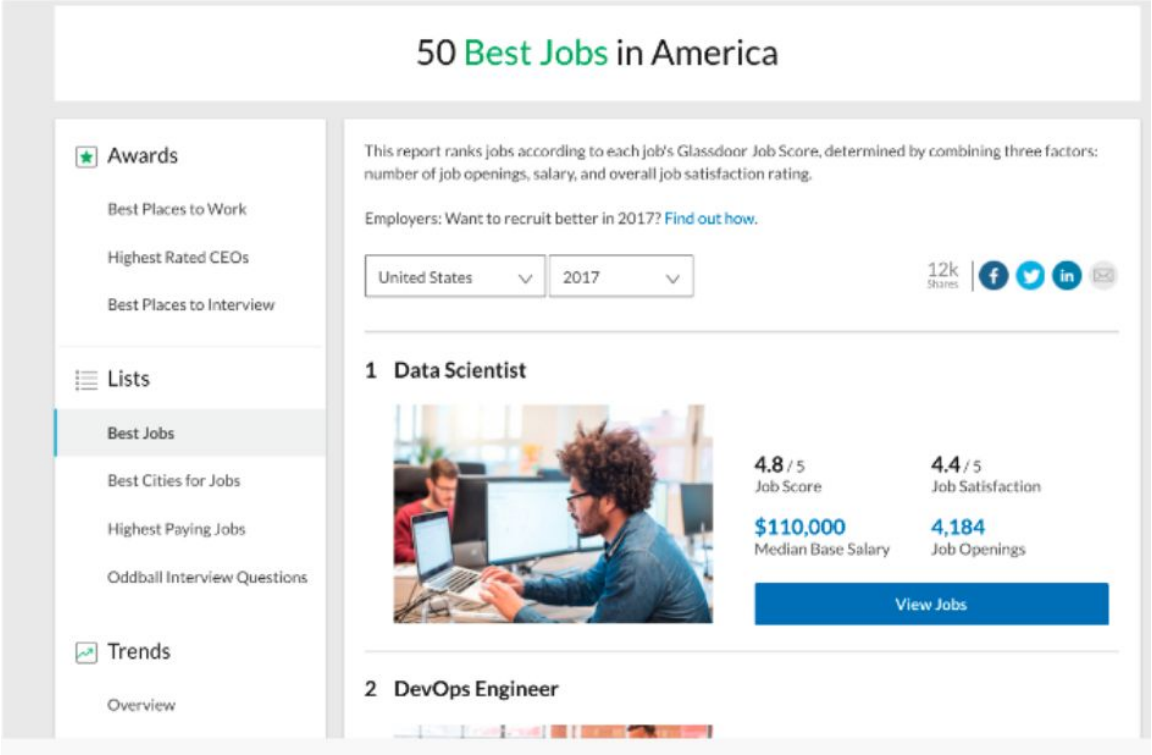
Introducción

75.06 / 95.58 Organización de Datos

Ciencia de Datos / Data Science



Los científicos de datos, las "estrellas" del mercado laboral que rechazan ofertas todas las semanas

Trabajan con el análisis de grandes volúmenes de datos; reciben al menos 3 propuestas laborales por semana y tienen sueldos altos; las universidades…

LANACION.COM.AR

15%

projected increase in data science jobs by 2020

110,000

projected new jobs for data-driven decision makers by 2020

$80K

is the average salary for data science and analytics jobs

Thanos (Marvel character) Avengers: Infinity War (2018 movie) +2

Did Thanos use Data Science to find the infinity stones?

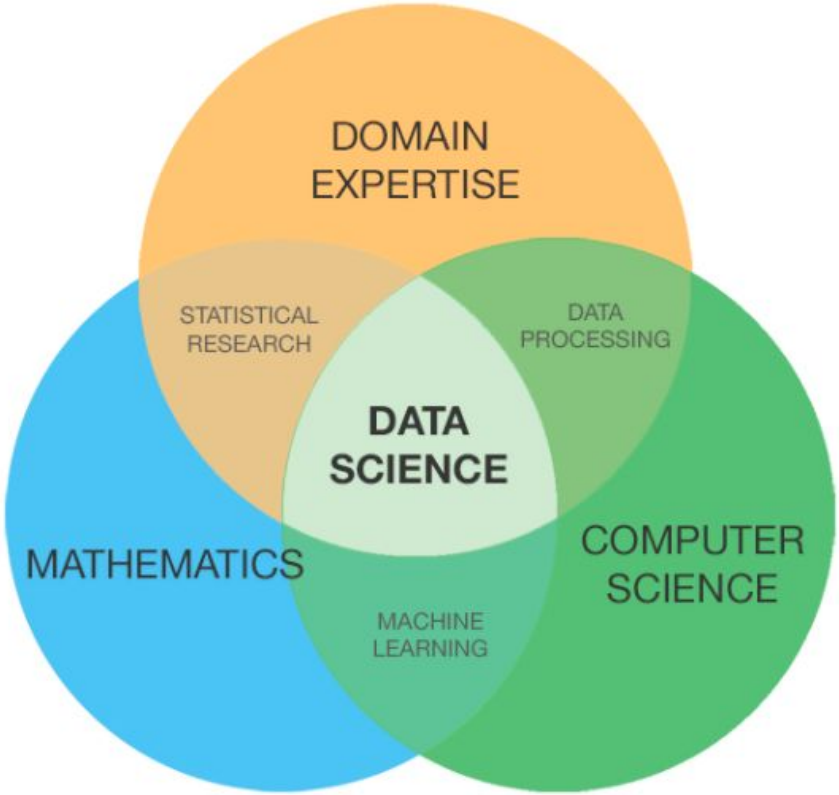
Answer Follow · 9 Request

Luis, can you answer this question?

People are searching for a better answer to this question.

Answer

Ciencia de Datos / Data Science



Roles

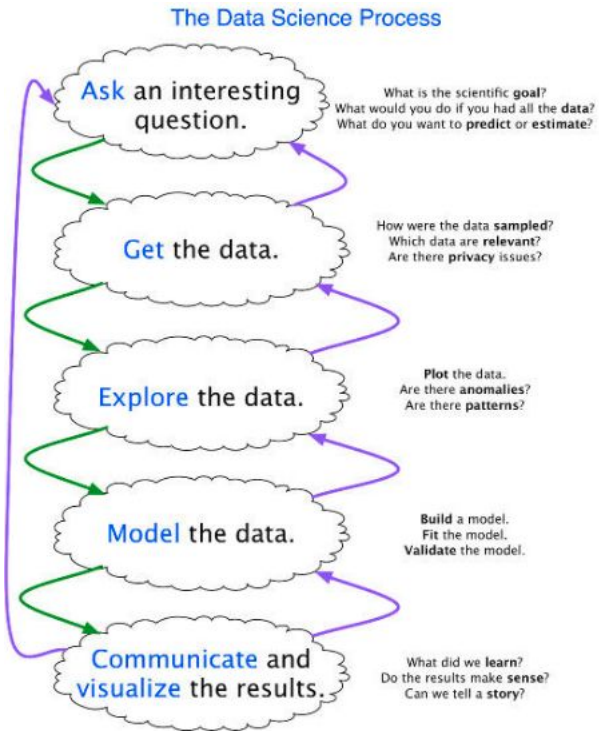
- Data Analyst
- Data Scientist
- Data Engineer
- Machine Learning Engineer

Áreas de Conocimiento

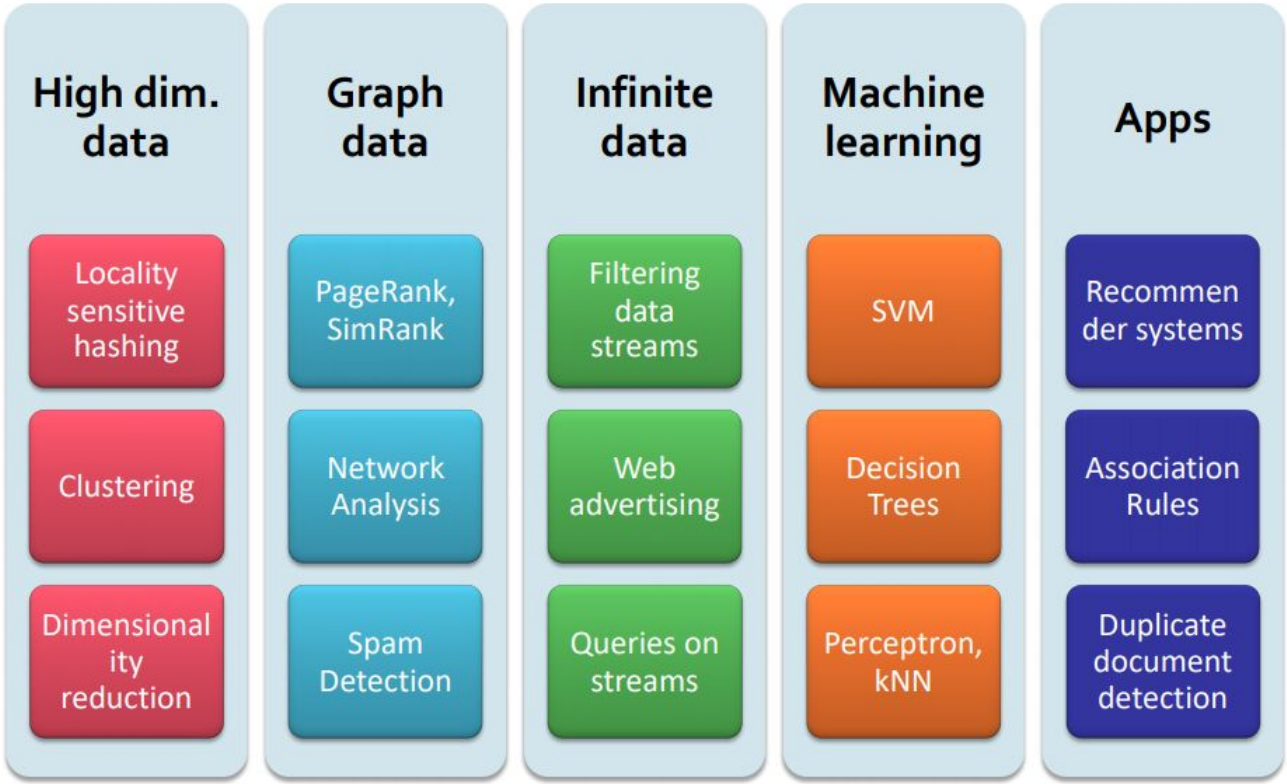
- Big Data, procesamiento distribuido (Spark, Databricks, etc)
- Machine Learning
- Deep Learning
- Análisis de Redes Sociales
- Sistemas de Recomendación
- Procesamiento de Lenguaje Natural
- Visualización de Datos
- Algoritmos de Streaming
- Bases de Datos, SQL
- No SQL (Cassandra, Mongo, etc)

Data Science: El Proceso

- Hacer una Pregunta Interesante
- Conseguir los Datos
- Explorar los Datos
- Modelar los Datos
- Comunicar y Visualizar los Resultados



Temas


High dim. data
Locality sensitive hashing
Clustering
Dimensionality reduction
Graph data
PageRank, SimRank
Network Analysis
Spam Detection
Infinite data
Filtering data streams
Web advertising
Queries on streams
Machine learning
SVM
Decision Trees
Perceptron, kNN
Apps
Recommender systems
Association Rules
Duplicate document detection

Temas a Desarrollar hoy

- El peor paper de la historia
- La ecuación más peligrosa de la historia
- Correlación vs Causa
- Biases
- Big Data
- Streaming
- Formatos de Datos
- Algo sobre herramientas

Data Science en Acción

Rayid Ghani

- Datos de TODOS los votantes
- Modelos de Machine Learning para predecir 3 cosas:
  - Probabilidad de votar
  - Probabilidad de votar a Obama
  - Probabilidad de cambiar de opinión

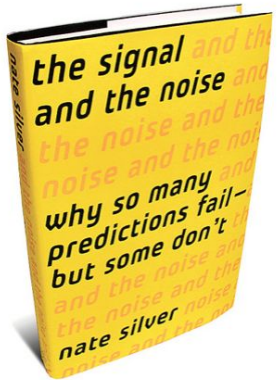
“Nate Silver won the election” – Harvard Business Review

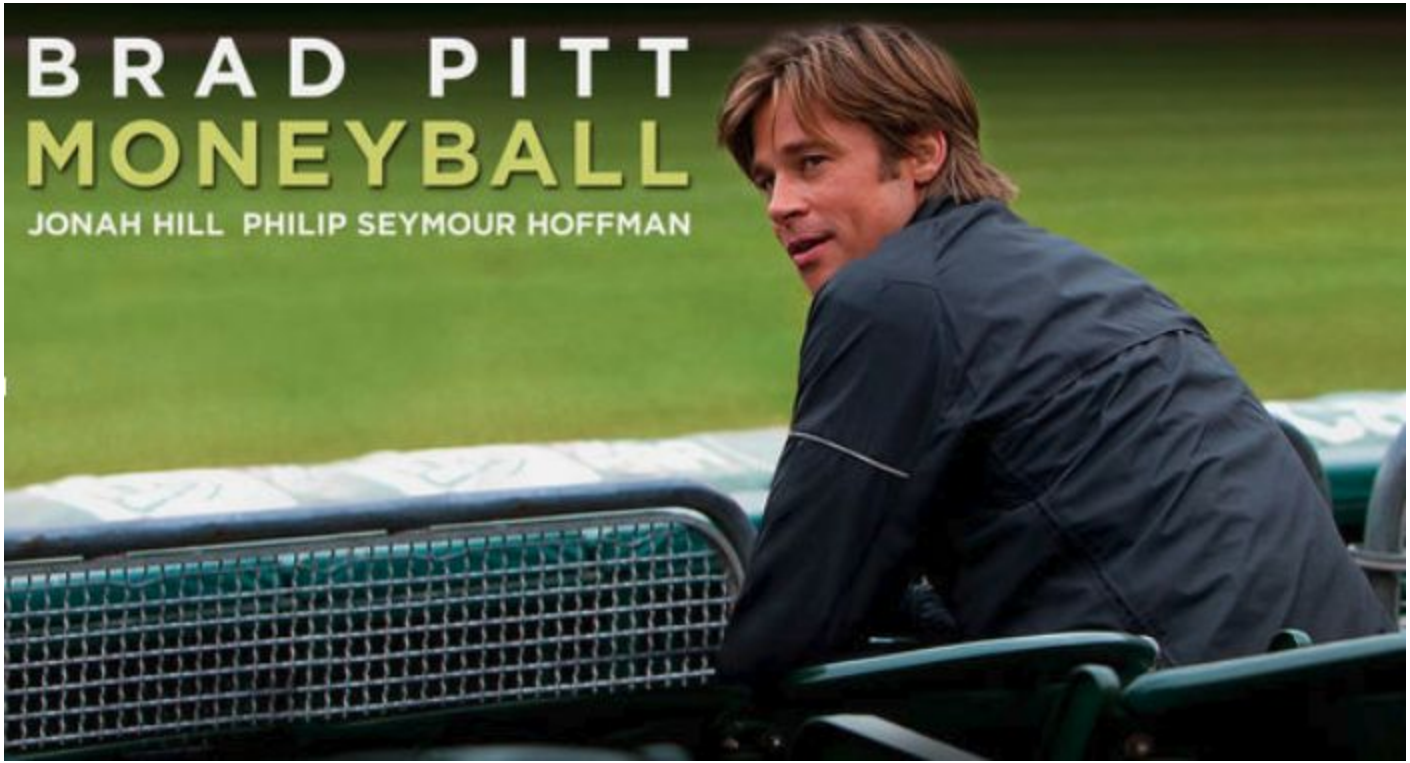
William Chen @wzchen 11 Nov

There is no such thing as missing data, only data that Nate Silver has not chosen to reveal to you. #natesilverfacts

Retweeted by Rodrigo Aldecoa and 1 other

Expand Reply Retweeted Favorite More




BRAD PITT
MONEYBALL
JONAH HILL PHILIP SEYMOUR HOFFMAN

No todos son ejemplos felices...

