How to Lie with Statistics

Ali Akbar Septiandri

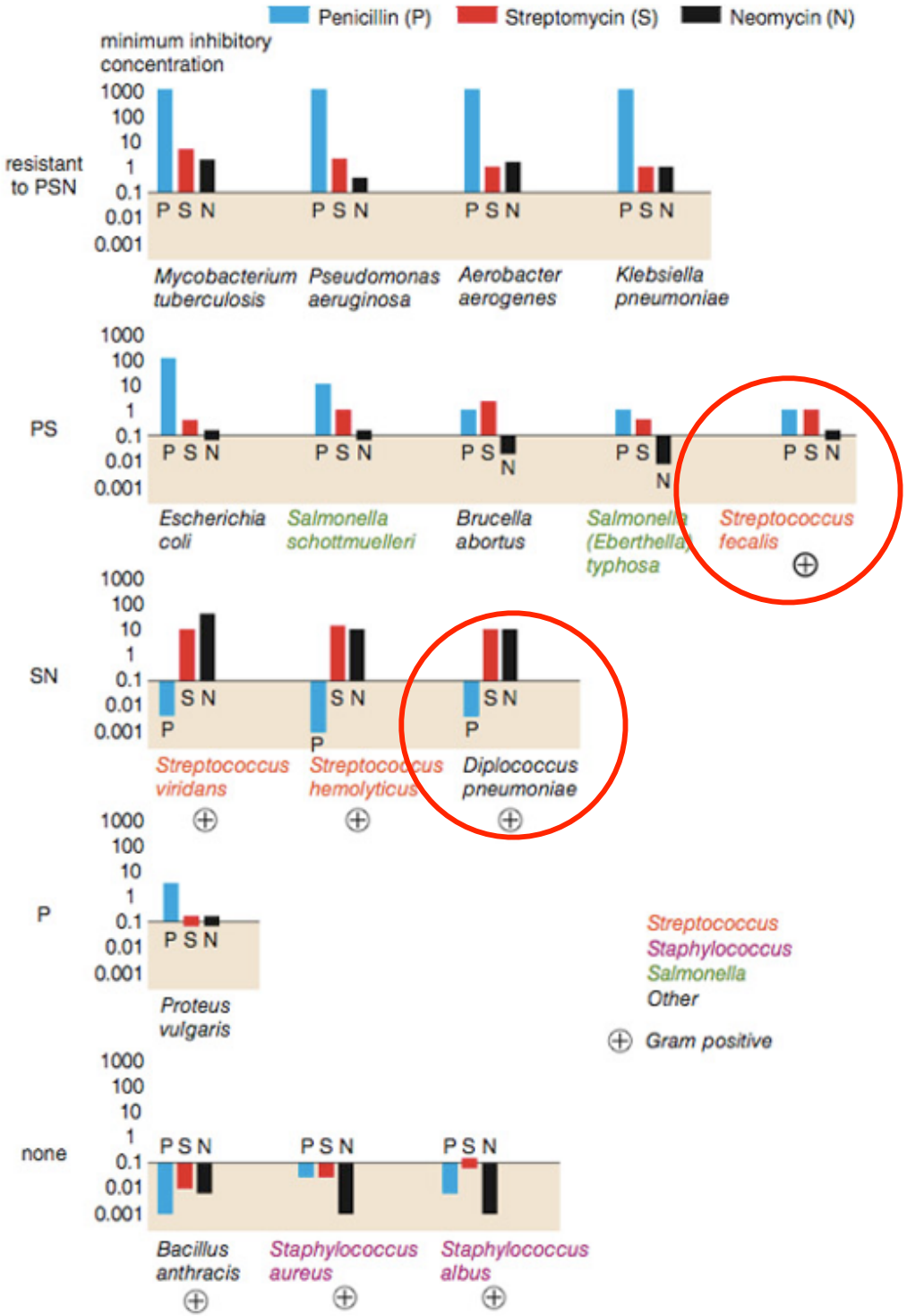
Universitas Al Azhar Indonesia

March 19, 2019

"There are three kinds of lies:
lies, damned lies, and statistics."

Visualisasi

Sebagai analis data yang baik,
penting sekali untuk memvisualisasikan data.
Mengapa?



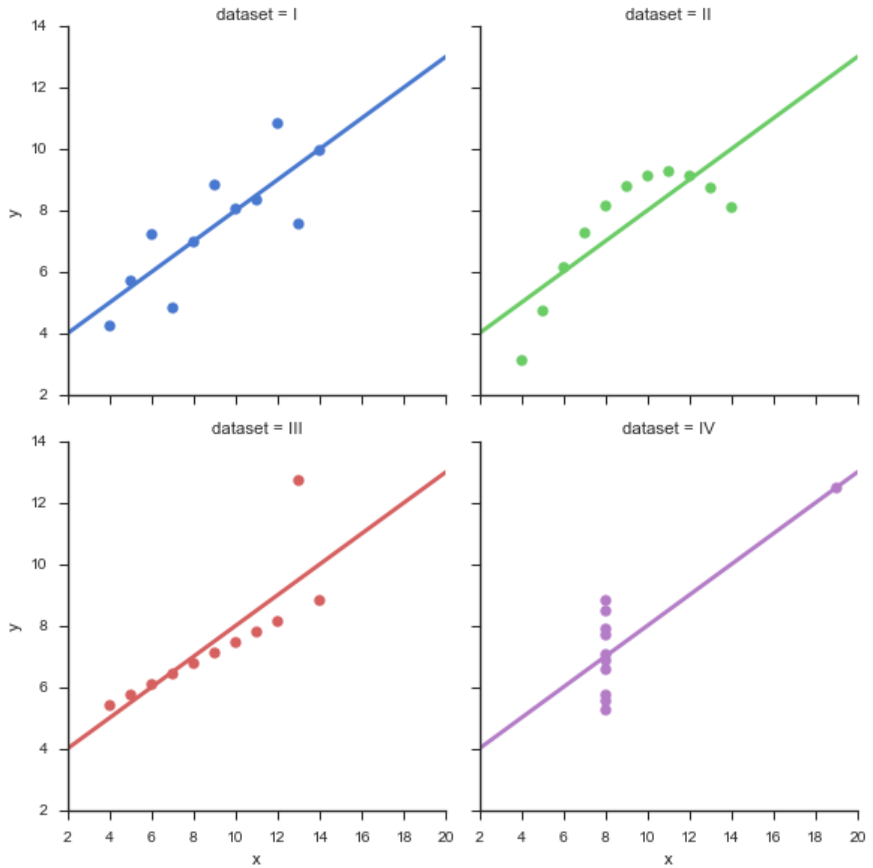
How do the bacteria compare?

Not a streptococcus!
(realized ~30 years later)

Really a streptococcus!
(realized ~20 years later)

Wainer & Lysen, "That's funny..."
American Scientist, 2009
Adapted from Brian Schmotzer

