What is **Data**?

Why is it **Science**?

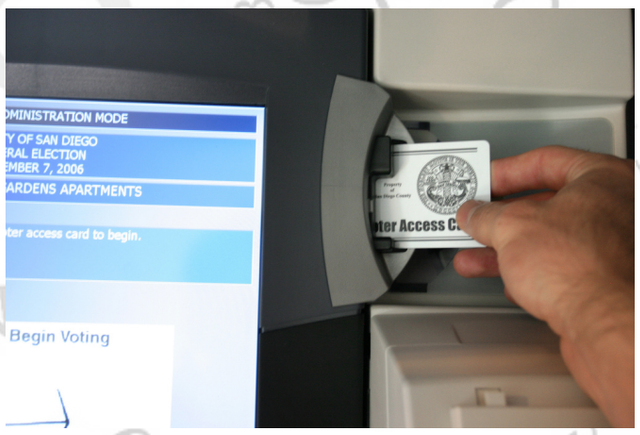
Data is everywhere…

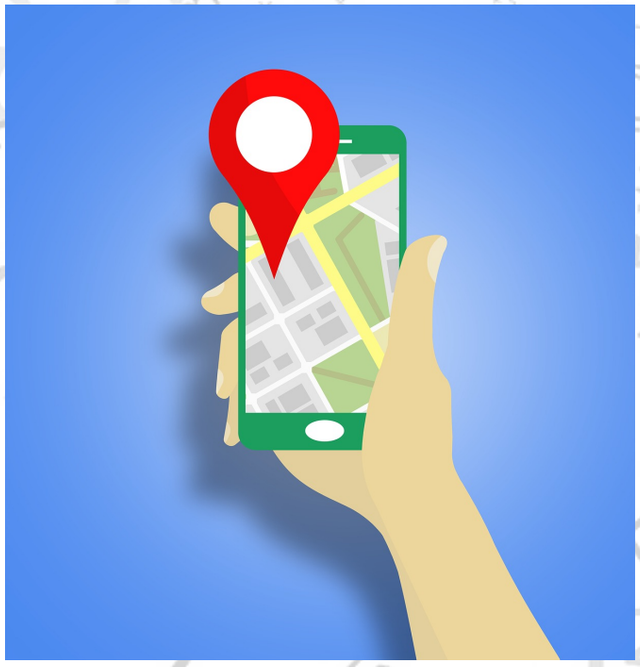
ebay

VISA

Booking.com

amazon





Meaning from data is **money**…

Harvard Business Review

DATA

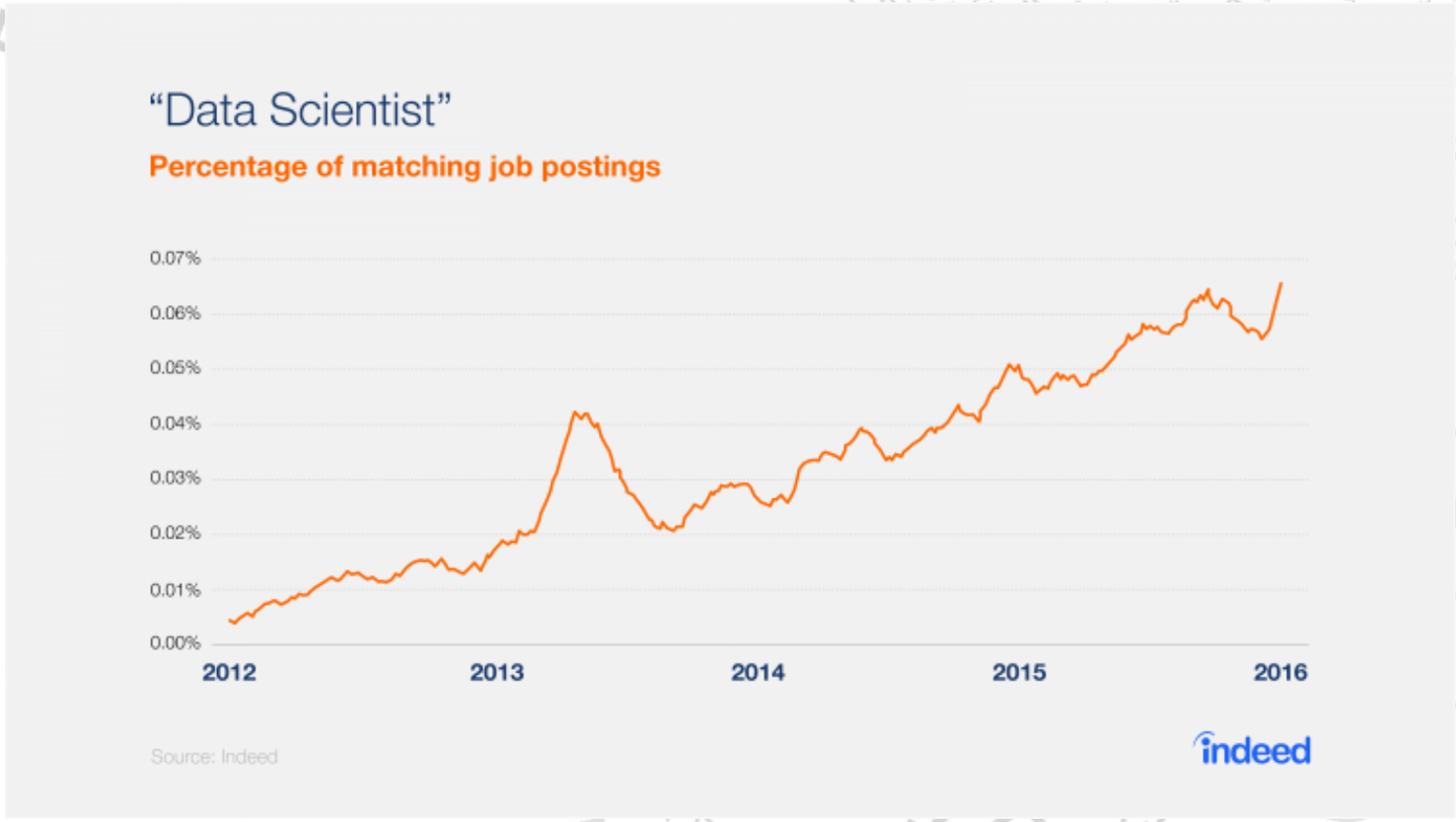
Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