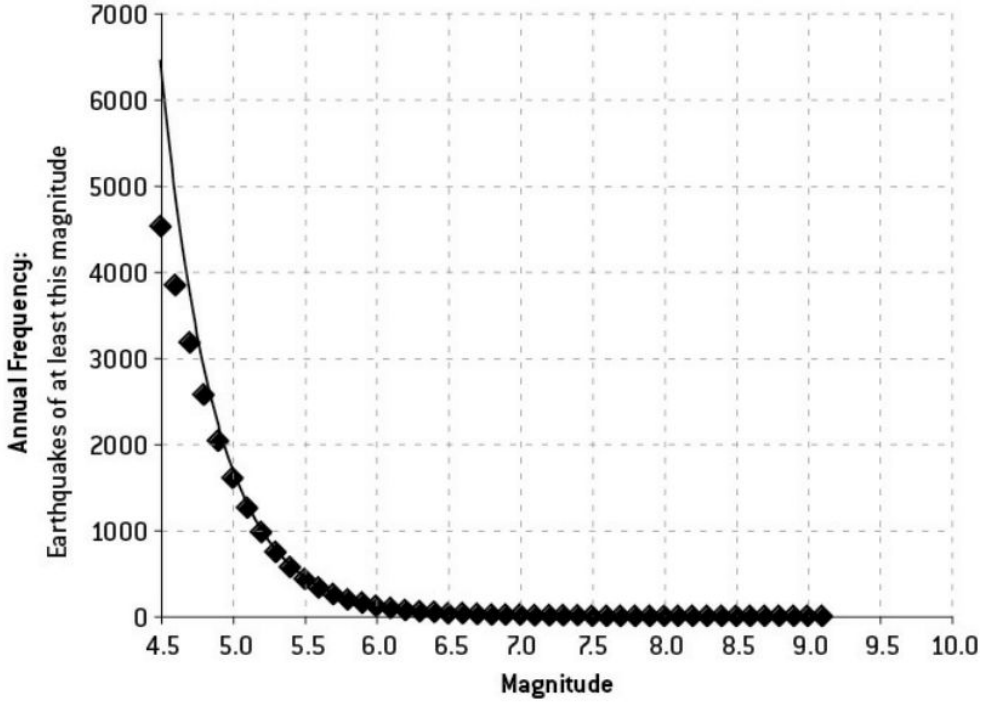

7000
6000
5000
4000
3000
2000
1000
0
Annual Frequency:
Earthquakes of at least this magnitude
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
8.5
9.0
9.5
10.0
Magnitude

FIGURE 5-7A: TŌHOKU, JAPAN EARTHQUAKE FREQUENCIES
JANUARY 1, 1964–MARCH 10, 2011

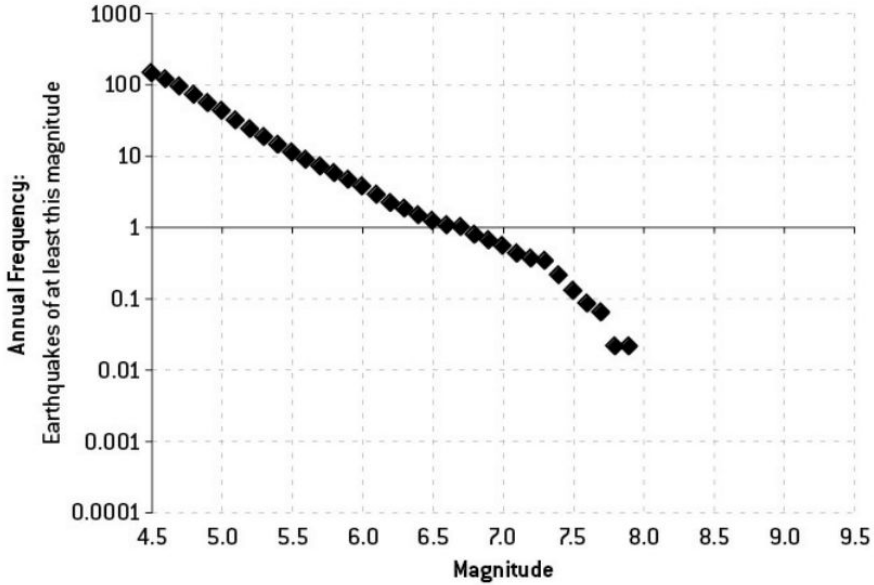


FIGURE 5-7C: TŌHOKU, JAPAN EARTHQUAKE FREQUENCIES CHARACTERISTIC FIT

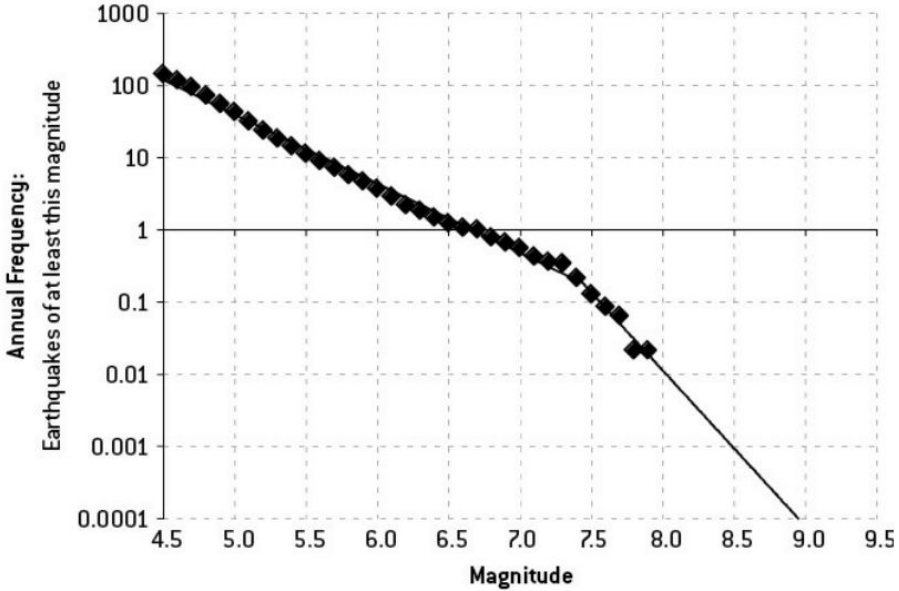
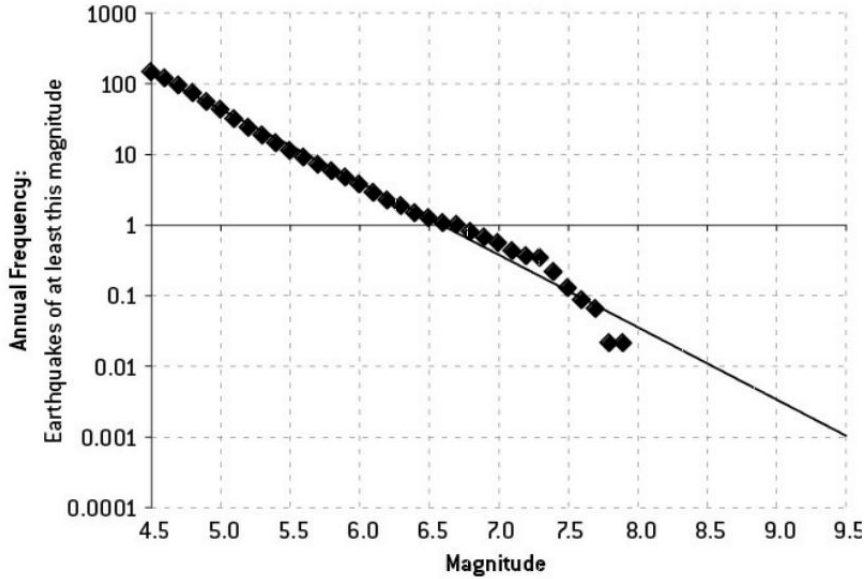


FIGURE 5-7B: TŌHOKU, JAPAN EARTHQUAKE FREQUENCIES
GUTENBERG-RICHTER FIT




NO.362
NO.352

Mail Online

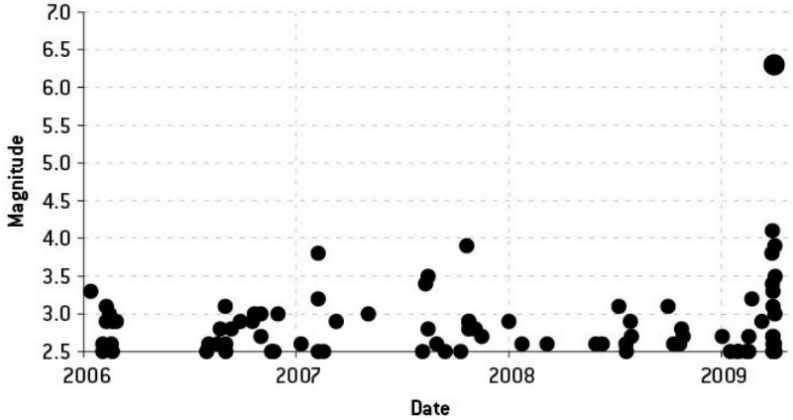
Home | News | U.S. | Sport | TV&Showbiz | Australia | Femail | Health | Science | Money |

Latest Headlines | News | Arts | Headlines | Pictures | Most read | News Board | Wires

Scientists jailed for manslaughter because they did not predict deadly earthquake in Italy which killed 309 people have been cleared

- Town of L'Aquila was struck by quake, which measured 6.3 on Richter Scale
- Hundreds were killed and thousands were left homeless in 2009 disaster
- Scientists visited town days before but concluded there was little risk
- They were sentenced to six years each in prison following 2012 trial
- Some of the Italy's most respected seismologists were among those jailed

FIGURE 5-4A: EARTHQUAKES NEAR L'AQUILA, ITALY
JANUARY 1, 2006–APRIL 6, 2009



El Peor Paper de la Historia

Proceedings of the National Academy of Sciences of the United States of America

Evidence on the impact of sustained exposure to air pollution on life expectancy from China's Huai River policy

Yuyu Chen[a,1], Avraham Ebenstein[b,1], Michael Greenstone[c,d,1,2], and Hongbin Li[e,1]

This paper's findings suggest that an arbitrary Chinese policy that greatly increases total suspended particulates (TSPs) air pollution is causing the 500 million residents of Northern China to lose more than 2.5 billion life years of life expectancy. The quasi-experimental empirical approach is based on China's Huai River policy, which provided free winter heating via the provision of coal for boilers in cities north of the Huai River but denied heat to the south. Using a regression discontinuity design based on distance from the Huai River, we find that ambient concentrations of TSPs are about 184 $\mu g/m^3$ [95% confidence interval (CI): 61, 307] or 55% higher in the north. Further, the results indicate that life expectancies are about 5.5 y (95% CI: 0.8, 10.2) lower in the north owing to an increased incidence of cardiorespiratory mortality. More generally, the analysis suggests that long-term exposure to an additional 100 $\mu g/m^3$ of TSPs is associated with a reduction in life expectancy at birth of about 3.0 y (95% CI: 0.4, 5.6).

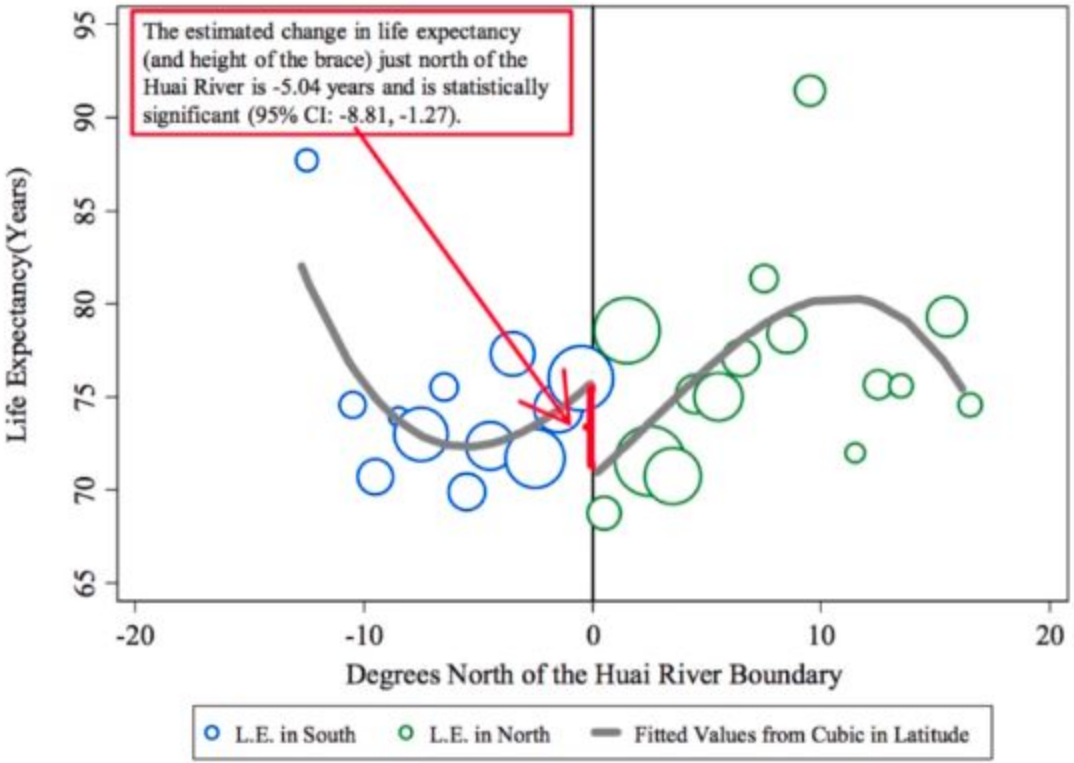


Fig. 3. The plotted line reports the fitted values from a regression of life expectancy on a cubic in latitude using the sample of DSP locations, weighted by the population at each location.

The New York Times

Asia Pacific

Pollution Leads to Drop in Life Span in Northern China, Research Finds

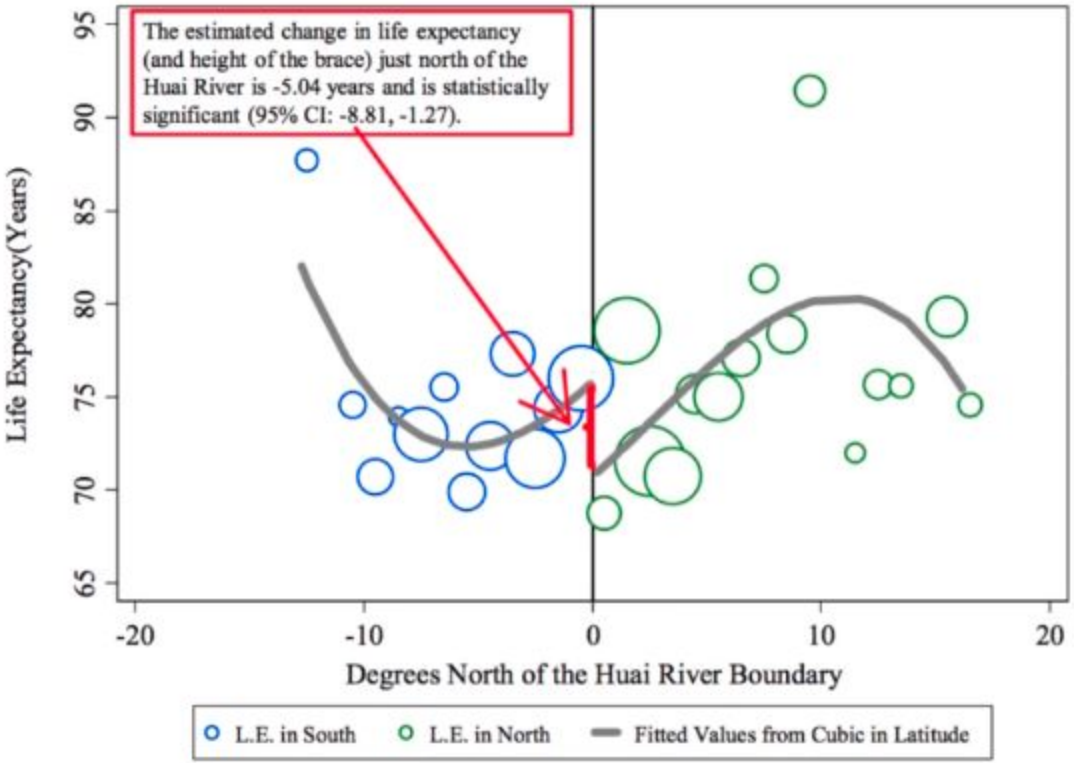


Fig. 3. The plotted line reports the fitted values from a regression of life expectancy on a cubic in latitude using the sample of DSP locations, weighted by the population at each location.