Visualisasi vs Summary Statistics

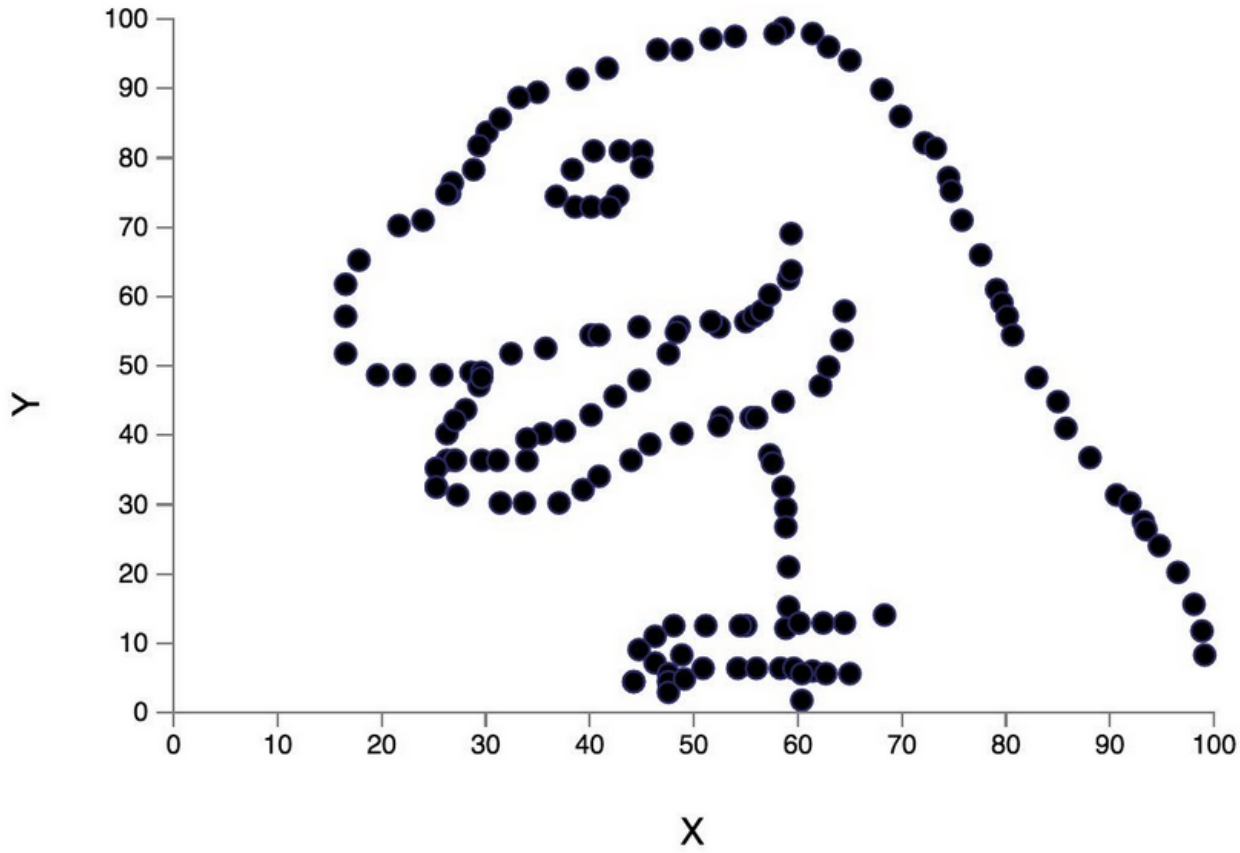


Gambar: *The infamous Anscombe's quartet*

Keempat data tersebut punya
$\mu_x, \mu_y, \sigma_x, \sigma_y, N, \rho$ yang sama!

Don't trust summary statistics.
Always visualize your data!

N = 157 ; X mean = 50.7333 ; X SD = 19.5661 ; Y mean = 46.495 ; Y SD = 27.2828 ;
Pearson correlation = -0.1772

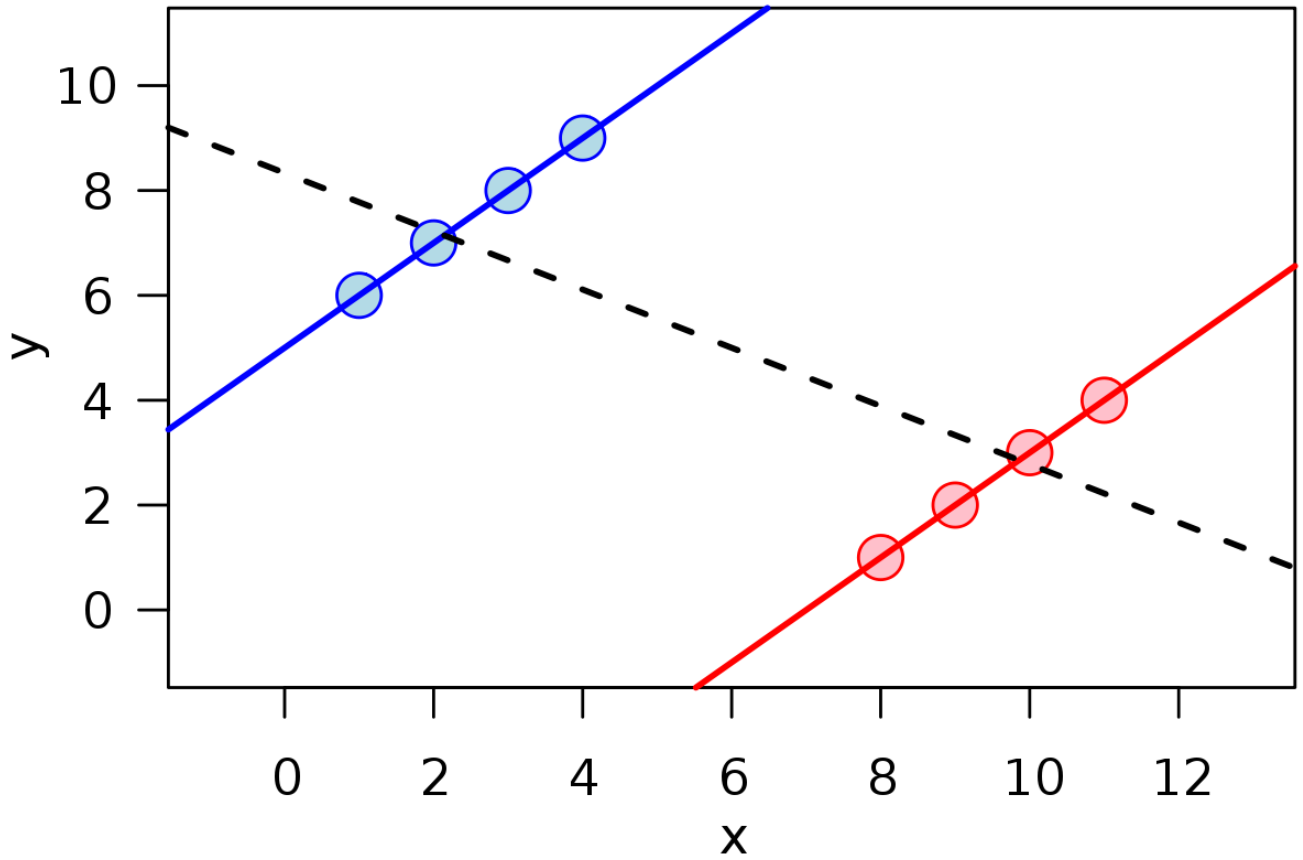


Alberto Cairo @albertocairo · 15 Aug 2016

Don't trust summary statistics. Always visualize your data first robertgrantstats.co.uk/drawmydata.html pic.twitter.com/5j94Dw9UAf

