什么是数据科学

姚远

北京大学

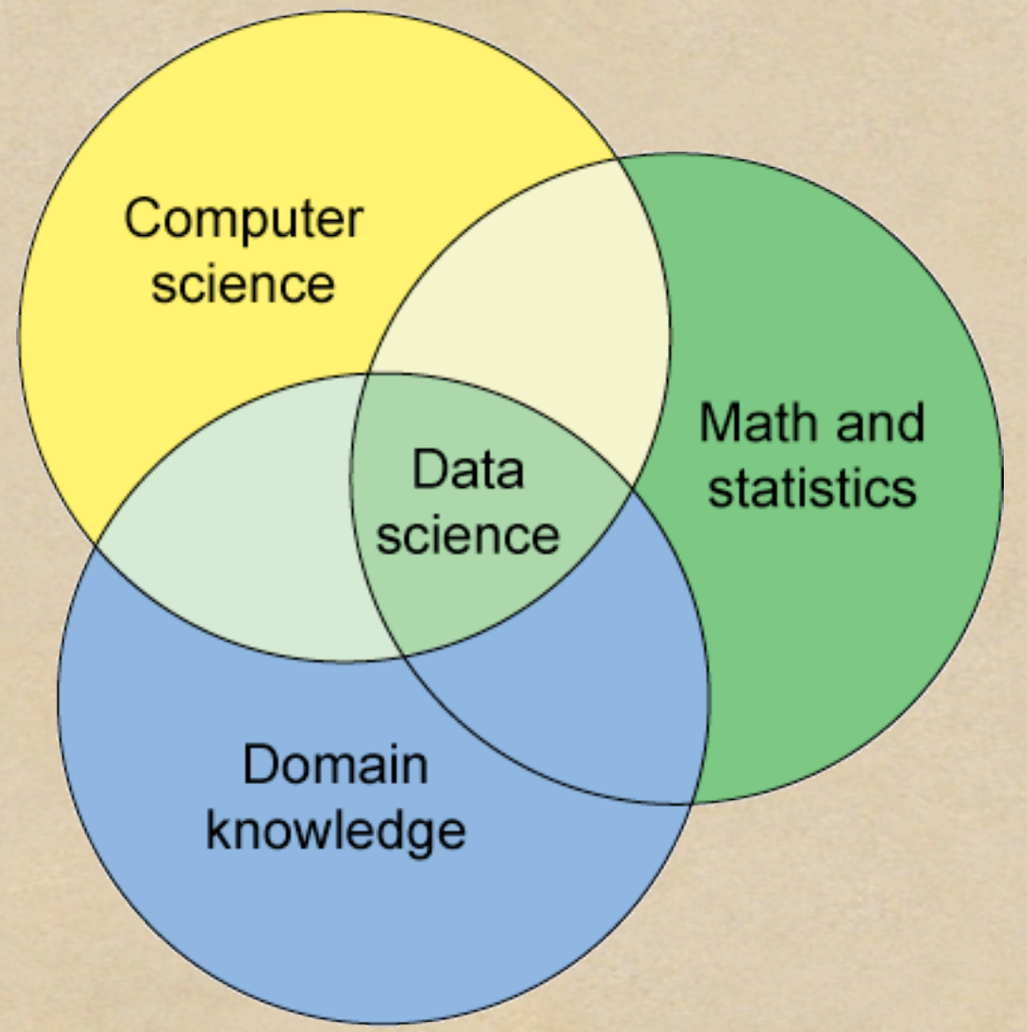
What's Data Science

- ***Data science*** is the study of the generalizable extraction of knowledge from data.

- Reference: wikipedia

- Dhar, V. (2013). "Data science and prediction". *Communications of the ACM* **56** (12): 64.

Data Science is highly interdisciplinary



http://www.ibm.com/developerworks/jp/opensource/library/os-datascience/figure1.png

Where the term from...