Breaking News….

Dave Harris @davidjayharris ·Mar 7

Remember that Chinese “criminal faces” computer vision paper?

Turns out that their model is a smile detector. The “non-criminal” photos came from promotional headshots on the web (where people smile) and the “criminal” photos were official ID photos (where they don’t)

Arthur Charpentier @freakonometrics

"Why would this possibly be? There’s a glaringly obvious explanation... As one smiles, the corners of the mouth spread out and the upper lip straightens. Try it yourself in the mirror" (composite faces for criminals (left) and…

Show this thread

La Ecuación más Peligrosa de la Historia

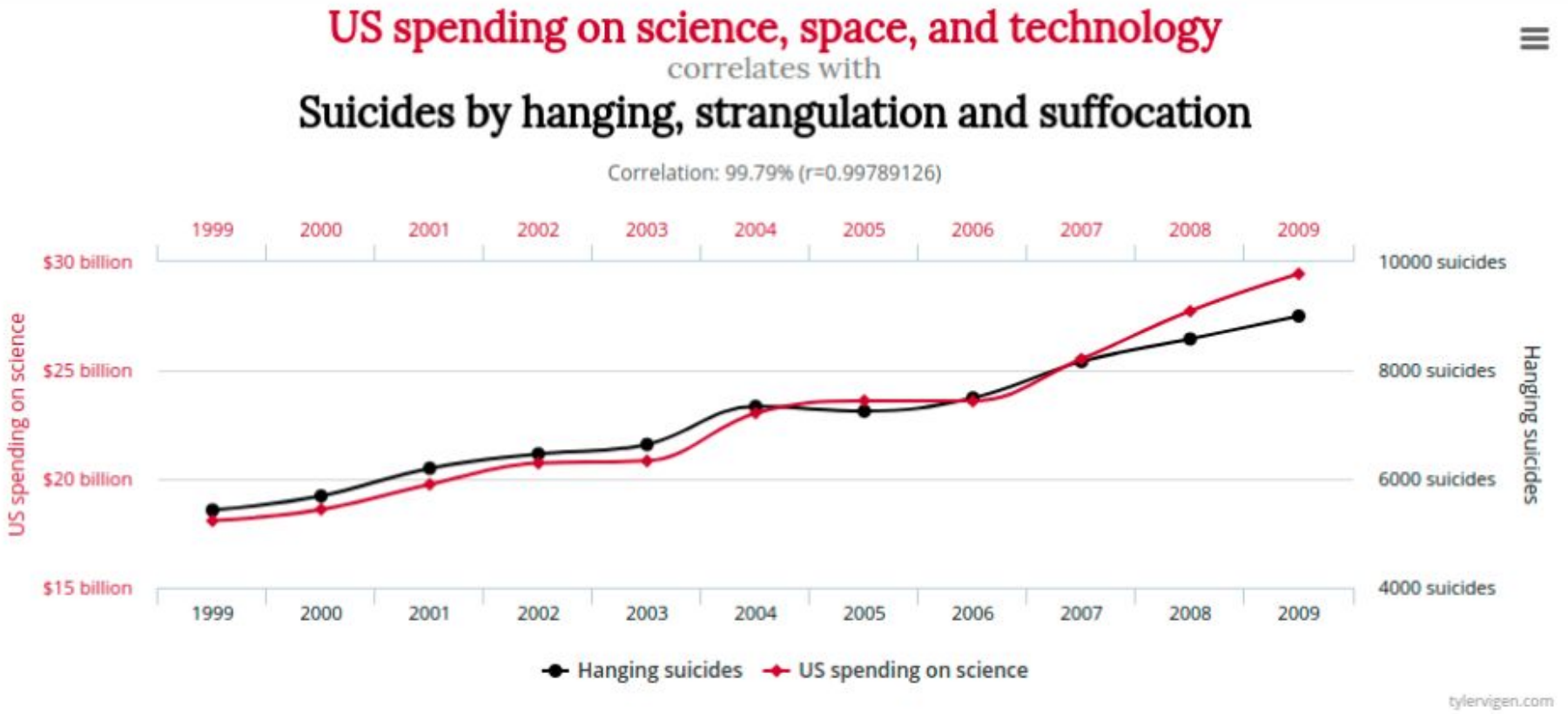
$$\sigma(\bar{x}) = \frac{\sigma}{\sqrt{n}}$$

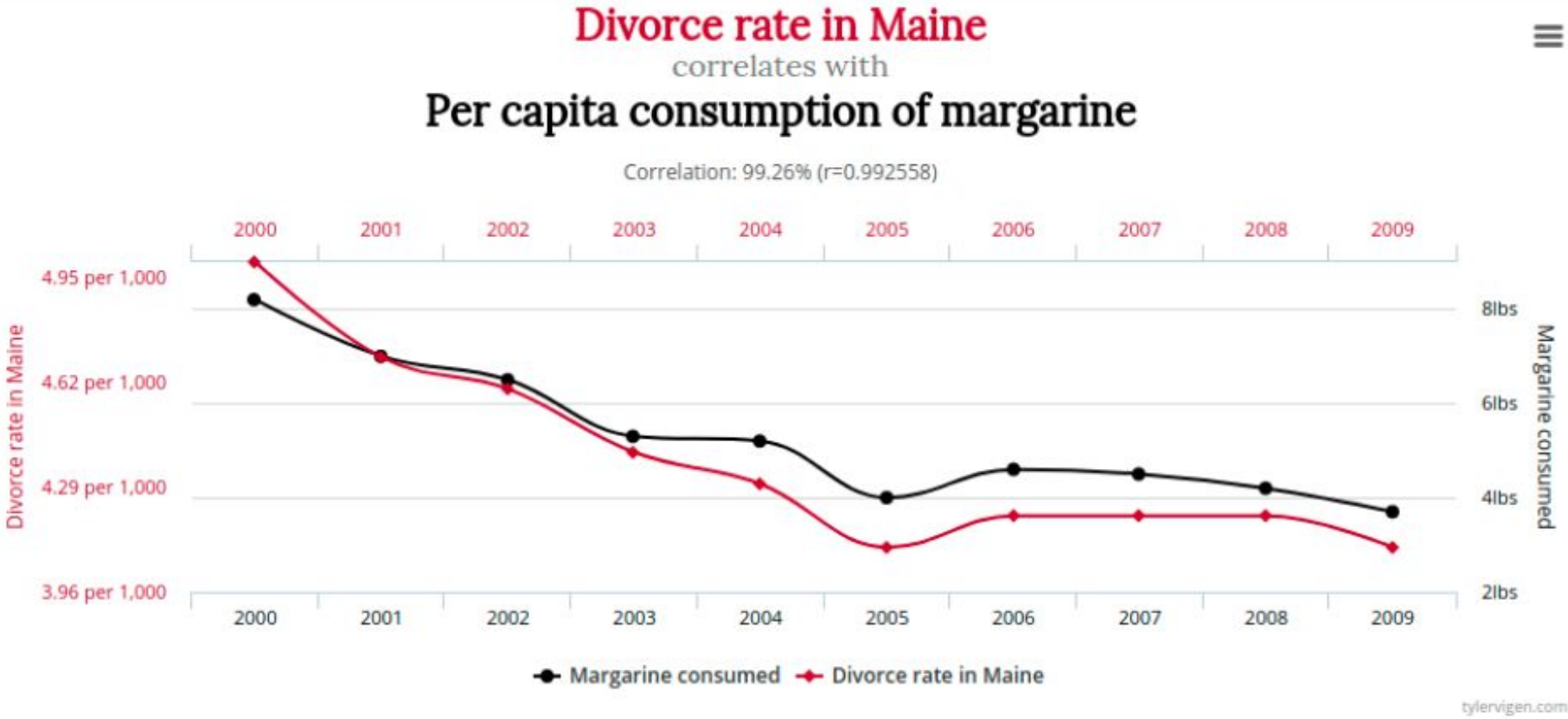
Information on Automobile Accident Rates in 20 Cities

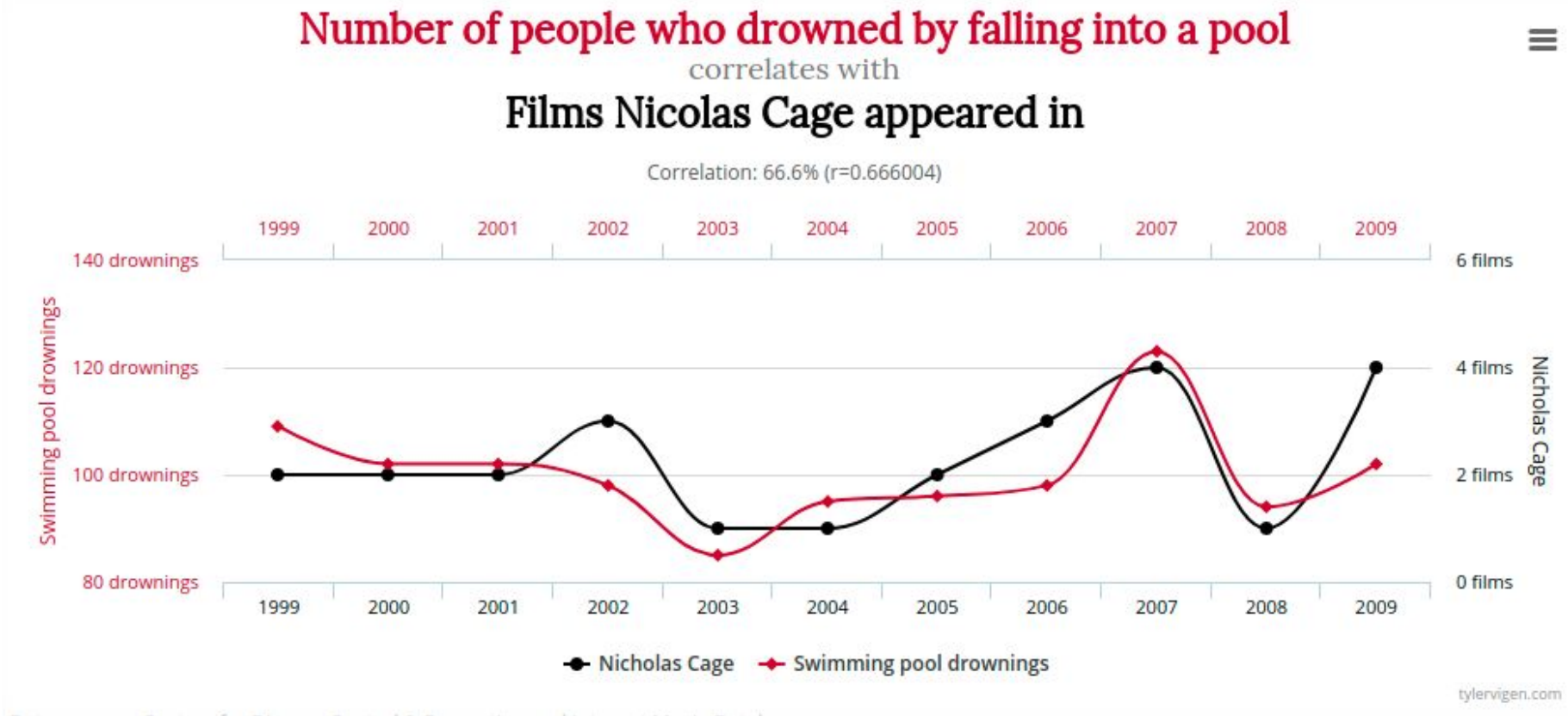
City	*State*	*Population rank*	*Population*	*Number of years between accidents*
Ten safest				
Sioux Falls	South Dakota	170	133,834	14.3
Fort Collins	Colorado	182	125,740	13.2
Cedar Rapids	Iowa	190	122,542	13.2
Huntsville	Alabama	129	164,237	12.8
Chattanooga	Tennessee	138	154,887	12.7
Knoxville	Tennessee	124	173,278	12.6
Des Moines	Iowa	103	196,093	12.6
Milwaukee	Wisconsin	19	586,941	12.5
Colorado Springs	Colorado	48	370,448	12.3
Warren	Michigan	169	136,016	12.3
Ten least safe				
Newark	New Jersey	64	277,911	5.0
Washington	DC	25	563,384	5.1
Elizabeth	New Jersey	189	123,215	5.4
Alexandria	Virginia	174	128,923	5.7
Arlington	Virginia	114	187,873	6.0
Glendale	California	92	200,499	6.1
Jersey City	New Jersey	74	239,097	6.2
Paterson	New Jersey	148	150,782	6.5
San Francisco	California	14	751,682	6.5
Baltimore	Maryland	18	628,670	6.5

Delaware And Oregon Have The Highest Rate Of Sex Offenders Per Capita

Correlación vs Causa


US spending on science, space, and technology
correlates with
Suicides by hanging, strangulation and suffocation
Correlation: 99.79% (r=0.99789126)
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
$30 billion
$25 billion
$20 billion
$15 billion
US spending on science
10000 suicides
8000 suicides
6000 suicides
4000 suicides
Hanging suicides
Hanging suicides
US spending on science
tylervigen.com


Divorce rate in Maine
correlates with
Per capita consumption of margarine
Correlation: 99.26% (r=0.992558)
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
4.95 per 1,000
4.62 per 1,000
4.29 per 1,000
3.96 per 1,000
Divorce rate in Maine
8lbs
6lbs
4lbs
2lbs
Margarine consumed
Margarine consumed
Divorce rate in Maine
tylervigen.com


Number of people who drowned by falling into a pool
correlates with
Films Nicolas Cage appeared in
Correlation: 66.6% (r=0.666004)
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
140 drownings
120 drownings
100 drownings
80 drownings
Swimming pool drownings
6 films
4 films
2 films
0 films
Nicholas Cage
Nicholas Cage
Swimming pool drownings
tylervigen.com

Big Data

SundayReview | NEWS ANALYSIS

The Age of Big Data

By STEVE LOHR FEB. 11, 2012

GOOD with numbers? Fascinated by data? The sound you hear is opportunity knocking.