13 895 984

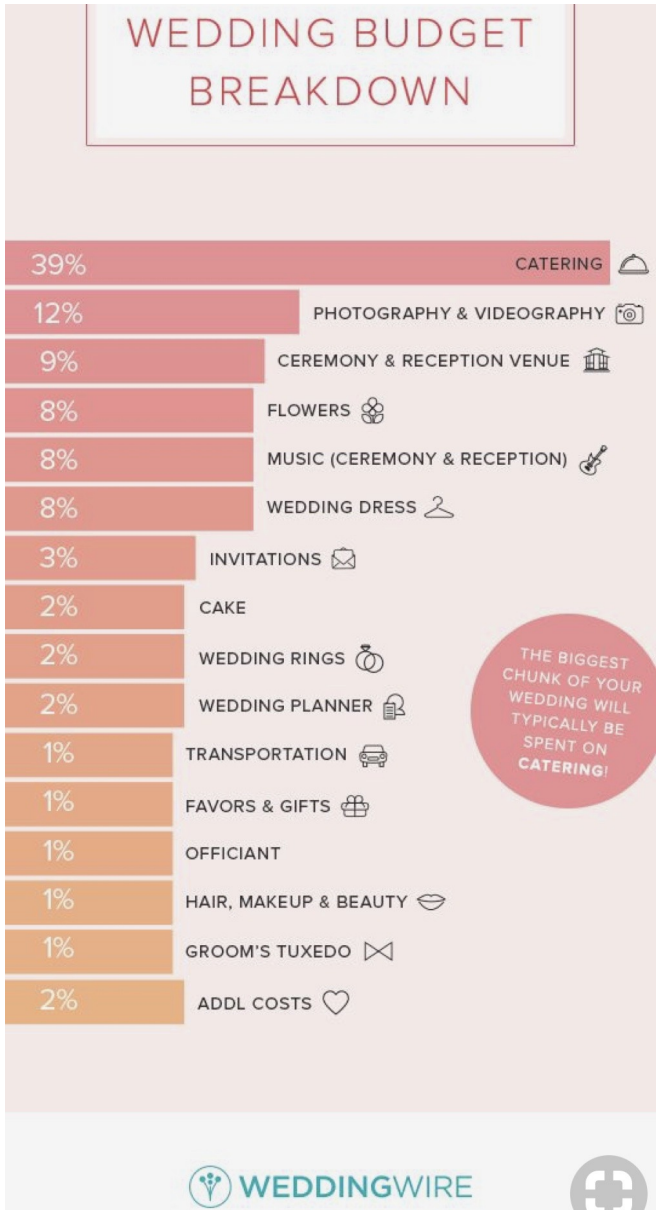
Simpson's Paradox

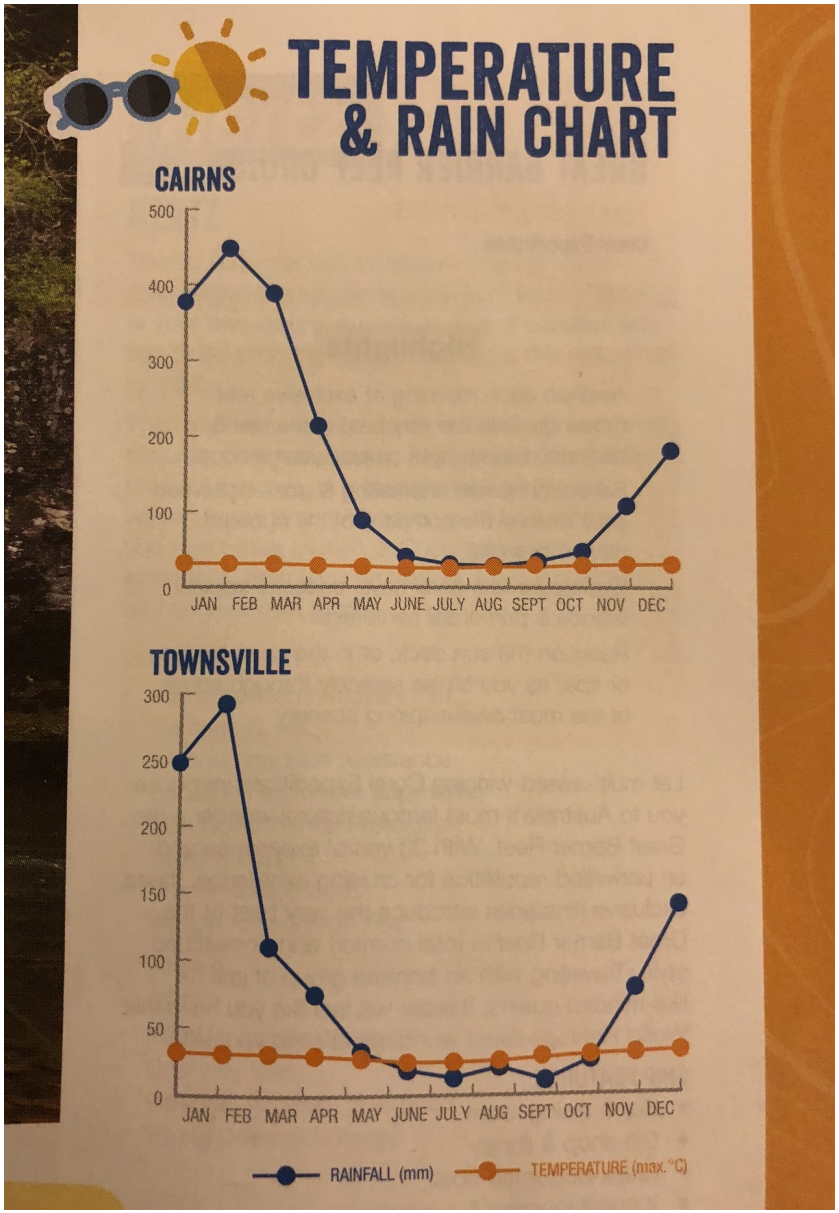


Gambar: Kesalahan dalam interpretasi pola karena tidak dibagi per grup

Dengan visualisasi, bukan berarti bebas masalah.

Kita tetap bisa berbohong!


WEDDING BUDGET BREAKDOWN
39% CATERING
12% PHOTOGRAPHY & VIDEOGRAPHY
9% CEREMONY & RECEPTION VENUE
8% FLOWERS
8% MUSIC (CEREMONY & RECEPTION)
8% WEDDING DRESS
3% INVITATIONS
2% CAKE
2% WEDDING RINGS
2% WEDDING PLANNER
1% TRANSPORTATION
1% FAVORS & GIFTS
1% OFFICIANT
1% HAIR, MAKEUP & BEAUTY
1% GROOM'S TUXEDO
2% ADDL COSTS
THE BIGGEST CHUNK OF YOUR WEDDING WILL TYPICALLY BE SPENT ON CATERING!
WEDDINGWIRE


TEMPERATURE
& RAIN CHART
CAIRNS
500
400
300
200
100
0
JAN FEB MAR APR MAY JUNE JULY AUG SEPT OCT NOV DEC
TOWNSVILLE
300
250
200
150
100
50
0
JAN FEB MAR APR MAY JUNE JULY AUG SEPT OCT NOV DEC
RAINFALL (mm)
TEMPERATURE (max. °C)

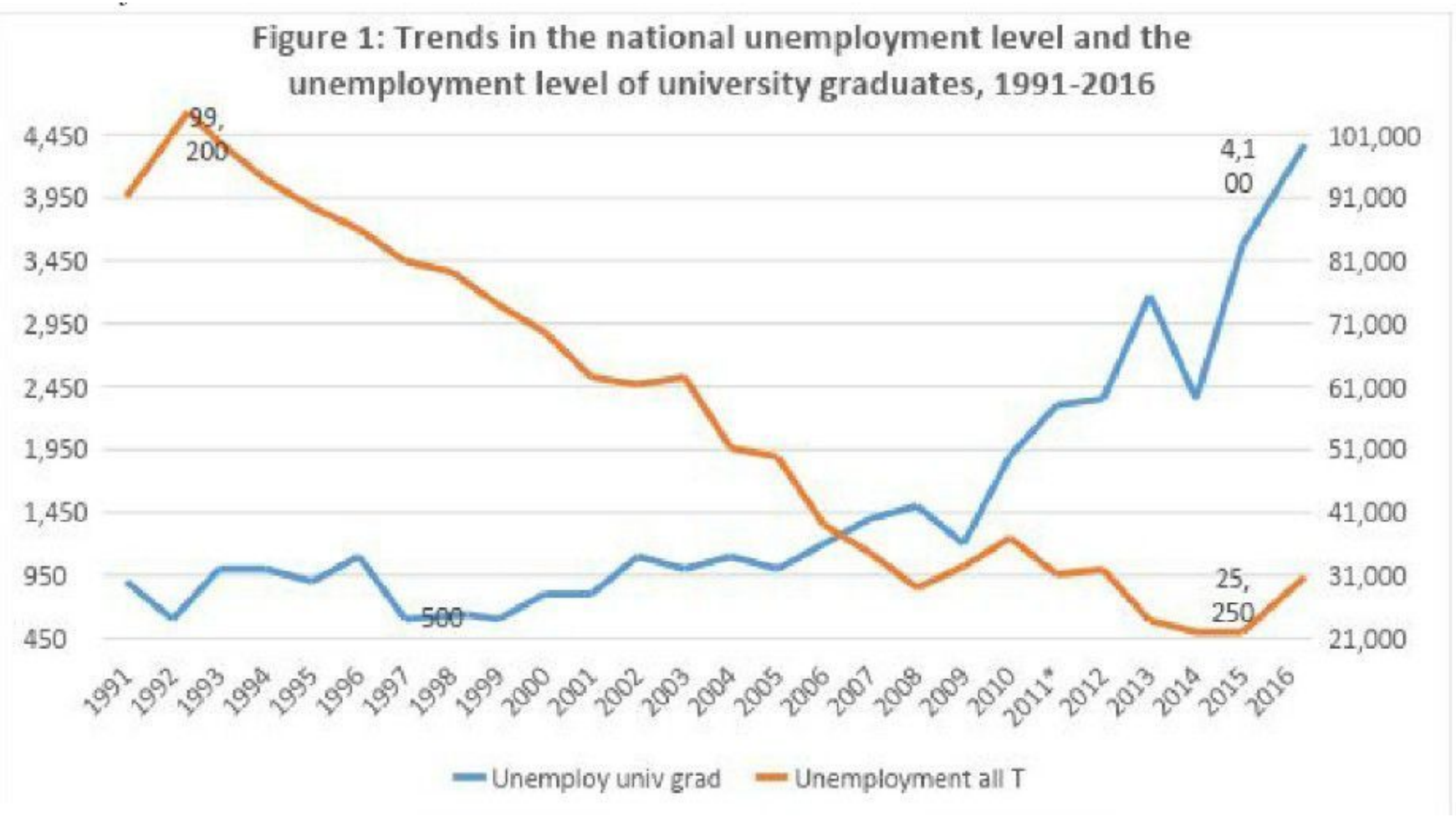
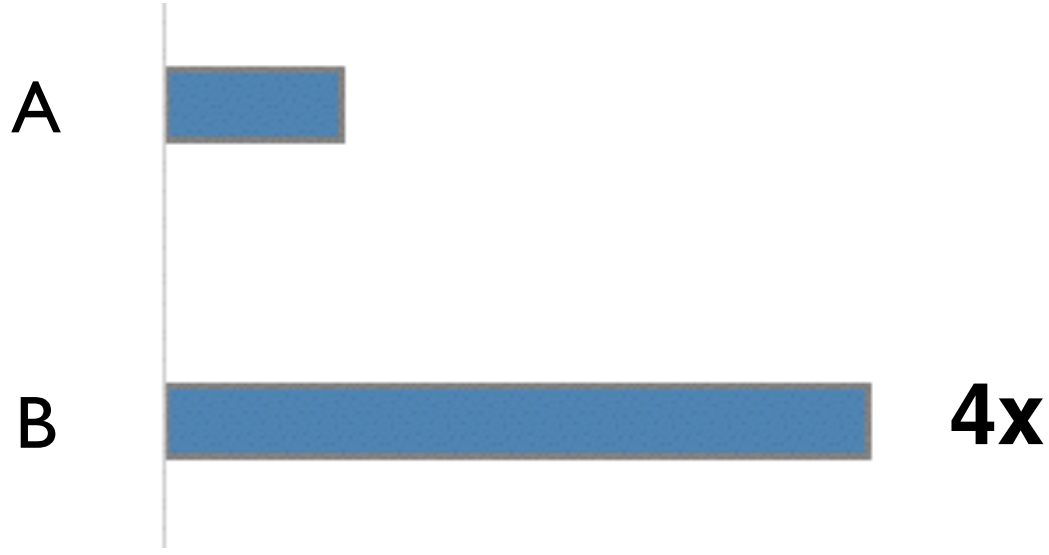