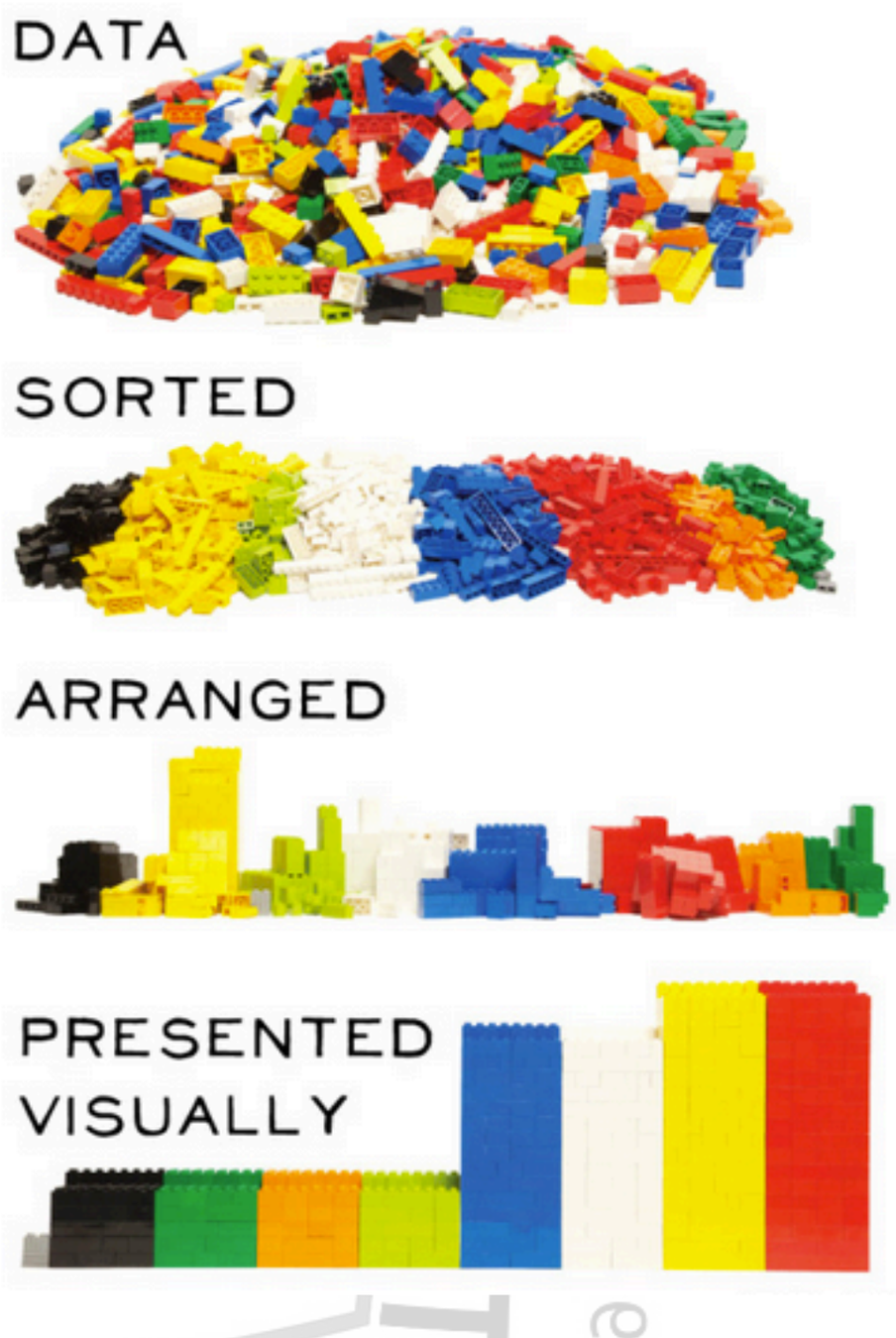
FROM THE OCTOBER 2012 ISSUE

DSPT
DATA SCIENCE PORTUGAL

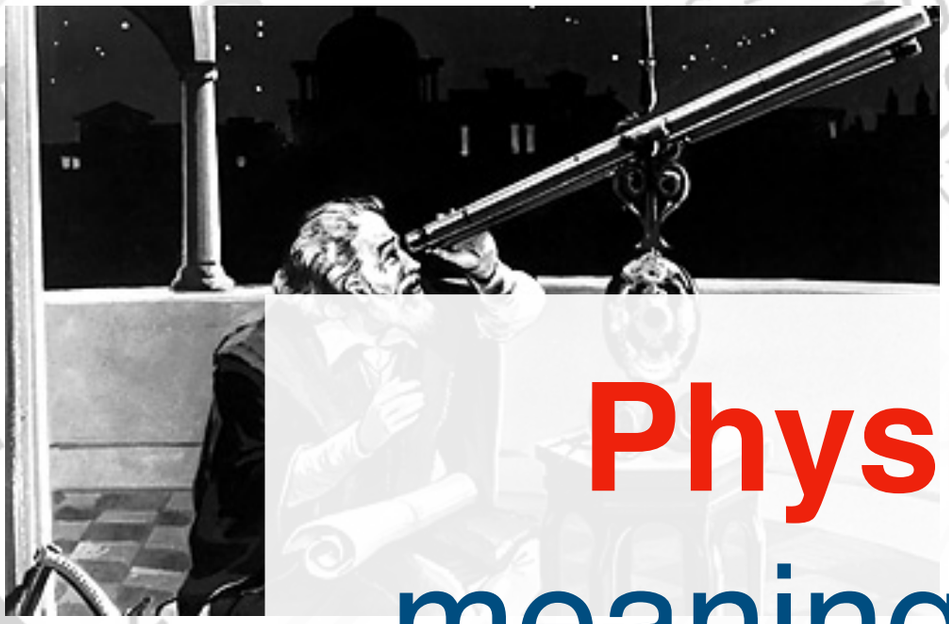
DSPA
DATA SCIENCE PORTUGUESE ASSOCIATION

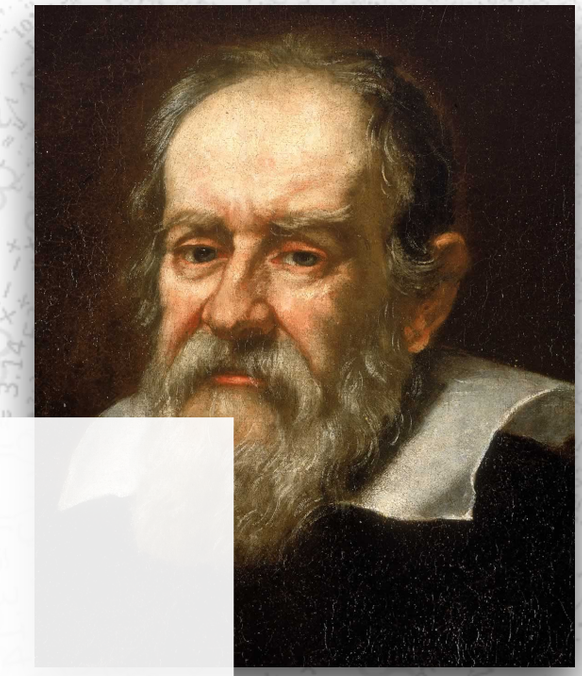


Challenges…

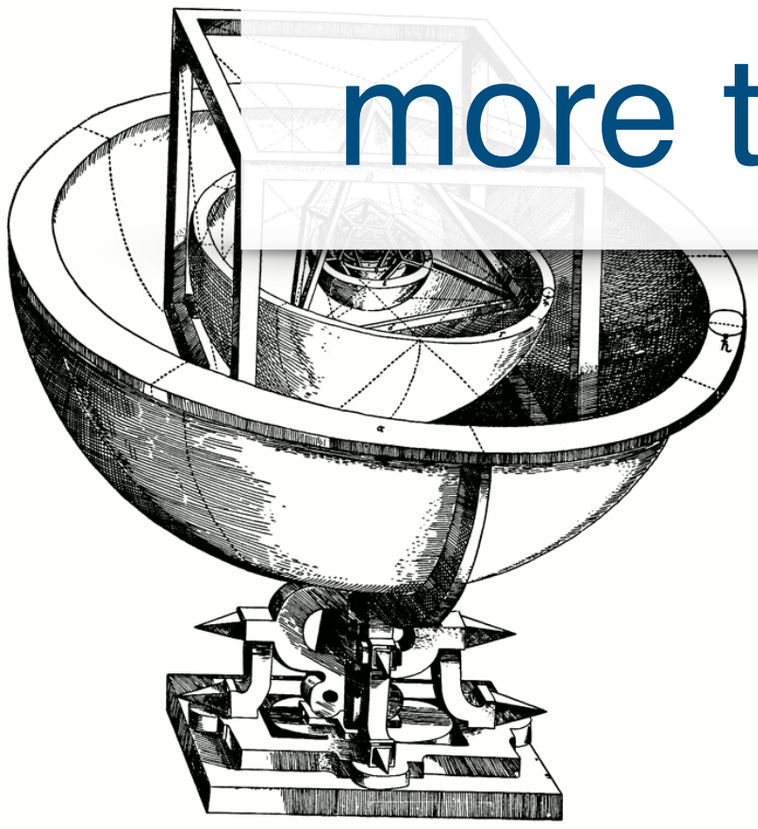


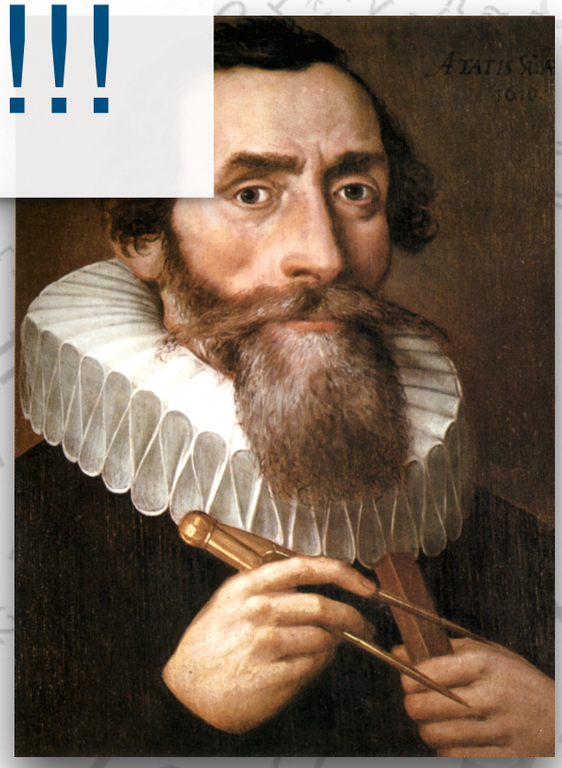
Why **us**?