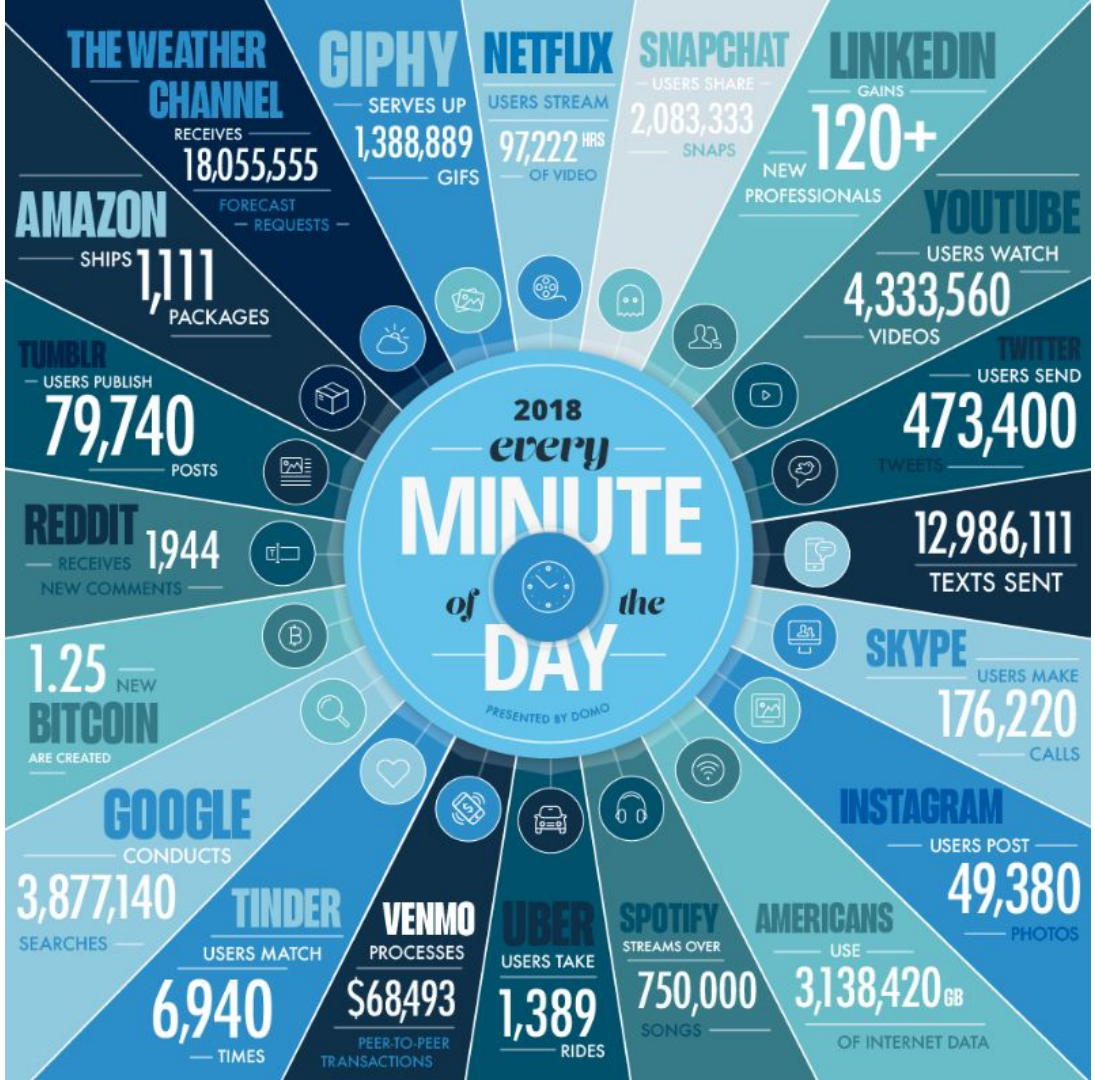
Mo Zhou was snapped up by I.B.M. last summer, as a freshly minted Yale M.B.A., to join the technology company's fast-growing ranks of data consultants. They help businesses make sense of an explosion of data — Web traffic and social network comments, as well as software and sensors that monitor shipments, suppliers and customers — to guide decisions, trim costs and lift sales. "I've always had a love of numbers," says Ms. Zhou, whose job as a data analyst suits her skills.

To exploit the data flood, America will need many more like her. A report last year by the McKinsey Global Institute, the research arm of the consulting firm, projected that the United States needs 140,000 to 190,000 more workers with "deep analytical" expertise and 1.5 million more data-literate managers, whether retrained or hired.

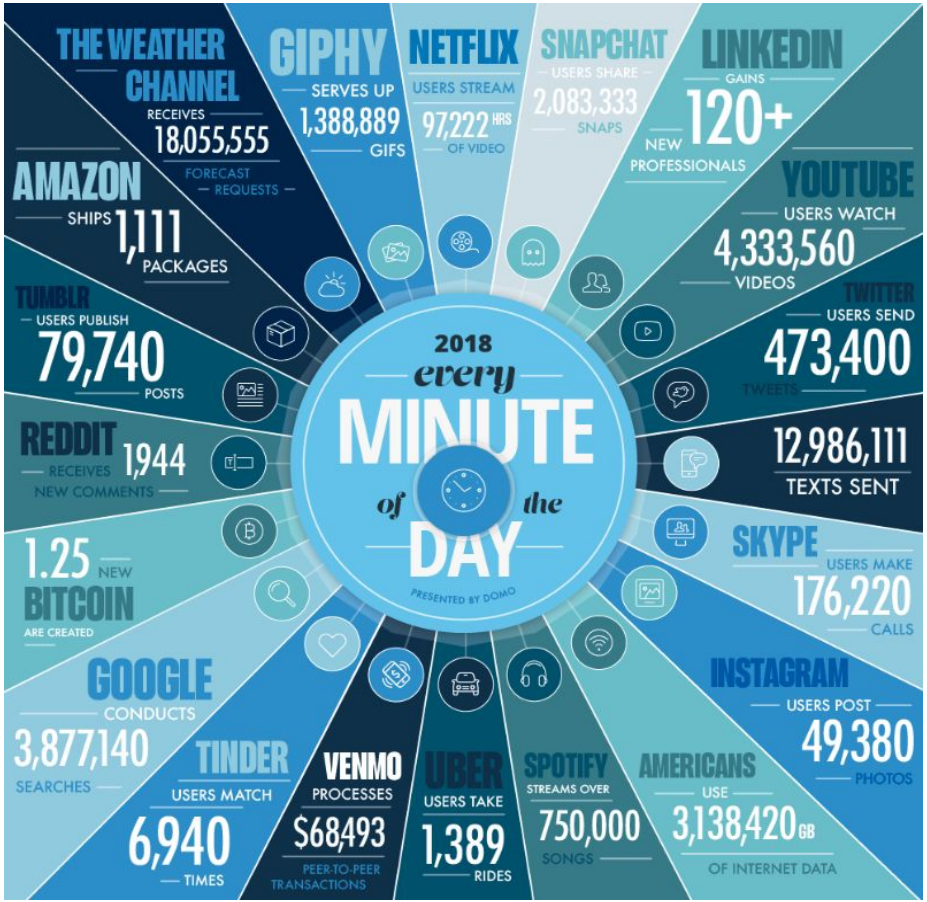

POPULAR SCIENCE
THE FUTURE NOW
THE CONTROL CENTERS
Using Data to Feed the World, Solve Cold Cases, Battle Malware, Predict Our Fate
OFFICER ALGORITHM
Can a Crime Be Prevented Before It Begins?
NEW WAYS OF SEEING
A Gallery of Extraordinary Infographics
PLUS
Juan Enriquez Reprograms Life
James Gleick Unsplits the Bit
AND Lawrence Weschler Questions the Cloud
SPECIAL ISSUE
DATA IS POWER
HOW INFORMATION IS DRIVING THE FUTURE


4 September 2008 www.nature.com/nature £10
THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE
nature
THE BITER BIT
Viral infections for viruses
TROPICAL CYCLONES
The strong get stronger
BLACK HOLE PHYSICS
A new window on the Galactic Centre
BIG DATA
NATUREJOBS
Minnesota musings
SCIENCE IN THE PETABYTE ERA

“25 años de datos sobre 250 millones de personas y nadie puede hacer nada”

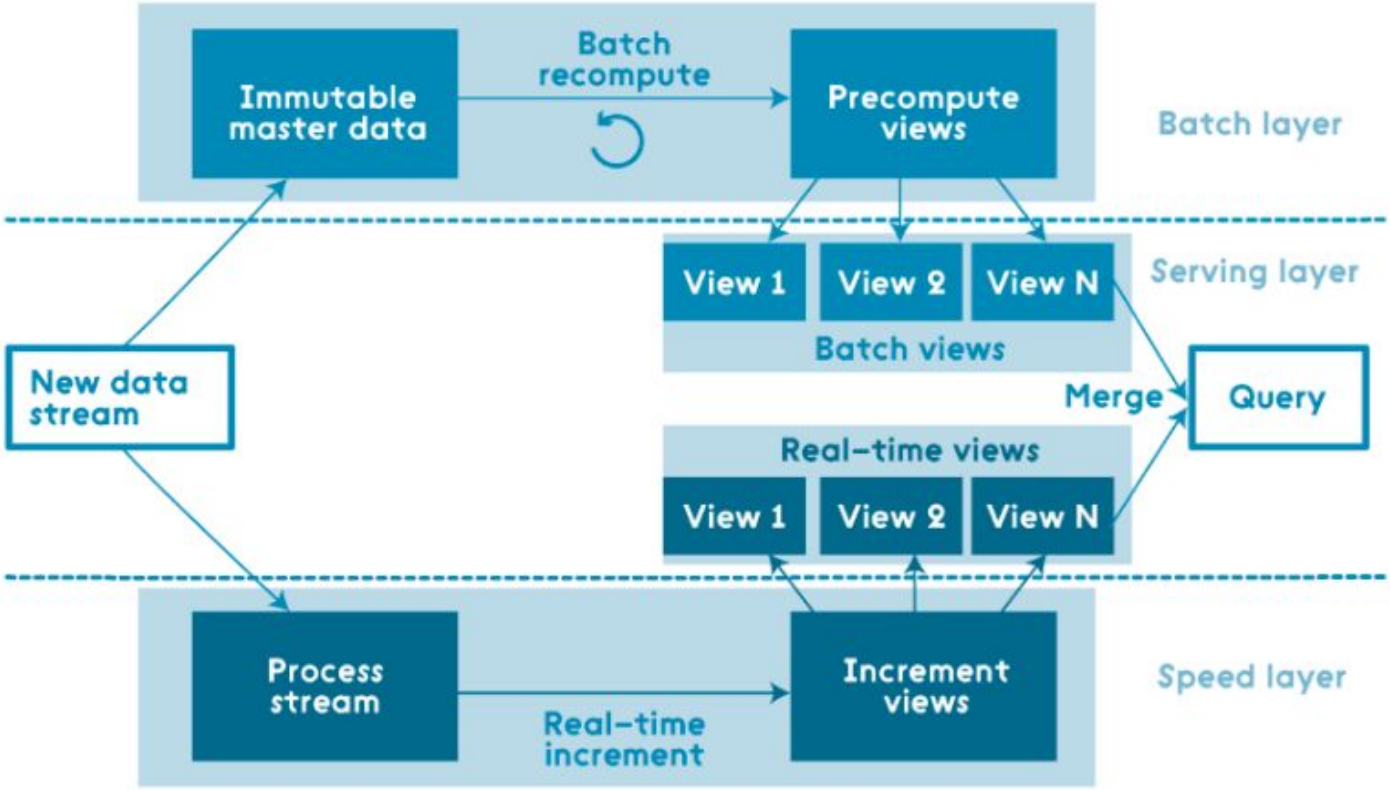

2018
every
MINUTE
of the
DAY
PRESENTED BY DOMO
THE WEATHER CHANNEL
RECEIVES
18,055,555
FORECAST REQUESTS
GIPHY
SERVES UP
1,388,889
GIFS
NETFLIX
USERS STREAM
97,222 HRS
OF VIDEO
SNAPCHAT
USERS SHARE
2,083,333
SNAPS
LINKEDIN
GAINS
120+
NEW PROFESSIONALS
AMAZON
SHIPS 1,111
PACKAGES
YOUTUBE
USERS WATCH
4,333,560
VIDEOS
TUMBLR
USERS PUBLISH
79,740
POSTS
TWITTER
USERS SEND
473,400
TWEETS
REDDIT
RECEIVES 1,944
NEW COMMENTS
12,986,111
TEXTS SENT
1.25 NEW
BITCOIN
ARE CREATED
SKYPE
USERS MAKE
176,220
CALLS
GOOGLE
CONDUCTS
3,877,140
SEARCHES
INSTAGRAM
USERS POST
49,380
PHOTOS
TINDER
USERS MATCH
6,940
TIMES
VENMO
PROCESSES
$68,493
PEER-TO-PEER TRANSACTIONS
UBER
USERS TAKE
1,389
RIDES
SPOTIFY
STREAMS OVER
750,000
SONGS
AMERICANS
USE
3,138,420 GB
OF INTERNET DATA


every MINUTE of the DAY
Skype USERS MAKE 110,040 CALLS
UBER PASSENGERS TAKE RIDES 694
Facebook USERS LIKE 4,166,667 POSTS
TWITTER USERS SEND 347,222 TWEETS
YOUTUBE USERS UPLOAD 300 HOURS OF NEW VIDEO
Instagram USERS LIKE 1,736,111 PHOTOS
PINTEREST USERS PIN 9,722 IMAGES
APPLE USERS DOWNLOAD 51,000 APPS
Netflix SUBSCRIBERS STREAM 77,160 HOURS OF VIDEO
reddit USERS CAST 18,327 VOTES
AMAZON RECEIVES 4,310 UNIQUE VISITORS
VINE USERS PLAY 1,041,666 VIDEOS
Tinder USERS SWIPE 590,278 TIMES
SNAPCHAT USERS SHARE 284,722 SNAPS
Buzzfeed USERS VIEW 34,430 VIDEOS


2018 every MINUTE of the DAY
PRESENTED BY DOMO
THE WEATHER CHANNEL RECEIVES 18,055,555 FORECAST REQUESTS
GIPHY SERVES UP 1,388,889 GIFS
NETFLIX USERS STREAM 97,222 HRS OF VIDEO
SNAPCHAT USERS SHARE 2,083,333 SNAPS
LINKEDIN GAINS 120+ NEW PROFESSIONALS
AMAZON SHIPS 1,111 PACKAGES
YOUTUBE USERS WATCH 4,333,560 VIDEOS
TUMBLR USERS PUBLISH 79,740 POSTS
TWITTER USERS SEND 473,400 TWEETS
REDDIT RECEIVES 1,944 NEW COMMENTS
12,986,111 TEXTS SENT
1.25 NEW BITCOIN ARE CREATED
SKYPE USERS MAKE 176,220 CALLS
GOOGLE CONDUCTS 3,877,140 SEARCHES
INSTAGRAM USERS POST 49,380 PHOTOS
TINDER USERS MATCH 6,940 TIMES
VENMO PROCESSES $68,493 PEER-TO-PEER TRANSACTIONS
UBER USERS TAKE 1,389 RIDES
SPOTIFY STREAMS OVER 750,000 SONGS
AMERICANS USE 3,138,420 GB OF INTERNET DATA