Figure 1: Trends in the national unemployment level and the unemployment level of university graduates, 1991-2016
4,450
3,950
3,450
2,950
2,450
1,950
1,450
950
450
101,000
91,000
81,000
71,000
61,000
51,000
41,000
31,000
21,000
99,200
500
4,100
25,250
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011*
2012
2013
2014
2015
2016
Unemploy univ grad
Unemployment all T

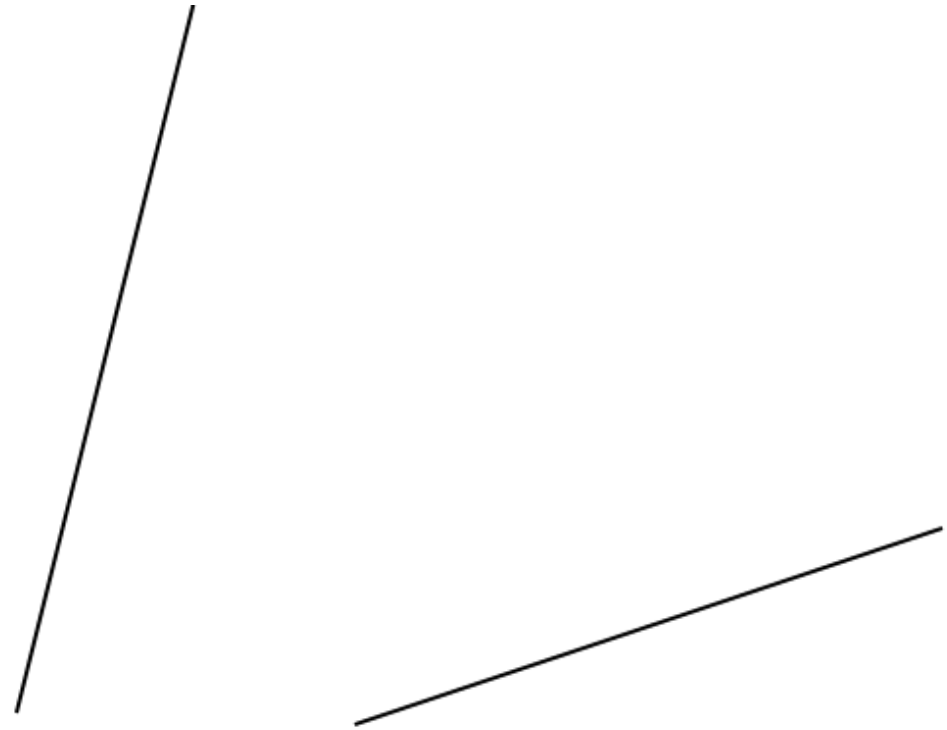
Gambar: Dapat digunakan untuk keuntungan politik. Sumber: @ainunnajib

Visualisasi Efektif

How much longer?



How much steeper slope?

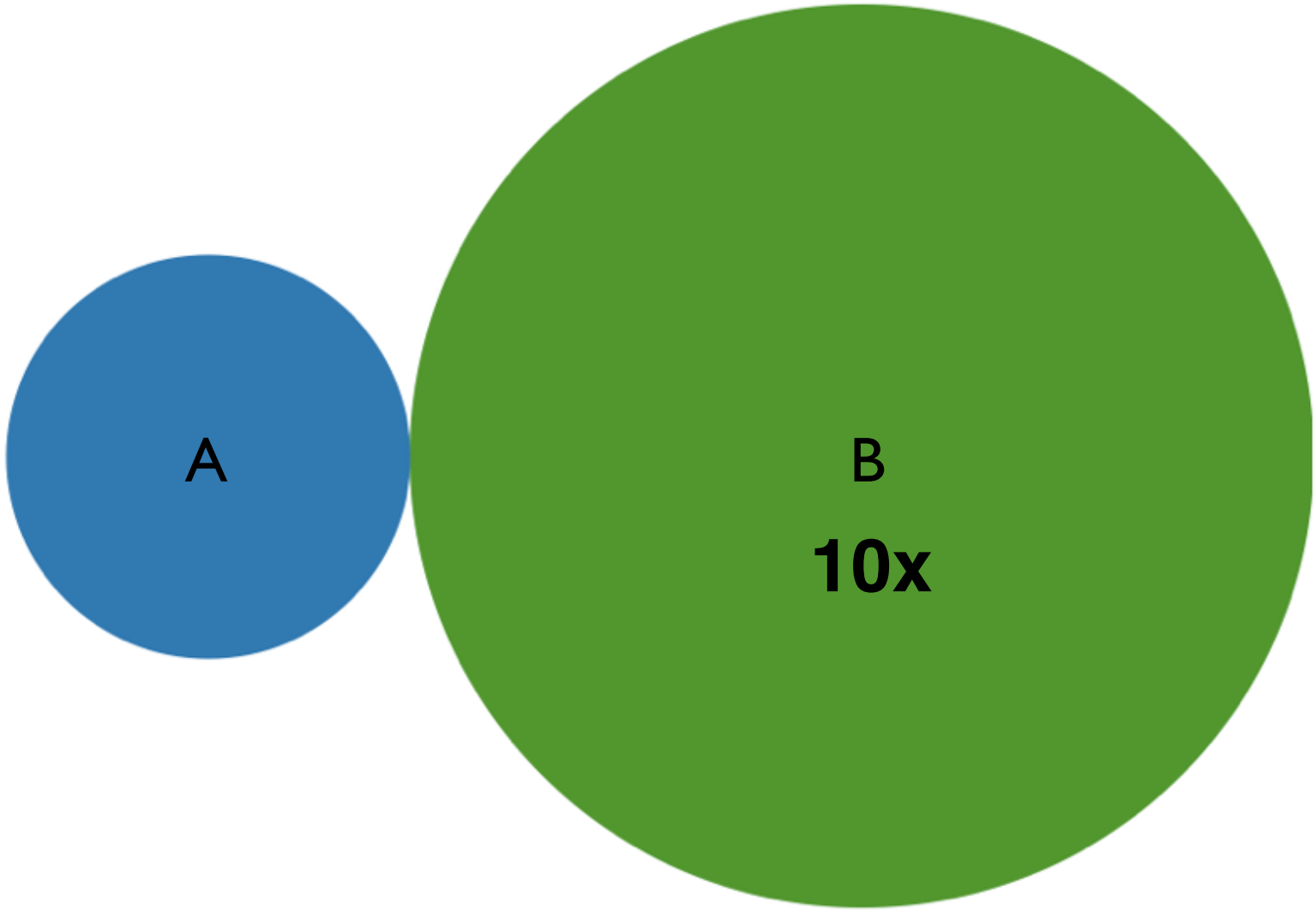


A

B

4x

How much larger area?



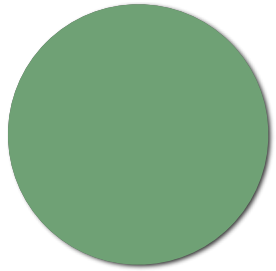
How much darker?