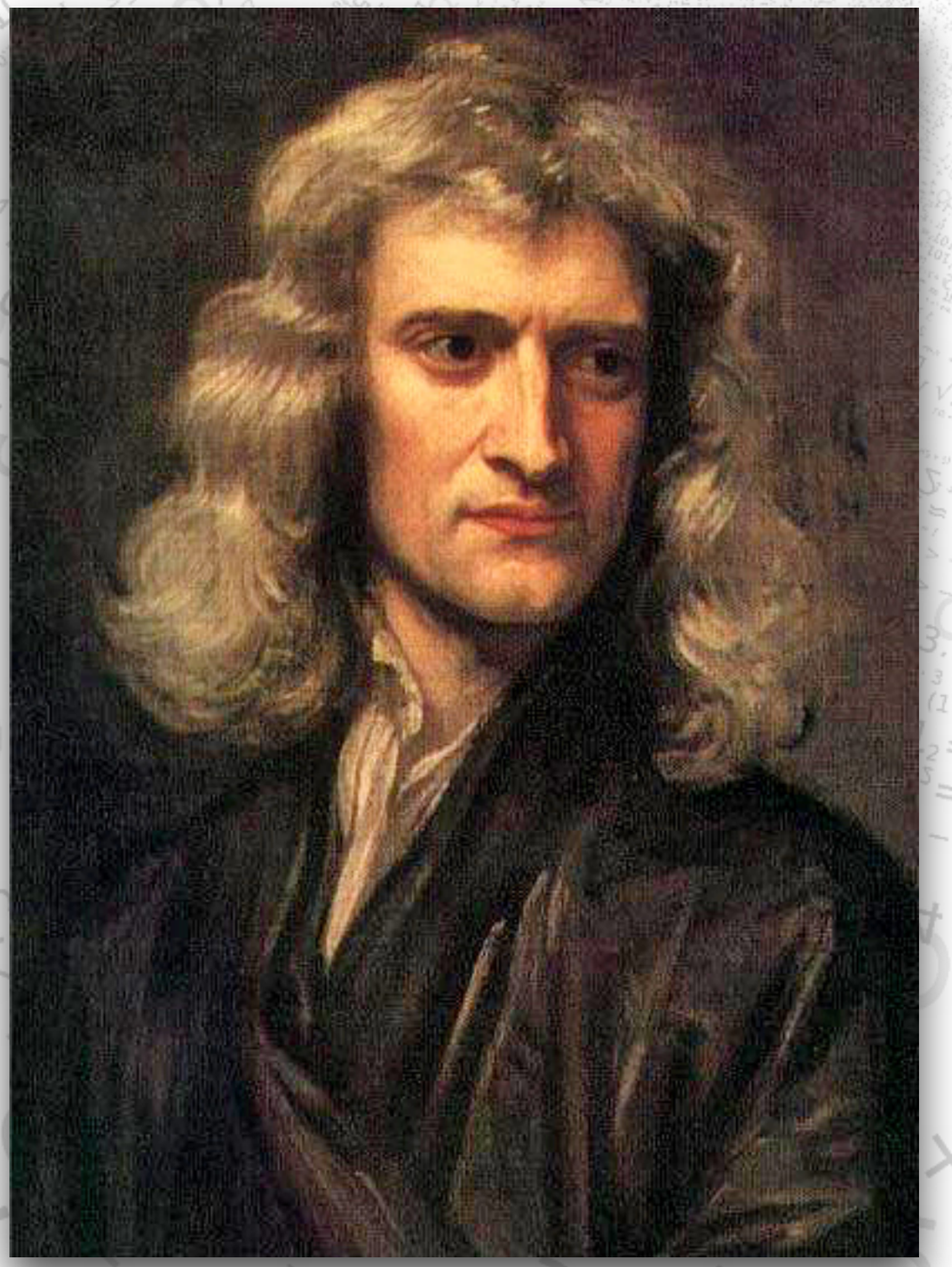
Physics: Taking meaning from data for more than **500 years**!!!