Data Lakes

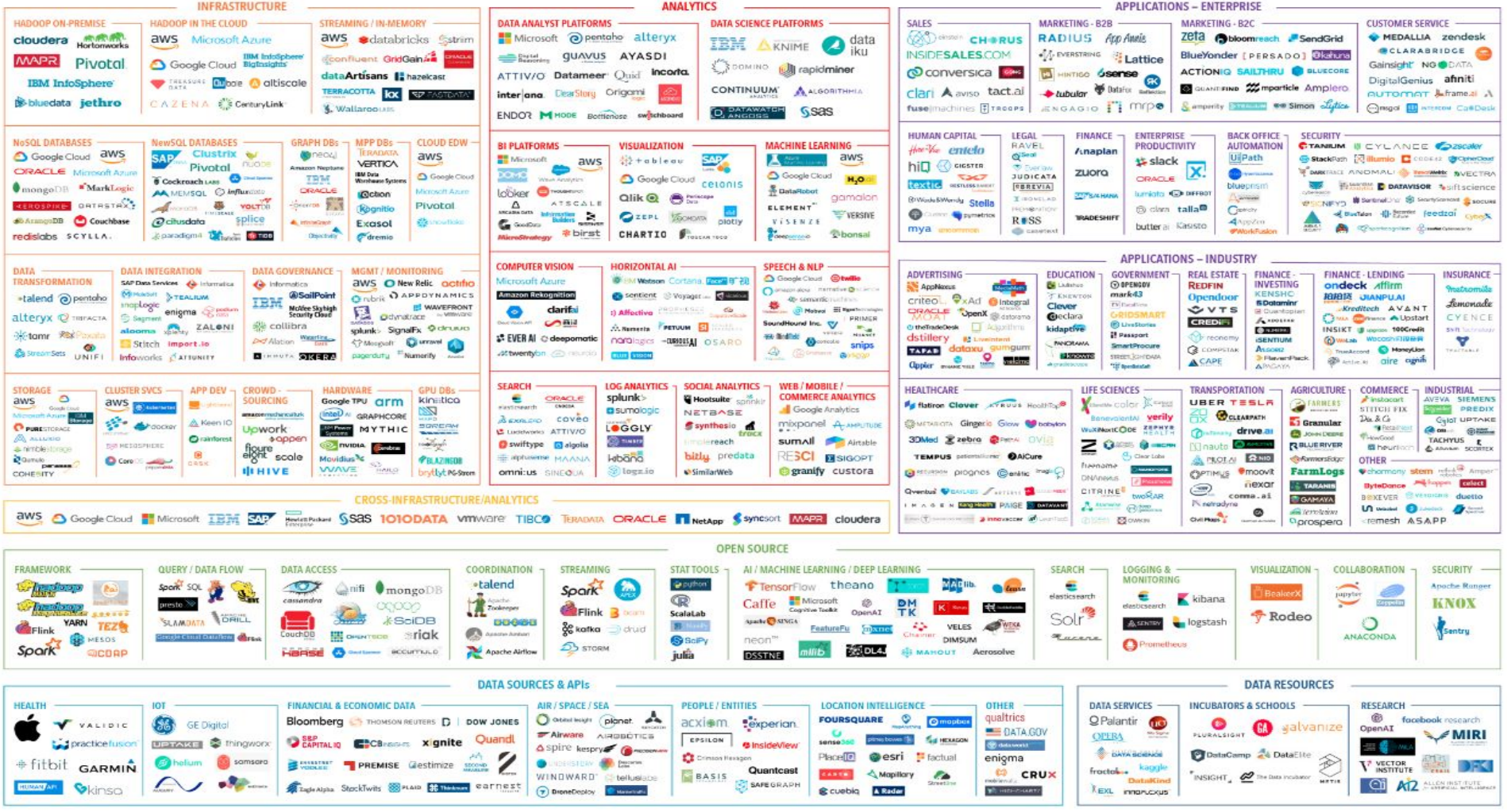
Arquitecturas Lambda



Las 6 “V” de Big Data

- Volumen
- Velocidad
- Variedad
- Veracidad
- Valor
- Valencia

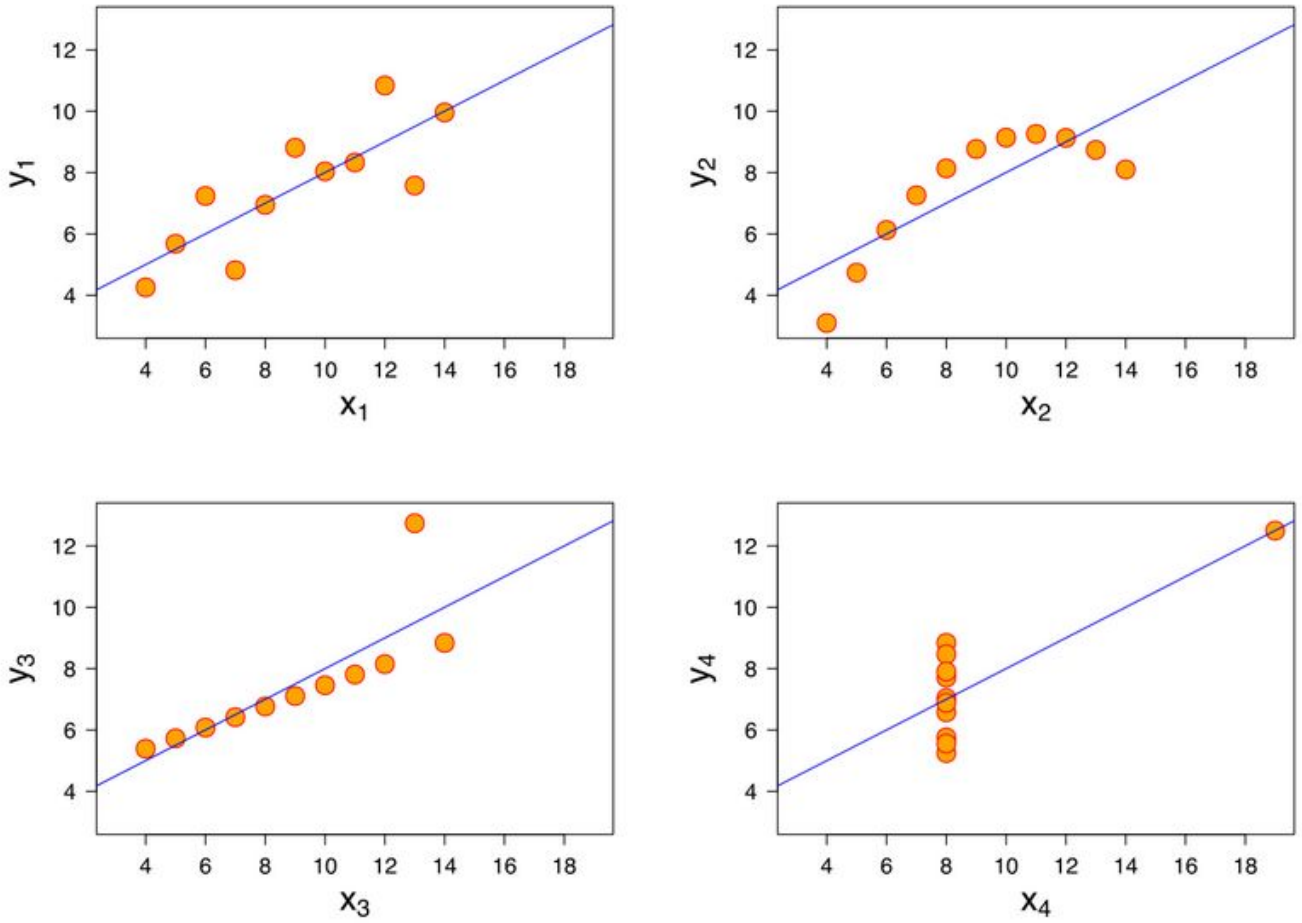
BIG DATA & AI LANDSCAPE 2018



Visualización de Datos

Anscombe's Quartet: Raw Data

	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
	8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
	13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
	9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
	11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
	14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
	6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
	4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
	12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
	7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
	5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89
mean	9.0	7.5	9.0	7.5	9.0	7.5	9.0	7.5
var.	10.0	3.75	10.0	3.75	10.0	3.75	10.0	3.75
corr.	0.816		0.816		0.816		0.816	


y1
x1
y2
x2
y3
x3
y4
x4
4
6
8
10
12
14
16
18

Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812-1813.

Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite.

Paris, le 20 Novembre 1869.

Les nombres d'hommes présents sont représentés par les largeurs des zônes colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en travers des zônes. Le rouge désigne les hommes qui entrent en Russie, le noir ceux qui en sortent. _____ Les renseignements qui ont servi à dresser la carte ont été puisés dans les ouvrages de M.M. Chiers, de Ségur, de Fezensac, de Chambray et le journal inédit de Jacob, pharmacien de l'Armée depuis le 28 Octobre.

Pour mieux faire juger à l'oeil la diminution de l'armée, j'ai supposé que les corps du Prince Jérôme et du Maréchal Davoust qui avaient été détachés sur Minsk et Mohilow et ont rejoint vers Orscha et Witebsk, avaient toujours marché avec l'armée.

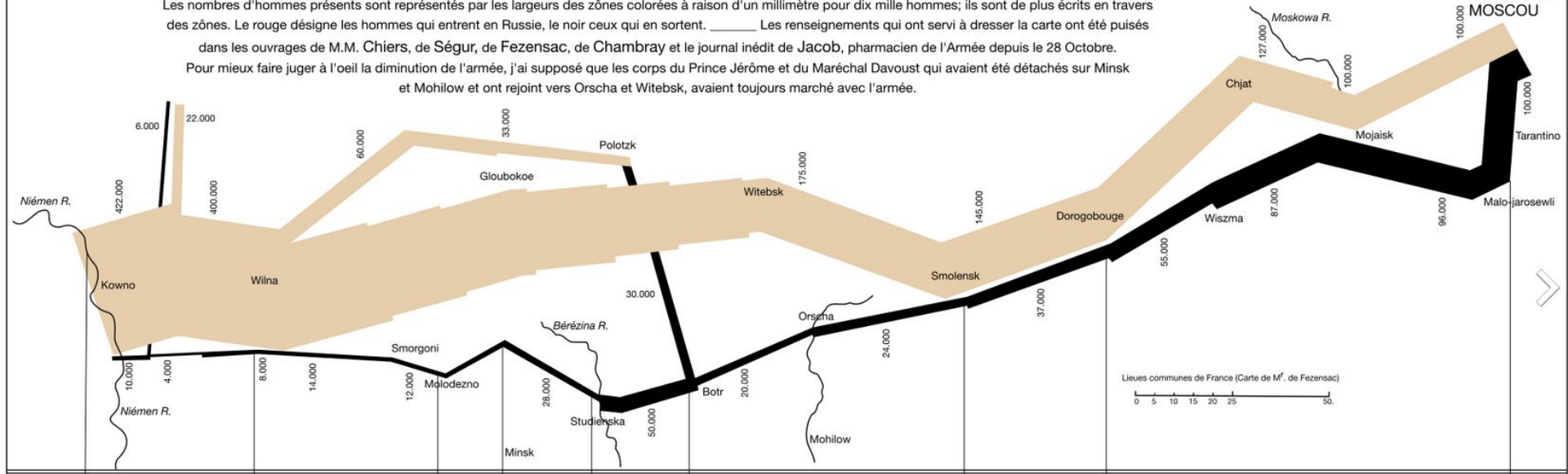
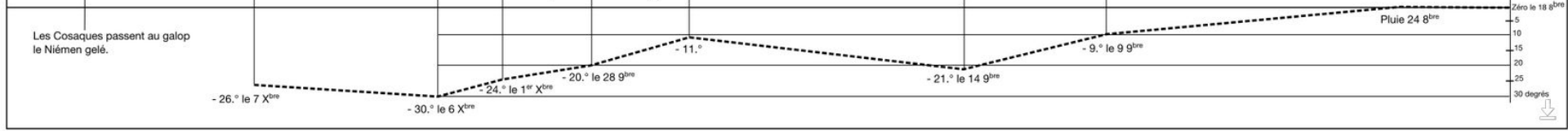


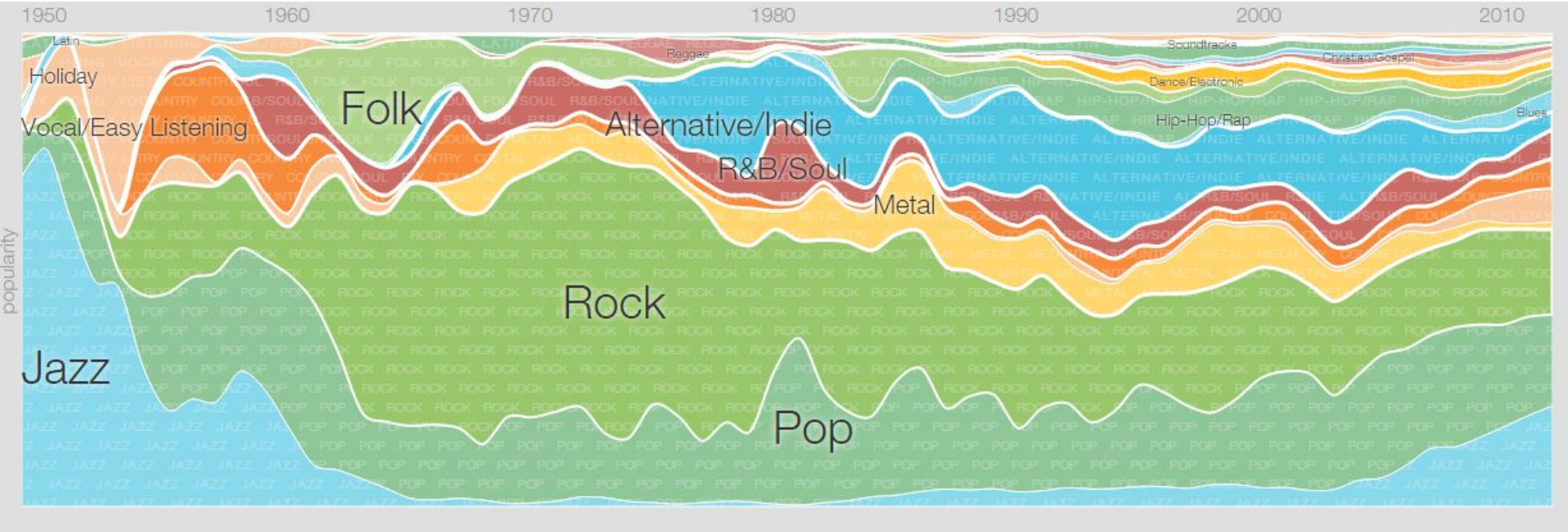
TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.