A

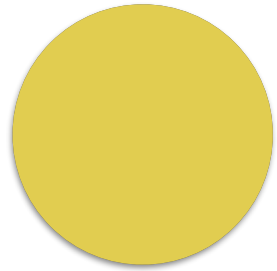
B

2x

How much bigger value?

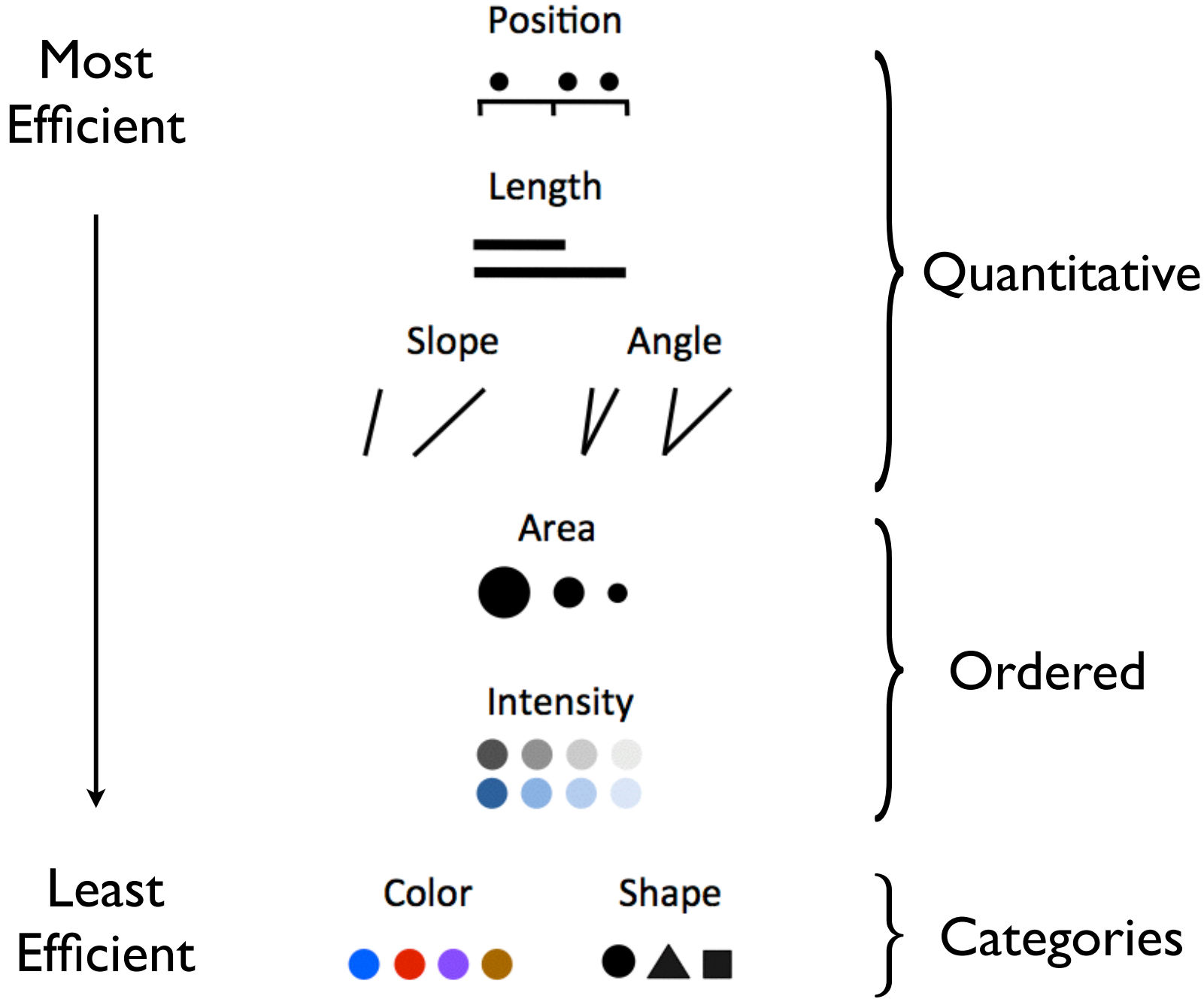


A



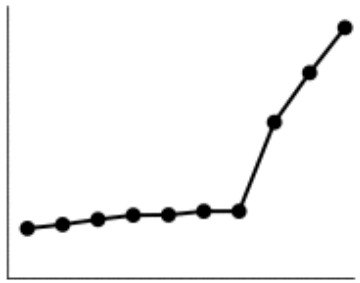
B

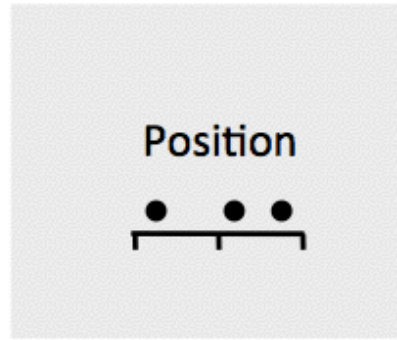
4x

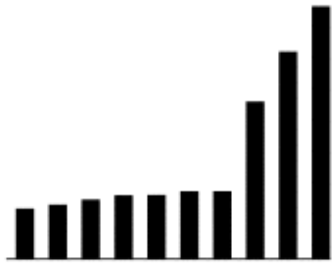


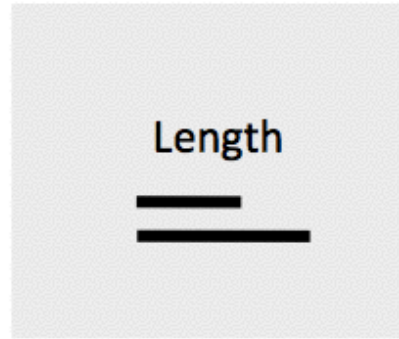
C. Mulbrandon
VisualizingEconomics.com

Most Effective

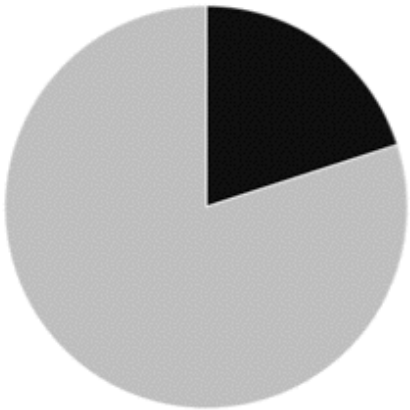


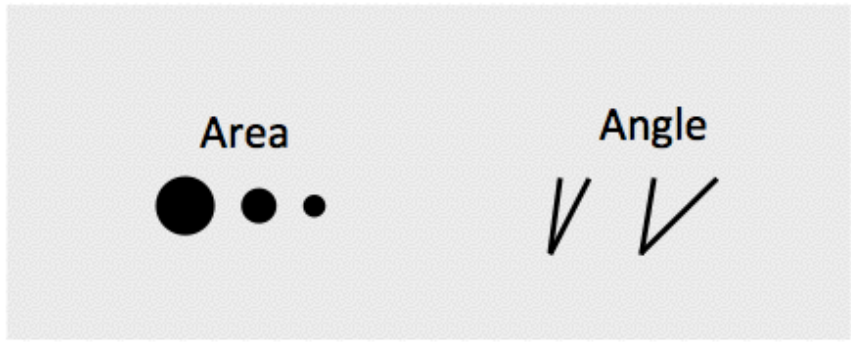


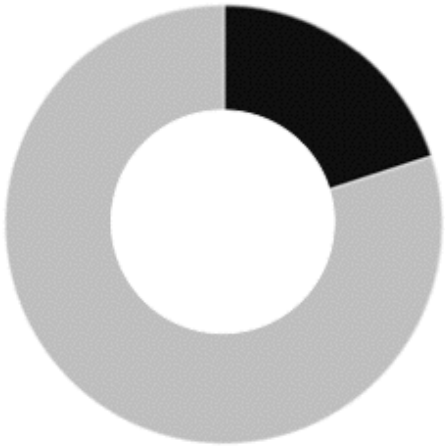


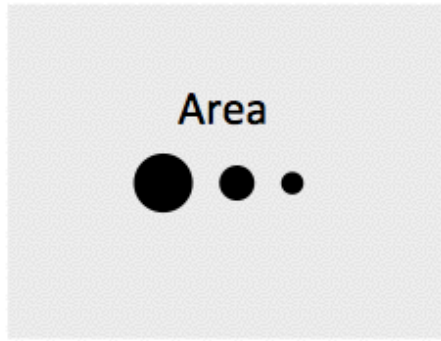


Less Effective



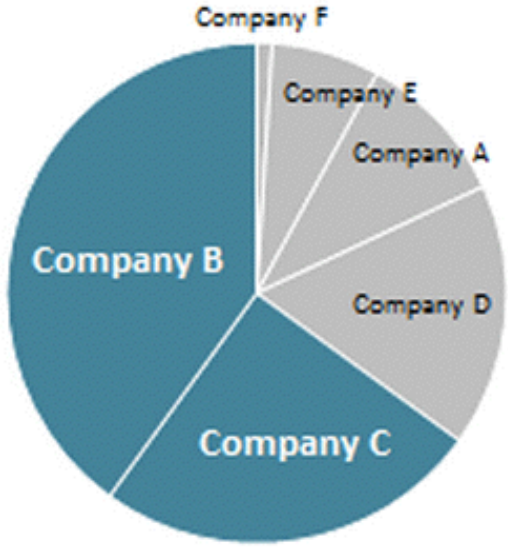


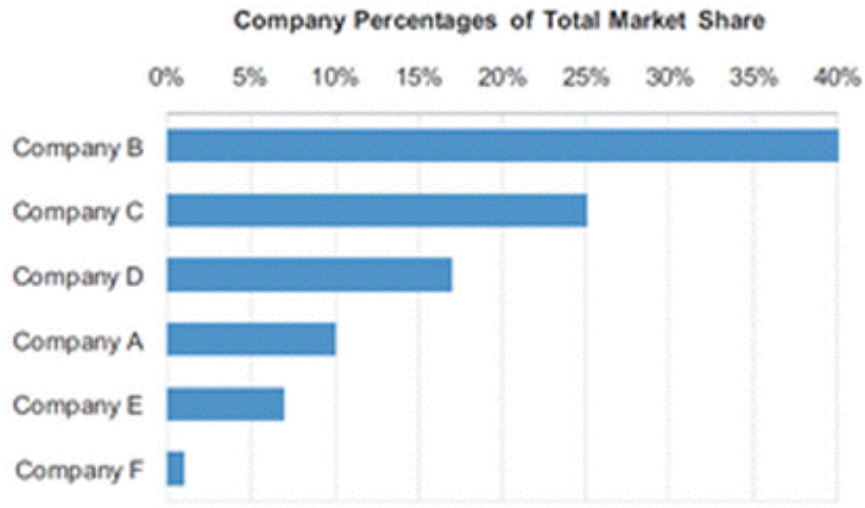




Pie vs. Bar Charts