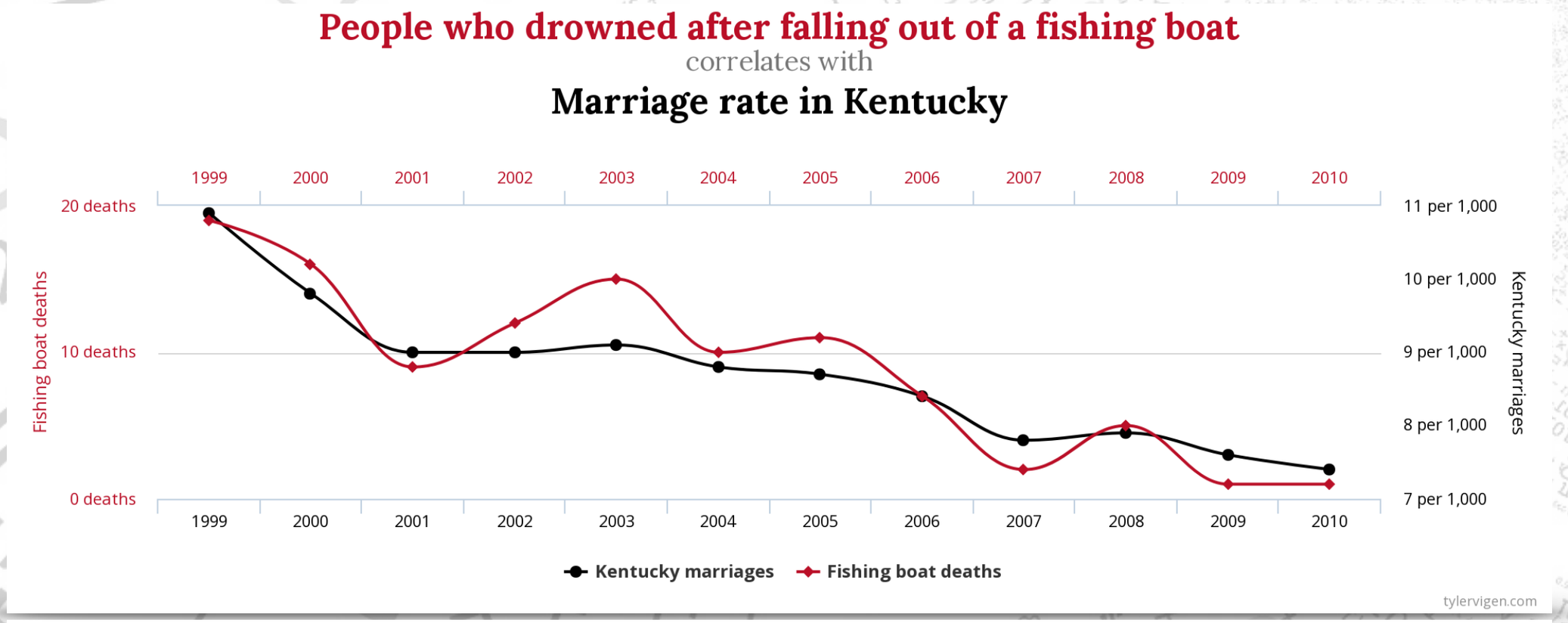
Johannes Kepler
1571-1630

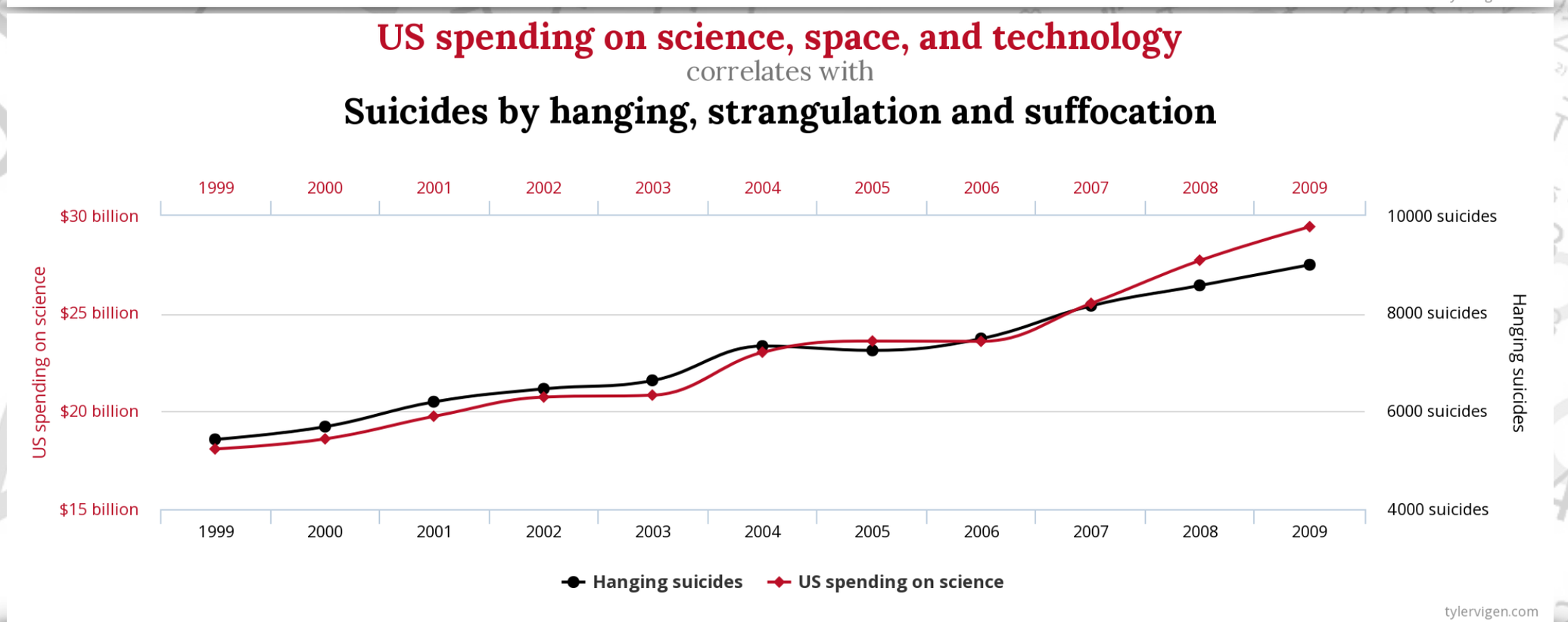
The **power** of the **model**…



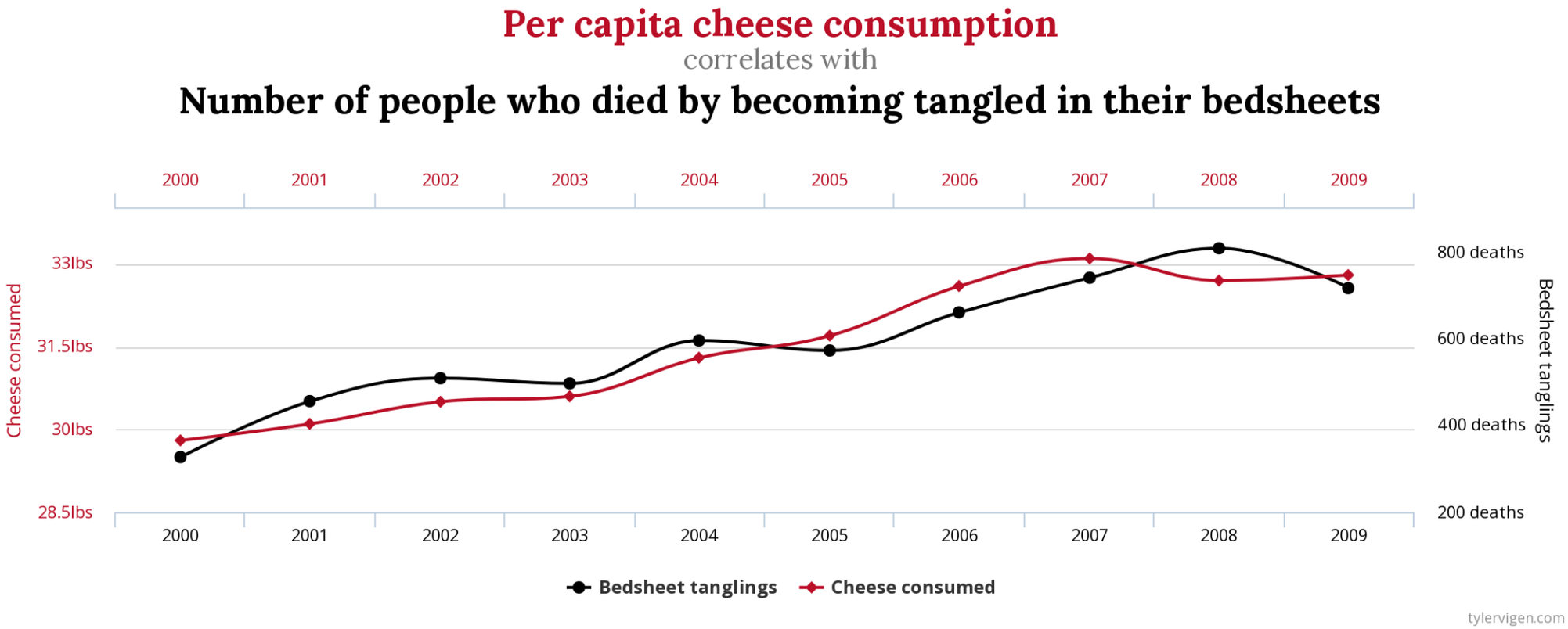
Isaac Newton
1643-1727

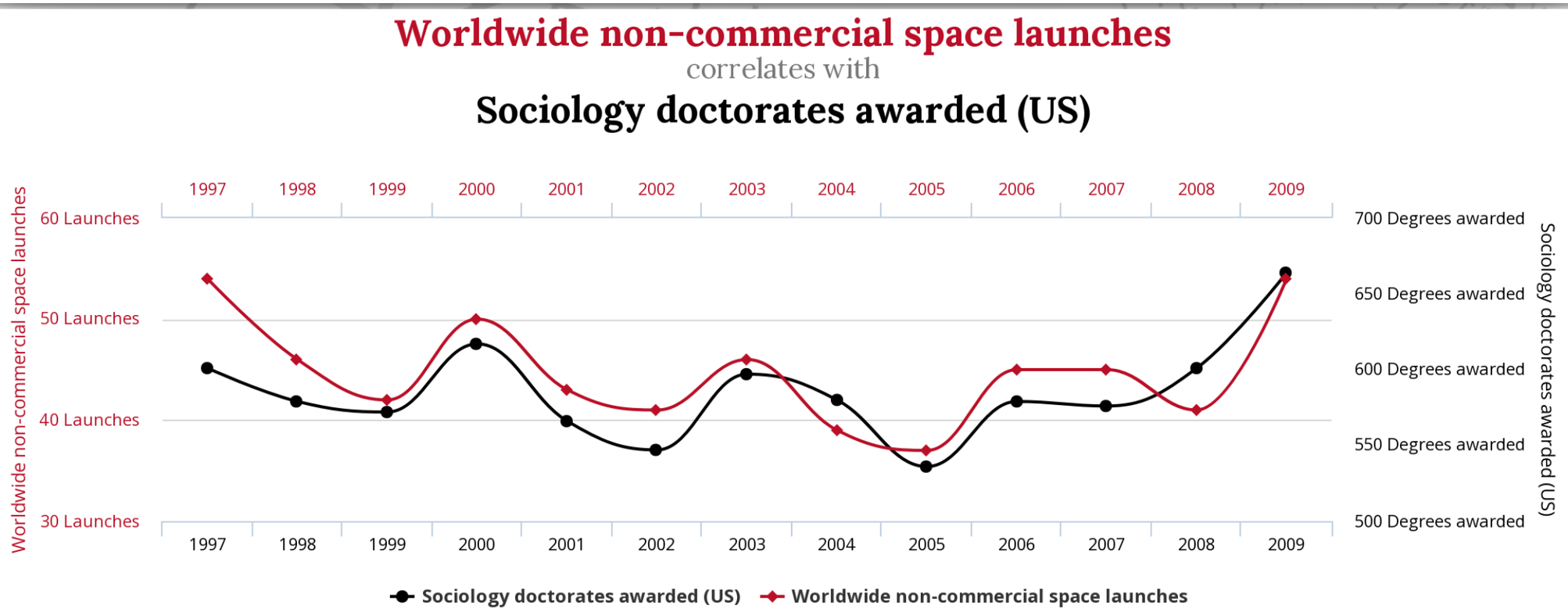
Causality . . .


People who drowned after falling out of a fishing boat
correlates with
Marriage rate in Kentucky
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
20 deaths
10 deaths
0 deaths
Fishing boat deaths
11 per 1,000
10 per 1,000
9 per 1,000
8 per 1,000
7 per 1,000
Kentucky marriages
Kentucky marriages
Fishing boat deaths
tylervigen.com


US spending on science, space, and technology
correlates with
Suicides by hanging, strangulation and suffocation
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

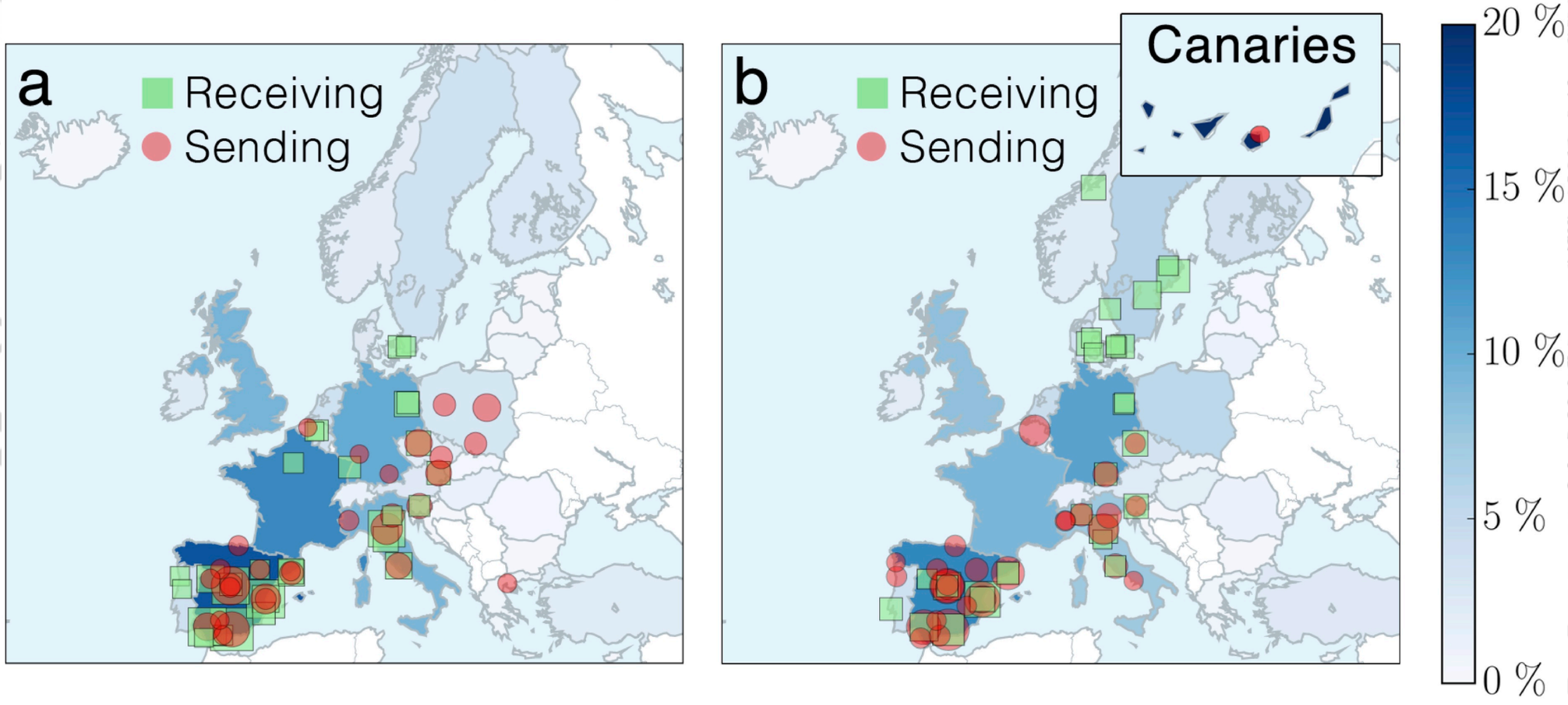
Causality . . .