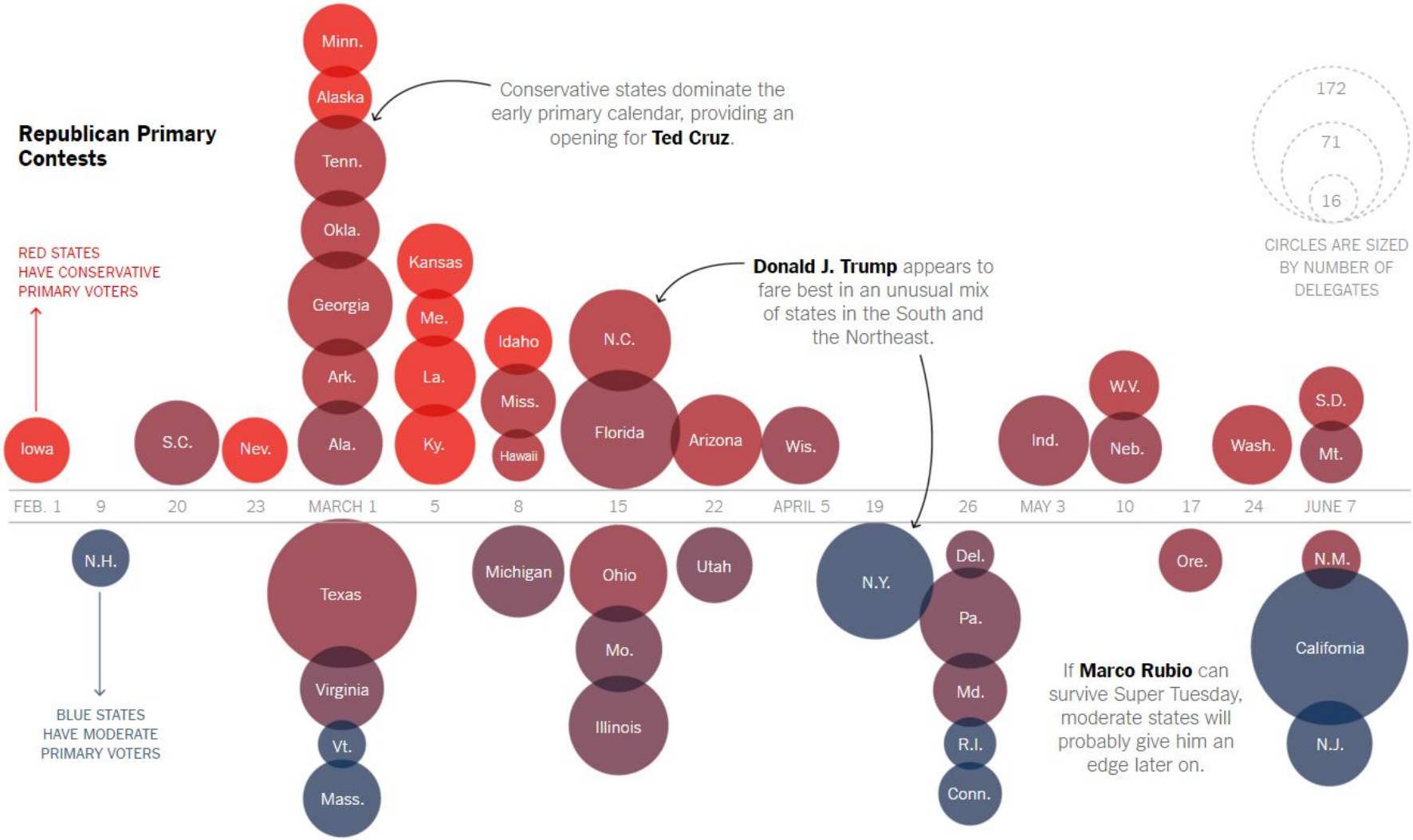


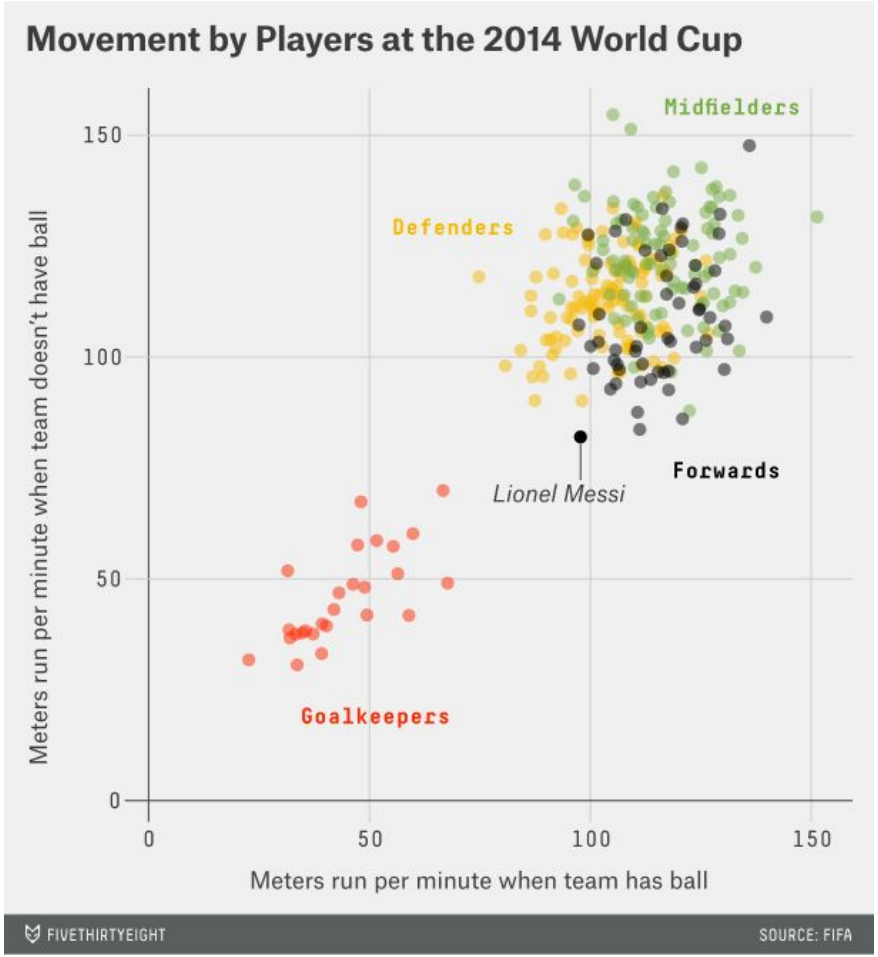
Autog. par Regnier, 8. Pas. St Marie St Gen à Paris.

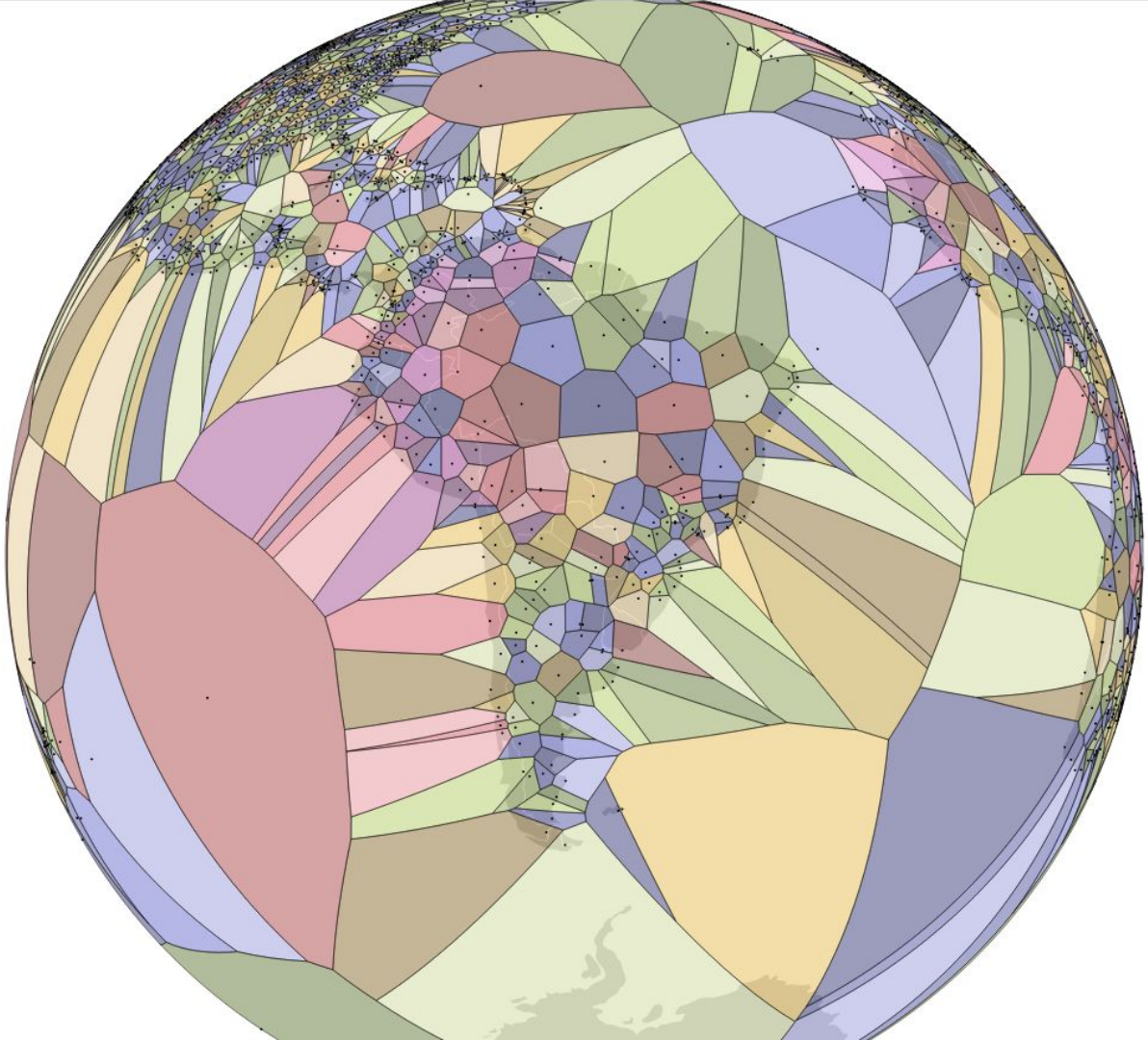
[Vectorization CC-BY-SA martingrandjean.ch 2014]

Imp. Lith. Regnier et Dourdet.

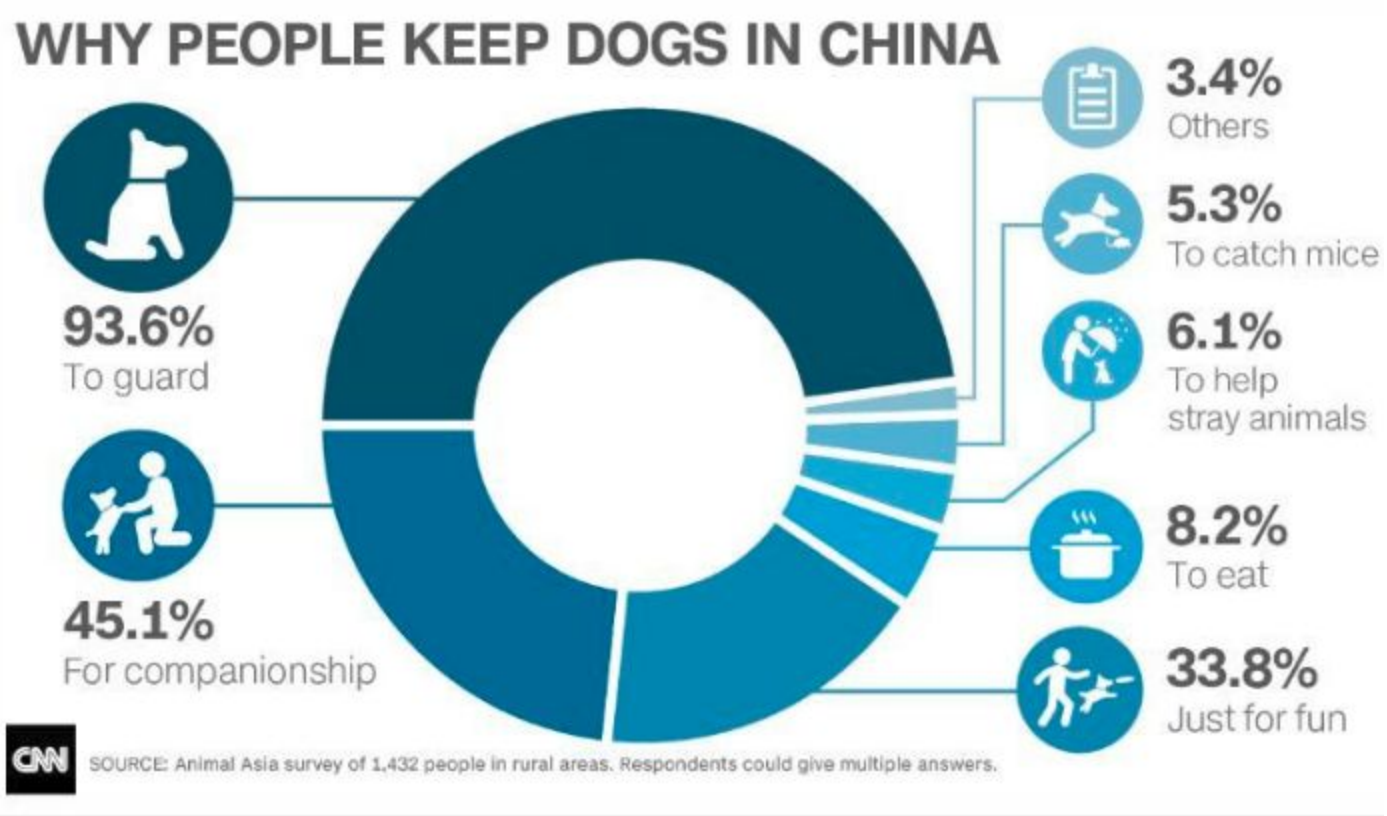

1950
1960
1970
1980
1990
2000
2010
popularity
Latin
Holiday
Vocal/Easy Listening
Folk
Reggae
Alternative/Indie
R&B/Soul
Metal
Rock
Jazz
Pop
Soundtracks
Christian/Gospel
Dance/Electronic
Hip-Hop/Rap
Blues


Republican Primary Contests
Conservative states dominate the early primary calendar, providing an opening for Ted Cruz.
Donald J. Trump appears to fare best in an unusual mix of states in the South and the Northeast.
172
71
16
CIRCLES ARE SIZED BY NUMBER OF DELEGATES
RED STATES HAVE CONSERVATIVE PRIMARY VOTERS
BLUE STATES HAVE MODERATE PRIMARY VOTERS
If Marco Rubio can survive Super Tuesday, moderate states will probably give him an edge later on.
Iowa
N.H.
S.C.
Nev.
Ala.
Ark.
Georgia
Okla.
Tenn.
Alaska
Minn.
Texas
Virginia
Vt.
Mass.
Kansas
Me.
La.
Ky.
Idaho
Miss.
Hawaii
Michigan
N.C.
Florida
Ohio
Mo.
Illinois
Arizona
Utah
Wis.
N.Y.
Del.
Pa.
Md.
R.I.
Conn.
Ind.
W.V.
Neb.
Ore.
Wash.
S.D.
Mt.
N.M.
California
N.J.
FEB. 1
9
20
23
MARCH 1
5
8
15
22
APRIL 5
19
26
MAY 3
10
17
24
JUNE 7

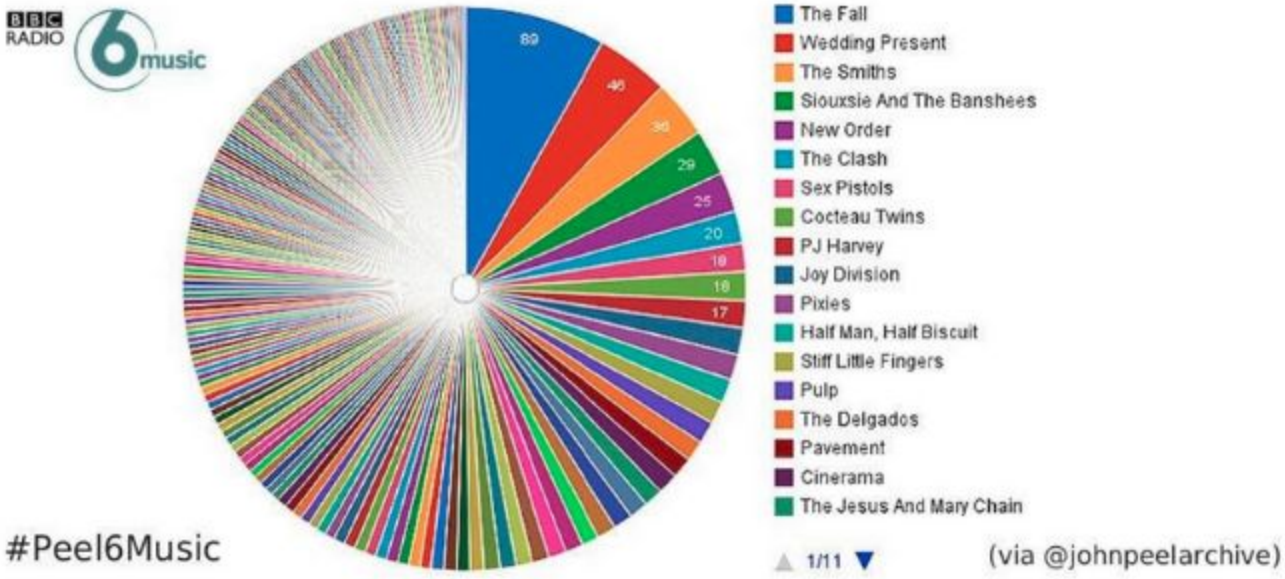

Movement by Players at the 2014 World Cup
Midfielders
Defenders
Forwards
Lionel Messi
Goalkeepers
150
100
50
0
0
50
100
150
Meters run per minute when team doesn't have ball
Meters run per minute when team has ball
FIVETHIRTYEIGHT
SOURCE: FIFA



El horror...