65% of the market is controlled by companies B and C





Least Effective

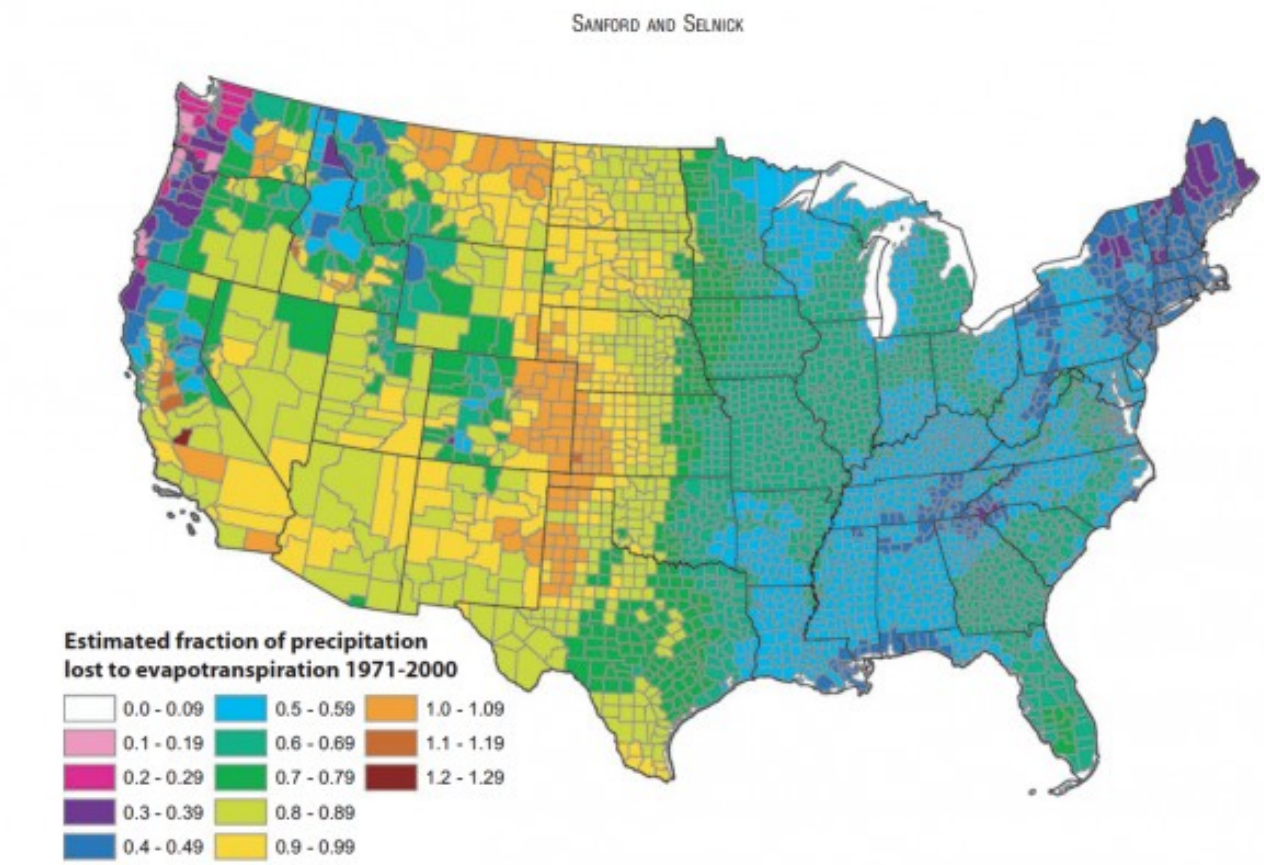


FIGURE 13. Estimated Mean Annual Ratio of Actual Evapotranspiration (ET) to Precipitation (P) for the Conterminous U.S. for the Period 1971-2000. Estimates are based on the regression equation in Table 1 that includes land cover. Calculations of ET/P were made first at the 800-m resolution of the PRISM climate data. The mean values for the counties (shown) were then calculated by averaging the 800-m values within each county. Areas with fractions >1 are agricultural counties that either import surface water or mine deep groundwater.