Per capita cheese consumption
correlates with
Number of people who died by becoming tangled in their bedsheets
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
33lbs
31.5lbs
30lbs
28.5lbs
Cheese consumed
800 deaths
600 deaths
400 deaths
200 deaths
Bedsheet tanglings
Bedsheet tanglings
Cheese consumed
tylervigen.com


Worldwide non-commercial space launches
correlates with
Sociology doctorates awarded (US)
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
60 Launches
50 Launches
40 Launches
30 Launches
Worldwide non-commercial space launches
700 Degrees awarded
650 Degrees awarded
600 Degrees awarded
550 Degrees awarded
500 Degrees awarded
Sociology doctorates awarded (US)
Sociology doctorates awarded (US)
Worldwide non-commercial space launches
tylervigen.com

Causality: Gender bias in Erasmus

Female *Male*



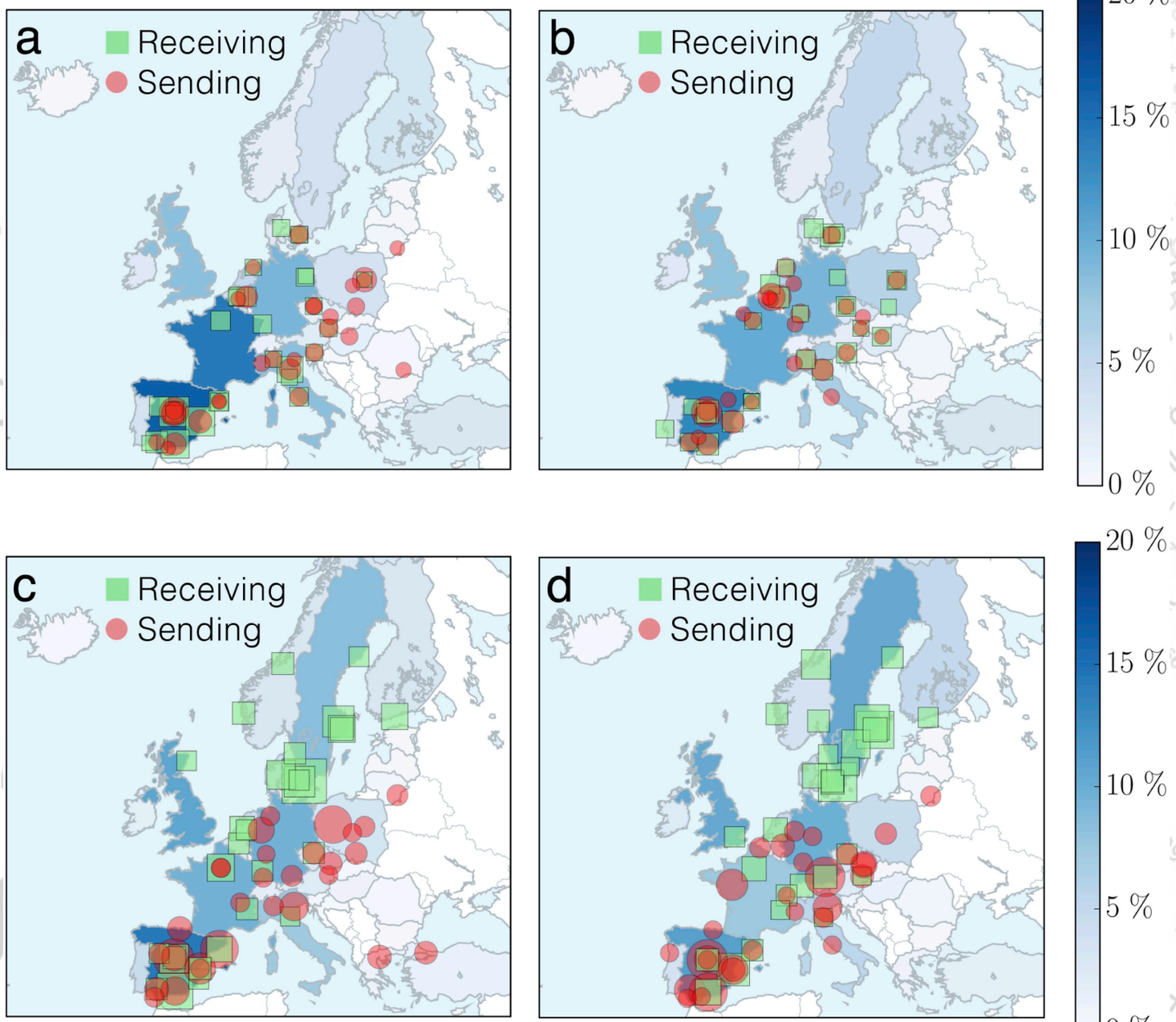
PLoS One **11**, e0149514 (2016).