WHY PEOPLE KEEP DOGS IN CHINA
93.6%
To guard
45.1%
For companionship
3.4%
Others
5.3%
To catch mice
6.1%
To help stray animals
8.2%
To eat
33.8%
Just for fun
CNN
SOURCE: Animal Asia survey of 1,432 people in rural areas. Respondents could give multiple answers.

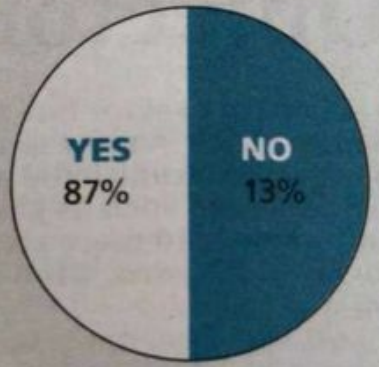
John Peel's most played artists in his Festive 50s



WEEKLY POLL

WE ASKED:

Would you pay more for your milk to help dairy farmers?

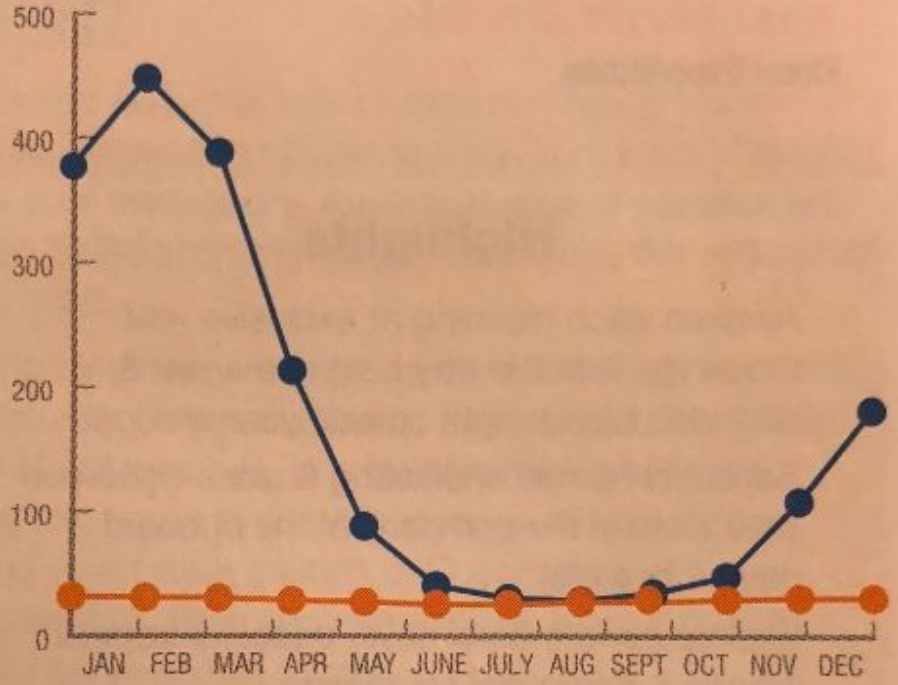


THIS WEEK'S QUESTION:

Do you welcome the growth of solar farms in Cumbria?

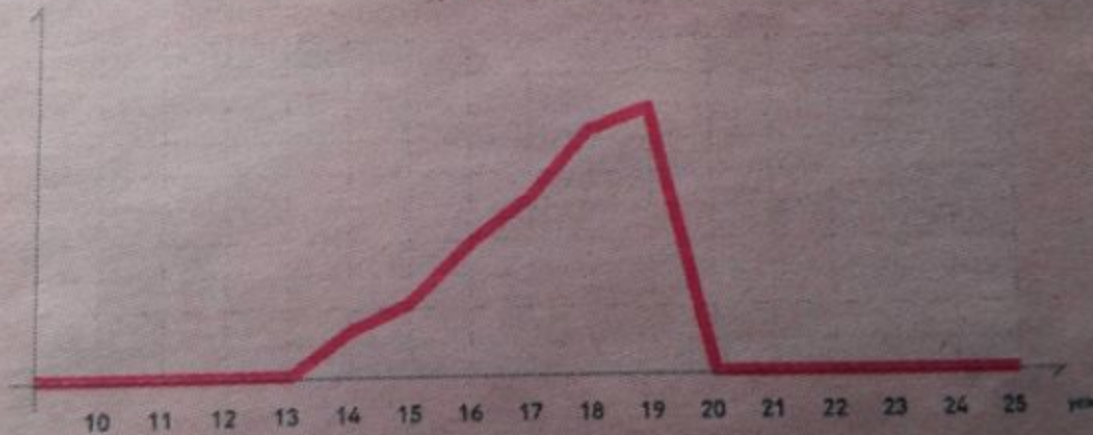
TEMPERATURE & RAIN CHART

CAIRNS

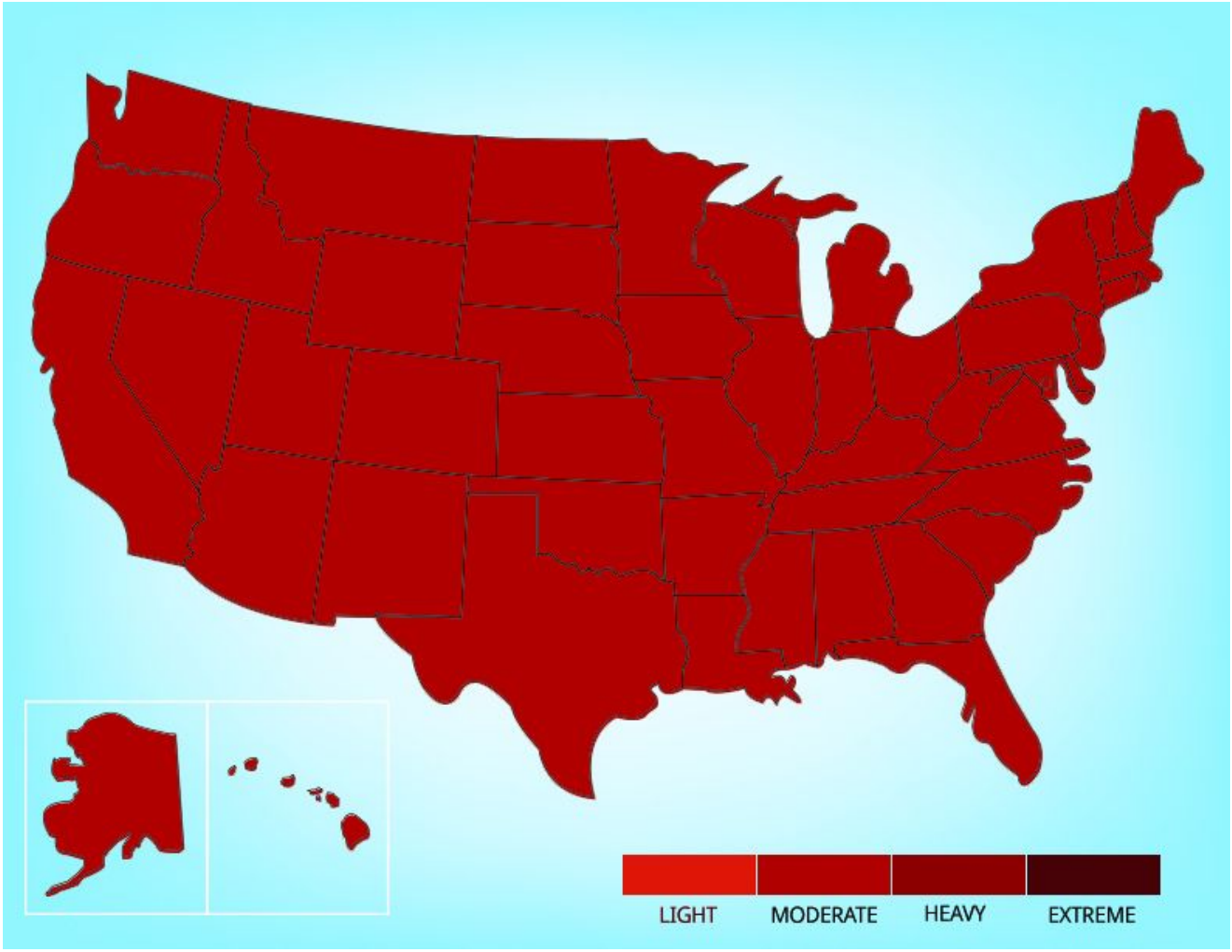


metro.us/truthfacts

STATISTICS SHOW THAT TEEN PREGNANCIES DROP DRASTICALLY AFTER THE AGE OF 20



BY WULFF & MORGENTHALER/TRUTHFACTS.COM


LIGHT
MODERATE
HEAVY
EXTREME


THIS IS WHY
WE CAN'T HAVE NICE THINGS
quickmeme.com

Formatos de Datos

JSON

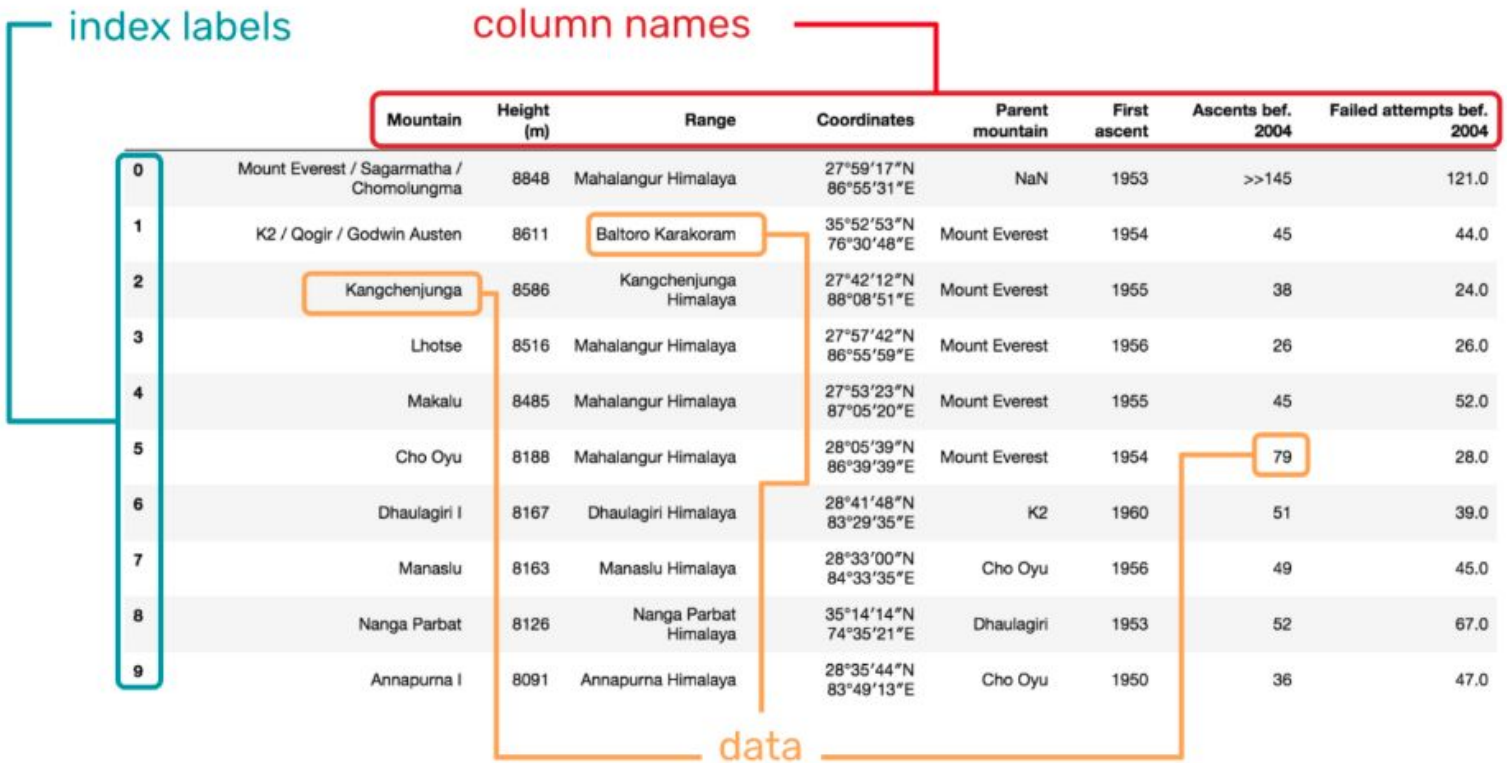
```
{"materia":"Organización de Datos",
 "creditos":6,
 "docentes": [ {"nombre": "Luis Argerich",
                "cargo": "profesor"},
               {"nombre": "Natalia Golmar",
                "cargo": "JTP"}
             ]
}
```

Matrices Dispersas: CSR

$$\begin{pmatrix} 1 & 0 & -3 \\ 4 & 0 & 1 \\ 0 & 0 & 5 \end{pmatrix}$$

```
{"elements" : [1,-3,4,1,5]
"columns" : [0,2,0,2,2]
"first_in_rows" : [0,2,4]}
```