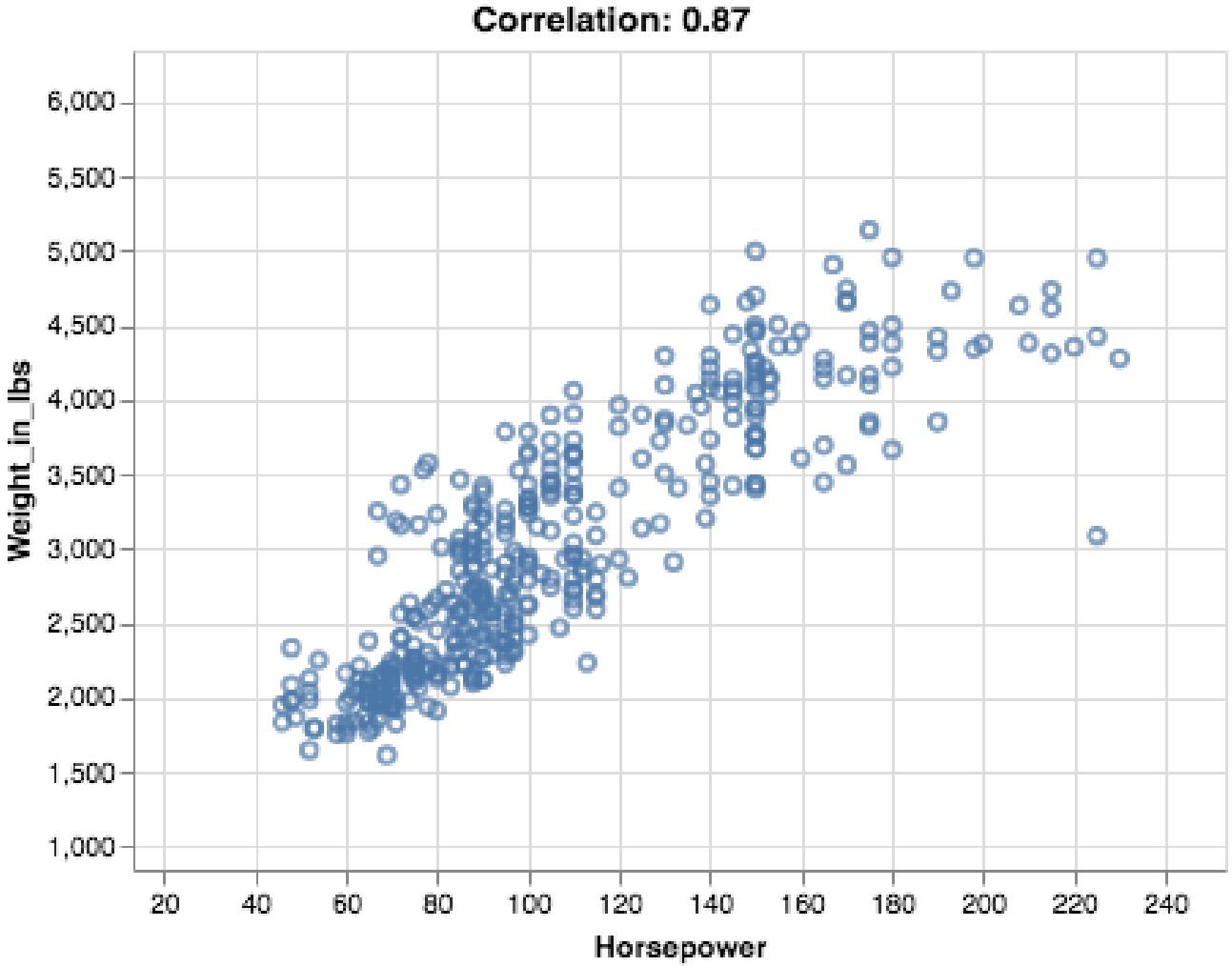
Cliff Mass

CS109 Data Science
Lecture 3: Exploratory Data Analysis

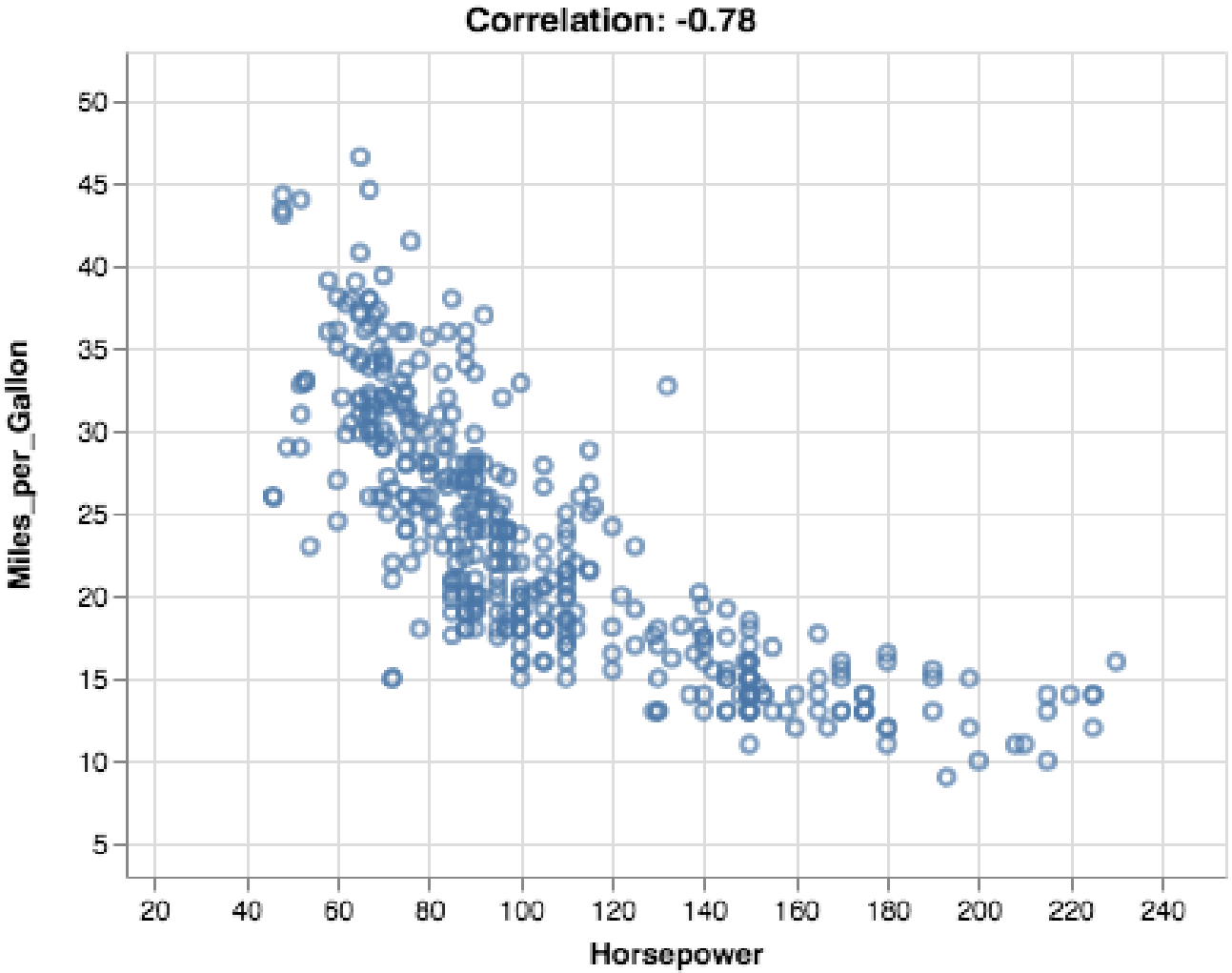
Korelasi

Korelasi

- Korelasi mengukur hubungan linear antarvariabel
- Nilai korelasi ada di rentang [-1, 1]
- Korelasi $= 1$ berarti berbanding lurus sempurna
- *Correlation doesn't imply causation!*

