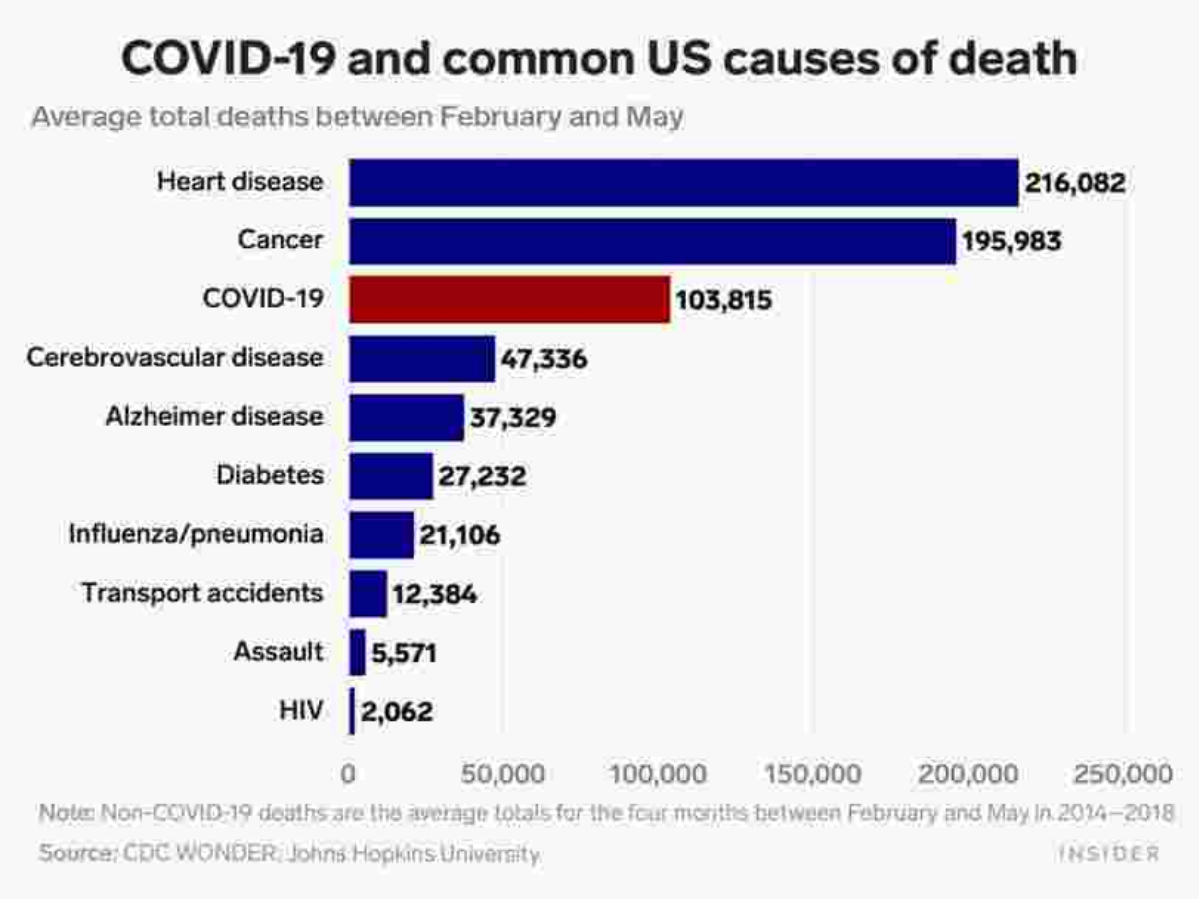