Causality: Gender bias in Erasmus

Female *Male*

Social Sciences

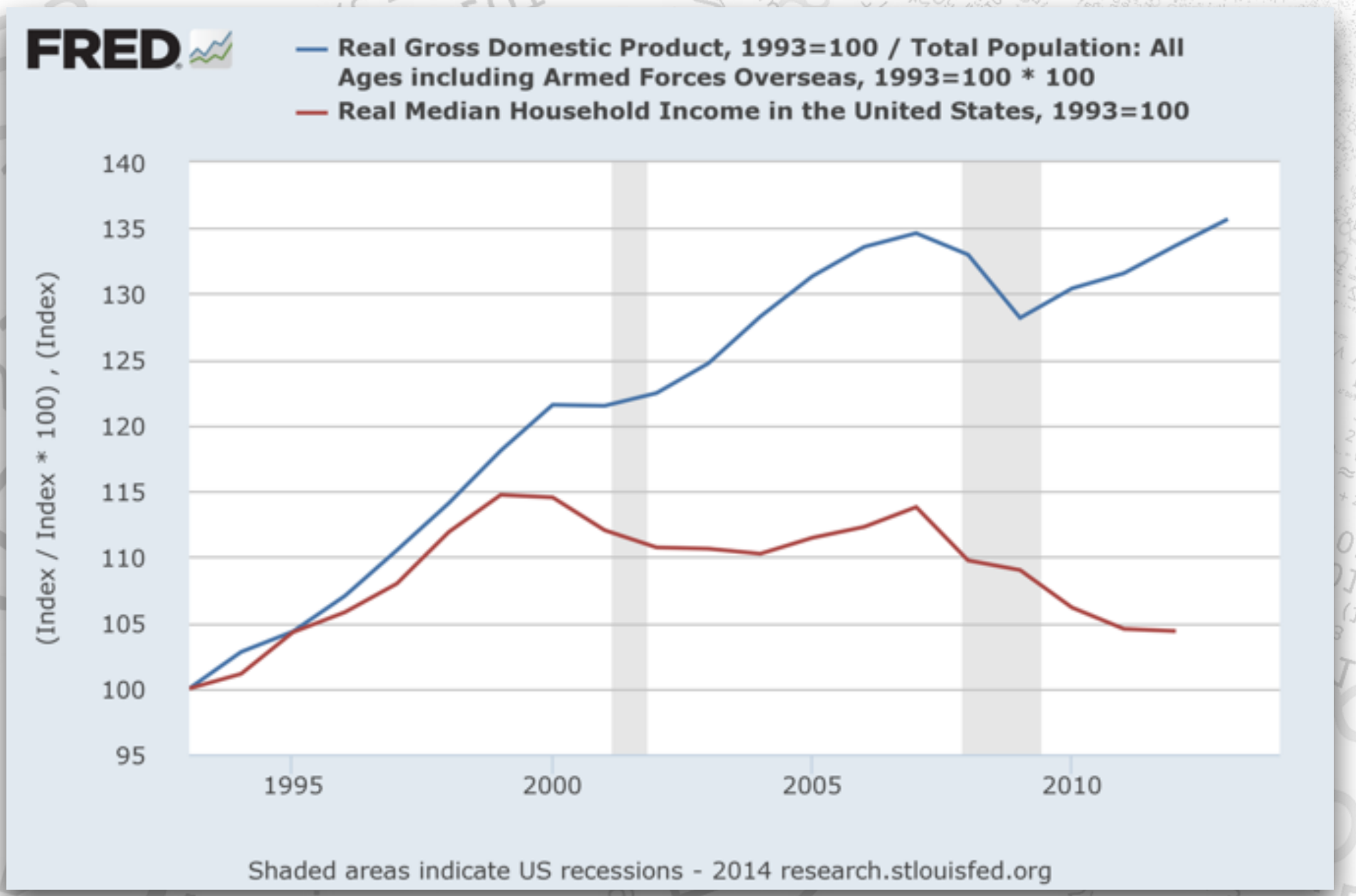
Natural Sciences



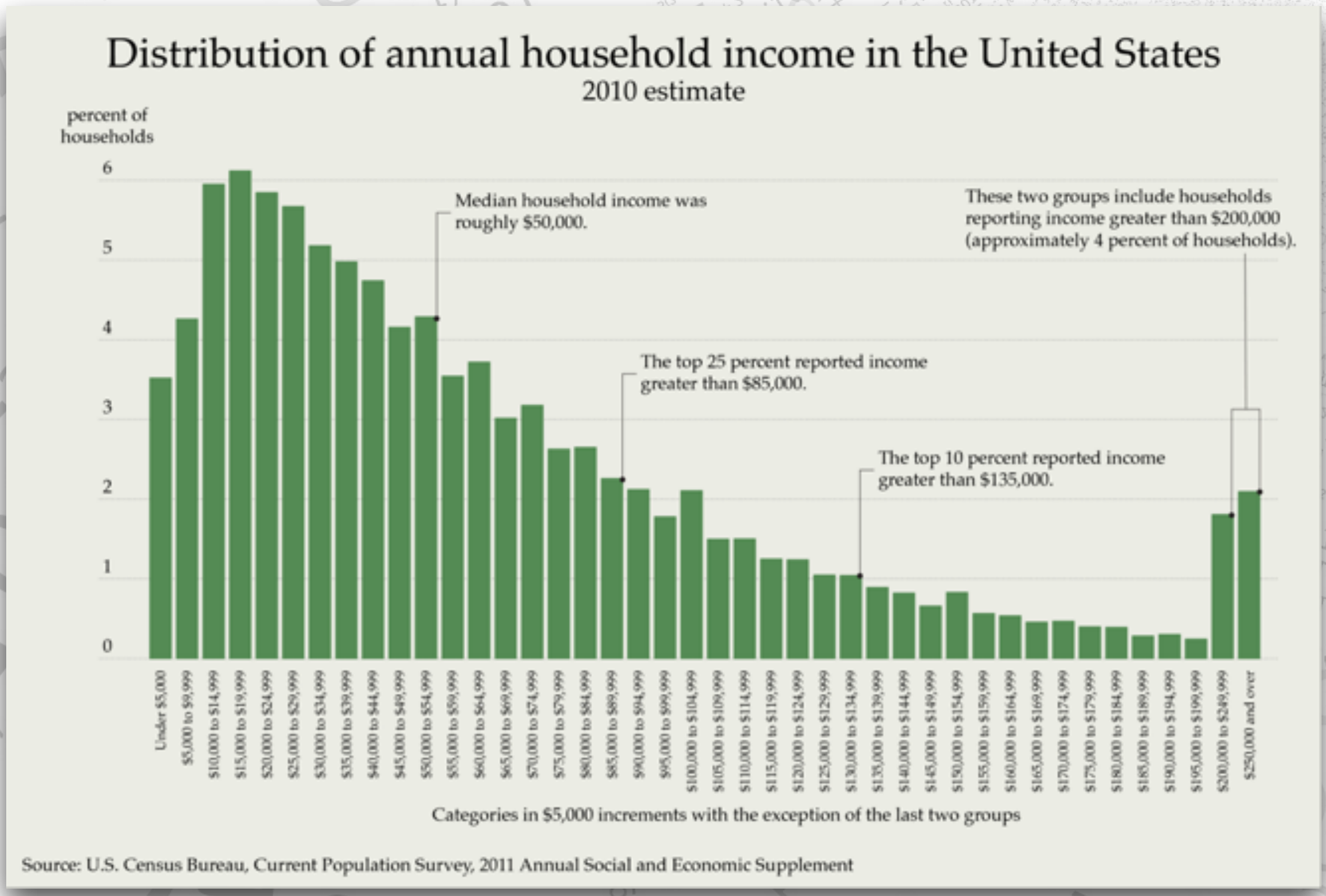
PLoS One **11**, e0149514 (2016).

Is the **average** meaningful?

Average vs Median

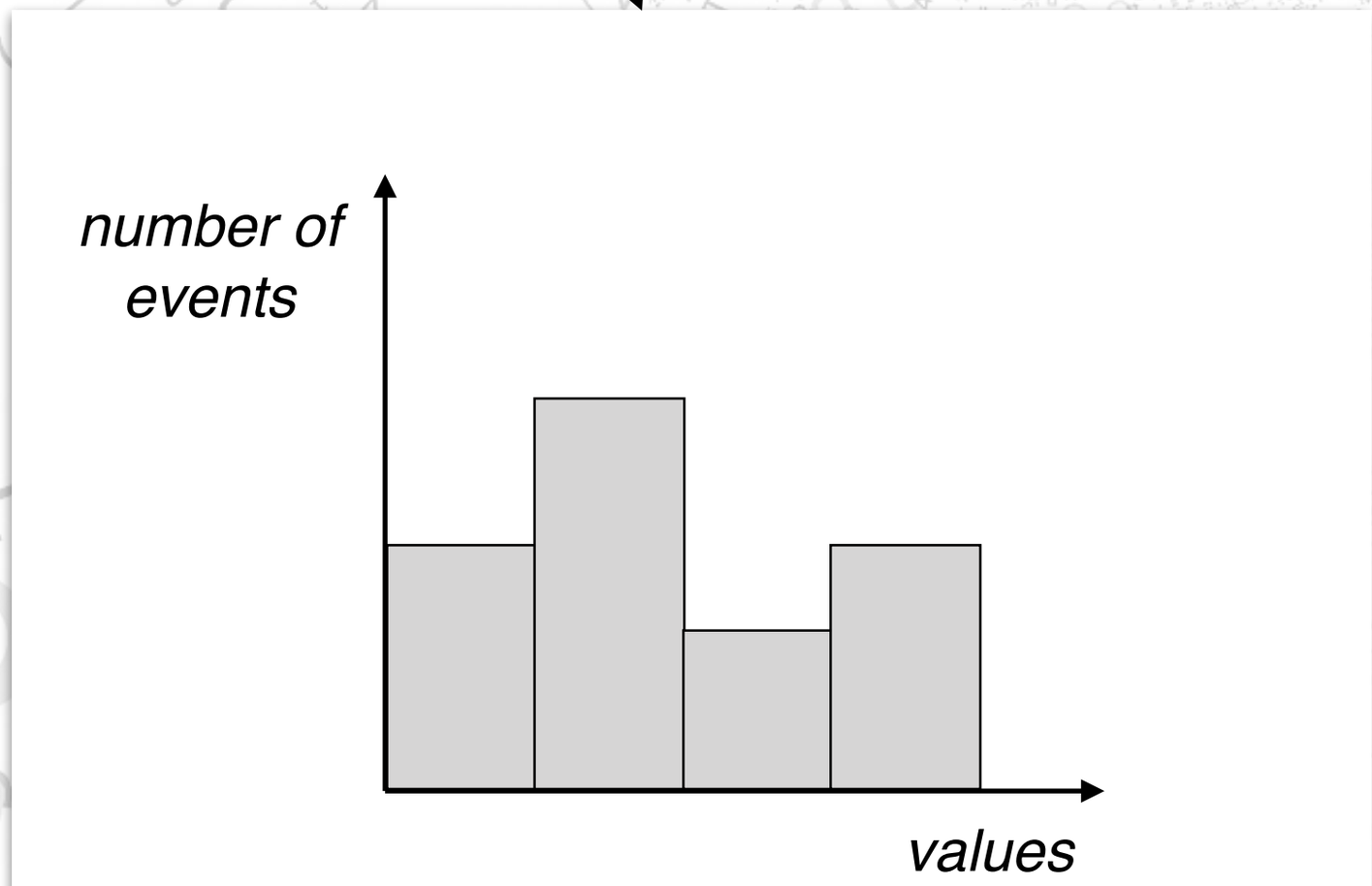


Histograms…


Distribution of annual household income in the United States
2010 estimate
percent of households
6
5
4
3
2
1
0
Median household income was roughly $50,000.
The top 25 percent reported income greater than $85,000.
The top 10 percent reported income greater than $135,000.
These two groups include households reporting income greater than $200,000 (approximately 4 percent of households).
Under $5,000
$5,000 to $9,999
$10,000 to $14,999
$15,000 to $19,999
$20,000 to $24,999
$25,000 to $29,999
$30,000 to $34,999
$35,000 to $39,999
$40,000 to $44,999
$45,000 to $49,999
$50,000 to $54,999
$55,000 to $59,999
$60,000 to $64,999
$65,000 to $69,999
$70,000 to $74,999
$75,000 to $79,999
$80,000 to $84,999
$85,000 to $89,999
$90,000 to $94,999
$95,000 to $99,999
$100,000 to $104,999
$105,000 to $109,999
$110,000 to $114,999
$115,000 to $119,999
$120,000 to $124,999
$125,000 to $129,999
$130,000 to $134,999
$135,000 to $139,999
$140,000 to $144,999
$145,000 to $149,999
$150,000 to $154,999
$155,000 to $159,999
$160,000 to $164,999
$165,000 to $169,999
$170,000 to $174,999
$175,000 to $179,999
$180,000 to $184,999
$185,000 to $189,999
$190,000 to $194,999
$195,000 to $199,999
$200,000 to $249,999
$250,000 and over
Categories in $5,000 increments with the exception of the last two groups
Source: U.S. Census Bureau, Current Population Survey, 2011 Annual Social and Economic Supplement