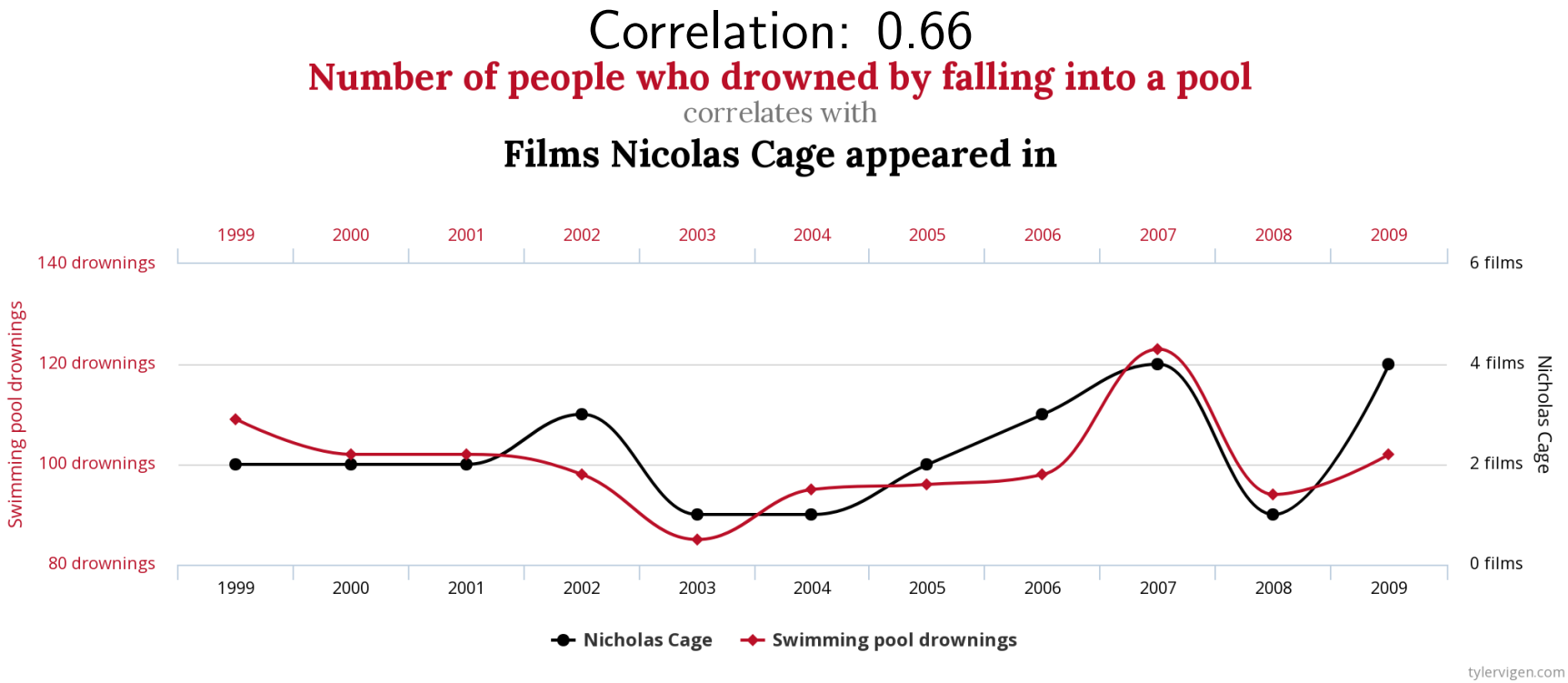


Gambar: Semakin besar tenaga kuda, mobil juga semakin berat



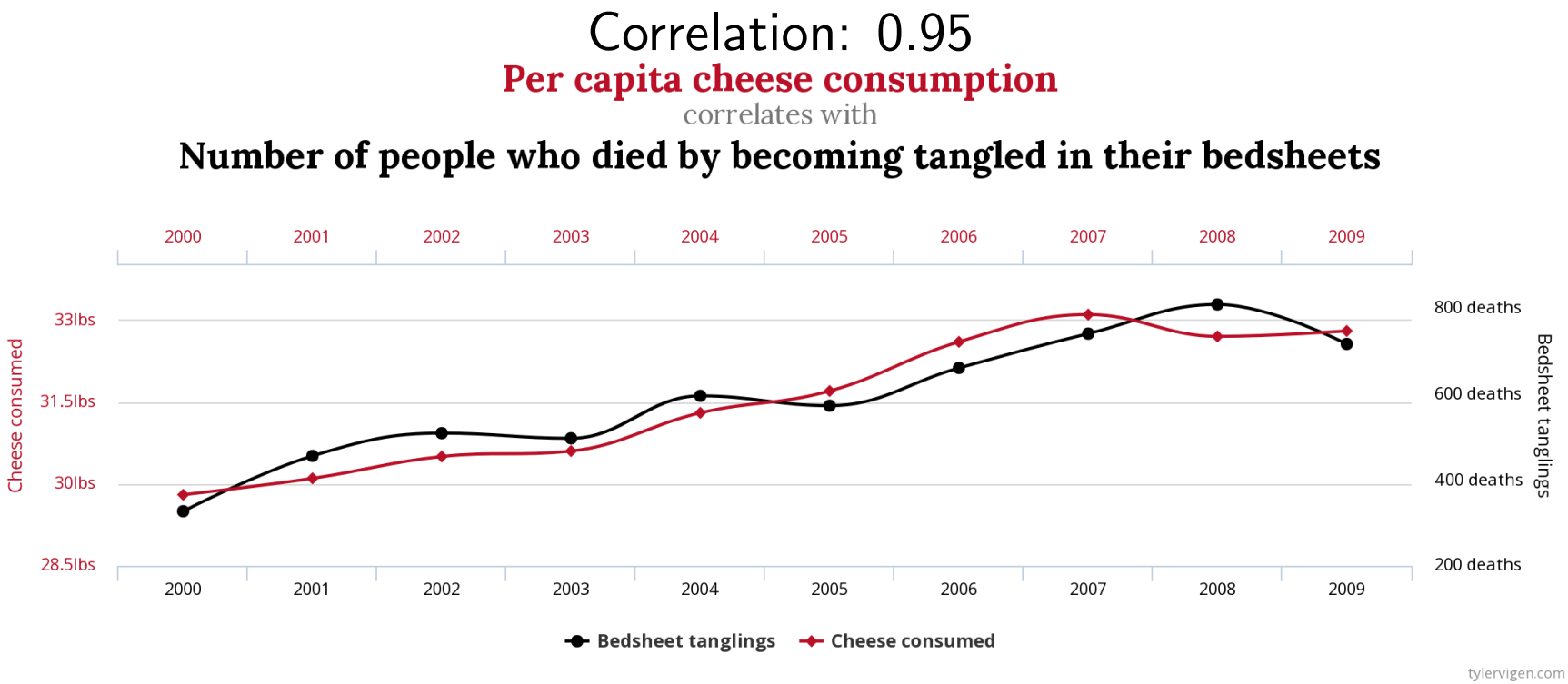
Gambar: Semakin besar tenaga kuda, semakin boros

Spurious Correlation 1

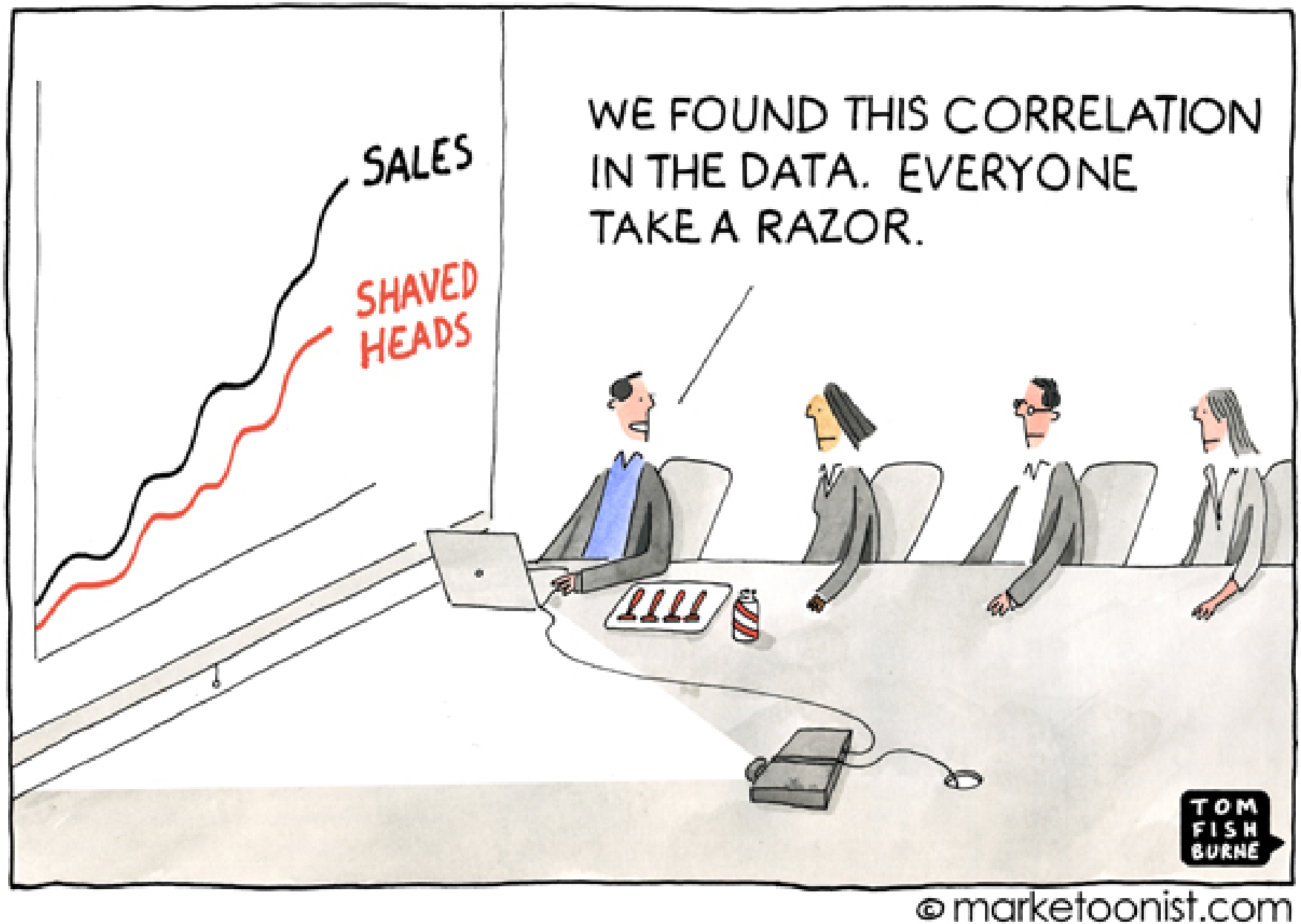


Gambar: Apakah kita harus meminta Nicolas Cage berhenti main film?

Spurious Correlation 2



http://www.tylervigen.com/spurious-correlations


SALES
SHAVED HEADS
WE FOUND THIS CORRELATION IN THE DATA. EVERYONE TAKE A RAZOR.
TOM FISH BURNE
© marketoonist.com

https://callingbullshit.org/

Pekan depan:

Kombinatorika

Terima kasih