t

Datos no estructurados



Exhaust data



FUENTE: http://www.pewinternet.org/2014/02/20/mapping-twitter-topic-networks-from-polarized-crowds-to-community-clusters

Armado del dataset

A diferencia de otras ciencias (física, geología, química, ...), el objeto de estudio no viene dado por la Naturaleza. Es un producto humano.

Como tal, los dataset son resultados de procesos y decisiones. Hay algo de abstracción, para arrancar, como mínimo.

En esas decisiones, se eligen los atributos, etc. (p.ej. para los cumpleaños elegimos poner el día, mes y año, pero no la hora).

En general, puede pasar que uno no tiene acceso a las personas u organizaciones que lo armaron, pero es importante tener conciencia de que se trata de un proceso humano.

Práctica

En este módulo:

- Vemos los pasos del proceso CRISP-DM en notebooks.
- Presentamos algunos conceptos de Machine Learning para el módulo 3.
- Ganamos experiencia en el tratamiento y análisis de datos de manera hands-on

```
1   long,lat,nro_registro,comuna,calle_nombre,calle_altura,calle_chap
    nombre_cientifico,ancho_acera,estado_plantera,ubicacion_plant
    nivel_plantera,diametro_altura_pecho,altura_arbol
2   -58.52246854385779,-34.635242540536105,248212,9,Alvarez Jonte Av.
    Lagerstroemia,4.4,Ocupada,Regular,Bajo nivel,4,2
3   -58.522629550526105,-34.6354294300277,248235,9,Alvarez Jonte Av.,
    Melia,4.4,Ocupada,Regular,A nivel,70,10
4   -58.522684484956706,-34.635493195266,248260,9,Alvarez Jonte Av.,6
    4.4,Ocupada,Regular,A nivel,73,10
5   -58.5228272952872,-34.635658962044204,248278,9,Alvarez Jonte Av.,
    Melia,4.4,Ocupada,Regular,A nivel,57,11
6   -58.5221122386389,-34.634885150978604,248675,9,Alvarez Jonte Av.,
    Melia,4,Ocupada,Regular,A nivel,58,9
7   -58.522164769565705,-34.6349430221208,248701,9,Alvarez Jonte Av.,
    Melia,4,Ocupada,Regular,A nivel,57,9
8   -58.5218837652316,-34.634633450096196,248619,9,Alvarez Jonte Av.,
    Lagerstroemia,4,Ocupada,Regular,Elevada,30,8
9   -58.521941356652896,-34.634696896668004,248632,9,Alvarez Jonte Av
    Melia,4,Ocupada,Regular,A nivel,77,8
10  -58.522055938054294,-34.634823126844694,248655,9,Alvarez Jonte Av
    Ficus,4,Ocupada,Regular,A nivel,35,6
11  -58.51345241543199,-34.6258097317006,290862,10,Alvarez Jonte Av.,
    Tilia,4.2,Ocupada,Regular,A nivel,25,6
12  -58.5131743311085,-34.6255628519808,290866,10,Alvarez Jonte Av.,5
    Lagerstroemia,4.2,Sobreocupada,Regular,A nivel,10,4
13  -58.513109978134395,-34.6255057199335,290867,10,Alvarez Jonte Av.
    Lagerstroemia,4.2,Sobreocupada,Regular,A nivel,8,3
14  -58.5130459627116,-34.6254488874498,290868,10,Alvarez Jonte Av.,5
    Lagerstroemia,4.2,Sobreocupada,Regular,A nivel,8,3
15  -58.512806797262705,-34.6252349281503,290871,10,Alvarez Jonte Av.
    Ficus,4.2,Ocupada,Ochava,Elevada,131,8
16  -58.50746009448001,-34.6206822126709,290611,10,Alvarez Jonte Av.,
```

Argentina programa 4.0