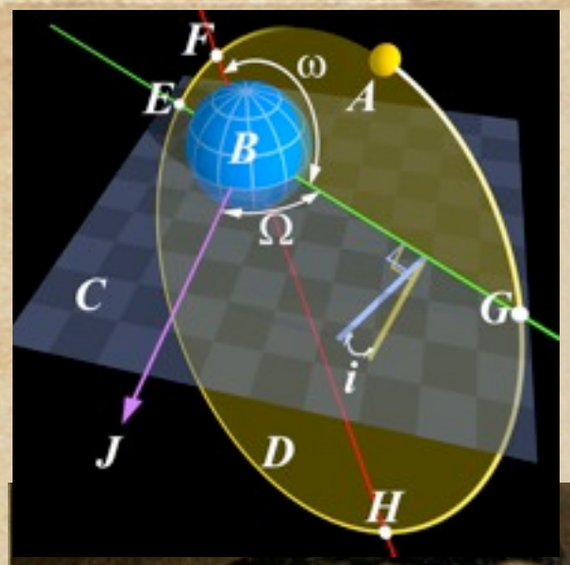
- 1960, Peter Naur, Computer Scientist
- 1972, John W. Tukey, Mathematician

  - It will still be true that there will be aspects of data analysis well called technology, but there will also be the hallmarks of stimulating science: intellectual adventure, demanding calls upon insight, and a need to find out "how things really are" by investigation and the confrontation

    of insights with experience. (Tukey's definition of 'Data Science'?)

- 1997, C. F. Jeff Wu (吴建福), Statistician

  - Statistics = Data Science?