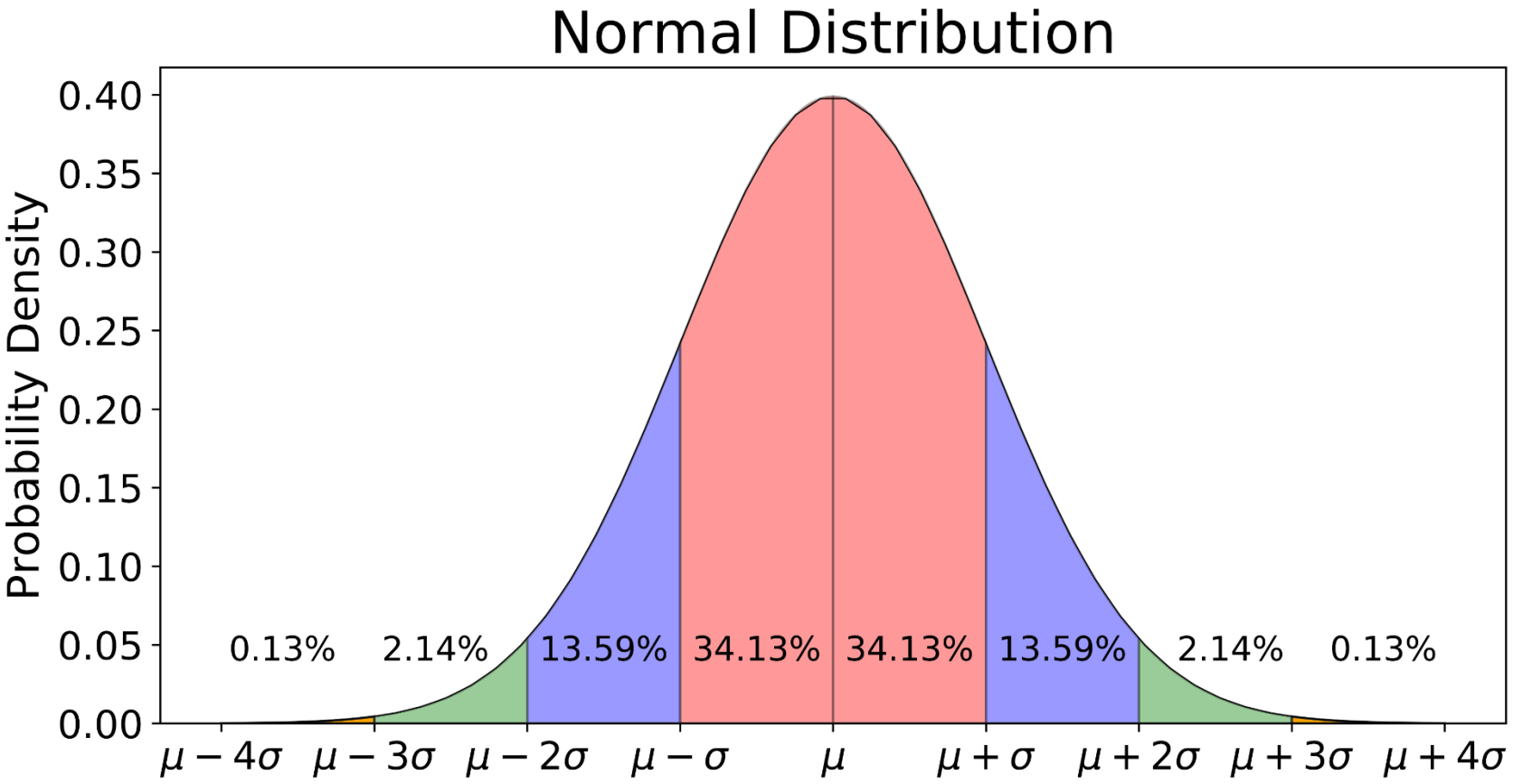
DATA COLLECTION AND VISUALIZATION