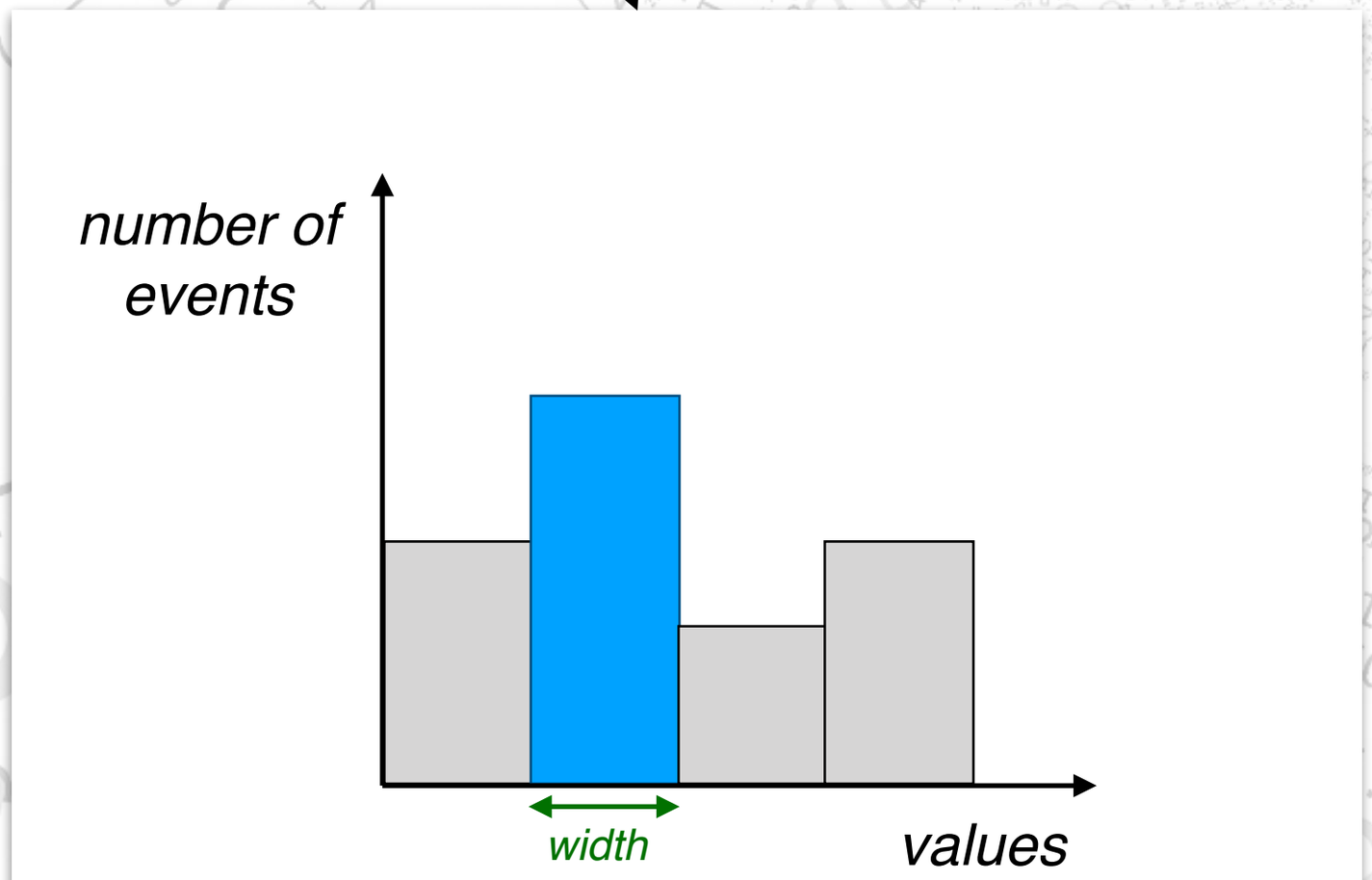
How to make a **histograms**…

6.39
4.62
3.13
4.96
0.09
0.19
8.88
0.59
1.74
0.24
7.58
…



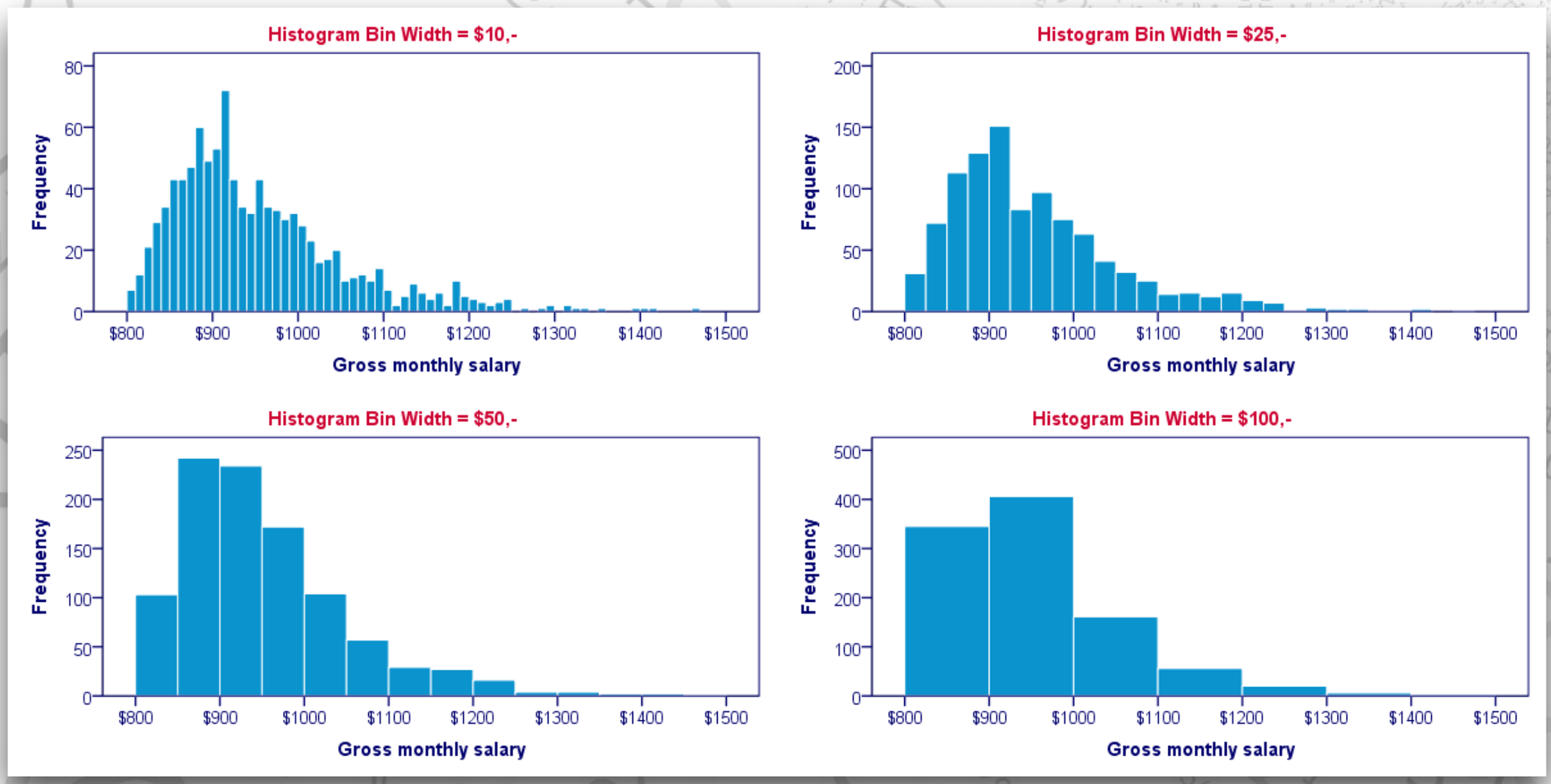
How to make a **histograms**…

6.39
4.62
3.13
4.96
0.09
0.19
8.88
0.59
1.74
0.24
7.58
…

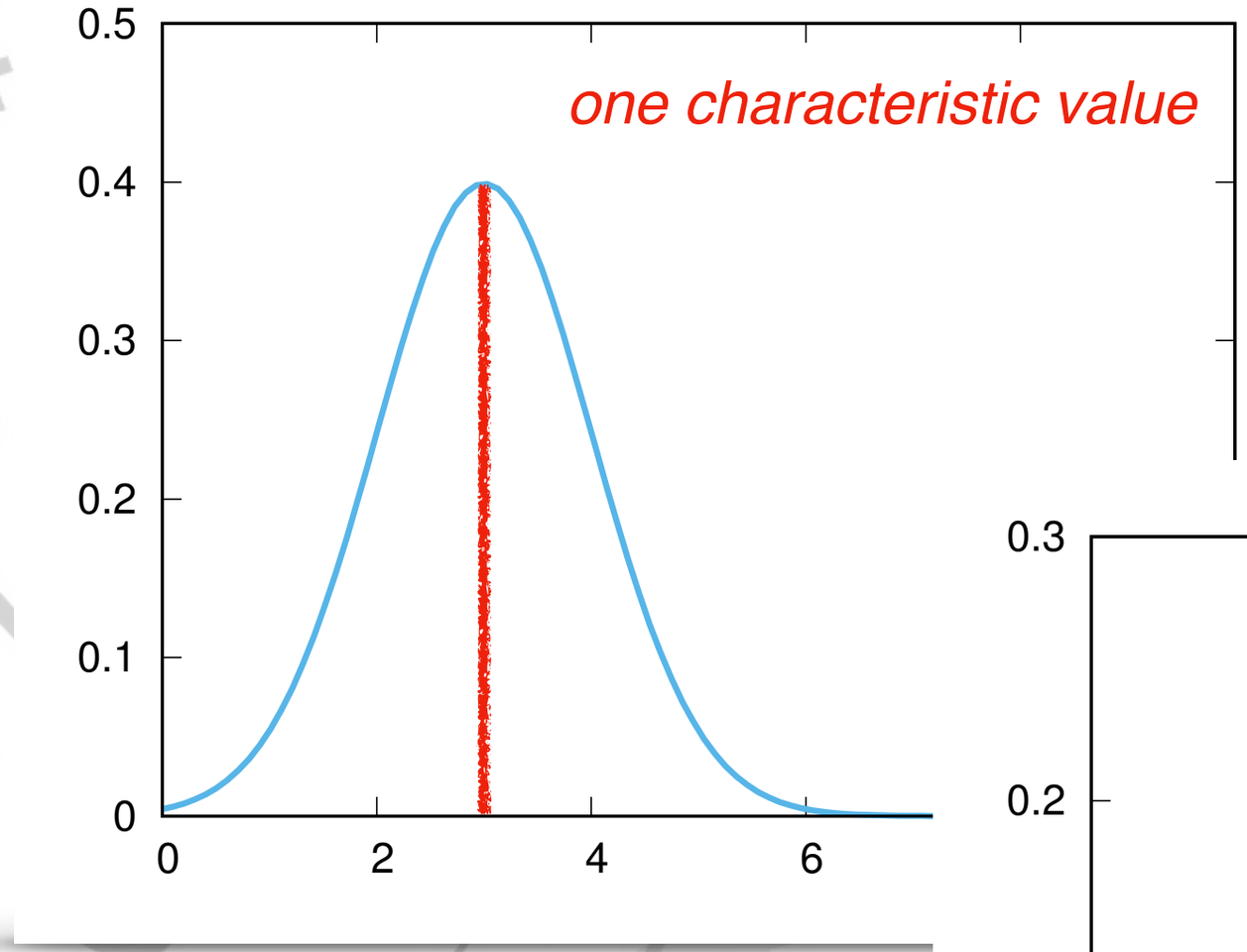


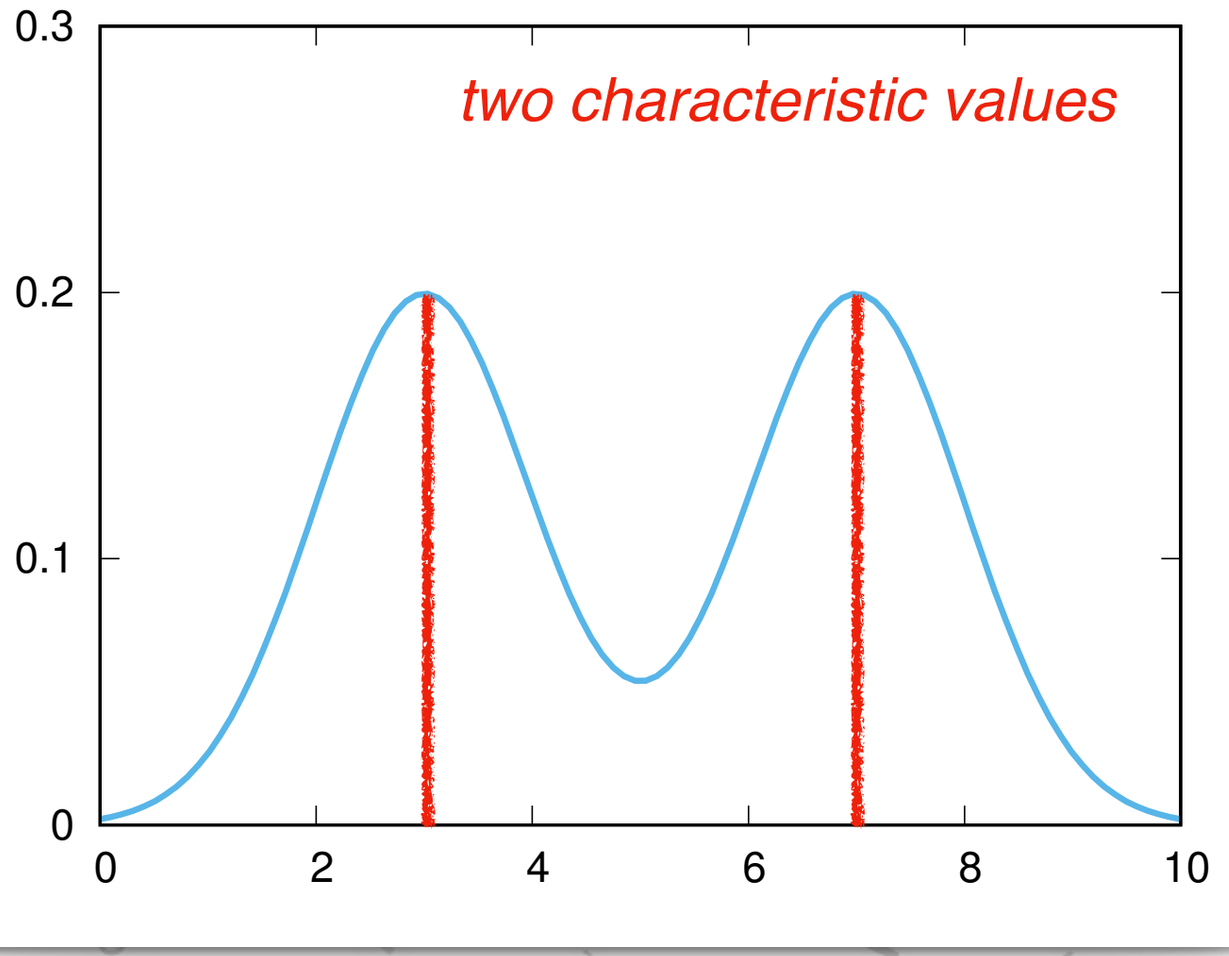
number of the bin = (int) (*value*/*width*)

How to make a **histograms**…