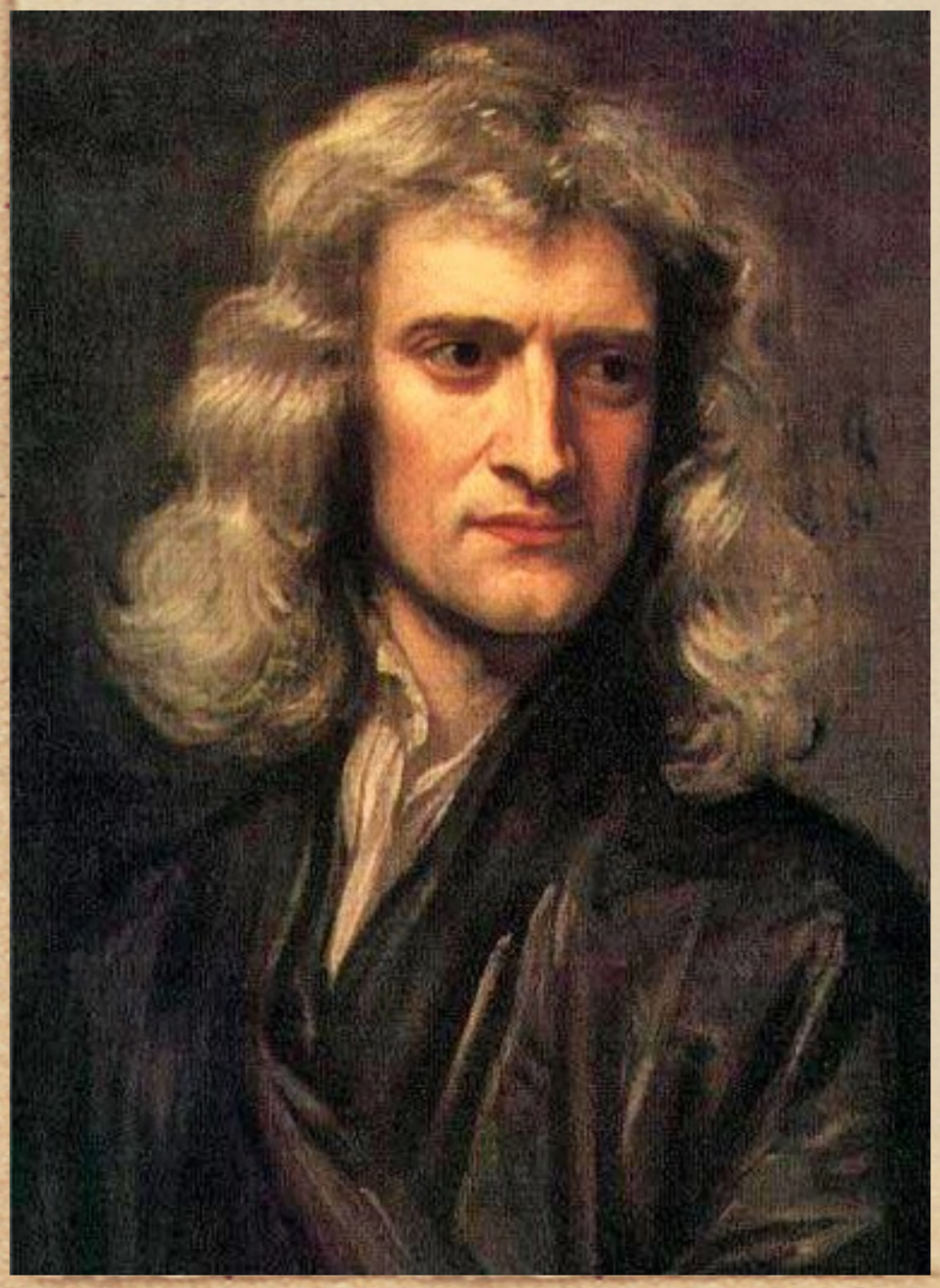
Johannes Kepler 1618

- 3 laws of planetary motion

Planet	Period (yr)	Average Distance (au)	T^2/R^3 (yr^2/au^3)
Mercury	0.241	0.39	0.98
Venus	.615	0.72	1.01
Earth	1.00	1.00	1.00
Mars	1.88	1.52	1.01
Jupiter	11.8	5.20	0.99
Saturn	29.5	9.54	1.00
Uranus	84.0	19.18	1.00
Neptune	165	30.06	1.00
Pluto	248	39.44	1.00

(NOTE: The average distance value is given in astronomical units where 1 a.u. is equal to the distance from the earth to the sun - 1.4957×10^{11} m. The orbital period is given in units of earth-years where 1 earth year is the time required for the earth to orbit the sun - 3.156×10^{7} seconds.)

Issac Newton's learning

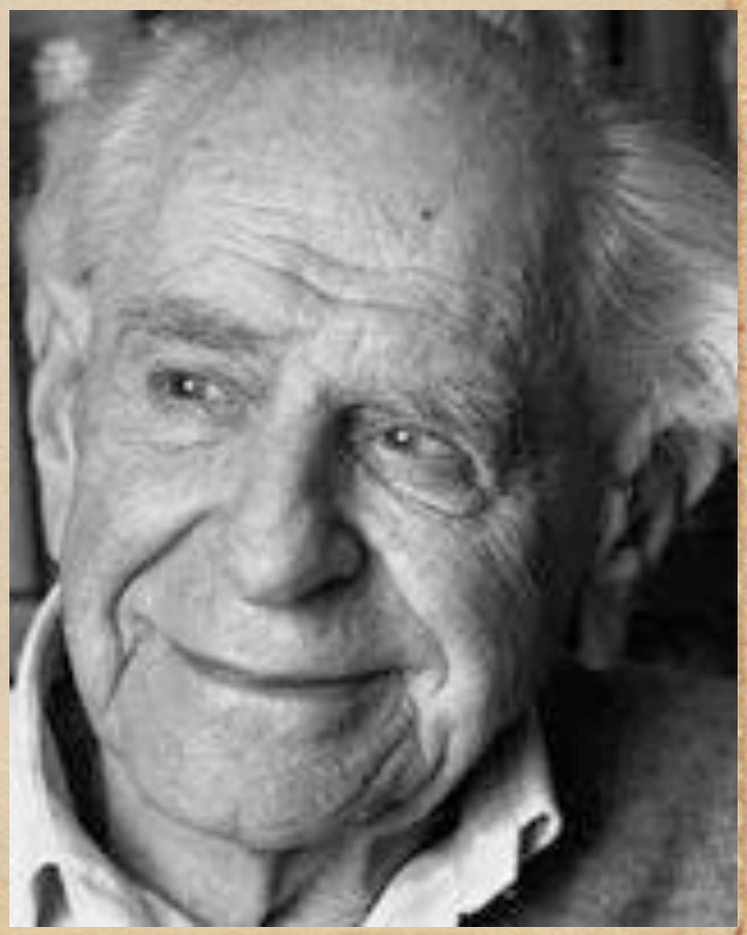


- Force f = m a
- Gratívity f = G Mm/a
- => Kepler's law

Karl Popper 1950s

- Falsifiability = Science vs. Pseudoscience

- A theory in the empirical sciences can never be proven, but it can be falsified, meaning that it can and should be scrutinised by decisive experiments.



Occam's Razor

- The hypothesis has to be as simple as possible, but not simpler. (Einstein)