Pierre-Alexandre Balland

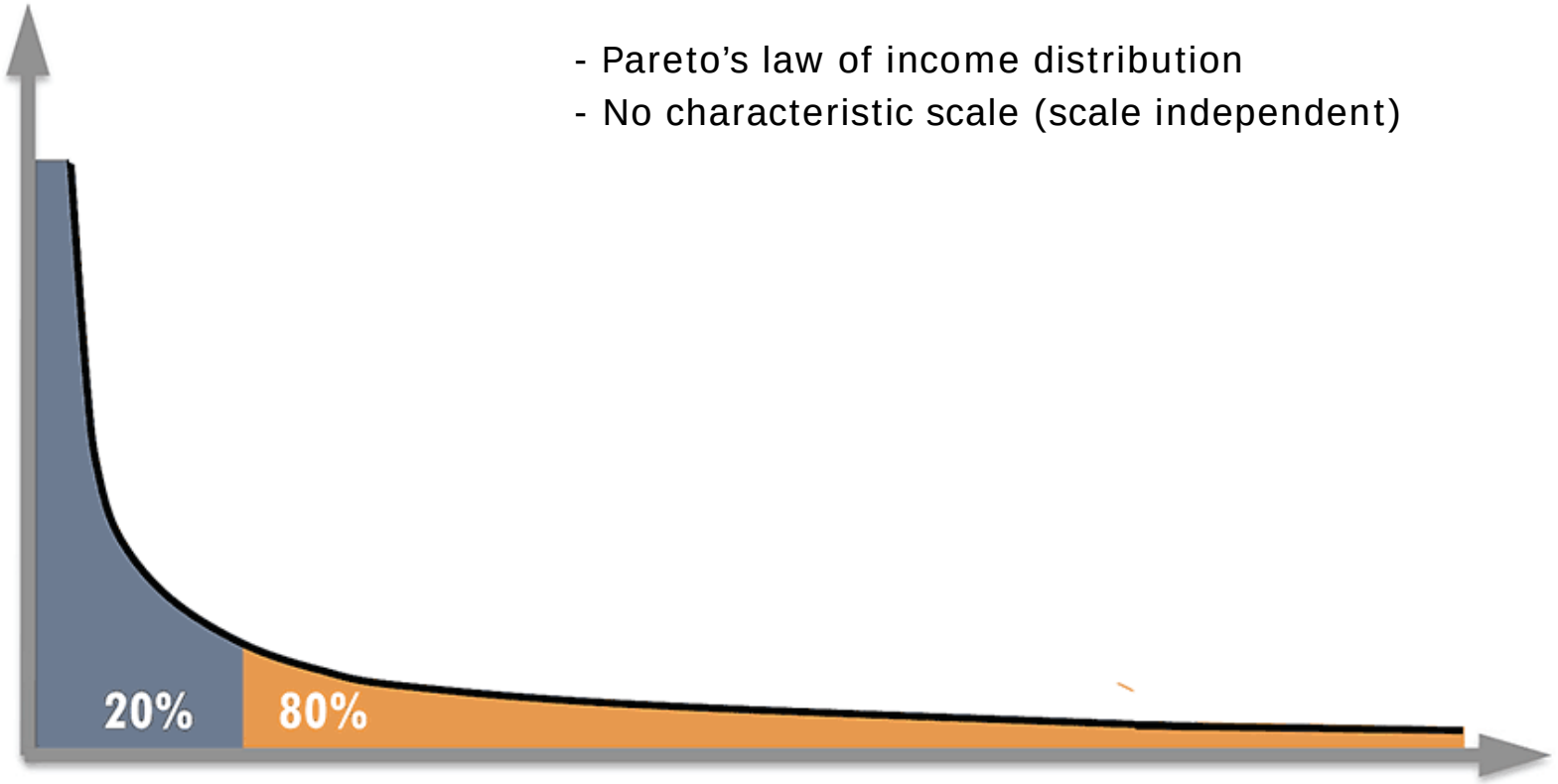
What is new in the DS revolution

What is the take-away?




Normal Distribution
Probability Density
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0.13%
2.14%
13.59%
34.13%
34.13%
13.59%
2.14%
0.13%
$\mu - 4\sigma$
$\mu - 3\sigma$
$\mu - 2\sigma$
$\mu - \sigma$
μ
$\mu + \sigma$
$\mu + 2\sigma$
$\mu + 3\sigma$
$\mu + 4\sigma$

Fat tails (power laws for instance)

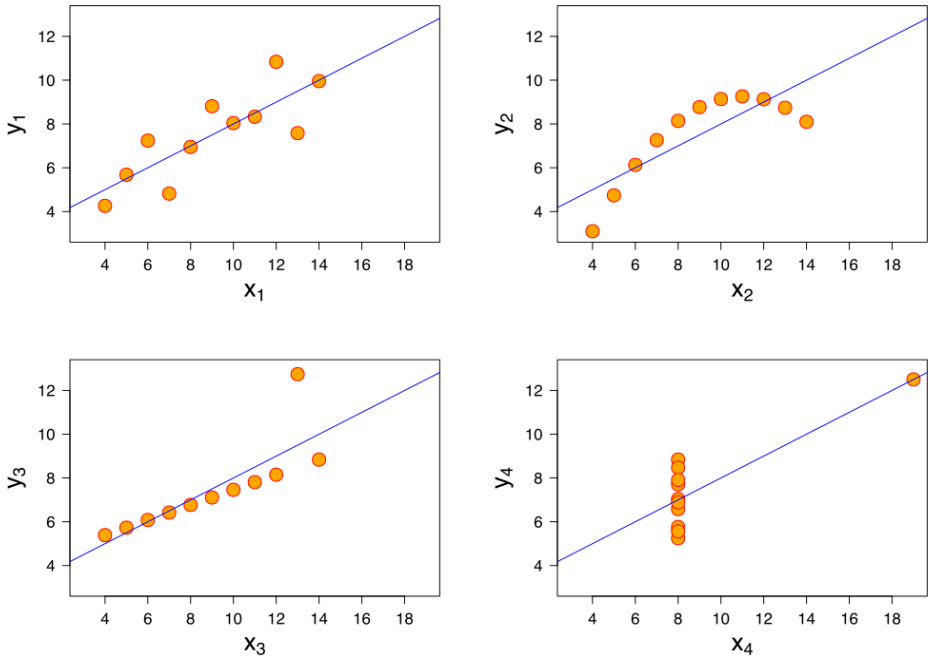


- Pareto’s law of income distribution
- No characteristic scale (scale independent)

Where is the data?

- Government sources: www.cbs.nl
- Intergovernmental organisations: www.oecd.org
- University datasets: www.dataverse.harvard.edu
- Data products: www.crunchbase.com
- Privately owned data
- Collect your own data: surveys
- Collect your own data: create a data harvesting product

Anscombe’s quartet



Mean of x = 9
Mean of y = 7.5
Correlation between x and y = 0.816
Linear regression line: y = 3.00 + 0.5x
R2 = 0.67