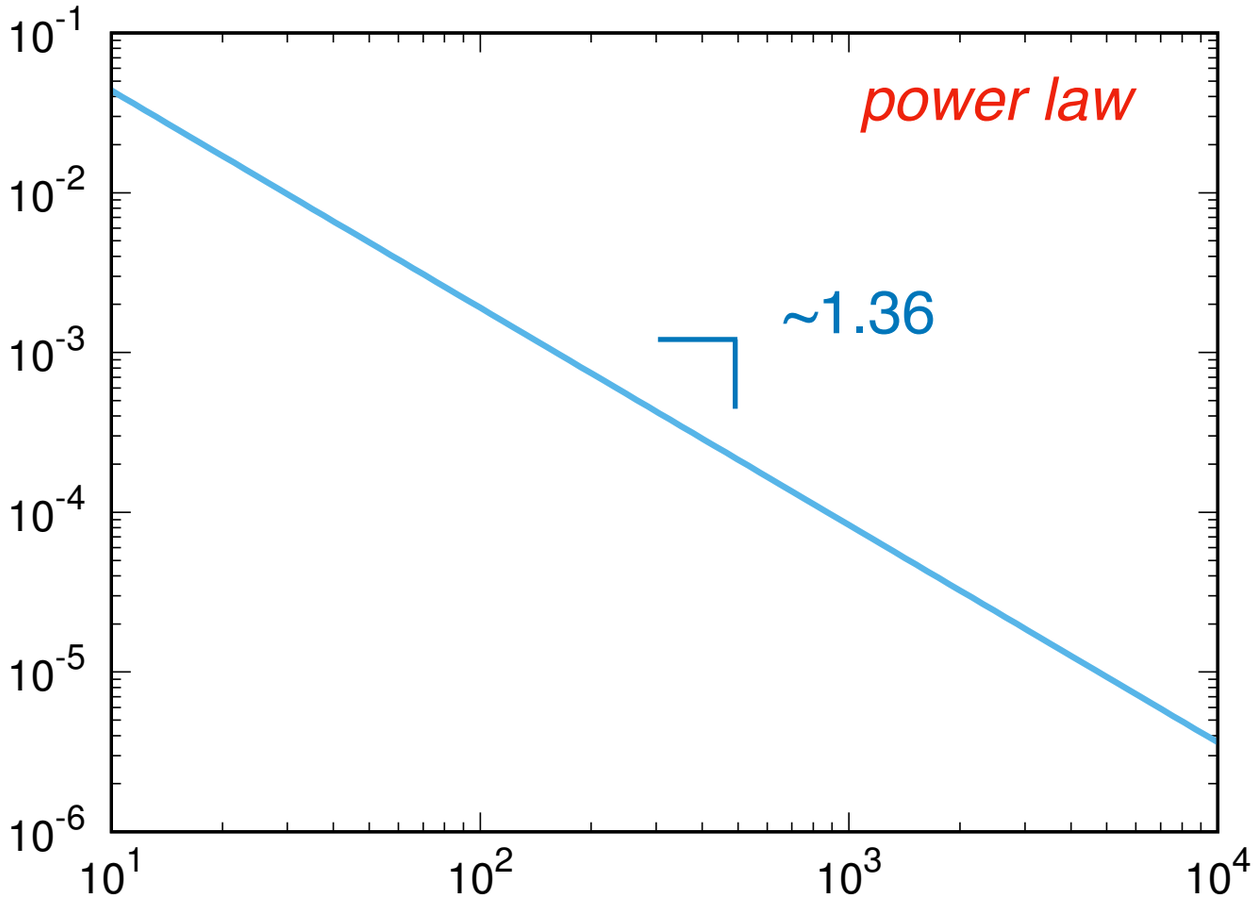


When does the **average** has a **meaning**?





When does the **average** has a **meaning**?



$$P(x) = Ax^{-\alpha}$$

$$< x >= \int_0^{x_{\max}} xP(x)\mathrm{d}x =$$

$$= A\int_0^{x_{\max}} x^{-\alpha+1}\mathrm{d}x$$

$$< x >= \frac{A}{2-\alpha}\left[x_{\max}^{2-\alpha}\right]$$

$\alpha < 2$:

$< x >$ *does not converge…*

$\alpha < 3$:

$< x^2 >$ *does not converge…*