DataFrames

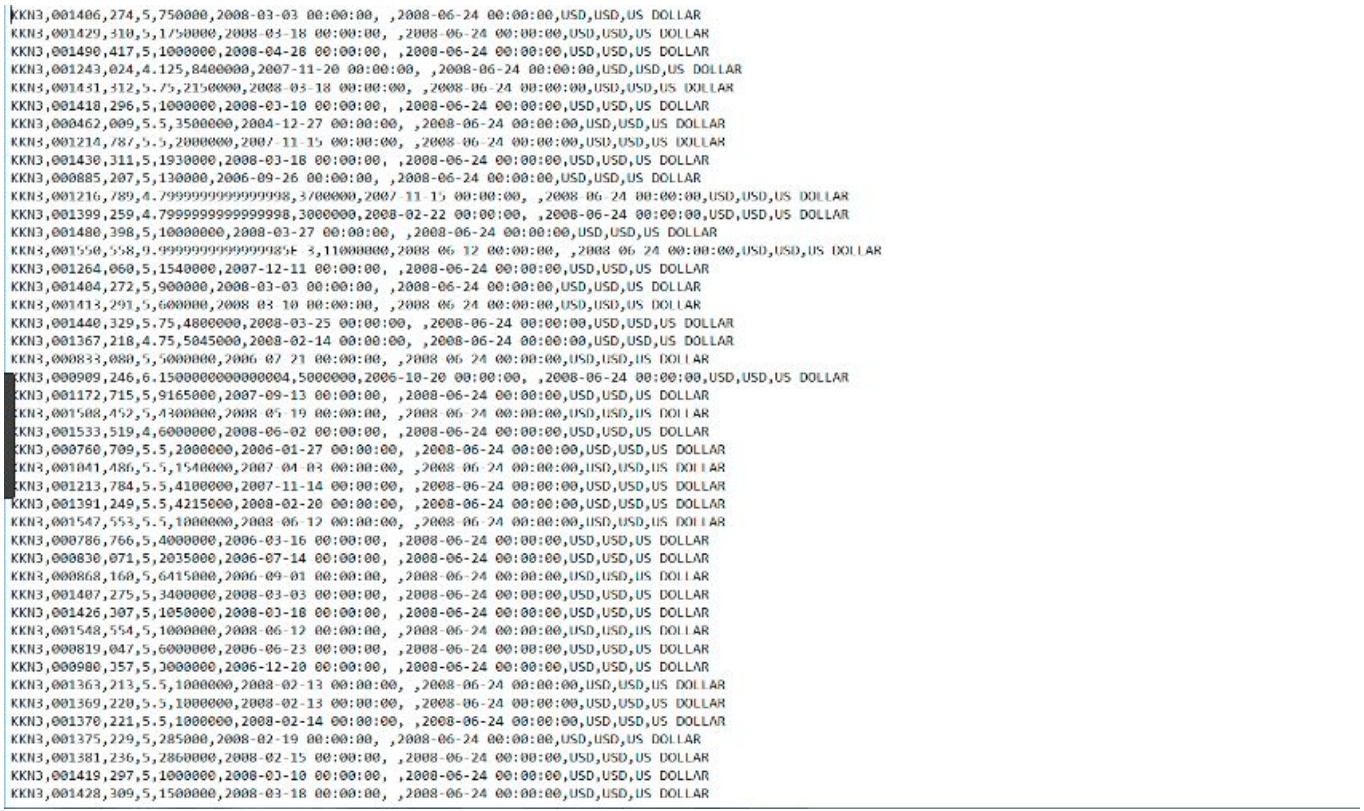


	Mountain	Height (m)	Range	Coordinates	Parent mountain	First ascent	Ascents bef. 2004	Failed attempts bef. 2004
0	Mount Everest / Sagarmatha / Chomolungma	8848	Mahalangur Himalaya	27°59′17″N 86°55′31″E	NaN	1953	>>145	121.0
1	K2 / Qogir / Godwin Austen	8611	Baltoro Karakoram	35°52′53″N 76°30′48″E	Mount Everest	1954	45	44.0
2	Kangchenjunga	8586	Kangchenjunga Himalaya	27°42′12″N 88°08′51″E	Mount Everest	1955	38	24.0
3	Lhotse	8516	Mahalangur Himalaya	27°57′42″N 86°55′59″E	Mount Everest	1956	26	26.0
4	Makalu	8485	Mahalangur Himalaya	27°53′23″N 87°05′20″E	Mount Everest	1955	45	52.0
5	Cho Oyu	8188	Mahalangur Himalaya	28°05′39″N 86°39′39″E	Mount Everest	1954	79	28.0
6	Dhaulagiri I	8167	Dhaulagiri Himalaya	28°41′48″N 83°29′35″E	K2	1960	51	39.0
7	Manaslu	8163	Manaslu Himalaya	28°33′00″N 84°33′35″E	Cho Oyu	1956	49	45.0
8	Nanga Parbat	8126	Nanga Parbat Himalaya	35°14′14″N 74°35′21″E	Dhaulagiri	1953	52	67.0
9	Annapurna I	8091	Annapurna Himalaya	28°35′44″N 83°49′13″E	Cho Oyu	1950	36	47.0

Tipos de Datos

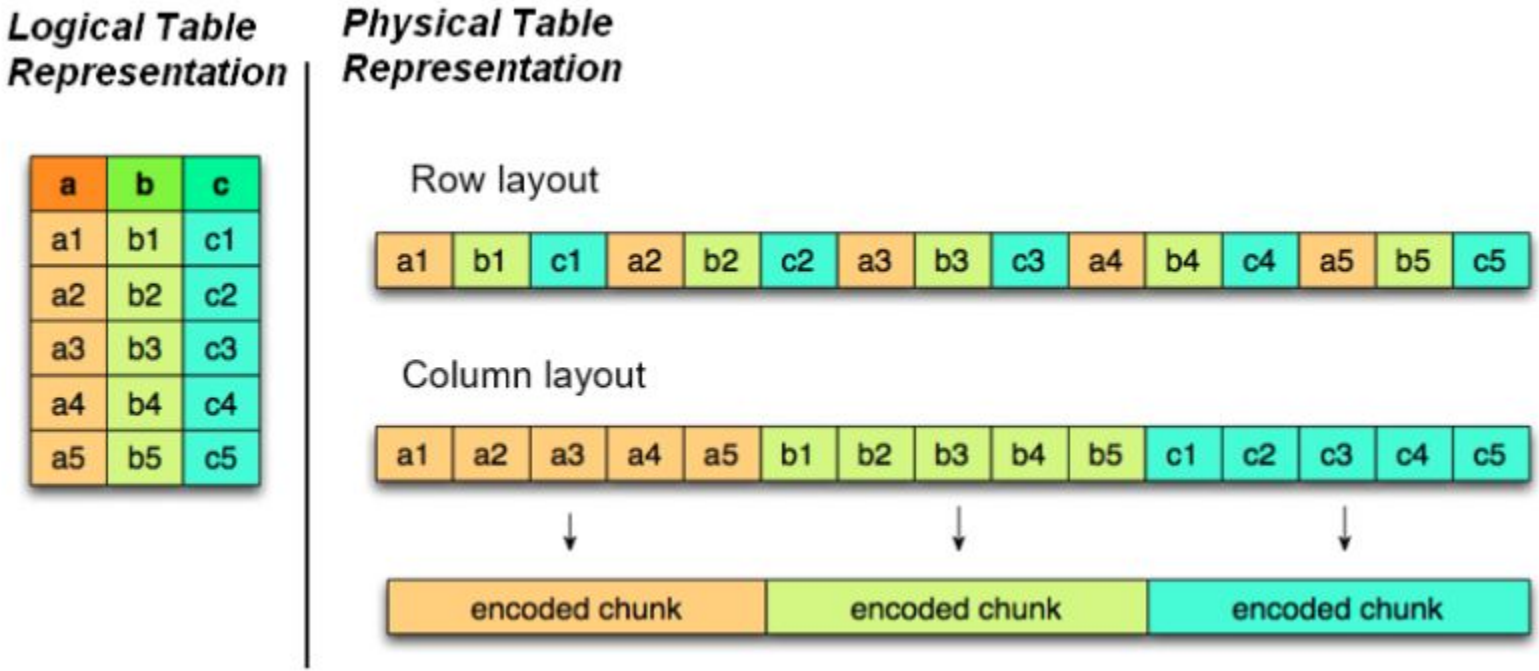
- Object
- Primitivos: int8, int16, int64, float32, float64
- Strings
- Categóricos
    - Ordinales vs no-ordinales

DataFrames: CSVs



```
KKN3,001406,274,5,750000,2008-03-03 00:00:00, ,2008-06-24 00:00:00,USD,USD,US DOLLAR
KKN3,001429,310,5,1750000,2008-03-18 00:00:00, ,2008-06-24 00:00:00,USD,USD,US DOLLAR
KKN3,001490,417,5,1000000,2008-04-28 00:00:00, ,2008-06-24 00:00:00,USD,USD,US DOLLAR
KKN3,001243,024,4.125,8400000,2007-11-20 00:00:00, ,2008-06-24 00:00:00,USD,USD,US DOLLAR
KKN3,001431,312,5.75,2150000,2008-03-18 00:00:00, ,2008-06-24 00:00:00,USD,USD,US DOLLAR
KKN3,001418,296,5,1000000,2008-03-10 00:00:00, ,2008-06-24 00:00:00,USD,USD,US DOLLAR
KKN3,000462,009,5.5,3500000,2004-12-27 00:00:00, ,2008-06-24 00:00:00,USD,USD,US DOLLAR
KKN3,001214,787,5.5,2000000,2007-11-15 00:00:00, ,2008-06-24 00:00:00,USD,USD,US DOLLAR
KKN3,001430,311,5,1930000,2008-03-18 00:00:00, ,2008-06-24 00:00:00,USD,USD,US DOLLAR
KKN3,000885,207,5,130000,2006-09-26 00:00:00, ,2008-06-24 00:00:00,USD,USD,US DOLLAR
KKN3,001216,789,4.7999999999999998,3700000,2007-11-15 00:00:00, ,2008-06-24 00:00:00,USD,USD,US DOLLAR
KKN3,001399,259,4.7999999999999998,3000000,2008-02-22 00:00:00, ,2008-06-24 00:00:00,USD,USD,US DOLLAR
KKN3,001480,398,5,10000000,2008-03-27 00:00:00, ,2008-06-24 00:00:00,USD,USD,US DOLLAR
KKN3,001550,558,9.9999999999999986E-3,11000000,2008-06-12 00:00:00, ,2008-06-24 00:00:00,USD,USD,US DOLLAR
KKN3,001264,060,5,1540000,2007-12-11 00:00:00, ,2008-06-24 00:00:00,USD,USD,US DOLLAR
KKN3,001404,272,5,900000,2008-03-03 00:00:00, ,2008-06-24 00:00:00,USD,USD,US DOLLAR
KKN3,001413,291,5,600000,2008-03-10 00:00:00, ,2008-06-24 00:00:00,USD,USD,US DOLLAR
KKN3,001440,329,5.75,4800000,2008-03-25 00:00:00, ,2008-06-24 00:00:00,USD,USD,US DOLLAR
KKN3,001367,218,4.75,5045000,2008-02-14 00:00:00, ,2008-06-24 00:00:00,USD,USD,US DOLLAR
KKN3,000833,080,5,5000000,2006-07-21 00:00:00, ,2008-06-24 00:00:00,USD,USD,US DOLLAR
KKN3,000909,246,6.1500000000000004,5000000,2006-10-20 00:00:00, ,2008-06-24 00:00:00,USD,USD,US DOLLAR
KKN3,001172,715,5,9165000,2007-09-13 00:00:00, ,2008-06-24 00:00:00,USD,USD,US DOLLAR
KKN3,001508,452,5,4300000,2008-05-19 00:00:00, ,2008-06-24 00:00:00,USD,USD,US DOLLAR
KKN3,001533,519,4,6000000,2008-06-02 00:00:00, ,2008-06-24 00:00:00,USD,USD,US DOLLAR
KKN3,000760,709,5.5,2000000,2006-01-27 00:00:00, ,2008-06-24 00:00:00,USD,USD,US DOLLAR
KKN3,001041,486,5.5,1540000,2007-04-03 00:00:00, ,2008-06-24 00:00:00,USD,USD,US DOLLAR
KKN3,001213,784,5.5,4100000,2007-11-14 00:00:00, ,2008-06-24 00:00:00,USD,USD,US DOLLAR
KKN3,001391,249,5.5,4215000,2008-02-20 00:00:00, ,2008-06-24 00:00:00,USD,USD,US DOLLAR
KKN3,001547,553,5.5,1000000,2008-06-12 00:00:00, ,2008-06-24 00:00:00,USD,USD,US DOLLAR
KKN3,000786,766,5,4000000,2006-03-16 00:00:00, ,2008-06-24 00:00:00,USD,USD,US DOLLAR
KKN3,000830,071,5,2035000,2006-07-14 00:00:00, ,2008-06-24 00:00:00,USD,USD,US DOLLAR
KKN3,000868,160,5,6415000,2006-09-01 00:00:00, ,2008-06-24 00:00:00,USD,USD,US DOLLAR
KKN3,001407,275,5,3400000,2008-03-03 00:00:00, ,2008-06-24 00:00:00,USD,USD,US DOLLAR
KKN3,001426,307,5,1050000,2008-03-18 00:00:00, ,2008-06-24 00:00:00,USD,USD,US DOLLAR
KKN3,001548,554,5,1000000,2008-06-12 00:00:00, ,2008-06-24 00:00:00,USD,USD,US DOLLAR
KKN3,000819,047,5,6000000,2006-06-23 00:00:00, ,2008-06-24 00:00:00,USD,USD,US DOLLAR
KKN3,000980,357,5,3000000,2006-12-20 00:00:00, ,2008-06-24 00:00:00,USD,USD,US DOLLAR
KKN3,001363,213,5.5,1000000,2008-02-13 00:00:00, ,2008-06-24 00:00:00,USD,USD,US DOLLAR
KKN3,001369,220,5.5,1000000,2008-02-13 00:00:00, ,2008-06-24 00:00:00,USD,USD,US DOLLAR
KKN3,001370,221,5.5,1000000,2008-02-14 00:00:00, ,2008-06-24 00:00:00,USD,USD,US DOLLAR
KKN3,001375,229,5,285000,2008-02-19 00:00:00, ,2008-06-24 00:00:00,USD,USD,US DOLLAR
KKN3,001381,236,5,2860000,2008-02-15 00:00:00, ,2008-06-24 00:00:00,USD,USD,US DOLLAR
KKN3,001419,297,5,1000000,2008-03-10 00:00:00, ,2008-06-24 00:00:00,USD,USD,US DOLLAR
KKN3,001428,309,5,1500000,2008-03-18 00:00:00, ,2008-06-24 00:00:00,USD,USD,US DOLLAR
```

Formatos de Datos: Columnares vs Tabulares



Datos Columnares

- Feather
- Arrow
- Parquet
- etc..

Tarea

Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing

Justin Matejka and George Fitzmaurice
Autodesk Research, Toronto Ontario Canada
{*first.last*}@autodesk.com

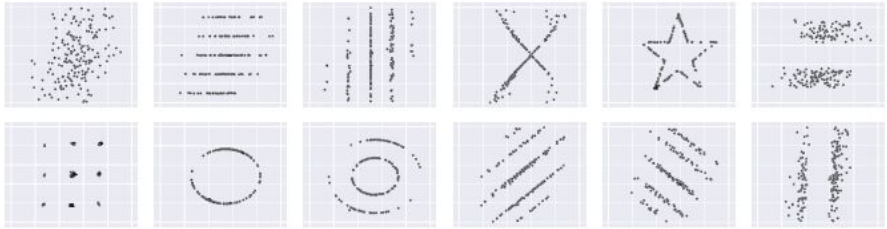


Figure 1. A collection of data sets produced by our technique. While different in appearance, each has the same summary statistics (mean, std. deviation, and Pearson's corr.) to 2 decimal places. ($\bar{x}$ =54.02, $\bar{y}$ = 48.09, sd_x = 14.52, sd_y = 24.79, Pearson's r = +0.32)

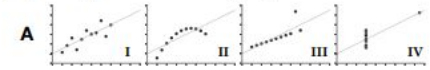
ABSTRACT

Datasets which are identical over a number of statistical properties, yet produce dissimilar graphs, are frequently used to illustrate the importance of graphical representations when exploring data. This paper presents a novel method for generating such datasets, along with several examples. Our technique varies from previous approaches in that new datasets are iteratively generated from a seed dataset through random perturbations of individual data points, and can be directed towards a desired outcome through a simulated annealing optimization strategy. Our method has the benefit of being agnostic to the particular statistical properties that are to remain constant between the datasets, and allows for control over the graphical appearance of resulting output.

INTRODUCTION

Anscome's Quartet [1] is a set of four distinct datasets each consisting of 11 (*x*,*y*) pairs where each dataset produces the same summary statistics (mean, standard deviation, and correlation) while producing vastly different plots (Figure 2A). This dataset is frequently used to illustrate the importance of graphical representations when exploring data. The effectiveness of Anscombe's Quartet is not due to simply having four different data sets which generate the same statistical properties, it is that four *clearly different* and *identifiably distinct* datasets are producing the same statistical properties. Dataset I appears to follow a somewhat noisy linear model, while Dataset II is following a parabolic distribution. Dataset III appears to be strongly linear, except for a single outlier, while Dataset IV forms a vertical line with the regression thrown off by a single outlier. In contrast, Figure 2B shows a series of datasets also sharing the same summary statistics as Anscombe's Quartet, however without any obvious underlying structure to the individual datasets, this quartet is not nearly as effective at demonstrating the importance of graphical representations.

While very popular and effective for illustrating the importance of visualizations, it is not known how Anscombe came up with his datasets [5]. Our work presents a novel method for creating datasets which are identical over a range of statistical properties, yet produce dissimilar graphics. Our method differs from previous by being agnostic to the particular statistical properties that are to remain constant between the datasets, while allowing for control over the graphical appearance of resulting output.



Permission to make digital or hard copies of all or part of this work for personal or


ONE DOES NOT SIMPLY
SAY WELCOME BUT.... HERE WE DO